Fall 1-18-2018

Worldwide Distribution of the Human Apolipoprotein E Gene - The Association between APOE, Subsistence, and Latitude

Tiffany S. Ho
CUNY Hunter College

Worldwide Distribution of the Human *Apolipoprotein E* Gene – the Association between *APOE*, Subsistence, and Latitude

by

Tiffany S. Ho

Submitted in partial fulfillment of the requirements for the degree of

Master of Arts in Anthropology, Hunter College

The City University of New York

2017

Thesis Sponsors:

December 01, 2017	Michael Steiper, PhD
Date	First Reader
December 01, 2017	Herman Pontzer, PhD
Date	Second Reader

DEDICATION

This thesis is dedicated to the memory of my grandma, Lam Yuet Mui, who has always inspired me with her hard work, sacrifice, and unconditional love. As an exceedingly intelligent, unconventional, and strong-willed woman – she inspired me to delve from the expected path and to pursue a graduate education in anthropology; watching her courageous fight against Alzheimer's Disease sparked my interest in the *apolipoprotein E* gene. Throughout moments of uncertainty and doubt while working on this thesis, I have looked to the memory of my grandma to continue onward.

ACKNOWLEDGEMENTS

I would like to begin by thanking my thesis advisor, Dr. Michael Steiper. Dr. Steiper helped me to gain the confidence that I needed to pursue this research and the many directions that it would ultimately take. He saw so many possibilities in this project, and his enthusiasm for the project fueled my own enthusiasm. Dr. Steiper spent a considerable amount of time guiding me through this project, and in the end, more so than anything else, it is he who helped to shape the direction that this project took. Because of his unfailing encouragement and guidance, I was able to finish this thesis — and subsequently, my M.A. in Anthropology.

I would also like to give special thanks to my second reader, Dr. Herman Pontzer, and to my graduate advisor, Dr. Yukiko Koga. Dr. Pontzer offered many insightful suggestions while writing my thesis. And Dr. Koga was instrumental in helping me forward throughout the graduate and thesis process. All of Dr. Pontzer and Dr. Koga's help and hard work made this thesis possible.

Thank you to my husband, Daniel Roses, for his support and faith in me as I completed my thesis. These past few years, we have experienced many life changes together. Throughout it all, his faith in me has never wavered, and he has pushed me to overcome challenges and refocus on my thesis.

Last but not least, I would like to thank my parents, Danny Ho and Lisa Cheung. My parents have a seemingly limitless capacity for generosity, kindness, and love. Throughout my years as an undergraduate and graduate anthropology student, they have encouraged and supported me. They did their best to minimize distractions so that

I could focus on continuing my education; their countless sacrifices and unconditional love have made it possible for me to pursue my passion and an education in anthropology. I could not have completed my education without them.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ABBREVIATIONS

ß coefficient	beta coefficient
1KGB	1000 Genomes Browser
A / arg	arginine
AD	Alzheimer's Disease
APOE	*apolipoprotein E* gene
apoE	apolipoprotein E protein
APOE1	*apolipoprotein* E1 allele
apoE1	apolipoprotein E3 isoform
APOE2	*apolipoprotein* E2 allele
apoE2	apolipoprotein E2 isoform
APOE3	*apolipoprotein* E3 allele
apoE3	apolipoprotein E3 isoform
apoE4	apolipoprotein E4 isoform
APOE3	*apolipoprotein* E3 allele
APOE4	*apolipoprotein* E4 allele
APOE5	*apolipoprotein* E5 allele
C / Cys	cysteine
CHD	coronary heart disease
CVH	cardiovascular disease
GWAS	genome wide association studies
Hardy Weinberg	HW

HDL	high-density lipoprotein
HDLC	high-density lipoprotein cholesterol
HDLR	high-density lipoprotein receptor
IBS	identity-by-state
IGSR	International Genome Sample Resource
LD	linkage disequilibrium
LDL	low-density lipoprotein
LDLC	low-density lipoprotein-cholesterol
LDLR	low-density lipoprotein receptor
MDS	multi-dimensional scaling
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NLM	National Library for Medicine
SNP	single-nucleotide polymorphism
VCF	variant call format

ABSTRACT

The human *apolipoprotein E* gene (*APOE*) plays an important role in metabolizing lipids, regulating plasma cholesterol, and maintaining biological function. Structural differences in *APOE* variants impact cholesterol absorption and health risk, so that alleles serve as biomarkers for numerous cardiovascular and neurological diseases (Lai 2015). Variant differences are determined by changes in two single nucleotide polymorphisms (SNPs), rs429358 and rs7412. Distribution of alleles varies across populations. Allele frequencies in populations have been shown to be associated with cultural and environmental factors, including subsistence strategy and latitude (Eisenberg 2010).

This study aims to provide a cross-population, genetic association study between *APOE*, subsistence strategy, and latitude. The objective of the study is to examine the roles that subsistence and latitude have in shaping *APOE* allele frequencies within populations. The study hypothesizes that E3 correlates with agriculture / post-agriculture and low latitude, and E4 correlates with non-agricultural and high latitude. The study further predicts that E2 is not linked to either subsistence or latitude.

To test these hypotheses, genotype data on 124 *APOE* SNPs, and subsistence and latitude data was compiled for 26 populations. The data were adjusted for population stratification, and remaining SNPs were tested for significance based on linkage between loci. Afterward, subsistence and latitude were first tested as independent variables for an association with each SNP / haplotype, then as covariates.

Results on the associations between *APOE* and subsistence and latitude were mixed. SNPs rs429358 and rs7412 were confirmed to be significant in determining *APOE* variation. Association results on each SNP showed a link between rs429358 and subsistence, and latitude, as well as between rs7412 and latitude – but not between rs7412 and subsistence. Association results on haplotypes confirmed the hypothesis that subsistence and latitude each play a role in *APOE* distribution – although this role lessened when considering the other variable. When subsistence and latitude were treated as independent variables, E3 showed an association with both subsistence and latitude. Yet, the correlation between E3 and subsistence disappeared when latitude was a covariate. Further, while E4 was confirmed to be associated with subsistence, this association decreased when latitude was a covariate.

The study also confirmed the subsistence hypotheses, with E3 linked to post-agriculture (when subsistence was an independent variable) and E4 linked to non-agriculture. However, the study refuted the latitude hypotheses by showing a reverse association than predicted, with E3 being associated with high latitude and E4 being associated with low latitude. Also, contrary to the hypotheses, E2 was shown to be associated with both subsistence and latitude. In summary, results from the study support an association between *APOE,* subsistence, and latitude; however, the results do not support the direction of association between specific *APOE* alleles and these variables.

1 INTRODUCTION

1.1 Apolipoprotein E

The human *apolipoprotein E* gene has been extensively studied by biomedical researchers for the role that it plays in regulating plasma cholesterol, repairing cells, and maintaining body function (Liu 2015). *APOE* has three codominant alleles: (1) E2, (2) E3, and (3) E4, resulting from changes in SNPs rs429358 and rs7412 (Utermann et al. 1980, Mahley et al. 2009, Villeneuve et al. 2014). Structural changes of these alleles affect function (Liu 2015). Changes impact major organs, including the liver, kidneys, and brain (Hu et al. 2011, Liu 2015), thereby affecting health risk and susceptibility to disease. Distribution of *APOE* alleles varies across populations, and variants serve as biomarkers for cardiovascular and neurological diseases (Lai 2013, Villeneuve et al. 2014). Research has also yielded findings showing that cultural and environmental factors, including subsistence strategy and latitude, are strongly associated with allele frequency (Corbo and Scacchi 1999, Eisenberg et al. 2010).

Both subsistence and latitude have been linked to *APOE* frequencies (Corbo and Scacchi 1999, Eisenberg et al. 2010). A non-agricultural (foraging, pastoralism, and horticulture) and agricultural / post-agricultural (agriculture, industrialism) population divide exists between E3 and E4 frequencies (Corbo and Scacchi 1999, Benyshek and Watson 2006, Trumble et al. 2017). A north-south population gradient of these two alleles also exists (Gerdes et al. 1992, Lucotte et al. 1997, Singh et al. 2006, Zhang et al. 2010, Hu et al. 2011). E2 has not been shown to have a link with subsistence or latitude (Corbo and Scacchi 1999). Findings on an association between *APOE*,

subsistence, and latitude have widespread implications for future studies on the gene and disease.

1.2 Aims

This study seeks to gain further insight into the roles that subsistence strategies and latitude play in determining *apolipoprotein E* allele distribution within populations. The study reviews biomedical research on *APOE*, subsistence, and latitude. Additionally, the study examines the association between *APOE* and these factors by compiling data from the 1000 Genomes Browser (1KGB), merging these datasets in VCFtools, and conducting statistical testing through PLINK software program.

The objective of this study is to examine the relationship between *APOE*, subsistence, and latitude in human populations. Structural differences between alleles play a key role in the regulation of plasma cholesterol (Mahley et al. 2009), with E2 and E3 demonstrating decreases in cholesterol and E4 demonstrating an increase in cholesterol (Liu 2015). Coupled with subsistence and latitude, which possibly impact cholesterol needs (Corbo and Scacchi 1999, Eisenberg et al. 2010), functional changes in alleles can be either advantageous or deleterious within populations.

Based on allele function and previous findings (reviewed below), the study hypothesizes a negative correlation between *APOE* and subsistence, whereby E4 (prevalent with non-agriculture), decreases with agriculture / post-agriculture – with the inverse being true for E3. The study also hypothesizes a positive correlation between *APOE* and latitude, whereby E4 increases with latitude – in direct contrast to E3

(reviewed below). E2 is not predicted to be linked to either subsistence or latitude (See TABLE 1.2).

TABLE 1.2 *Apolipoprotein E* Association Hypotheses

HAPLOTYPE	SUBSISTENCE*	LATITIUDE**
E2	No Correlation	No Correlation
E3	Agriculture, Industrialism	Low Latitude
E4	Foraging, Pastoralism, Horticulture	High latitude

*negative correlation **positive correlation

1.3 Layout

This thesis consists of six chapters. Chapter 1 begins with a summary section that briefly introduces *apolipoprotein E* and findings from previous studies. The second section of this chapter then discusses the research objectives for the study. The current section (1.3 Layout) provides an outline of the thesis chapters and subchapters.

Chapter 2 delves into the background of the *apolipoprotein E* gene and protein. The first two subchapters discuss the role that *apolipoprotein E* plays in regulating plasma cholesterol and how the structure of *apolipoprotein E* variants impact function. The third subchapter explains the importance of *APOE* polymorphisms in research on health risk and disease. The last two subchapters discuss findings from previous association studies, which have linked *APOE* frequencies to both subsistence strategies and latitude.

The remainder of the thesis details the five phases of the study: (1) Data Collection (2) Compact Datasets (3) Data Management and Summary Statistics (4)

Population Stratification, and (5) Association. Chapter 3 describes the materials and methods, including statistical analyses. This chapter outlines Phases 1 through 5.

Chapters 4 and 5 are divided into two parts. In Chapter 4, the first two subchapters describe the compiled data, as well as the pruned dataset. The remaining three subchapters examine the results of Phase 4 and Phase 5. Results include those of an indep-pairwise LD test and of a Hardy-Weinberg (HW) significance test, as well as, findings of association tests for *APOE* SNPs rs429358 and rs7412, and for *APOE* haplotypes E2, E3, and E4. Chapter 5 follows up with a discussion of results and theories regarding findings from previous studies. This chapter also offers possible theories for findings.

Finally, Chapter 6 concludes the thesis. The first section gives a brief summary of the study hypotheses and results. The second section discusses the strengths and limitations. The final section points out potential areas for future research.

2 BACKGROUND

2.1 Apolipoprotein E Gene / Protein

Discovered by V. Shore and B.G. Shore in 1973, and named by Gerd Utermann in 1975 (Liu 2015), the human *apolipoprotein E* (*APOE*) gene has been shown to be a biomarker for cardiovascular, and neurological diseases – leading to numerous association studies of the gene (Fullerton et al. 2010). *APOE* is located on chromosome 19 of the human genome. Found at 19q13.32, the gene spans base pairs 44,905,749 to 44,909,395 on the chromosome. *APOE* is composed of three introns and four exons, totaling 3,597 base pairs (Lai 2013). The *apolipoprotein E* gene encodes the lipid-binding protein of the same name, apolipoprotein E (apoE).

