TESTING NATURAL SELECTION ON POLYGENIC TRAIT-ASSOCIATED ALLELES IN ANATOLIA USING NEOLITHIC AND PRESENT-DAY HUMAN GENOMES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

EVRİM FER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
BIOINFORMATICS

JULY 2019

Approval of the thesis:

TESTING NATURAL SELECTION ON POLYGENIC TRAIT-ASSOCIATED ALLELES IN ANATOLIA USING NEOLITHIC AND PRESENT-DAY HUMAN GENOMES

Submitted by EVRİM FER in partial fulfillment of the requirements for the degree of **Master of Science in the Department of Bioinformatics, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics, METU** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics, METU** _____

Assoc. Prof. Dr. Mehmet Somel
Supervisor, **Biological Sciences Dept., METU** _____

Examining Committee Members:

Prof. Dr. Ergi Deniz Özsoy
Biological Sciences Dept., Hacettepe University _____

Assoc. Prof. Dr. Mehmet Somel
Biological Science Dept., METU _____

Asst. Prof. Dr. Aybar Can Acar
Health Informatics Dept., METU _____

Asst. Prof. Dr. Füsun Özer
Anthropology Dept., Hacettepe University _____

Assoc. Prof. Dr. Nurcan Tunçbağ
Health Informatics Dept., METU _____

Date: **29.07.2019**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : EVRİM FER

Signature : _____

ABSTRACT

TESTING NATURAL SELECTION ON POLYGENIC TRAIT-ASSOCIATED ALLELES IN ANATOLIA USING NEOLITHIC AND PRESENT-DAY HUMAN GENOMES

Fer, Evrim

MSc., Department of Bioinformatics

Supervisor: Assoc. Prof. Dr. Mehmet Somel

July 2019, 63 pages

The Neolithic transition, which started approximately 10,000 year ago in west Eurasia and introduced sedentary life style and food production, led to major shifts in the human diet. Previous studies have reported strong selection signals on genes related to processing of plant-based diets (Buckley et al., 2017; Harris et al., 2019) or the consumption of dairy products (Schlebusch et al., 2013). With the advent of archeogenomics studies, genetic signatures of such adaptations have also been supported using DNA data from ancient populations (Mathieson et al., 2015). In this study, polygenic adaptations in Anatolia after the Neolithic transition were investigated by comparing Neolithic and modern-day genome sequence data. First, we chose 40 mainly polygenic traits previously subject to selection studies. For 6651 single nucleotide polymorphisms (SNPs) associated with these traits, we compared the genetic distance between Neolithic Anatolian (n=36) and present-day Anatolian (n=16) individuals, measured using the F_{ST} statistic, with SNPs in evolutionary neutral regions. Then, frequency changes of alleles that elevating phenotypes were studied, to test for a common direction of allele-frequency change affecting these traits. Finally, a population branch statistic (*PBS*) approach was applied to detect adaptation signals specific to the modern-day Anatolia in comparison to Neolithic Anatolia and an outgroup population. We found that the frequency of alleles related to GWAS traits broadly linked to lipid metabolism to be more differentiated between Neolithic and present-day Anatolia, than neutrally expected. Directionality analyses also suggested that such traits might have been driven by selection. Consistently, the genes showing the highest differentiation along the modern Anatolia branch in the *PBS* analysis were frequently associated with lipid metabolism. Our results imply that lipid metabolism-related traits may have been subject to selective pressures in the last 10,000 years.

Keywords: Ancient DNA, Neolithic Transition, Polygenic Adaptation, F_{ST}, Population Branch Statistic

ÖZ

ANADOLU'DA NEOLİTİK VE GÜNÜMÜZ İNSAN GENOMLARININ KARŞILAŞTIRILMASIYLA POLİGENİK ÖZELLİKLERLE İLİŞKİLİ ALELLER ÜZERİNDE DOĞAL SEÇİLİM TESPİTİ

Fer, Evrim

Yüksek Lisans, Biyoenformatik Bölümü

Tez Yöneticisi: Doç. Dr. Mehmet Somel

Temmuz 2019, 63 sayfa

Yaklaşık 10,000 yıl önce batı Avrasya'da başlayan Neolitik Geçiş de yerleşik yaşam şeklini ve insan beslenmesindeki büyük değişimlerden birini insan hayatına katmıştır. Birçok çalışmada, modern popülasyonlar üzerinden bitkisel beslenmeye dayalı (Buckley vd., 2017; Harris vd., 2019), hayvan evcilleştirmesi sonucunda süt ürünlerinin kullanımına dayalı (Schlebusch vd., 2013) genlerde adaptasyon tespit edilmiştir. Gelişen arkeogenomik çalışmaları sayesinde, bu adaptasyonlar antik DNA analizleriyle desteklenmektedir (Mathieson vd., 2015). Bu çalışmada, Neolitik sonrası Anadolu'daki poligenik adaptasyonlar, Neolitik ve modern DNA karşılaştırmasıyla incelenmiştir. İlk olarak, beslenme, bağışıklık ve başka kompleks özellikleri içeren 40 poligenik özellik belirlenmiştir. Bu özelliklerle ilişkili 6651 tek nükleotit polimorfizmi (TNP) için Neolitik (n=36) ve günümüz (n=16) Anadolu bireylerinin arasında genetik uzaklıklar F_{ST} istatistiği kullanılarak hesaplanmış ve genomda seçilimden direkt etkilenmeyen bölgelerle karşılaştırılmıştır. Daha sonra, bu özelliklere artırıcı etkisi bulunan risk alelleri belirlenmiş ve frekans değişimleri incelenerek ortak bir doğrultunun varlığına bakılmıştır. Son olarak, Neolitik ve modern Anadolu ve bir dış popülasyon kullanılarak popülasyon dal istatistiği yöntemi ile genom boyu seçilim analizi yapılmış ve modern Anadolu dalına özgü seçilim sinyalleri tespit edilmeye çalışılmıştır. Sonuç olarak, yağ metabolizmasıyla ilişkili alel frekanslarının nötral bölgelere göre önemli ölçüde değiştiği görülmüş ve doğrultu analizi de bu özelliklerde seçilime bağlı bir değişim olabileceğini göstermiştir. Ayrıca genom-boyu analizinde modern Anadolu dalında yüksek farklılık gösteren bölgelerin lipid metabolizmasıyla ilgili olabileceği bulunmuştur.