The apoE protein is a 34-KDA polypeptide composed of 299 amino acids, and synthesized primarily in the liver and macrophages (Mahley et al 2009, Villeneueve et al. 2014). As a part of a class of six lipid-binding proteins, apoE's main function is in regulating plasma cholesterol. As a component of both low density lipoproteins (LDLs) and high density lipoproteins (HDLs), apoE helps in the uptake of cholesterol-rich lipoproteins through low density lipoprotein receptors (LDLRs) and high density lipoprotein receptors (HDLRs). These receptors then transport and release cholesterol through the bloodstream (Liu 2015).

HDL cholesterol (HDLC) is transported back to the liver through reverse cholesterol transport (RCT) (Liu 2015), and the cholesterol is metabolized to be used in the production of bile for digestion and as a building block for cells (Villeneuve et al. 2014). In contrast, LDL cholesterol (LDLC), is quickly absorbed into the macrophages,

resulting in atherosclerosis (Liu 2015) and increasing risk for diseases such as cardiovascular disease (CVD) (Lai 2013) and Alzheimer's Disease (AD) (Ma et al. 2016). Thus, apoE's role in metabolizing HDLC and LDLC is key in maintaining body function (Villeneuve et al. 2014).

2.2 Apolipoprotein E Variants

Apolipoprotein E exists in three codominant alleles: (1) E2 (which codes for the apoE2 isoform), (2) E3 (which codes for the apoE3 isoform), and (3) E4 (which codes for the apoE4 isoform) (Villeneuve et al. 2014). Several other gene variations also exist, including E1 and E5; however, these mutations are rare and exist in less than 1% of the human population (Corbo and Scacchi 1999, Liu 2015).

Allele differences result from nucleotide substitutions in two non-synonymous single nucleotide polymorphisms (SNPs), rs429358 and rs7412 (Bekris et al. 2008), located on exon 4 of the apolipoprotein gene. Alleles vary based on a change of nucleotide bases between thymine (T) and cysteine (C) in each SNP. Changes in nucleotide bases take the form of T-rs429358/T-rs7412 in E2, T-rs429358/C-rs7412 in E3, and C-rs429358/C-rs7412 in E4 (Aucan et al. 2004) (See TABLE 2.2.1). Together, these haplotypes allow for three homozygous genotypes: (1) E2/E2, (2) E3/E3, and (3) E4/E4, as well as three heterozygous genotypes: (1) E2/E3, (2) E2/E4, and (3) E3/E4 (Villeneuve et al. 2014).

Variations in *APOE* alleles have a significant impact on apoE phenotypes. Nucleotide substitutions in the gene lead to amino acid substitutions in the protein.

Changes are expressed in three isoforms: (1) apoE2 (coded for by E2), (2) apoE3 (coded for by E3), and (3) apoE4 (coded for by E4). ApoE isoforms are noted by a single amino acid substitution between arginine (arg) and cysteine (cys) at residues 112 and 158 on the protein. Substitutions take the form of cys112/cys158 in apoE2, cys112/arg158 in apoE3, and arg112/arg158 in apoE4 (Aucan et al. 2004) (See TABLE 2.2.1). Amino acid substitutions cause structural changes that impact apoE function.

TABLE 2.2.1 Apolipoprotein E Substitutions and Variants

HAP	Nucleotide Base Substitution		Isoform	Amino Acid Substitution		GENO	Variations	
	rs429358	rs7412		112	158		HAP1	HAP2
E2	T	T	ApoE2	Cys	Cys	E2/E2	TT	TT
						E2/E3	TT	TC
						E2/E4	TT	CC
E3	T	C	ApoE3	Cys	Arg	E3/E3	TC	TC
						E3/E4	TC	CC
E4	C	C	ApoE4	Arg	Arg	E4/E4	CC	CC

Structural differences in apoE isoforms lead to differences in the rate of cholesterol absorption by each variant, thereby affecting cholesterol levels in the bloodstream (Mahley et al. 2009, Liu 2015). Differences are seen in lipoprotein and lipoprotein receptor preference, whereby isoforms have differential affinity for these molecules and receptors (Saito et al. 2001). ApoE2 demonstrates a marked affinity for HDLs, in contrast to apoE4's affinity for LDLs. These changes impact the ability of apoE to bind with LDLRs, which transport artery-clogging LDLC. The bonds between apoE4 and LDLRs are 50 – 100 times stronger than those between apoE2 and LDLRs. In

contrast to apoE2 and apoE3, apoE3 is the most neutral of the isoforms, due to its preference for HDLs and its contrasting affinity for LDLRs (Liu 2015).

Isoform differences in lipoprotein and lipoprotein receptor preference influence plasma concentration of apoE in individuals, which negatively correlates to clearance rate of lipoproteins (Han et al. 2016, Liu 2015). Individuals with the apoE2 phenotype show the highest concentration of apoE (Corbo and Scacchi 1999, Liu 2015), as well as the slowest uptake rate of lipoproteins. Slow uptake of lipoproteins ultimately leads to slower absorption of cholesterol (Liu 2015), resulting in lower overall cholesterol in the bloodstream (Corbo and Scacchi 1999, Hu et al. 2011). In comparison, individuals with the apoE4 phenotype have a lower concentration of plasma apoE (Corbo and Scacchi 1999, Liu 2015) but a high clearance rate, leading to quick cholesterol absorption (Liu 2015) and an increase in overall cholesterol levels (Corbo and Scacci 1999, Hu et al. 2011, Liu 2015) (See TABLE 2.2.2). Overall cholesterol levels play an important role in health and fitness; the impact of plasma cholesterol is seen in the disease risk of E2, E3, and E4 carriers (Gonzalez 2016).

TABLE 2.2.2 *APOE* Isoforms in Cholesterol Metabolism

Isoform	Lipoprotein Preference	LDLR Affinity	Concentration*	Clearance Rate**
ApoE2	HDL	Low	High	High
ApoE3	HDL	High	Medium	Medium
ApoE4	LDL	High	Low	Low

*E2>E3>E4 **E4>E3>E2

Susceptibility to disease varies by genotype, and risk is cumulative with each allele present (Corbo and Scacchi 1999). While E3 carriers tend towards moderate levels of plasma cholesterol, carriers of E2 and E4 show marked differences in their concentrations of HDLC and LDLC. Studies have shown that while E2 carriers have lower overall cholesterol levels, they also have higher concentrations of HDLC, which serve as a protection against CVD and neurological diseases such as AD (Lai 2015). In contrast, E4 has been associated with higher concentrations of LDLC, which leads to an increased risk of disease – particularly for homozygotes. For example, E4 carriers are more likely to develop AD than E2 carriers, and the odds of developing AD increases by 20% for homozygous carriers (Villeneuve et al. 2014).

TABLE 2.2.3 *APOE* Genotypes and Odds Ratio for Alzheimer's Disease (AD)

GENO	OR (< 1 are protective against AD; > 1 are susceptible to AD)
E2/E2	0.20
E2/E3	0.20
E2/E4	2.60
E3/E3	1.00
E3/E4	3.20
E4/E4	14.90

Studies have shown that a single copy of E4 leads to lower neurological fitness; two copies of E4 decreases fitness further. 50% of AD patients carry the E4 allele, and homozygous carriers have a 20% increase in likelihood of developing AD (Villeneuve et al. 2014).

2.3 Apolipoprotein E Single Nucleotide Polymorphisms

Genetic variation studies have extensively examined *APOE* and its association with common CVD and neurological diseases (Burns et al. 1972, Gonzalez 2016, Zhang

et al. 2010). However, while literature extensively covers rs429358 and rs7412, an association has also been found between disease and other *APOE* SNPs (Bekris et al. 2008, Bizarro et al. 2009, Masoodi et al. 2012, Ma et al. 2016, Limon-Sztencel et al. 2016). *APOE* includes several polymorphic SNPs, each of which has been shown to have causal effects on gene variation and risk to disease. An association test between *APOE* and AD revealed seven highly-polymorphic SNPs that have been correlated with AD: rs121918398, rs12198394, rs41382345, rs11542041, rs11542040, rs1142034, and rs11083750 (Masoodi et al. 2012).

APOE-correlated diseases have also been linked to rs405509 on the promoter region of the gene (Ma et al 2016), as well as to rs449467 and rs769446 (Bizzarro et al. 2009). These *APOE* SNPs serve as biomarkers for CVD and neurological diseases, including hypertension, coronary heart disease (CHD), and dementia (Limon-Sztencel et al. 2016). Thus, the association between *APOE* polymorphisms and environmental factors that impact cholesterol intake, basal metabolic rate (BMR), and cholesterol needs, has become an important study in biomedical research (Eisenberg et al. 2010).

The myriad of association studies on *APOE*, disease, and environment, have revealed surprising findings about *APOE* SNP correlation and the complexity of the gene. Causal effects on gene variation are intricate and likely the contribution of more than the two commonly-studied SNPs, rs429358 and rs7412 (Bizarro et al. 2009, Ma et al. 2016, Limon-Sztencel et al. 2016). This complexity highlights the importance of *APOE* variations and suggests the need for additional research on other gene

polymorphisms, as molecular studies yield new discoveries on the relationship between the gene, disease – and environment.

2.4 Association with Subsistence

Due to findings of a correlation between *APOE* and disease, *APOE* is amongst the most widely examined polymorphisms in genetic association studies (Fullerton et al. 2010). Studies have expanded to include research on the environmental variables that play key roles in the distribution of *APOE* alleles within and between populations (Eisenberg et al. 2010). While *APOE* alleles serve as biomarkers for disease, seemingly-deleterious allele variants may also have evolutionary advantages, causing them to be selected for in certain populations (Benyshek and Watson 2006, Corbo and Scacchi 1999, Trumble et al. 2017). Notable factors, including subsistence patterns (See TABLE 2.4.1), play a large role in the selection of gene variants within populations (Hancock et al. 2010).

Variance in subsistence strategies may account for the differences in allele frequencies between populations. Why specific alleles are more prevalent in certain populations is unknown. Theories suggest that subsistence strategies impact cholesterol intake (Corbo and Scacchi 1999) and energy expenditure (Boone 2002), which in turn, affect cholesterol needs. However, these theories remain unsubstantiated (Benyshek et al. 2006, Pontzer et al. 2012, Raichlen et al. 2017).

TABLE 2.4.1 Subsistence Strategies and Explanations

Category	Subsistence Strategy	Explanation
Non-Agricultural	Foraging	Plant collection, hunting, fishing
	Pastoralism	Domestication of animals
	Horticulture	Crop production on periodically-cultivated soil, left fallow after periods of use
Agricultural / Post-Agricultural	Agriculture	Crop production on permanently-cultivated land
	Industrialism	Standardized crop production

While only a handful of populations continue to rely on foraging as a subsistence strategy, this strategy has left a biological imprint on the human genome (Di Rienzo and Hudson 2005). Foraging and other non-agricultural populations demonstrate high rates of E4 (Corbo and Scacchi 1999). In the African Pygmies, a foraging people, E4 maintains a frequency of 40%, which is the highest known frequency of E4 in human populations. Another African people, the Khoisan, are also amongst the human populations with the highest frequency of E4, at 30%. Like the Pygmies, the Khoisan rely on foraging, as well as pastoralism (Corbo and Scacchi 1999, Fullerton et al. 2000). In contrast, E3 has been shown to be higher in agricultural / post-agricultural populations, suggesting that the gradual replacement of foraging with agriculture and industrialism gave rise to increase rates of E3 over time. Agricultural and industrialist European populations have comparatively high E3 frequencies, with Sardinians showing the highest frequency at 90% (Corbo and Scacchi 1999).

Furthermore, E3 is the highest frequency allele in all populations, including in non-agricultural populations with high E4 frequencies. And E4 retains a presence, albeit

at lower rates, even in agricultural / post-agricultural populations. This suggests that E4 was the ancestral allele, and E3 arose as a mutation that has gradually replaced E4 as the dominant allele over time (Corbo and Scacchi 1999, Fullerton et al. 2000, Di Rienzo and Hudson 2005). E2 frequencies remain low throughout all human population, with non-agricultural and agricultural / post-agricultural populations having similar frequencies (Corbo and Scacchi 1999) (See TABLE 2.4.2 and TABLE 2.4.3) – suggesting no selection. These studies support the premise that subsistence plays a role in the selection of some *APOE* variants and may account for allele differences between populations. However, theories as to why some alleles demonstrate a stronger association with subsistence than others remain unclear. This study seeks to confirm the hypotheses that E4 is linked with non-agriculture, and E3 is linked with agriculture / post-agriculture.

TABLE 2.4.2 Frequency of *APOE* Haplotypes in Human Populations (Liu 2015).

Allele	**FREQUENCY**
E2	0.05 – 0.10
E3	0.65 – 0.70
E4	0.15 - 0.20

TABLE 2.4.3 Frequency of *APOE* Genotypes in Human Populations(Liu 2015).

GENO	**FREQUENCY**
E2/E2	0.60
E2/E3	0.12
E2/E4	0.02
E3/E3	0.60
E3/E4	0.23

E4/E4	0.002

2.5 Association with Latitude

Like subsistence, latitude has become an increasingly researched variable in *APOE* studies. Theories suggest that the mechanism behind the association between *APOE* and latitude is climate, whereby extreme climates at specific latitudes and a subsequent increase in energy expenditure increase cholesterol needs (Eisenberg et al. 2010). As with subsistence theories, these theories require further research; further, findings on the direction of association between *APOE* and latitude have been mixed (Gerdes et al. 2006, Han et al. 2011). While most studies reveal a north-south gradient (Zhang et al. 2010, Han et al. 2011), other studies have found a curvilinear association with latitude (Eisenberg et al. 2010). Nonetheless, from these studies, it can be determined that latitude plays a role in the selection of *APOE* variants.

The impact of latitude on *APOE* distribution is demonstrated in several studies (Eisenberg et al. 2010, Hu et al. 2011). A positive correlation exists between E4 and latitude in Chinese populations. The Harbin population, located at 45° N, has an E4 frequency of 17.5%; this is compared to the lower latitude Haikou population, located at 20° N, which has an E4 frequency of a 6.5% (Hu et al. 2011). A positive correlation is also found in European populations, with Nordic, German, and Scottish populations in northern Europe demonstrating higher E4 frequencies than those of Swiss, Spanish,

and French populations in lower latitudes (Gerdes et al. 1992, Corbo and Scacchi 1999).

The highest E4 frequencies in Europe are found in the Saami populations located in northern countries, including Sweden, Finland, Norway, and parts of Russia. These populations have an overall E4 frequency of 0.310. Inversely, at a 0.64 frequency rate, these populations have the lowest E3 frequency amongst sampled European populations (Corbo and Scacchi 1999, Eisenberg et al. 2010). In comparison, the Sardinians, as the southernmost European population studied, have the lowest E4 frequency at 0.052 and the highest E3 frequency at 0.898 (Eisenberg et al. 2010).

A similar correlation has been found throughout Asia (Zhan et al. 2015, Han et al. 2016) and North America (Fullerton et al. 2000), in addition to a smaller, but still present, correlation in Africa (Zhan et al. 2015). E2 frequencies are low throughout all studied human populations, with the high-latitude Saami and lower latitude Sardinians having the same E2 frequency of 0.050 (Corbo and Scacchi 1999), suggesting that latitude plays no role in the presence of this particular allele. However, given previous research, latitude is likely associated with the other *APOE* variants. This study aims to examine the *APOE* latitude gradient by testing the hypotheses that E3 is correlated with low latitude, and E4 is correlated with high latitude.

3 Methods

3.1 Phase 1 – Collecting Data

Phase 1 focused on compiling data. Genotype data included SNP data, population codes, and allele frequencies that had been compiled during Phase 3 of the 1000 Genomes Project and archived in the National Center for Biotechnology Information (NCBI). Phenotype data included subsistence and latitude data (See TABLE 3.1.1).

TABLE 3.1.1 Compiling Data

Purpose	Steps	Data Compiled
Compile Genotype Data	1	
	1A	Genetic Sequences
	1B	Population Codes
		Allele Frequencies
Compile Phenotype Data	2	Subsistence
	3	Latitude

Genotype data was acquired in variant call format (VCF) through a NCBI database, the 1000 Genomes Browser (1KGB). This was located on the center website (accessed at https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/?assm=GCF_000001405.25). Samples were comprised of genotype sequences and allele frequencies from 124 loci on the *apolipoprotein E* gene, located on chromosome 19 of the human genome. Data from 2,535 individuals was obtained.

The study also included phenotype data on the subsistence strategies of each sampled population. Information on subsistence patterns was gathered through anthropological literature and published information (Benyshek et al. 2006, Hancock et al. 2010). Based on literature, each population was catalogued as having a subsistence strategy of: (1) foraging, (2) pastoralism, (3) horticulture, (4) agriculture, or (5) industrialism.

Foraging, pastoralism, and horticulture were associated with nomadic lifestyles, non-mechanized tools, and low fat diets. In comparison, agriculture and industrialism were associated with sedentary lifestyles, high-fat diets, and greater mechanization (Binford 1990, Ember 2014). While the majority of sampled populations were agricultural or industrial (Benyshek et al. 2006, Hancock et al. 2010), several populations used a mix of strategies. For the purposes of this study, these strategies were simplified into a logistics / binary variable (Purcell et al. 2007) and further classified into two groups: (1) non-agricultural, and (2) agricultural / post-agricultural (See TABLE 3.1.2).

TABLE 3.1.2 Subsistence Strategies in Human Populations

Code	Category	Subsistence Strategy	Lifestyle	Explanation	Diet / Lifestyle	Geographical Location
1	Non-Agricultural	Foraging	Non-Sedentary	Non-mechanized	Low-Fat	Central / S Africa, NE Asia, NE Australia
		Pastoralism				E Africa, throughout Asia
		Horticulture				NW South America, SE Asia, NE Australia

2	Agricultural / Post-Agricultural	Agriculture	Sedentary	Some Mechanization	High-Fat	Throughout
		Industrialism		Mechanized	High-Fat	Throughout (seen in developed nations)

After gathering subsistence information, a separate dataset with population data for each sampled individual was acquired through The International Genome Sample Resource (IGSR) at http://www.internationalgenome.org/category/population/. In addition to population codes, which delimited the population origin of each individual, the dataset included the geographic location for each population. The latitude of each population was revealed by mapping the population's location in Google Earth, Version 7.3.0.3832 (accessed at https://www.google.com/earth/) (Keyhole, Inc. 2001). In this study, latitude was treated as a quantitative variable in PLINK (Purcell et al. 2007).

3.2 Phase 2 – Creating Compact Datasets

In Phase 2, compact datasets were created from the data compiled in the first phase. Data was first converted from VCF files into flat files that were compatible with software used in later phases of the study. These files were then merged, first with population data, then with subsistence and latitude data (See TABLE 3.2.1).

TABLE 3.2.1 Creating Compact Datasets

Purpose	Step	Output Files
File Conversion	1	Genotype Data
	1A	MAP

	1B	PED
Merging Files	2	Population Codes (PED)
	3	Phenotype Data (PED)
	3A	Subsistence
	3B	Latitude

Data for this study, downloaded from the 1KGB, was stored in compressed files VCF files (Danecek et al. 2011). The datasets were converted into flat file formats that could be used with bioresearch software in the next three phases of the study. Afterward, genotype and phenotype data were merged into single, compact datasets. These compact datasets would allow for quicker parsing of data.

Data from 124 typed SNPs was converted from VCF files into flat files through VCFtools, Version 4.2, (accessed at http://vcftools.sourceforge.net/) (Danecek et al. 2011). VCFtools produced two output files: (1) a MAP file for genotype data, and (2) a PED file for phenotype data. The MAP file for this study provided summary statistics for the information stored in the PED file, with four columns listing chromosome, biomarker (locus), genetic distance, and genome position (See Step 1A).

The correlating PED file had six columns: (1) family ID (initially listed as copies the family IDs) (2) sample ID, (3) paternal ID, (4) maternal ID, (5) sex, and (6) affection. Phenotype columns were followed by the genotypes of each individual. Paternal ID, maternal ID, sex, and affection were not variables in this study. As such, these columns were left as unknown, coded for by 0 (See Step 1B). Affection was later replaced with subsistence pattern and latitude, according to the variable being tested (See Step 3).

Step 1A) MAP File

19	rs565734271	0	45408766
19	rs72654466	0	45408786
19	rs405509	0	45408836
19	rs567707344	0	45408860
19	rs537380977	0	45408915

Step 1B) PED File, Original

HG02485	HG02485	0	0	0	0	C	T	C
HG02489	HG02489	0	0	0	0	C	T	C
HG02052	HG02052	0	0	0	0	T	T	T
HG02053	HG02053	0	0	0	0	C	T	T
HG02054	HG02054	0	0	0	0	C	T	C

Column 1 lists copies of the individual IDs (shown in Column 2); this list is later replaced by a list of population codes.
Columns 3, 4, 5, and 6 list paternal ID, maternal ID, sex, and affection; these variables are not used in the study and are left as unknown.

The PED file was merged with several other phenotype datasets through R, Version 3.3.3 (accessed at https://www.r-project.org/) (R Core Team 2013). The file was first merged with a file on population codes, which was acquired from the IGSR and replaced the family ID column in the original PED file. (See Step 2).

Step 2) Merged Population Codes (PED)

ACB	HG02485	0	0	0	0	C	T	C
ACB	HG02489	0	0	0	0	C	T	C
ACB	HG02052	0	0	0	0	T	T	T
ACB	HG02053	0	0	0	0	C	T	T
ACB	HG02054	0	0	0	0	C	T	C

Column 1 lists population codes, which have replaced the copies of individual IDs in the original PED file, shown in Step 1.

Two compact copies of this PED file were created; each file was combined with the dataset of an independent variable. Having two copies of the PED file allowed for

separate testing in Phase 5 of the study, whereby the variables, subsistence and latitude, were individually tested against *APOE*. With these copies, the dataset totaled three files: (1) a MAP file and (2) PED files that were almost identical, differing only in the sixth column.

The first of these files was merged with subsistence data; data in the subsequent file was coded as 1 for non-agricultural subsistence strategies and 2 for agricultural / post-agricultural subsistence strategies (See Step 3A). The second PED file was merged with latitude data (See Step 3B). This phenotype data replaced the affection column in the original PED file.

Step 3A) Final PED File – Copy 1, Subsistence

ACB	HG02485	0	0	0	2	C	T	C
ACB	HG02489	0	0	0	2	C	T	C
ACB	HG02052	0	0	0	2	T	T	T
ACB	HG02053	0	0	0	2	C	T	T
ACB	HG02054	0	0	0	2	C	T	C

Column 6 lists subsistence data, which has been coded for non-agriculture and agriculture / post-agriculture.

Step 3B) Final PED File – Copy 2, Latitude

ACB	HG02485	0	0	0	13	C	T	C
ACB	HG02489	0	0	0	13	C	T	C
ACB	HG02052	0	0	0	13	T	T	T
ACB	HG02053	0	0	0	13	C	T	T
ACB	HG02054	0	0	0	13	C	T	C

Column 6 lists latitude data.

3.3 Phase 3 – Data Management and Summary Statistics

Data for this study was analyzed through PLINK, Version 1.07 (accessed at http://zzz.bwh.harvard.edu/plink/) (Purcell et al. 2007), which has five key functions: (1) data management, (2) summary statistics, (3) population stratification, (4) association analysis, and (5) identity-by-descent estimation. Phase 3 was comprised of data management and summary statistics and broken up into three steps: (1) file conversion, (2) genotyping rate test, and (3) allele frequency test (See TABLE 3.3.1).

TABLE 3.3.1 Managing and Summarizing Data

Purpose	Steps	Output Files
Data Management	1	Convert Files
	1A	FAM
	1B	BIM – 2 copies: 1) subsistence, 2) latitude
Summary Statistics	2	Genotyping Rate
	2A	Missing Loci Data
	2B	Missing Individual Data
	3	Allele Frequency

Prior to analysis, the MAP and PED files from Phase 2 were converted into binary files through PLINK. PLINK produced three binary files: (1) a BIM file (converted from the MAP file) for genotype data (See Step 1A), and (2) two FAM files (converted from the PED files in Phase 1) for phenotype data (See Step 1B).