Anahtar Sözcükler: Antik DNA, Neolitik Geçiş, Poligenik Adaptasyon, F_{ST}, Popülasyon Dal İstatistiği

To my family

ACKNOWLEDGMENTS

First, I would like to express my gratitude to my advisor Assoc. Prof. Dr. Mehmet Somel for his endless support and continuous patience. It was a big chance to start my academic career under his guidance since he always tries his best to help and enlightens the problems throughout the work with his never ending energy.

Also, I would like to thank my thesis committee members: Prof. Dr. Ergi Deniz Özsoy, Asst. Prof. Dr. Aybar Can Acar, Asst. Prof. Dr. Füsun Özer and Assoc. Prof. Dr. Nurcan Tunçbağ, for accepting to be in my committee and sharing their time to evaluate my study despite their strict schedules. Also, their precious ideas and contributions to improve the study were really appreciated. I also want to thank Asst. Prof. Dr. Can Alkan and his master student Ezgi Ebren for sharing data of Turkish Genome Project which has an important part in this study. I would like to thank Assoc. Prof. Dr. Ömer Gökçümen and Asst. Prof. Dr. Pavlos Pavlidis for sharing their knowledges on evolutionary biology and population genetics and their contributions.

I would also like to thank all my friends and colleagues from Informatics Institute who are Elif Bozlak, Cansu Demirel, Cansu Dinçer, Meriç Kınalı, Muazzez Çelebi Çınar, Ali Çınar, Gökçe Senger and Alperen Taciroğlu. They made the best master years for me with their support, kindness and sharing their ideas about life, science and any other stuff in both intellectual and humorous ways. I am really grateful to know them and spend these years with their friendships. Also, I would like to thank my dear colleagues from Compevo who are Erinç Yurtman, Reyhan Yaka and Zeliha Gözde Turan for sharing their knowledges on this field and their cheerful personalities even in the most tiresome moments. They always have a patience to cheer me up although I could not respond them as energetic as they are. I would like to thank Dilek Koptekin for her endless patience and supportive helps throughout the study, and also Ahmet Yetkin Alıcı, Hamit İzgi, Meriç Erdolu to share their time to give suggestions on the work and all other group members who listen and give ideas to make a more sophisticated study in each regular group meetings. I also send my special thanks to our department secretary Hakan Güler for his patience and helps that making this hard processes a little bit easier.

I owe special thanks to my family: to my mother Şenay Fer for always being by my side and giving her all to make everything possible for me to focus only on my education; and my father Kadri Metin Fer who always believes in me and supports me even though he is at distant, and my brother Yetkin Fer who is always interested in what I study and encourages me and especially my sister İstem Fer who always shares her academic experiences to enlighten my career path while giving thoughtful advices and encouragements.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

aDNA	Ancient DNA
BAM	Binary Alignment Map
BCE	Before Common Era
BMI	Body Mass Index
BP	Before Present
bp	Base Pair
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry
cM	Centi Morgan
CNV	Copy Number Variation
DNA	Deoxyribo Nucleic Acid
EST	Expressed Sequence Tags
EHH	Extented Haplotype Homozygosity
F_{ST}	Fixation Index
gVCF	Genomic Variant Call Format
GWAS	Genome Wide Association Studies
HKA	Hudson-Kreitman-Aguadé
iHS	Integrated Haplotype Score
IL	Interleukin
Kbp	Kilo Base Pair
LWK	Luhya in Webuye, Kenya
MHC	Major Histocompatability Complex
MK	McDonald-Kreitman
NRE	Neutral Region Explorer
OMIM	Online Mendelian Inheritance in Man
PBS	Population Branch Statistic
SNP	Single Nucleotide Polymorphism
TSI	Toscani in Italia
TSV	Tab Separated Values
VCF	Variant Call Format
YRI	Yoruba in Ibadan, Nigeria
XP-EHH	Cross Population Extended Haplotype Homozygosity