Step 1A) BIM File – Genotype Data

```
19          rs565734271    0          45408766    A          G
19          rs72654466     0          45408786    G          C
19          rs405509       0          45408836    T          G
19          rs567707344    0          45408860    A          T
```

```
19          rs537380977    0          45408915     T          C
```

Step 1B) FAM File, Subsistence

```
ACB         HG01914      0          0          0          2
ACB         HG01985      0          0          0          2
ACB         HG01986      0          0          0          2
ACB         HG02013      0          0          0          2
ACB         HG02051      0          0          0          2
```

Subsistence data is shown in the last column.

Step 1B) FAM File, Latitude

```
ACB         HG01914      0          0          0          13
ACB         HG01985      0          0          0          13
ACB         HG01986      0          0          0          13
ACB         HG02013      0          0          0          13
ACB         HG02051      0          0          0          13
```

Latitude data is shown in the last column.

After the data was converted into binary files, the dataset was pruned for later testing in PLINK. To ensure more accurate results for later analysis, the data was filtered based on two premises: (1) genotyping rate and (2) minor allele frequency.

In the genotyping test, loci and individuals with missing genotypes were eliminated from the study. The parameters of the test were left at the default, MIND > 0.1 (Purcell et al. 2007), so that the threshold for missing data was 10%. The genotyping rate test produced two output files for: (1) loci with missing samples (See Step 2A) and (2) individuals with missing genotypes (See Step 2B); individuals and loci with more than 10% of their data missing were eliminated from the dataset.

Step 2A) Missing Loci Data

CHR	SNP	N_MISS	N_GENO	F_MISS
19	rs565734271	0	2535	0
19	rs72654466	0	2535	0
19	rs405509	0	2535	0
19	rs567707344	0	2535	0
19	rs537380977	0	2535	0

Step 2B) Missing Individual Data

FID	ID	MISS_PHENO	N_MISS	N_GENO	F_MISS
ACB	HG02485	N	0	124	0
ACB	HG02489	N	0	124	0
ACB	HG02052	N	0	124	0
ACB	HG02053	N	0	124	0
ACB	HG02054	N	0	124	0

The remaining data was analyzed for allele frequencies, using the default parameters of MAF > 0.05 (Purcell et al. 2007). The output file showed the major and minor allele for each *APOE* locus, as well as revealed the number of alleles observed for each locus and the minor allele frequency. Based on the parameters, only loci with minor allele frequencies above 0.05 continued to be used in the study.

3.4. Phase 4 – Population Stratification and Significance

In Phase 4, data was corrected for population stratification through three filters. First, an independent-pairwise (“indep-pairwise”) test was conducted. Then results were subject to a Hardy-Weinberg test, whereby significant SNPs were flagged for further study. Afterward, an IBS test was conducted, and sampled individuals were clustered into homogenous groups (See TABLE 3.4.1).

TABLE 3.4.1 Adjusting for Population Stratification

Steps	Test
1	Indep-Pairwise
2	HWE
3	IBS / Clustering

To account for population structuring, which results in genotyping errors, an indep-pairwise test was used to eliminate loci that were affected by linkage disequilibrium (LD) (Purcell et al. 2007). The indep-pairwise test was based on a linear regression model of $r^2 = 1 - SS_{res} / SS_{tot}$ (Li 1997, Buzdugan et al. 2016). The parameters of the analysis were set so that $r^2 > 0.5$, with SNPs having r^2 values greater than 0.5 showing LD and pruned from the dataset.

Using a 50 SNP window that shifted across 10 SNPs at a time, the indep-pairwise function tested the 124 sampled SNPs for r^2 values that were greater than 0.5. Based on those parameters, the test produced two output files: (1) loci not affected by LD, and (2) loci affected by LD.

SNPs affected by LD were removed from the study, and SNPs not affected by LD were extracted from the complete dataset for testing. Remaining loci were subject to a HW test, which was based on the algorithm, $p^2 = AA + aa$ (Li 1997). Results showing highly significant *p* values are often indicative of genotyping errors that result from population structuring, and these are often pruned from large genome-wide dataset.

However, in multi-allelic genes such as *APOE*, results may instead be indicative of natural selection – thereby, showing SNPs with modestly significant *p* values as important markers and retaining them in the study. Given that *APOE* is a multi-allele biomarker based on two particular SNPs, rs429358 and rs7412 (Wang 2011, Villeneuve

et al. 2014, Buzdugan et al. 2016), only these two SNPs were retained for association tests.

Using genotype data from the remaining loci, sampled individuals were examined through an IBS test, which analyzed similarity between all possible pairs of sampled individuals. The default controls, ppc1e-3, were used, so that each group would have at least one case and at least one control. Based on these parameters, genotypes of sampled individuals were examined for identical nucleotide segments at the typed SNPs (Buzdugan et al. 2016), and genetic distance was calculated between individuals. Sampled individuals were then accordingly clustered into homogenous groups, which also generated several MDS covariates that were used to control for population stratification in association tests (Purcell et al. 2007). The first two covariates were included when testing for association in later phases of the study.

3.5 Phase 5 – Association

In the final phase of the study, the data were subject to two sets of association tests through PLINK (See TABLE 3.5). All tests included the first two MDS covariates from the IBS test to control for population stratification. The *APOE* alleles were tested both as two separate SNPs and then as haplotypes against subsistence and latitude as independent variables. The purpose of these tests was to determine whether an association existed between *APOE* and each variable, as well as examine the direction of association (for example, is non-agriculture associated with E2, E3, or E4?)

Logistical association tests were conducted for subsistence, which was treated as an “affection” or a binary trait. Variants with odds ratios less than 1 were inferred as being correlated to non-agriculture; those with odds ratios greater than 1 were inferred to be associated with agriculture / post-agriculture. For latitude, which was treated as a continuous / quantitative trait, linear association tests were conducted. Variants with negative beta coefficients (ß coefficients) (less than 0) were determined to be linked with low latitude; those with positive ß coefficients (greater than 0) were determined to be linked with high latitude (See TABLE 3.5.1).

Afterward, logistic models were conducted with subsistence and latitude as covariates. The purpose of these tests was to determine whether each variable had a modifying effect on *APOE*. Did the strength of association with *APOE* change when both variables were examined together? The parameters of each association test were set to $p > 0.05$, whereby SNPs / haplotypes with p values of less than 0.05 were determined to be statistically significant in subsistence and latitude.

TABLE 3.5.1 *APOE* Association Tests

Test Number	Loci	Variable	Association Test
1A	SNP	Subsistence	Logistic
1B		Latitude	Linear
1C		Subsistence / Latitude as Covariates	Logistic
2A	Haplotype	Subsistence	Logistic
2B		Latitude	Linear
2C		Subsistence / Latitude as Covariates	Logistic

TABLE 3.5.2 Parameters for *APOE* Association Tests

Association		Logistic Association		Linear Association	
Significance	**P Value**	**SUB**	**OD**	**LAT**	**BETA**
Yes	< 0.05	1	< 1	Low	< 0
No	> 0.05	2	> 1	High	> 0

4 RESULTS

4.1 Compiled Data

Data for the study was collected through three sources: (1) the 1KGB for genetic data, (2) anthropological literature on subsistence data, and (3) Google Earth for latitude data. A search for *APOE* on the NCBI's 1KGB yielded several VCF files containing genotype data for 124 loci from *APOE*. For each of the 124 loci, 2,535 individuals were genotyped. VCF files contained the 124 SNP markers, individual IDs of the 2,535 individuals, and genotypes for each individual at the 124 loci.

Population data was also found through the 1KGB. The data showed that the 2,535 individuals sampled came from 26 populations, for which allele frequencies were included (See TABLE 4.1.2). Sample sizes ranged from 65 to 112 per population, with an average of 98 individuals sequenced. Population data, comprised of geographical locations, showed that populations engaged in a variety of subsistence strategies and were widely-dispersed through Africa, Asia, Europe, North America, and South America.

A literature search revealed that populations participated in various subsistence strategies. While many subsistence strategies, excepting pastoralism, were present in the total sample, there was little variety in the methods that groups used to acquire food. Variation was further eliminated by treating subsistence as a binary variable, whereby populations were further categorized into two subsistence methods: (1) non-agricultural and (2) agricultural / post-agricultural.

Amongst the sampled populations, the majority tended towards agricultural / post-agricultural subsistence strategies (Hancock et al. 2010). Half of the populations

included in the study were categorized as agricultural or industrialist (Benyshek et al. 2006). Of the remaining populations, only three were categorized as non-agricultural (Hancock et al. 2010). The Gambian population demonstrated a subsistence method of horticulture. And two other groups, the Esan and the Luhya, engaged in multiple subsistence methods. The Esan were determined to have a mixed foraging and agriculturalist strategy, while the Luhya were categorized as a mixed horticulturalist and agriculturalist population. No pastoralist populations were included in the sample (Benyshek et al. 2006).

Following a literature search on subsistence, locations of groups, acquired through the 1GSR, were mapped in Google Earth. Findings showed that with the exception of two populations, the majority of sampled populations were situated north of the equator. Of the exceptions, only the Peruvian population of Lima were located below the equator at 12° S, and the Luhya of Webuye, Kenya were located directly on the equator. Nonetheless, despite the lack of populations sampled from the southern hemisphere, sampled populations were widely-dispersed, ranging between low and mid latitudes, with a few high latitude populations. The geographic location of these groups ranged from the southernmost populations, the Peruvians at 12° S and the Luhya at 0°, to the northernmost population, the Finnish at 61° N (See TABLE 4.1.1).

TABLE 4.1.1 Subsistence Strategies and Latitude of Sampled Populations

Population	POP	Subsistence	SUB	Location	LAT
African Caribbean	ACB	Industrialism	2	Bahamas	13 ° N
African American	ASW	Industrialism	2	U.S.A.	32° N
Bengali	BEB	Agriculture	2	Bangledash	23° N
Chinese Dai	CDX	Agriculture	2	Xishuang, China	22° N
European American	CEU	Industrialism	2	Utah, U.S.A.	39° N
Northern Chinese Han	CHB	Industrialism	2	Beijing, China	39° N
Southern Chinese Han	CHS	Industrialism	2	Guangzhou, China	23° N
Colombian	CLM	Industrialism	2	Medellin, Colombia	6° N
Esan	ESN	Foraging, Agriculture	1	Nigeria	9° N
Finish	FIN	Industrialism	2	Finland	61° N
British	GBR	Industrialism	2	Great Britain	53° N
Gujarti Indian	GIH	Agriculture	2	Houston, T.X., U.S.A.	29° N
Gambian	GWD	Horticulture	1	Western Division, Gambia	13° N
Iberian Spanish	IBS	Industrialism	2	Iberian Peninsula, Spain	40° N
Indian Telegu	ITU	Agriculture	2	U.K.	55° N
Japanese	JPT	Industrialism	2	Tokyo, Japan	35° N
Khin	KHV	Agriculture	2	Ho Chi Minh City, Vietnam	10° N
Luhya	LWK	Horticulture, Agriculture	1	Webuye, Kenya	0°
Mende	MLS	Agriculture	2	Sierra Leone	8° N
Mexican American	MXL	Industrialism	2	L.A., C.A., U.S.A.	34° N
Peruvian	PEL	Agriculture	2	Lima, Peru	12° S
Punjabi	PJL	Agriculture	2	Lahore, Pakistan	31° N
Puerto Rican	PUR	Industrialism	2	Puerto Rico	18° N
Tamil	STU	Agriculture	2	Siri Lanka	55° N
Toscani	TSI	Industrialism	2	Tuscany, Italy	43° N
Yoruba	YBI	Agriculture	2	Ibadan, Nigeria	7° N