CHAPTER 1

1.INTRODUCTION

After the out-of-Africa dispersal starting about 60-100,000 years ago (Demenocal & Stringer, 2016), human populations started to colonize various environments in the new continents they encountered, from cold climates to deserts or low altitudes to high mountains. Despite of all the difficulties that they encountered, humans could adapt and occupied even extreme environments (Jeong & Di Rienzo, 2014). In addition to changes in the physical environment, cultural and dietary transitions have also occurred in time. In human history, there are two major diet-shifts, which are the farming diets first starting in the Neolithic Period (~10,500 years ago before present in west Eurasia) and processed diets starting with industrialization approximately around 1850 CE (Adler et al., 2013). During the Neolithic, humans adapted to farming, stock breeding and sedentary life, which are thought to have led to diverse physiological and morphological changes, including vitamin deficiencies and changes in tooth and facial bone morphologies, as well as the introduction of new pathogens due to close intimacy to animals and the easy spread of pathogens in complex large communities (Armelagos, Goodman, & Jacobs, 1991; Larsen, 2006; Latham, 2013).

In the last 15 years, we saw major developments in genomics technologies that today enable the generation and analysis of high-throughput data from many individuals at the same time. This advanced technology is used in a wide range of fields, including disease diagnosis (Lefterova, Suarez, Banaei, & Pinsky, 2015), for the detection of variant association between phenotypic traits and diseases (Luo, Boerwinkle, & Xiong, 2011), or in population genetics to explore indirect signatures of evolutionary mechanisms (Fumagalli, Vieira, Linderoth, & Nielsen, 2014). Using high-throughput sequences of modern populations and statistical tests, previous studies have been compiling a comprehensive list of selection signatures involving human adaptive traits. For instance, there are notable studies identifying signatures of recent selection, such as lactose tolerance in adulthood in Europe that was likely a response to the widespread consumption of milk (Schlebusch et al., 2013); convergent evolution of light

skin pigmentation in Europeans (Wilde et al., 2014) and East Asians (Deng & Xu, 2018); high altitude adaptation in Andean (Bigham et al., 2010), Ethiopian (Alkorta-Aranburu et al., 2012) and Tibetian (Xu et al., 2011; Yi et al., 2010) highlanders.

Ancient DNA, meanwhile, adds another perspective to human population genetic studies, by revealing demographic histories of past populations, admixture of modern human populations and archaic hominins, describing relatedness between the individuals found at the same excavation sites, as well as shedding light on complex selective processes (Slatkin & Racimo, 2016). With the archeogenomics revolution, adaptations to new conditions can now be detected in various populations using ancient DNA. For instance, the genotypes of derived alleles for light skin pigmentation and lactose persistence were observed in ancient European populations and an increase in their frequencies was shown by comparison of pre- and post-Neolithic periods' populations (Allentoft et al., 2015; Gamba et al., 2014; Mathieson et al., 2015; Olalde et al., 2015). In addition, genome-wide selection scan analysis using ancient genomic data also revealed that alleles associated with fatty acid metabolism (Buckley et al., 2017; Harris et al., 2019; Mathieson et al., 2015; S. Mathieson & Mathieson, 2018), immunity gene clusters including interleukin (Gelabert, Olalde, De-Dios, & Civit, 2017) and major histocompatibility complex (MHC) (Mathieson et al., 2015), and vitamin D production (Mathieson et al., 2015) to be subject to positive selection. However, those studies mainly focused on West European or East Asian populations. Comprehensive selection analyses to detect adaptation and adaptive phenotypic changes are still missing for most world populations, including Anatolian (or generally, west Asian) populations.

The main aim of this study was detection of adaptive phenotypic changes in Anatolia over the last ~10,000 years. Specifically, I aimed to investigate polygenic adaptation signatures in the genomes of Anatolian populations taking advantage of ancient DNA data. For this goal, variants associated with 40 phenotypic traits were collected and their alleles determined in publicly available Neolithic Anatolian and modern Turkish individuals' genomes. For these variants, the genetic distance between ancient and modern Anatolia was calculated using the F_{ST} statistic and significantly differentiated traits were determined in comparison to SNPs in neutral regions. Then, for each trait, the frequencies of alleles with an increasing effect on that trait were compared between ancient and modern samples to test if there is a common direction of frequency change for the alleles with the same effect. Finally, another F_{ST}-based method called the population branch statistic (*PBS*) was used to detect specific genes that might be have differentiated on the modern Anatolia branch, indicative of positive selection.

In Chapter 2, I first explain the mechanisms that change allele frequencies in populations. I go on to review current approaches for selection detection and provide examples from human genome analyses. Lastly, I discuss the usage of variant-trait associations estimated in genome-wide association studies (GWAS).

In Chapter 3, the methodology of the present study is outlined starting from sample collection and trait-associated SNP collection, direction analysis of the alleles, and selection analysis using F_{ST}-based methods.

In Chapter 4, I present my results while discussing their contribution to selection studies in Anatolia. The significantly changed traits according to both F_{ST} and directionality analysis results are mostly associated with lipid metabolism. *PBS* analysis also reveals that there is a selection signal on chromosome 12 where the genes associated with cholesterol are located.

In Chapter 5, I discuss my results and how much they are consistent with the literature. The limitations and possible further analyses are also mentioned in this chapter.

CHAPTER 2

2.LITERATURE REVIEW

2.1. Main Mechanisms Driving Allele Frequency Changes

Allele frequency is the quantity measuring how much an allele is common or rare in a population (Futuyma, 2013). In small-scale evolution processes, changes in allele frequencies within and between the populations are mainly caused by the three major evolutionary forces, which are genetic drift, gene flow and natural selection, where mutations are the source of new alleles (Andrews, 2010). As a null hypothesis, the Hardy-Weinberg model describes how genotype frequencies should appear given allele frequencies, when the drivers of evolution do not act on populations (Bergstrom & Dugatkin, 2012). If the observed genotype frequencies violate the Hardy-Weinberg model and/or when allele frequencies are directly found to change over time, the problem becomes inferring which evolutionary forces may have contributed to this violation.

Genetic drift. Neutral drift is one of these processes causing random fluctuations in allele frequencies in a population, representing a sampling effect and occurring independent of natural selection (Bergstrom & Dugatkin, 2012). As the allele frequencies fluctuate randomly, some alleles may be fixed (reach frequency 100%) and others lost (reach frequency 0%) over time, a process that leads to decrease in heterozygosity in the population (Bergstrom & Dugatkin, 2012; Nielsen & Slatkin, 2013). Thus, the frequency of a new allele can increase or even become fixed in a population without natural selection or migration taking effect (Jobling, Hurles, & Tyler-Smith, 2004). As genetic drift occurs stochastically, allele frequency change will follow different paths in different populations. For example, imagine two sister populations, recently isolated, that both carry *A* and *a* alleles for a locus at some point. After certain number of

generations, the frequency of *A* allele may go to fixation in one population if the individuals with the *A* allele produce more offspring due to random factors independent of the allele itself and *A* fixes, while the individuals with the *a* allele reproduce more in the other population leading to fixation of the *a* allele in this latter population. As a result, these sister populations will genetically diverge from each other due to random allele frequency changes (Bergstrom & Dugatkin, 2012; Nielsen & Slatkin, 2013). The magnitude of genetic drift varies according to population size (Jobling et al., 2004). In small populations, an allele can be fixed or lost very rapidly, while alleles may persist at similar frequencies for a long time in large populations as illustrated in Figure 2.1.

Kimura's neutral theory of molecular evolution (Kimura, 1983) proposes that majority of the genomic variation is shaped by genetic drift and positive selection is playing a minor role in molecular evolution and in shaping overall molecular diversity (Booker, Jackson, & Keightley, 2017; Nielsen & Slatkin, 2013).

Today, the neutral theory maintains a central role in molecular evolution research, and neutral models representing drift are the main null hypothesis in selection studies (Nielsen & Slatkin, 2013). According to this approach, if the observed data does not obey to neutral hypothesis, which can be tested using a number of statistical tests, only then can more complicated demographic processes (e.g. admixture) or natural selection be considered a plausible alternative (Booker et al., 2017; Nielsen & Slatkin, 2013; Vitti, Grossman, & Sabeti, 2013).

Gene flow. New alleles can be introduced into a population by migration from another population carrying different allele frequencies. Gene flow only occurs when migrants reproduce and contribute to the gene pool of the new population (Jobling et al., 2004). Gene flow is capable of changing allele frequencies of sub-populations leading to a loss of differentiation among populations due to increased genomic similarity (Bergstrom & Dugatkin, 2012; Jobling et al., 2004; Nielsen & Slatkin, 2013).

Notably, genetic drift and gene flow are processes that affect all loci in the genome in roughly the same way, whereas natural selection will affect different loci more or less independently, depending on their fitness effect (Futuyma, 2013).

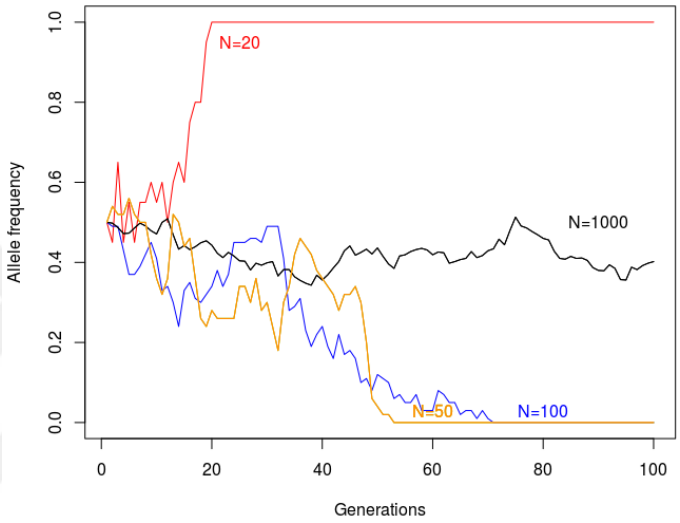

Figure 2.1. The result of simulations showing neutral evolution at biallelic loci in populations of different population sizes, with allele frequencies each starting from 0.5 and evolving under the effect of genetic drift.

Natural selection. In natural selection models, certain genotypes that have higher or lower chances of survival and reproduction, or fitness, are more frequently or less frequently passed to the next generations and increase or decrease in the population over time (Vitti et al., 2013). Mutations that create alleles with harmful effects on the organism and reduce fitness are subject to negative selection, also called as purifying selection (Thiltgen, dos Reis, & Goldstein, 2017). On the other hand, alleles that increase fitness undergo positive selection (Vitti et al., 2013). Thus, the frequency of an advantageous allele is expected to eventually go to fixation in a population due to positive directional selection, all else being equal (Vitti et al., 2013). These types of natural selection can act on 'de novo' as well as standing variants, and are expected to decrease variation in the population over time due to selective sweep effect which is the genetic hitchhiking of genomic background near to a strongly selected allele (Nielsen & Slatkin, 2013).