TABLE 4.1.2 *APOE* Allele Frequencies of Sampled Populations

POP	SUB	LAT	rs429358		rs7412	
			T	C	C	T
ACB	2	13°N	0.7448	0.2552	0.9249	0.0751
ASW	2	32°N	0.7951	0.2049	0.8750	0.1250
BEB	2	23°N	0.9128	0.0872	0.9477	0.0523
CDX	2	22°N	0.8978	0.1022	0.8987	0.1022
CEU	2	39°N	0.8232	0.1768	0.9343	0.0657
CHB	2	39°N	0.8981	0.1019	0.8932	0.1068
CHS	2	23°N	0.9429	0.0571	0.9238	0.0762
CLM	2	6°N	0.8457	0.1543	0.9149	0.0851
ESN	1	9°N	0.7576	0.2424	0.9495	0.0505
FIN	2	61°N	0.8131	0.1869	0.9293	0.0707
GBR	2	53°N	0.8242	0.1758	0.9231	0.0769
GIH	2	29°N	0.9515	0.0485	0.9612	0.0388
GWD	1	13°N	0.7257	0.2743	0.8673	0.1327
IBS	2	40°N	0.8598	0.1402	0.9439	0.0561
ITU	2	55°N	0.9167	0.0833	0.9510	0.0490
JPT	2	35°N	0.9183	0.0817	0.9519	0.0481
KHV	2	10°N	0.9091	0.0909	0.8283	0.1717
LWK	1	0°	0.6212	0.3788	0.9545	0.0455
MLS	2	8°N	0.7412	0.2588	0.8588	0.1412
MXL	2	34°N	0.9141	0.0859	0.9531	0.0469
PEL	2	12°S	0.9412	0.0588	0.9941	0.0059
PJL	2	31°N	0.9167	0.0833	0.9635	0.0365
PUR	2	18°N	0.8942	0.1058	0.9519	0.0481
STU	2	55°N	0.8676	0.1324	0.9559	0.0441
TSI	2	43°N	0.8972	0.1028	0.9533	0.0467
YBI	2	7°N	0.7639	0.2361	0.8935	0.1065
AVG:			**0.8494**	**0.1506**	**0.9249**	**0.0751**

4.2 Pruned Dataset

Phase 3 and Phase 4 of the study were focused on filtering the dataset. Individuals / loci with missing data and SNPs affected by population stratification were eliminated from study. SNPs not found to be statistically significant to the *APOE* gene were also eliminated from study. The data was subject to several filters, including a genotyping rate test, allele frequency test, indep-pairwise test, HW test, and IBS analysis.

A genotyping rate test, with parameters set at MIND > 0.1, yielded results that showed no missing data. For every one of the 2,535 individuals sampled, genotype data was available at all 124 of the loci tested. With this test, no individuals or SNPs were eliminated from study.

Following a genotyping test, an allele frequency test, with parameters of 0.01, resulted in 116 SNP being removed from the study. This left eight biomarkers for further testing. The remaining SNPs included rs405509 (located on the promoter region of *APOE*), rs440446, rs769450 (which has also been associated with apolipoprotein A5), rs769449, rs1081105, and rs877973. Minor allele frequencies for these SNPs ranged from 0.4736 in rs405509 (highest) to 0.1598 in rs877973 (lowest). Rs429358 and rs7412 were also included, with minor allele frequencies of 0.1507 and 0.07475, respectively (See TABLE 4.2.1).

TABLE 4.2.1 Minor Allele Frequencies of *APOE* SNPs

CHR	SNP	A1	A2	MAF
19	rs405509	T	G	0.4736
19	rs440446	C	G	0.3746

19	rs769450	A	G	0.3268
19	rs429358	C	T	0.1507
19	rs7412	T	C	0.0747
19	rs769449	A	G	0.0658
19	rs1081105	C	A	0.0297
19	rs877973	A	C	0.0159

4.3 Linkage Disequilibrium and Significance of *APOE* SNPs

Data was adjusted for population stratification through an indep-pairwise test. An indep-pairwise test on the eight remaining locus followed the parameters of $r^2 > 0.5$. SNPs that did not pass the parameters of the test were listed in an output file. A second output file listed five remaining SNPs, which had passed the indep-pairwise test. These SNPs were rs769450 at position 45410444, rs1081105 at position 45412955, and rs877973 at position 45409283 – as well as SNPs rs429358 and rs7412 (See TABLE 4.3.1). These SNPs were statistically determined to be in linkage equilibrium and remained in the dataset, which was further pruned through a HW test.

TABLE 4.3.1 Indep-Pairwise Results for *APOE*

CHR	SNP	POS	A1	A2
19	rs769450	45410444	A	G
19	rs429358	45411941	C	T
19	rs7412	45412079	T	C
19	rs1081105	45412955	C	A
19	rs877973	45409283	A	C

After the indep-pairwise test, a HW test was conducted on remaining loci. Using the parameters, $p > 0.05$, the test further reduced the list of biomarkers to be tested for

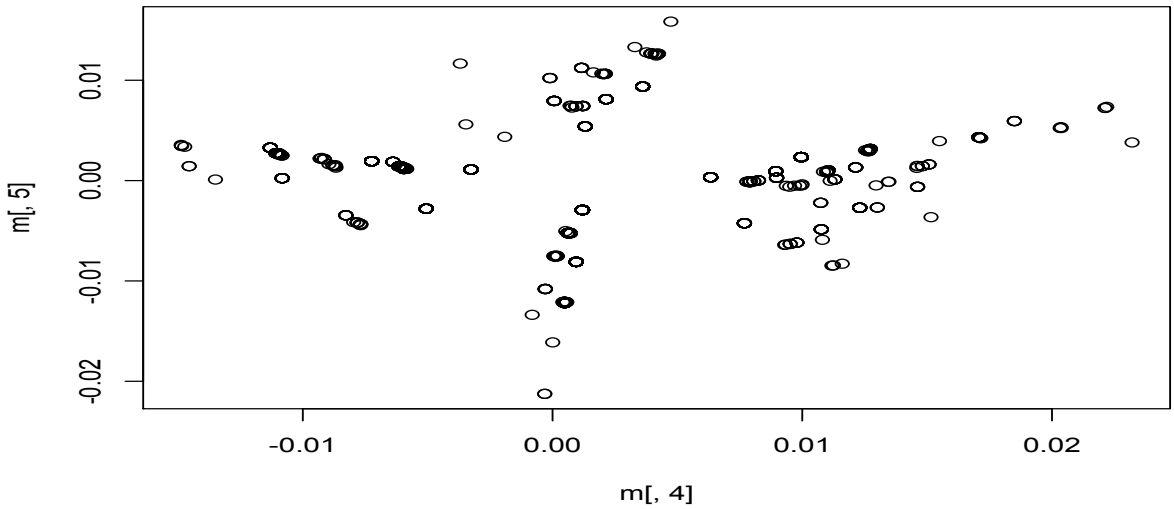
association with subsistence and latitude. HW results yielded two SNPs, rs429358 and rs7412, which *p* values of 0.02945 for rs429358 and 0.001484 for rs7412 (See 4.3.2). Given that research has linked these SNPs to *APOE* alleles (Wang 2011, Buzdugan et al. 2016), as well as, the modestly significant *p* values, it is unlikely that HW findings were the result of genotyping errors. Thus, these SNPs were not removed from the analysis based on the HW test. Because rs529358 and rs7512 are precisely the SNPs relevant to determining *APOE* variants (Villeneuve et al. 2014), only these two SNPs were retained for the association analysis to simplify the remaining analyses.

TABLE 4.3.2 HWE Results for *APOE*

CHR	SNP	A1	A2	GENO	O(HET)	E(HET)	P
19	rs429358	C	T	72/620/1843	0.2446	0.2560	0.02945
19	rs7412	T	C	14/351/2170	0.1385	0.1383	0.00148

Finally, an IBS test was conducted on remaining SNPs. Using the control, ppc1e-3 (Purcell et al. 2007), genetic distance was calculated between sampled individuals. Individuals were then clustered into homogenous groups (See FIGURE 4.3.3), based on similarity of their nucleotide sequences at rs429358 and rs7412 (Buzdugan et al. 2016). Clustering also generated several MDS covariates, which adjusted for population stratification. Based on IBS results, the first two covariates were included in association tests.

FIGURE 4.3.3 Sampled Data, Adjusted for Population Stratification



4.4. Association of *APOE* SNPs

SNPs rs429358 and rs7412 were tested for a correlation with: (1) subsistence, (2) latitude, and (3) subsistence and latitude as covariates. In all tests, the first two MDS covariates were included to adjust for population stratification. A logistical association test for subsistence showed that rs429358 was significantly associated with this phenotype. The odds ratio for rs429358, 0.5217, suggests that the dominant allele, C, is associated with non-agriculture (See TABLE 4.4.1).

TABLE 4.4.1 Logistic Association Test for *APOE* SNPs and Subsistence

SNP	A1	A2	Test	OR	STAT	P	Significance
rs429359	C	T	SUB	0.5217	-7.4000	1.36e-13	Yes
rs7412	T	C	SUB	1.0140	0.1012	0.9194	No

A linear association test for latitude showed a significant association with both SNPs. For rs429358, the negative clinked the C allele to low latitude. Similarly, the negative ß coefficient for rs7412 linked T to low latitude (See TABLE 4.4.2).

TABLE 4.4.2 Linear Association Test for *APOE* SNPs and Latitude

SNP	A1	A2	Test	BETA	STAT	P	Significance	Correlation
rs429359	C	T	LAT	-4.327	-6.598	5.067e-11	Yes	Low Lat
rs7412	T	C	LAT	-3.171	3.455	0.0005603	Yes	Low Lat

However, when subsistence and latitude were treated as covariates, association results changed significantly. In the covariate model, rs429358 had a *p* value of 0.01697 for subsistence and 4.955e-78 for latitude. From these values, it was inferred that each variable, subsistence and latitude, had a mitigating effect on the association between rs429358 and the other variable. While still present, the strength of association between rs429358 and subsistence, and between rs429358 and latitude, decreased.

For rs7412, controlling for latitude caused the SNP to become highly significant in its association with subsistence (p = 9.396e-83). With an odds ratio of 1.135, this demonstrated that the dominant allele, T, was associated with agriculture / post-agriculture when testing for subsistence and latitude together. In this model, rs7412 continued to be associated with latitude (See TABLE 4.4.3).

TABLE 4.4.3 Covariate Association Model for *APOE* SNPs, rs429358 and rs7412

SNP	A1	A2	Test	OR	STAT	P	Significance	Correlation
rs429359	C	T	COVAR	0.7758	-2.387	0.01697	Yes	1, High Lat
rs7412	T	C	COVAR	1.1350	19.270	9.396e-83	Yes	2, High Lat

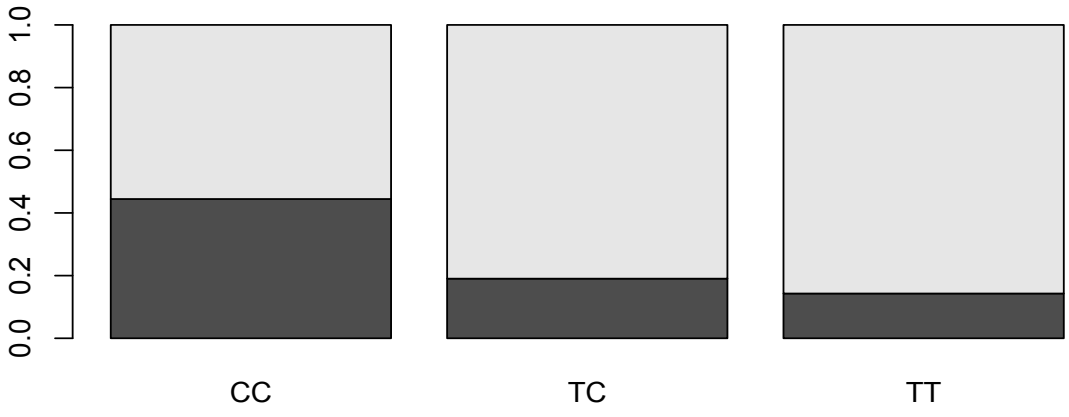
4.5 Association of *APOE* Haplotypes

The same set of association tests was done for the haplotypes: E2, E3, and E4. In each test, population stratification was accounted for by including the first two MDS covariates resulting from the IBS test in Phase 4. Both E3 and E4 were associated with subsistence (See TABLE 4.5.1). E3 was shown to be correlated with agriculture / post-agriculture. In contrast, E4, with an odds ratio less than 1, was shown to be correlated with non-agriculture (See FIGURE 4.5.2).