A third general type of natural selection is called balancing selection, which differs from positive and negative selection by maintaining both alleles in the population in the long term. In one such model, the heterozygote genotype has higher fitness than both homozygotes, a process called over-dominant selection,

causing both alleles to remain in the population (Jobling et al., 2004). Another type of balancing selection that preserves variation is negative frequency-dependent selection, which arises when the rare genotype has higher fitness. Through frequency dependent selection, an intermediate allele frequency will be maintained in the long term, achieved when fitness becomes high at low genotype frequencies, and vice versa (Vitti et al., 2013).

One of main interests of molecular population genetics and evolutionary biology is to reveal which type of evolutionary mechanism acts on the genome and results with allele frequency change (Hancock & Di Rienzo, 2008). Numerous statistical methods have been developed to differentiate those three mechanisms from each other (Thiltgen et al., 2017). Most of these are specialized in detecting positive selection signatures, since these are the basis and evidence of adaptation; moreover positive directional selection increases intra and interspecific population variability when the populations are adapted to different selection regimes (Nielsen, 2005).

2.2. Widely Used Methods to Detect Selection in Genomic Data

2.2.1. Detection of selection in macroevolutionary events

Approaches for detecting selection from genomic data can be divided into two modes, selection at the macroevolutionary scale and at the microevolutionary scale. Macroevolutionary changes are investigated between species and higher levels of taxonomic classification, rather than within species (Vitti et al., 2013). The methods developed for studying species divergence include *Ka/Ks* ratio test for the genic regions (Hurst, 2002), the McDonald-Kreitman (MK) test (McDonald & Kreitman, 1991) and the Hudson-Kreitman-Agaudé (HKA) test (Hudson, Kreitman, & Aguadé, 1987).

The *Ka/Ks* ratio, also known as the d_N/d_S ratio, compares the nonsynonymous substitution rate (d_N, i.e. substitutions that code for a different amino acid) and synonymous substitution rate (d_S, i.e. substitutions that code for the same amino acid) from aligned genomic regions of different species (Nielsen & Slatkin, 2013; Vitti et al., 2013a). Note that synonymous substitutions here are assumed to be neutrally evolving. Under neutral conditions, where nonsynonymous substitutions would also be neutrally evolving, this ratio is supposed to be equal to 1 ($d_N/d_S=1$) while natural selection violates this expectation (Kryazhimskiy & Plotkin, 2008). A d_N/d_S ratio greater than 1 ($d_N/d_S>1$) means that for the peptide in question, the nonsynonymous substitution rate was greater than the synonymous rate, even though the former nonsynonymous substitutions change the amino acid sequence and protein functionality. An interpretation of this

pattern is that such new nonsynonymous alleles may have been subject to positive selection due to their advantageous effect and fixed faster than neutral synonymous substitutions (Hurst, 2002). In contrast, a ratio lower than 1 (d_N/d_S <1) indicates negative selection, since protein changing mutations are less likely to remain in the genome and the protein sequence is kept as it is (Hurst, 2002). Unfortunately, this method is only able to detect selection on coding regions and recurrent selection events on a locus; it has no power to detect selection on a single mutation or in a noncoding region (Nielsen & Slatkin, 2013).

Another method is the McDonald-Kreitman (MK) test, which calculates the same nonsynonymous rate to synonymous rate ratio for both between and within species differences (Vitti et al., 2013). Under neutrality, the ratio of polymorphic and fixed nonsynonymous sites are expected to be equal to that ratio at synonymous sites (Egea, Casillas, & Barbadilla, 2008; Nielsen & Slatkin, 2013). If the ratio at fixed sites exceeds the ratio at polymorphic sites, this shows a positive selection signal, while the opposite pattern is evidence for negative selection (Egea et al., 2008; Nielsen & Slatkin, 2013). For example, Le Corre et al. (2002) compared the fixed and polymorphic sites for *FRI* gene that is responsible for flowering time between *Arabidopsis thaliana* and *Arabidopsis lyrata* species and showed that first exon of this gene has high heterogeneity between the species and might be under a functional selection (Le Corre, Roux, & Reboud, 2002). One potential disadvantage of MK test might be that the number of segregating (i.e. polymorphic) sites depends on the sample size. For example, a slightly deleterious allele can be counted as fixed in a small sample while it contributes as a polymorphism in a large sample (Walsh & Lynch, 2018).

Similar to the MK test, the HKA test also compares polymorphic and fixed positions across two loci for both intra and interspecies divergence (Vitti et al., 2013). Unlike the MK test, calculation of the expected number of fixed and segregating sites depends on the effective population size, sample size and divergence time (Nielsen & Slatkin, 2013). As an example of the HKA method, Nachman and Crowell (2000) tested the statistical significance of the genetic variation in two introns of the *Duchenne muscular dystrophy* (*Dmd*) gene in human populations in comparison to chimpanzee. Using the HKA test, they showed that there is a decrease in variability around intron 7 that could be due to a selective sweep (Nachman & Crowell, 2000).