TABLE 4.5.1 Logistic Association Test for *APOE* Haplotypes and Subsistence

HAP	rs429358	rs7412	Test	OR	STAT	P	Significance	Correlation
E2	T	T	SUB	1.03	0.058	0.809	No	N/A
E3	T	C	SUB	1.63	38.7	4.92e-10	Yes	2
E4	C	C	SUB	0.525	52.8	3.76e-13	Yes	1

FIGURE 4.5.2 *APOE* Haplotype by Subsistence



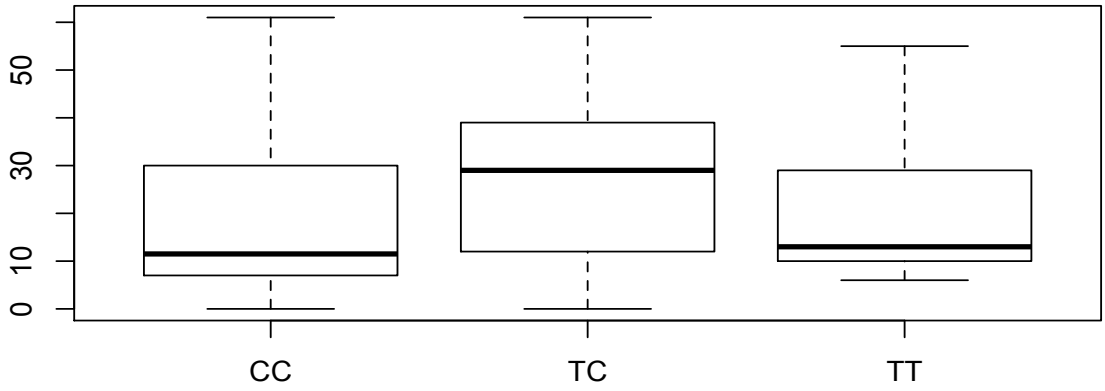
*Dark Grey = Non-Agriculture **Light Grey = Agriculture / Post-Agriculture

A linear association test for latitude showed that all three haplotypes are correlated with latitude (See TABLE 4.5.3). The β coefficients for E2, E3, and E4 were 3.05, 4.32, and -4.28, respectively showing that E2 and E4 are correlated with low latitude and that E3 is correlated with higher latitude (See FIGURE 4.5.4).

TABLE 4.5.3 Linear Association Test for *APOE* Haplotypes and Latitude

HAP	rs429358	rs7412	Test	BETA	STAT	P	Significance	Correlation
E2	T	T	LAT	-3.05	10.8	0.00103	Yes	Low Lat
E3	T	C	LAT	4.32	59.3	1.93e-14	Yes	High Lat
E4	C	C	LAT	-4.28	42	1.1e-10	Yes	Low Lat

FIGURE 4.5.4 *APOE* Haplotype by Latitude



In the final association test, subsistence and latitude were treated as covariates. Thus, the results were adjusted to consider the mitigating influence of each variable. When corrected for latitude, E2 was significantly associated with post-agriculture while

E4 was significantly associated with non-agriculture (See TABLE 4.5.3). For E3, the correlation disappeared, as findings showed that the main determinant for that haplotype is latitude.

TABLE 4.5.5 Covariate Association Model for *APOE* Haplotypes, E2, E3, and E4

HAP	rs429358	rs7412	Test	OR	STAT	P	Significance	Correlation
E2	T	T	COVAR	1.41	4.71	0.0299	Yes	Post-Agriculture
E3	T	C	COVAR	1.07	0.535	0.4650	No	-
E4	C	C	COVAR	0.775	5.63	0.0177	Yes	Non-Agriculture

5 DISCUSSION

Study results yielded mixed findings. HW analysis showed a clear link between rs429358 and rs7412, as well as, demonstrated similarities between observed and expected heterozygosity for both SNPs – indicating natural selection acting on these two SNPs and also confirming that *APOE* is a multi-allelic biomarker (Wang 2011) determined by these SNPs (Mahley et al. 2009). After accounting for population structuring, preliminary association tests on these *APOE* SNPs demonstrated a link between rs429358 and both subsistence and latitude, as well as a link between rs7412 and latitude.

Association tests on *APOE* haplotypes E2, E3, and E4 supported the subsistence hypothesis of a link between E4 and non-agriculture (See TABLE 5.0.1), even when correcting for latitude and population stratification. However, evidence for this link is weak, given the small sample size of non-agricultural populations in the study (see CHAPTER 6.2). When correcting for latitude, there was no link between E3 and agriculture / post-agriculture. These results also did not support the latitude hypotheses of a link between E3 and low latitude and E4 and high latitude. Rather, the study found a negative correlation between E3 and high latitude, and between E4 and low latitude. This is inconsistent with previous studies that demonstrate link between E3 and low latitude and E4 and higher latitude (Lucotte et al. 1999, Hu et al. 2011). Further, association tests yielded results demonstrating a negative correlation between E2, agriculture / post-agriculture, and low latitude (See TABLE 5.0.2). Thus, the hypothesis

that E2 is not connected to either subsistence or latitude was refuted. When corrected for latitude, results linked E2 to agriculture / post-agriculture.

TABLE 5.0.1 Association of *APOE*, Subsistence, and Latitude (Hypotheses)

rs429358	rs7412	HAP	SUB*	LAT**
T	T	E2	2	Low
T	C	E3	2	Low
C	C	E4	1	High

*Negative association **Positive association

TABLE 5.0.2 Association of *APOE*, Subsistence, and Latitude (Results)

rs429358	rs7412	HAP	SUB	LAT
T	T	E2*	2	Low
T	C	E3**	N/A	High
C	C	E4*	1	Low

*Association with subsistence and latitude **Association with latitude only
APOE haplotypes show a negative correlation with both subsistence and latitude.

Research has offered several theories regarding direction of association between *APOE* and both subsistence and latitude, whereby E3 demonstrates neutral selection while E4 demonstrates positive selection (Corbo and Scacchi 1999, Di Rienzo and Hudson 2005, Raichlen and Alexander 2014). With regards to subsistence, theories include the “thrifty allele” hypothesis and the energy expenditure hypothesis. In the “thrifty allele” hypothesis, researchers theorize that food shortages amongst foraging populations led to a selection for E4 (Corbo and Scacchi 1999). In the energy expenditure hypothesis researchers theorize that higher activity levels amongst foragers led to higher energy costs, whereby E4 was selected for (Boone 2002).

Both hypotheses have been challenged and ultimately disproven. Studies demonstrate similar quantities of food between foraging and agricultural populations

(Benyshek et al. 2006). Other studies have also shown that while foraging societies have higher activity levels, their energy expenditure is similar to those of sedentary lifestyles (Pontzer et al. 2012, Raichlen et al. 2017).

Nonetheless, results from this study demonstrate that E3 is common in all populations, regardless of subsistence strategy. These findings, as well as, a low *p* value for E3 ($p = 0.465$, when adjusted for latitude and population stratification) support theories of E3 as having neutral selection in subsistence. In comparison, E4 is shown as having a significant *p* value ($p = 0.0177$, when adjusted for latitude and population stratification), supporting theories of positive selection with regards to subsistence.

Theories on latitude are more limited, but follow similar reasoning as theories regarding subsistence, whereby higher energy expenditure due to extreme climates has resulted in the selection of E4. In these findings, while an association exists between *APOE* and latitude, this association is curvilinear rather than showing a north / south gradient. Additionally, these findings show that climate, rather than latitude, results in the selection of *APOE* variants (Eisenberg et al. 2010). Thus, latitude becomes a correlation, rather than a causal effect, of *APOE* gene frequencies.

P values from this study show that E3 ($p = 1.193e\text{-}14$) increases with latitude, and E4 ($p = 1.1e\text{-}10$) decreases with latitude. In addition to being inconsistent with findings showing an opposite north / south gradient (Lucotte et al. 1997, Hu et al. 2011), these results are also inconsistent with findings showing a curvilinear association (Eisenberg et al. 2010).

However, taking into consideration the samples used for each study, differences in range of latitudes could account for inconsistent findings. A larger, globally-dispersed sample size in this study could have resulted in different north / south findings from previous studies, which focused on populations within a single country (Hu et al. 2011) or continent (Lucotte et al. 1997) – thereby, limiting the range in latitudes and resulting in a stronger association between *APOE* alleles and latitude.

Further, when considering the smaller sample size and limited latitude range in comparison to research showing curvilinear findings, results from this study may not be inconsistent with those findings. This is given that the sample data in this study ranges from low to mid latitude populations, with the highest latitude populations falling in the mid latitude range. This would have resulted in a E3 / high latitude, E4 / low latitude association, rather than in a curvilinear association – whereby E3 increases at mid-latitude (represented as high latitude in this study) and E4 increases at low latitude and high latitude (less represented in this study) (Eisenberg et al. 2010).

Given a larger sample, with more high latitude populations, it is possible that the results would have shown higher E4 frequencies at those latitudes – demonstrating a curvilinear finding consistent with the previous study. Such findings would have indicated support for the theory of latitude as a correlation and climate as a causal effect of *APOE* frequencies (Eisenberg et al. 2010).

6 CONCLUSION

6.1 Final Outcome

The study sought to examine SNPs, rs429358 and rs7412, as key determinants of the human *APOE* gene (Mahley et al. 2009, Villeneuve et al. 2014), as well as, analyze the associations between *APOE*, subsistence (Corbo and Scacchi 1999, Fullerton et al. 2000), and latitude (Eisenberg et al. 2010, Han et al. 2011). Based on prior research (Corbo and Scacchi 1999), the study hypothesized that E2 is not associated with subsistence or latitude, E3 is correlated with agriculture / post-agriculture and low latitude, and E4 is correlated with non-agriculture and high latitude (See TABLE 5.0.1).

In summary, the study demonstrates the significance of rs429359 and rs7412 in *APOE* variation, and shows a clear link between *APOE*, subsistence and latitude. However, the study offers limited support for the hypotheses, confirming the subsistence hypotheses and refuting the latitude hypotheses.

6.2 Strengths and Limitations

The main strength of this study was in its breadth and its aim to study the interaction between subsistence and latitude in the context of *APOE*. This study included sequences of 124 *APOE* SNPs, from 26 populations – each with a wide range of latitudes.

While biomedical research has focused primarily on SNPs, rs429358 and rs7412, this study sought to avoid deductive bias resulting from a reliance on

assumption, and to move beyond a focus on rs429358 and rs7412. Instead, the study aimed to follow an inductive process that would confirm / reject rs429358 and rs7412 as significant in *APOE* association. A second aim was to examine other polymorphic *APOE* SNPs by including a larger sample size of loci. The resulting dataset was comprised of 124 loci, including rs429358 and rs7412, as well as rs11542041 and other SNPs, which have been demonstrated to be significant in *APOE* and AD association studies (Masoodi et al. 2012). Including a larger dataset also allowed SNPs to be tested for linkage disequilibrium in a later phase of the study, thereby eliminating non-significant SNPs, as well as taking into consideration *APOE*'s role as multi-allelic biomarker.

The study also included a large sample size of populations, which were widely-dispersed throughout low and mid latitudes (while most populations were in the northern hemisphere, latitudes ranged from 6°N to 61°N, and also included latitudes of 0° and 12°S). Phenotype data was collected with the aim of providing a diverse sample size. Data was collected based on the theory that diversity in subsistence strategies and latitude would also lead to diversity in allele frequencies between populations. Having a large and diverse sample size allowed for a more accurate analysis of human genetic variability.

This analysis was furthered by the inclusion of subsistence as a covariate. Association results with subsistence and latitude as independent variables differed from those of subsistence and latitude as covariates. These findings demonstrate that human variability is complex and affected by several factors, rather than having a linear causal link with a single factor.

A major limitation of this study was using a binary analysis on subsistence. Upon initial review, populations showed a large range of subsistence strategies. When categorized into five groups (foraging, pastoralism, horticultural, agriculture, and industrialism), thirteen populations had a strategy of industrialism, and ten populations had a strategy of agriculture. The sample also included one horticulturalist population and two mixed subsistence populations.

However, when categorizing populations into two groups (non-agriculture and agriculture / post-agriculture), 23 of the 26 populations engaged in agricultural / post-agricultural subsistence strategies. Only three populations engaged in non-agricultural strategies. A binary analysis of subsistence eliminated variation between populations and further decreased the sample size of non-agricultural populations. Thus, results of an association between *APOE* and subsistence were limited by a small sample size (demonstrated in the higher *p* values / lower strength of association between *APOE* and subsistence compared to between *APOE* and latitude).

A second limitation of this study was that despite the inclusion of globally-dispersed populations, the sample data continued to have limited variability in latitude. Genotype data acquired from 1000 Genomes was comprised of mostly low and mid latitude populations, with only a few high latitude populations. This resulted in a stronger link between E3 and high latitude, and between E4 and low latitude. Such findings were inconsistent with previous studies (Lucotte et al. 1997, Eisenberg et al. 2010, Hu et al. 2011) and provided for a less accurate analysis.

Finally, this study failed to address other variables associated with either subsistence or latitude, that may also play a key role in shaping the distribution of *APOE* in human populations. Both subsistence and latitude bring selective challenges, including extreme climate, high cell turnover, and elevated metabolic rate – resulting in increased cholesterol needs that influence *APOE* selection (Eisenberg et al. 2010). Failing to measure and test for an association of these variables led to the inability to form theories as to the nature of the association between *APOE*, subsistence, and latitude. The study was able to discern that an association existed but not why that association existed, nor why specific variants were selected for over others. Thus, the study became limited in its depth and examination of the complexity of human genetic variation.

6.3 Further Research

Given that study results are contrary to those of previous studies, with regards to the direction of association between *APOE* and latitude, further research is needed to offer comparisons that would confirm / refute findings. Future studies should include a wider range of latitudes between populations, including low, mid, and high latitude populations.

Also, while the study has shown the complexity of human genetic variation and the modifying roles that environmental / cultural factors have on *APOE* selection, further exploration is needed for a more in-depth understand the nature of those associations. Comparative studies would allow for clarification on previous theories and the formation

of new theories as to the selective factors, accompanying subsistence and latitude, which influence *APOE* distribution.

APPENDIX – TABLE HEADINGS

A1	allele 1 / dominant allele / major allele
A2	allele 2 / minor allele
BETA	beta coefficient
CHR	chromosome
E(HET)	expected heterozygosity
FREQ	frequency
GENO	genotype
HAP	haplotype
LAT	latitude
MAF	minor allele frequency
OD	odds ratio
O(HET)	observed heterozygosity
P	*p* value
POP	population code
POS	position
SNP	single-nucleotide polymorphism
SUB	subsistence code

REFERENCES

1. Aucan, C., A. J. Walley, and A. V. S. Hill. 2012. Common apolipoprotein E polymorphisms and risk of clinical malaria in the Gambia. Journal of Medical Genetics 41(1):21–24.

2. Baschetti, Riccardo. 1999. Evolution, Cholesterol, and Low-Fat Diets. Circulation 99(1):264-167.

3. Beaglehole, R., M. A. Foulkes, I. A. Prior, and E. F. Eyles. 1980. Cholesterol and mortality in New Zealand Maoris. The British Medical Journal 280(6210):285-287.

4. Benyshek, D. C. and J. T. Watson. 2006. Exploring the thrifty genotype's food shortage assumptions: a cross-cultural comparison of ethnographic accounts of food security among foraging and agricultural societies. American Journal of Physical Anthropology 131(1):120-126.

5. Berkenstadt, M., S. Shiloh, G. Barkai, M. B. Katznelson, B. Goldman. 1999. Perceived personal control (PPC): a new concept in measuring outcome of genetic counseling. American Journal of Medical Genetics 82(1):53–59.

6. Bekris, Lynn M., Steven P. Millard, Nichole M. Galloway, Simona Vuletic, John J. Albers, Ge Li, Douglas R. Galasko, Charles DeCarli, Martin R. Farlow, Chris M. Clark, Joseph F. Quinn, Jeffrey A. Kaye, Gerard D. Schellenberg, Debby Tsuang, Elaine R. Peskind, and Chang-En Yu. 2008. Multiple SNPs Within and Surrounding the Apolipoprotein E Gene Influence Cerebrospinal Fluid Apolipoprotein E Protein Levels. Journal of Alzheimer's Disease 13(3):255 266.

7. Binford, Lewis R. Mobility, Housing, and Environment: A Comparative Study. 1990. Journal of Anthropological Research 46(2):119-152.

8. Bizarro, Alessandra, Davide Seripa, Adele Aciarri, Maria Giovanna Matera, Francesco Danilo Tiziano, Christina Brahe, and Carlo Masullo. 2009. The complex interaction between APOE promoter and AD: an Italian case–control study. European Journal of Genetics 17:938–945.

9. Boone, J. L. 2002. Subsistence strategies and early human population history: an evolutionary ecological perspective. World Archaeology 34(1):6-25.

10. Burns-Cox, C. J., Y. H. Chong, and R. Gillman. 1972. Risk factors and the absence of coronary heart disease in aborigines in West Malaysia. British Heart Journal 34(9):953-958.

11. Buzdugan, Laura, Markus Kalisch, Arcadi Navarro, Daniel Schunk, Ernst Fehr, and Peter Buhlmann. 2016. Assessing statistical significance in multivariable genome wide association analysis. Bioinformatics 32 (13):1990-2000.

12. Callaway, Ewen. 2017. Genome studies attract criticism. Nature 546:463.

13. Chen, G. Chi, Luke S. S. Guo, Robert L. Hamilton, Virginia Gordon, E. Glenn Richards, and John P. Kane. 1984. Circular Dichroic Spectra of Apolipoprotein E in Model Complexes and Cholesterol-Rich Lipoproteins: Lipid Contribution. Biochemistry 23:6530-6538.

14. Corbo, Rosa Maria and Renato Scacchi. 1999. Apoplipoprotein E (APOE) allele distribution in the world. Is APOE*4 a 'thrifty' allele? Annals of Human Genetics 63(4):301-310.

15. Cordain, L., S. B. Eaton, J. Brand Miller, N. Mann, and K. Hill 2002. The paradoxical nature of hunter-gatherer diets: meat-based, yet non-atherogenic. European Journal of Clinical Nutrition 56(1):S42-52.

16. Danecek, Petr, Adam Auton, Gonacolo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group. 2011. VCFtools, Version 4.2. United Kingdom. URL http://vcftools.sourceforge.net/.

17. Danecek, Petr, Adam Auton, Gonacolo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group. 2011. The Variant Call Format and VCFtools. Bioinformatics 27(15):2156-2158.

18. Di Rienzo, Anna and Richard R. Hudson. 2005. An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends in Genetics 21(11):596-601.

19. Eaton, S. B, Konner, and M. Shostak. Stone agers in the fast lane: chronic degenerative diseases in evolutionary perspective. 1988. American Journal of Medicine 84(4):739-49.

20. Eisenberg, Dan T. A., Christopher W. Kuzawa, and M. Geoffrey Hayes. 2010 Worldwide allele frequencies of the apolipoprotein E gene: climate, local adaptations, and evolutionary history. American Journal of Physical Anthropology 143(1):100-111.

21. Elhaik, Erin, T. Tatarinova, D. Chebotarev, I. S. Piras, Calo C. Maria, A. De Montis, M. Atzori, M. Marini, S. Tofanelli, P. Francalacci, L. Pagani, C. Tyler-Smith, Y. Xue, F. Cucca, T. G. Schurr, J. B. Gaieski, C. Melendez, M. G. Vilar, A. C. Owings, R. Gomez, R. Fujita, F. R. Santos, D. Comas, O. Balanovsky, E. Balanovska, P. Zalloua, H. Soodyall, R. Pitchappan, A. Ganeshprasad, M. Hammer, L. Matisoo-Smith, R. S. Wells, and the Genographic Consortium. 2013. Geographic population structure of analysis of worldwide human populations infers their biographical origins. Nature Communications 5(3513).

22. Ember, Carol R. 2014. “Hunter-Gatherers” in C. R. Ember, ed. Explaining Human Culture. Human Relations Area Files, http://hraf.yale.edu/ehc/summaries/hunter gatherers, accessed [September 18, 2017].

23. Fiori, G., F. Facchini, D. Pettener, A. Rimondi, N. Battistini, and G. Bedogni. 2000. Relationships between blood pressure, anthropometric characteristics and blood lipids in high- and low-altitude populations from Central Asia. Annals of Human Biology 27(1):19-28.

24. Fullerton, S. M., A. G. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor, J. H., Stengård, and C. F. Sing. 2000. Apolipoprotein E Variation at the Sequence Haplotype Level: Implications for the Origin and Maintenance of a Major Human Polymorphism. The American Journal of Human Genetics 67(4):881–900.

25. Gerdes, Lars Ulrik, Christian Klausen, Inger Sihm, and Ole Faergeman. 1992. Apolipoprotein E polymorphism in a Danish population compared to findings in 45 other study populations around the World. Genetic Epidemiology 9(3):155 167.

26. Gonzalez, Bianca. 2016. Apolipoprotein E receptor and cholesterol alterations are pronounced in Alzheimer’s Disease cortical synapses. Ph.D. dissertation, Department of Nursing, University of California, Los Angeles.

27. Han, ShuYi, YiHui Xu, MeiHua Gao, YunShan Wang, Jun Wang, YanYan Liu,Min Wang, and XiaoQian Zhang. 2016. Serum apolipoprotein E concentration and polymorphism influence serum lipid levels in Chinese Shandong Han population. Medicine 95(50):e5639-e5644.

28. Han, Summer. 2010. Reconsidering the asymptotic null distribution of likelihood ratio tests for genetic linkage in multivariate variance components models under complete pleiotropy. Biostatistics 11(2):226-241.

29. Hancock, Angela M., David B. Witonsky, Edvard Ehler, Gorka Alkorta-Aranburu, Cynthia Beall, Amha Gebremedhin, Rem Sukernik, Gerd Utermann, Jonathan Pritchard, Graham Coop, and Anna Di Rienzo. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. 2010.

Proceedings of the National Academy of Sciences of the United States of America 107(2):8924–8930.

30. Hart, Steven H. Patrick Duffy, Daniel J. Quest, Asif Hossain, Mike A. Meiners, and Jean Pierre Kocher. 2016. VCF-Miner: GUI-based application for mining variants and annotations stored in VCF files. Briefings in Bioinformatics 17(2):346-351.

31. Hill, W. G. and A. Mäki-Tanila. 2015. Expected influence of linkage disequilibrium on genetic variance caused by dominance and epistasis on quantitative traits. Journal of Animal Breeding and Genetics 132(2):176-186.

32. Hu, Peng, Y.H. Qin, Cheng Xue Jing, and Peng Fei Du. 2011. Does the geographical gradient of APOE4 allele exist in China? A systemic comparison among multiple Chinese populations. Molecular Biology 38(1):489-494.

33. Hu, Peng, Yuan Han Qin, Feng Yi Lei, Juan Pei, Bo Hu, and Ling Lu. 2011. Variable Frequencies of Apolipoprotein E Genotypes and Its Effect on Serum Lipids in the Guangxi Zhuang and Han Children. International Journal of Molecular Sciences 12:5604-5615.

34. Keyhole, Inc. 2001. Google Earth, Version 7.3.0.3832. Mountain View, CA. URL https://www.google.com/earth/.

35. Klasen, Jonas R., Elke Barbez, Lukas Meier, Nikolai Meinshausen, Peter Buhlmann, Maarten Koornneef, Wolfgang Busch, and Korbinian Schneeberger. 2016. A multi-marker association method for genome-wide association studies without the need for population structure correction.

36. Konner, M. and S. B. Eaton. Paleothic nutrition: twenty-five years later. 2010. American Society for Parenteral and Enteral Nutrition 25(6):594-602.