2.2.2. Detection of selection in microevolutionary events

Selection analyses at microevolutionary scales mainly focus on variations within populations. Since a beneficial allele may reach high frequencies and even

fixation under positive selection in a short time, the methods to detect selection mostly investigate frequency changes of alleles but can also be linkage disequilibrium-based (Vitti et al., 2013). Some methods use the site frequency spectrum (SFS) (Nielsen, 2005) or specific statistics such as Tajima's D (Tajima, 1989), F_{ST} (Weir & Cockerham, 1984; Wright, 1965), and the population branch statistic (Yi et al., 2010); these statistics rely on the distribution of the allele frequencies in the population (Vitti et al., 2013). SFS-based methods, such as composite likelihood (Zhu & Bustamante, 2005), use information about allele frequencies across the genome. Under neutral conditions, alleles that segregate at high frequencies are rare, compared to alleles that segregate at low frequencies, and thus SFS profiles are usually right-tailed (Figure 2.2). However, if a locus is under purifying selection, alleles at this locus will segregate at lower frequencies than neutral, and SFS will be skewed to the left; a locus that has undergone directional selection will behave just the opposite, and advantageous alleles at such a locus will segregate at higher frequencies than neutral, and have a right-skewed SFS (Nielsen, 2005). This right-skewed SFS is seen more strongly when there is a selective sweep signal in the population. Figure 2.2 shows the expected site frequency spectrum model under neutrality, negative and positive selection and selective sweep, using 10 individuals and different selection coefficients (Nielsen, 2005; Nielsen & Slatkin, 2013). According to this, there is an excess for both positive selection and selective sweep at high frequencies in comparison to neutrality, while no frequency is observed at high frequency sites under negative selection, since the deleterious mutations cannot segregate in the population (Nielsen & Slatkin, 2013).

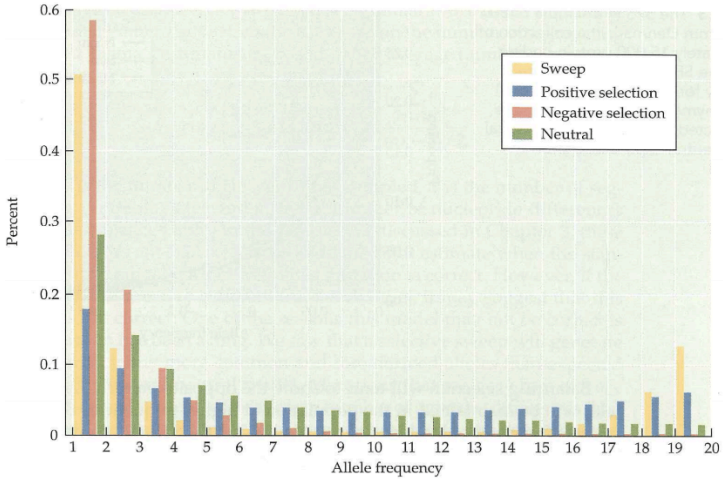

Figure 2.2. The expected site frequency spectrum of 20 chromosomes for 10 diploid individuals under neutrality, negative selection, positive selection and selective sweep models. The figure was taken from Nielsen & Slatkin, (2013).

Tajima's D is another method to test deviation of allele frequencies from neutrality. This method uses pairwise genome comparisons to calculate the average number of differences, called Tajima's estimator (θ_T) and number of segregating sites called Watterson's estimator (θ_W) (Nielsen & Slatkin, 2013; Tajima, 1989). Under neutrality, these two estimators should be equal to each other and Tajima's D will be 0. However, θ_W will be higher than θ_T when a locus undergoes a selective sweep or negative selection, since allele frequencies will be shifted toward lower values (with more singleton sites) and there will be few pairwise differences for a certain number of segregating sites (Korneliussen, Moltke, Albrechtsen, & Nielsen, 2013; Nielsen & Slatkin, 2013). On the other hand, negative Tajima's D values can be an indication of population expansion since the number of singletons will be increase with the increasing population size (Vitti et al., 2013), although this would be expected to affect the whole genome and not be locus-specific.

Evolutionary mechanisms that change allele frequencies may lead to genetic differentiation among populations. To measure such differentiation, Wright's F_{ST} (Wright, 1965) statistic was initially introduced, but later, different estimators such as Weir-Cockerham's F_{ST} (Weir & Cockerham, 1984), Nei's F_{ST} (Nei, 1986), Hudson's F_{ST} (Hudson et al., 1987) were developed. Wright and Weir-Cockerham F_{ST} are defined as ratio of the probability of heterozygosity of randomly selected gametes from one population to the probability for the total population, but total population is assumed as combination of two populations in Wright's F_{ST}, while the total population is referred to as the most recent common ancestral population in Weir-Cockerham's F_{ST} definition, thus it measures the ratio the variance between two populations to the variance in the ancestral population (Bhatia, Patterson, Sankararaman, & Price, 2013).

The F_{ST} score ranges between 0 and 1, where lower values indicate population similarity and higher values indicate differentiation. Sometimes negative F_{ST} values can be calculated but this occurs due to an overestimation of F_{ST} when the populations are not differentiated or weakly differentiated (Willing, Dreyer, & Van Oosterhout, 2012). In addition, F_{ST} can be found very high due to overestimation again when the sample sizes of the populations are very different (Bhatia et al., 2013). In selection analysis, the values in the upper tail of the F_{ST} distribution are sometimes considered as putatively under selection (Berg & Coop, 2014; Bhatia et al., 2013). To distinguish gene flow, genetic drift and selection for causing high population differentiation, the regions with high F_{ST} values are compared with neutral regions since gene flow and genetic drift affect whole genome equally (Futuyma, 2013; Nielsen & Slatkin, 2013).

For example, F_{ST} estimates calculated between modern populations using whole genomes are found to average 0.111 for European and East Asian populations

and 0.156 for European and West African populations (Bhatia et al., 2013), as it is known that human populations differentiated after the out-of-Africa process mainly by genetic drift (López, van Dorp, & Hellenthal, 2015). However, F_{ST} estimates calculated specific to a locus may be higher between these populations. For instance Bersaglieri et al., (2004) detected high F_{ST} scores around the *LCT* gene in comparison to genome-wide markers in European populations as a possible selection signal. Different evolutionary forces may shape F_{ST} distributions, such that F_{ST} differences may arise between genic and non-genic regions due to selection or F_{ST} may vary due to complex demographic history (Bhatia et al., 2013). Weir-Cockerham's and Nei's F_{ST} estimators are widely used (Bhatia et al., 2013) but in the present study Weir-Cockerham's estimator is used to calculate F_{ST} estimates of single SNPs, although there is a possibility that it may give high F_{ST} scores when the sample sizes of the populations are very different (Bhatia et al., 2013; Weir & Cockerham, 1984). Unlike Weir-Cockerham's F_{ST}, Nei's F_{ST} is not sensitive to sample size but gives overestimated values (Bhatia et al., 2013).

Pairwise F_{ST} can be used to measure which regions of the genome are most differentiated between two populations, and thus may be affected by natural selection; however, pairwise F_{ST} analysis does not reveal whether each population has diverged equally since their last common ancestor, or one population has diverged more, which could be expected under directional selection (Yi et al., 2010). To detect the specific branch that may have diverged more, a method based on F_{ST}, but including a third distant population, or outgroup, can be used. This method is called the population branch statistic (*PBS*) and is widely used in selection analysis (Fumagalli et al., 2015; Harris et al., 2019; Schlebusch et al., 2013; Yi et al., 2010). It was first used in the study by Yi et al. (2010) for detection of high altitude adaption-related alleles in the Tibetan population. To find which genes are responsible for this adaptation, the authors compared the Tibetan population with a close Han Chinese population and a distant Danish population using exome sequencing of 50 genes. Using *PBS* analysis they discovered the *EPAS1* gene that appears to have differentiated specifically in the Tibetan population, indicating positive selection, and has a significant role in hypoxia response (Yi et al., 2010). Again using *PBS*, another study detected selection on the ancestral *FADS* haplotype in Native Americans, and the selected allele was interpreted as causing inefficient fatty acid metabolism compared to other human populations (Harris et al., 2019). Since this method can detect selection successfully, it was used in this study to identify regions under selection on the modern Anatolia branch.

Linkage-based methods detect selection and recent selective sweep events using haplotype information (Vitti et al., 2013). Extended haplotype homozygosity (EHH) is one of these method that detects stretches of homozygosity around the

putatively selected allele since neighboring regions are affected by evolutionary forces due to linkage disequilibrium (Vitti et al., 2013; Zhong, Lange, Papp, & Fan, 2010). Similarly, cross-population extended haplotype homozygosity (XP-EHH) compares populations for the extent of haplotype lengths and takes into account variation in recombination rates between populations in detecting fixed sweeps (Vitti et al., 2013). Another method called integrated haplotype score (iHS) measures the amount of EHH for an ancestral allele relative to derived allele on a locus, within the same population (Voight, Kudaravalli, Wen, & Pritchard, 2006). Among them iHS has been suggested to be the strongest method to detect incomplete selective sweeps (Vitti et al., 2013).

2.3. Selection Studies Using Ancient DNA

Increasing numbers of high quality genomic datasets being published enables studying selection on various organisms. Using the methods explained in the previous section, many recent selection events can be effectively discovered. On the other hand, even though it is often hard to isolate and sequence DNA from highly degraded ancient organic material, massively parallel sequencing technology has allowed ancient genome studies to be successfully performed (Marciniak & Perry, 2017).