37. Lai, Lana Yin Hui. 2013. Association of Apolipoprotein E (APO E) Polymorphisms with the Prevalence of Metabolic Syndrome (METS): The National Heart, Lung, and Blood Institute Family Heart Study. M.A. dissertation, Department of Pharmacy, University Science of Malaysia.

38. Leonard, William R., J. Josh Snodgrass, and Marcia L. Robertson. 2010. “Evolutionary Perspectives on Fat Ingestion and Metabolism in Humans.” In Fat Detection: Taste, Texture, and Post Ingestive Effects, edited by J. P. Montmayeur and J. le Coutre. Boca Raton (FL): CRC Press/Taylor & Francis.

39. Li, Wen-Hsiung. 1997. Molecular Evolution. Sunderland, MA:Sinauer Associates, Inc., Publishers.

40. Limon-Sztencel, Anna, Beata S. Lipsak-Zietkiewicz, Magdalena Chmara, Wasag

Bartosz, Bidzan Leszek, Beata R. Godlewska, and Janusz Limon. 2016. The algorithm for Alzheimer risk assessment based on *APOE* promoter polymorphisms. Alzheimer's Research & Therapy 8:19

41. Lindeberg, S., P. Nilsson-Ehle, A. Terent, B. Vessby, and B. Schersten. 1994. Cardiovascular risk factors in a Melanesian population apparently free from stroke and ischaemic heart disease: the Kitava study. 1994. Journal of Internal Medicine 236(3):331-40.

42. Liu, Guodong, Xiang Liu, Pulin Yu, Qi Want, Hua Wang, Chenfang Li, Guangming Ye, Xiaoling Wu, and Chunling Tan. 2017. APOE gene polymorphism in long lived individuals from a central China Population. Scientific Reports 7(1):1-16.

43. Liu, Yifen. 2015. Clinical Study on Apolipoprotein E Distribution, Metabolism and Glycation. Ph.D. dissertation, Department of Medical and Human Sciences, University of Manchester.

44. Lucotte, Gerard, France Loirat, and Serge Hazout. 1997. Pattern of Gradient of Apolipoprotein E Allele *4 Frequencies in Western Europe. Human Biology 69(2):253-262.

45. Ma, C., Y. Zhang, X. Li, Y. Chen, J. Zhang, Z. Liu, K. Chen, and Z. Zhang. 2016. The TT allele of rs405509 synergizes with APOE ε4 in the impairment of cognition and its underlying default mode network in non-demented elderly. Current Alzheimer Research 13(6):708-717.

46. Marchini, Jonathan Marchini, Lon R Cardon, Michael S Phillips, and Peter Donnelly. 2004. The effects of human population structure on large genetic association studies. Nature Genetics 36(5):512-517.

47. Mahley Robert W., Karl H. Weisgraber, and Yadong Huang. 2009. Apolipoprotein E: structure determines function, from atherosclerosis to Alzheimer's disease to AIDS. The Journal of Lipid Research 50(SUPPL):S183-S188.

48. Masoodi, Tariq Ahmad, Sulaiman A. Al Shammari, May N. Al-Muammar, and Adel A. Alhamdan. 2012. Screening and Evaluation of Deleterious SNPs in APOE Gene of Alzheimer's Disease. Neurology Research International (2012):480609 480617.

49. Martin, George M. 1999. *APOE* alleles and lipophylic pathogens. Neurobiology of Aging 20:441-443.

50. Morris, Nathan J., Robert Elston, and Catherine M. Stein. Calculating Asymptotic Significance Levels of the Constrained Likelihood Ratio Test with Application to

Multivariate Genetic Linkage Analysis. 2009. Statistical applications in genetics and molecular biology 8(1).

51. NCBI Resource Coordinators. 2012. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research41:D8-D20.

52. Nordborg, M. and S. Tavare. 2002. Linkage disequilibrium: what history has to tell us. Trends in Genetics 8(2):83-90.

53. Nickerson, Deborah A., Scott L. Taylor, Stephanie M. Fullerton, Kenneth M. Weiss, Andrew G. Clark, Jari H. Stengard, Veikko Salomaa, Eric Boerwinkle, and Charles F. Sing. 2000. Sequence Diversity and Large-Scale Typing of SNPs in the Human Apolipoprotein E Gene. Genome Research 10:1532–1545.

54. Pickerell, Joseph K., Nick Patterson, Chiara Barbieri, Falko Berthold, Linda Gerlach, Tom Guldemann, Blesswell Kure, Sununguko Wata Mpoloka, Hirosi Nakagawa, Christfried Naumann, Mark Lipson, Po-Ru Loh, Joseph Lachance, Joanna Mountain, Carlos D. Bustamante, Bonnie Berger, Sarah A. Tishkoff, Brenna M. Henn, Mark Stoneking, David Reich, and Brigitte Pakendorf. 2012. The genetic prehistory of southern Africa. Nature Communications 3(1):1-6.

55. Pontzer, Herman. 2017. The Exercise Paradox. Scientific American 316:26-31.

56. Pontzer, H., D. A. Raichlen, B. M. Wood, M. Emery Thompson, S. B. Racette, A. Z. Mabulla, F. W. Marlowe. 2015. Energy expenditure and activity among Hazda hunter-gatherers. American Journal of Human Biology: the Official Journal of the Human Biology Council 27(5):628-637.

57. Posse de Chaves, Elena and Vasanthy Narayanaswami. 2008. Apolipoprotein E and cholesterol in aging and disease in the brain. Future Lipidol 3(5):505-530.

58. Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mary J. Daly, and Pak C. Sham. 2007. PLINK: Whole genome association analysis toolset, Version 1.07. Boston, MA. URL http://zzz.bwh.harvard.edu/plink/.

59. Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mary J. Daly, and Pak C. Sham. 2007. PLINK: A Tool Set for Genome Wide Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81(3):559-575.

60. Purcell, Shaun and S. Sham. Properties of structured association approaches to detecting population stratification. 2004. Human Heredity 58(2):98-107.

61. Qin, Huaizhen and Xiaofeng Zhu. 2012. Allowing for Population Stratification in Association Analysis. Methods in Molecular Biology 850:399-409.

62. Quaye, Lydia, Nicos Nicolaou, Scott Shane, and Massimo Mangino. 2012. A Discovery Genome-Wide Association Study of Entrepreneurship. International Journal of Development Science 6:127-135.

63. R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Version 3.3.3. Vienna, Austria. URL http://www.R project.org/.

64. Raichlen, David A. and Gene E. Alexander. 2014. Exercise, APOE genotype, and the evolution of the human lifespan. Trends Neuroscience 37(5):247-255.

65. Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumijan, S. F. Farhadian, R. Ward, and E. S. Lander. Linkage disequilibrium in the Human Genome. 2001. Nature 411(6834):199-204.

66. Rogers, Alan R. and Chad Huff. 2009. Linkage Disequilibrium Between Loci with Unknown Phase. Genetics 182(3):839-844.

67. Rosenthal, Samantha L., Michael Barmada, Xingbin Wang, F. Yesmin Demirci, and 2M. Ilyas Kamboh. 2014. Connecting the Dots: Potential of Data Integration to Identify Regulatory SNPs in Late-Onset Alzheimer's Disease GWAS Findings. Public Library of Science One 9(4):e95152-e95162.

68. Saito, Hiroyuki, Padmaja Dhanasekaran, Faye Baldwin, Karl H. Weisgraber, Sissel Lund-Katz, and Michael C. Phillips. 2001. Lipid Binding-induced Conformational Change in Human Apolipoprotein E: Evidence for Two Lipid-Bound States on Spherical Particles. The Journal of Biological Chemistry 276:40949-40954.

69. Schaid, Daniel J., Jason P. Sinnwell, and Gregory D. Jenkins. Regression Modeling of Allele Frequencies and Testing Hardy-Weinberg Equilibrium. 2012. Human Heredity 74(2):71–82.

70. Schwartz, Colin John, Allan J. Day, J. Andrew Peters, and John Royle Casley Smith. Serum cholesterol and phospholipid levels of Australian aborigines. 1957. The Australian Journal of Experimental Biology and Medical Science 35(5):449-56.

71. Shore, V. G. and Shore B. 1973. Heterogeneity of human plasma very low density lipoproteins. Separation of species differing in protein components. Biochemistry 12:502–507.

72. Singh, Puneetpal, Monica Singh, and Sarabjit Mastana. 2006. APOE distributions in world populations with new data from India and the UK. Annals of Human Biology 33(3):279-308.

73. Slatkin, M. 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Genetics Nature Rev Genet 9(6):477-485.

74. Stevens, Eric L., Greg Heckenberg, Elisha D. O. Robertson, Joseph D. Baugher, Thomas J. Downey, and Jonathan Pevsner. 2011. Inference of Relationships in Population Data Using Identity-by-Descent and Identity-by-State. Public Library of Science Genetics 7(9):e1002287-e1002287.

75. Tabas, Ira. 2002. Cholesterol in health and disease. The Journal of Clinical Investigation 110(5):583-590.

76. Templeton, Alan R. 2006. Population Genetics and Microevolutionary Theory. Hoboken, NJ:John Wiley & Sons, Inc.

77. Thomas, Duncan C. and John S. Witte. 2002. Point: Population Stratification: A Problem for Case-Control Studies of Candidate-Gene Associations? Cancer Epidemiology Biomarkers & Prevention 11(6):505-512.

78. Trumble, Benjamin C., Jonathan Stieglitz, Aaron D. Blackwell, Hooman Allayee, Bret Beheim, Caleb E. Finch, Michael Gurven, and Hillard Kaplan. 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.

79. Utermann, Gerd, Ulrich Langenbeck, Ulrike Beisiegel, and Wilfried Weber. 1980. Genetics of the apolipoprotein E system in man. The American Journal of Human Genetics 32:339-347.

80. Villeneuve, Sylvia, Diane Brisson, Natalie L. Marchant, and Daniel Gaudet. 2014. The potential applications of Apolipoprotein E in personalized medicine. Frontiers in Aging Neuroscience 6:154-165.

81. Visscher, P. M. 2006. A note on the asymptotic distribution of likelihood ratio tests to test variance components. Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies 9(4):490-495.

82. Walker, A. R. Cholesterol and mortality rates. 1980. The British Journal of Medicine. 280(6227):1320.

83. Wang, Tao. 2011. On coding genotypes for genetic markers with multiple alleles in genetic association study of quantitative traits. BioMed Central Genetics 12:82 103.

84. Wozniak, M. A., E. B. Faragher, J. A. Todd, K. A. Koram, E. M. Riley, and R. F. Itzhaki. 2003. Does apolipoprotein E polymorphism influence susceptibility to malaria? Journal of Medical Genetics 40:348-351.

85. Yu, Chang-En, Howard Seltman, Elaine R. Peskind, Nichole Galloway, Peter X. Zhou, Elizabeth Rosenthal, Ellen M. Wijsman, Debby W. Tsuang, Bernie Devlin, and Gerard D. Shellenberg. 2007. Comprehensive analysis of *APOE* and selected proximate markers for late-onset Alzheimer's Disease: patterns of linkage disequilibrium and disease/marker association. Genomics 89(6):655-665.

86. Zang, Yong, Wing Kam Fung, and Gang Zheng. 2010. Simple Algorithms to Calculate Asymptotic Null Distributions of Robust Tests in Case-Control Genetic Association Studies in R. Journal of Statistical Software 33(8):1-24.

87. Zhan, Xiu-Hiu, Guang-Cai Zha, Ji-Wei Jiao, Li-Ye Yang, Xiao-Fen Zhan, Jiang Tao Chen, Dong-De Xie, Urbano Monsuy Eyi, Rocio Apicante Matesa Maximo Miko Ondo Obono, Carlos Sala Ehapo, Er-Jia Wei, Yu-Zhong Zheng, Hui Yang, and Min Lin. 2015. Rapid identification of apolipoprotein E genotypes by high resolution melting analysis in Chinese Han and African Fang populations. Experimental and Therapeutic Medicine 9(2):469-475.

88. Zhang, Hong-Liang, Yi Yang, and Jiang Wu. 2010. Can prevalence of apolipoprotein E epsilon 4 allele explain the geographical variation of coronary heart disease mortality rates in Western Europe? European Journal of Epidemiology 25(12):897-910.