Many ancient genome studies focus on the demographic history of populations and their migrations. In addition, the combination of ancient and modern genome sequences is significantly helping with the detection and timing of human adaptations (Jeong & Di Rienzo, 2014). For example, signals of adaptation and convergent evolution of light skin pigmentation in European and East Asian populations had been long examined in several studies using modern genetic data (Lamason, 2005; Makova & Norton, 2005; Rana et al., 1999), but the origins of these mutations were difficult to pinpoint. Recently, using ancient genomics, an increase in the frequency of derived alleles for light skin pigmentation were shown in ancient European individuals, some starting during the Hunter-gatherer period while others starting during Early Neolithic (Mathieson et al., 2015). To track change in allele frequencies over time, some studies have genotyped ancient individuals from several time periods for the same variant or genomic region. For instance, (Gelabert et al., (2017) used 224 ancient Eurasian individuals from eight different periods, ranging from Upper Palaeolithic to the Post-Roman era, and detected genetic differences and selection signals on *IL-10* and *ATP2B4* genes that are related to immunity and resistance to malaria. These investigations based on different time series have also revealed the effect of dietary-shifts and life-style transitions, such as the Neolithic transition that introduced food production and thus reliance on specific plant or animal products. For example, adaptation to lactose tolerance at some

point after the start of milk consumption in European populations was discovered in modern populations by Bersaglieri et al., (2004). Bersaglieri et al. (2004), used 28,440 markers around *LCT* gene and compared F_{ST} of these markers with the genome-wide F_{ST} distribution using modern populations including European Americans, African Americans, East Asians, Scandinavians and other Europeans (Bersaglieri et al., 2004). They found high frequency differences around the *LCT* gene among populations and interpreted this as evidence of selection on this loci in Europe. Meanwhile, they estimated this selection should have occurred within the last 10,000 years. Recently, using genome data from 230 ancient Eurasian individuals, Mathieson et al. (2015) showed that the strong increase in Central / North Europe in the frequency of the derived allele for lactase persistence, rs4988235, on the *LCT* gene, occurred only within the last 4,000 years. In another example of possible diet-related selection, derived haplotypes on *FADS1* and *FADS2* genes, related to more efficient desaturase enzymes, and that had been previously reported to show signatures of selection in various populations including Europeans, Asians and Africans (Harris et al., 2019), were compared between the genomes of Bronze Age and present-day Europeans, which supported the view that this gene was under selection in Europe (Buckley et al., 2017). At the same time, Mathieson and Mathieson showed that selection on the *FADS1/FADS2* genes is more recent than the dairy-related adaptations on the *LCT/MCM6* locus and selection on *FADS1/FADS2* may have occurred after the late Bronze Age rather than directly after the Neolithic Period (Mathieson & Mathieson, 2018).

2.4. Use of SNP-Trait Associations from Genome-Wide Association Studies in Selection Scans

When a phenotype is determined by only a single gene or locus it is called a Mendelian trait (Jeong & Di Rienzo, 2014). However, many human traits are influenced by large number of loci, called polygenic traits (Berg & Coop, 2014). These include important physiological characters, such as metabolism or life-history traits, and some may have been shaped by directional selection. In recent decades, large numbers of associations between marker loci and complex traits, which are controlled by many different genes and the environment, have been discovered using genome-wide association studies (GWAS) (Buniello et al., 2019; Welter et al., 2014). Therefore, GWAS are an important source for the population and quantitative geneticist to study evolution and genetic diversity in the short and long time scales at the phenotypic level (Berg & Coop, 2014). However, studying locus-based selection of polygenic traits carries the disadvantage that, because many loci have small effect sizes on the trait, each locus usually gives only a weak selection signal. Because of this weakness, most of the time it is impossible to identify signatures of selection at an individual

locus against the whole genome background (Berg & Coop, 2014). Therefore, phenotype-based analysis, using the combined signal of all variants associated with a complex trait, is more powerful to detect selection in the genome than classical scans that focus on single loci.

In recent years, multiple studies have investigated polygenic adaptation combining GWAS data with population genomic datasets. These include studies focused on specific polygenic and Mendelian traits or identified traits as a result of genome-wide scans like body height (Berg & Coop, 2014; Zhang, Muglia, Chakraborty, Akey, & Williams, 2013), skin pigmentation (Berg & Coop, 2014; Mathieson et al., 2015), type 1 and type 2 diabetes (Zhang et al., 2013), body mass index (Berg & Coop, 2014; Zhang et al., 2013), obesity (Myles, Davison, Barrett, Stoneking, & Timpson, 2008), fatty acid measurement (Buckley et al., 2017; Harris et al., 2019), total cholesterol, HDL, LDL, triglycerides (Zhang et al., 2013), vitamin D (Mathieson et al., 2015), blood type (Gelabert et al., 2017), IL-10 (Gelabert et al., 2017) and MHC (Mathieson et al., 2015) and have identified signatures of positive selection in different populations for these traits.

Even though GWAS provides information about the genetic architecture of traits, the discovered variants may not be universal. There is huge heterogeneity for the variant-trait associations among populations (Berg & Coop, 2014; Myles et al., 2008; Wojcik et al., 2019). Especially, identified risk alleles may vary among populations since the frequencies of the specific alleles may differ. For example, an allele at low frequencies may appear not related to a trait in one population while the same allele, at higher frequency in another population, may show significant relation to the same trait in this latter group (Wojcik et al., 2019). Furthermore, differences in the genetic background may also influence associations. Therefore, risk alleles found in one population do not account for all humans (Myles et al., 2008). In consideration of this variation, it is important to choose variant associations based on the studied population and sample size when studying polygenic adaptations (Wojcik et al., 2019).

In the present study, we focused on several, mainly polygenic phenotypes that we hypothesized could be under selection in Anatolia since the Neolithic transition. We thus investigated the frequency changes of trait-associated alleles between Neolithic and present-day Anatolia populations. Using pairwise F_{ST} and F_{ST}-based *PBS* methods, the most strongly differentiated traits, representing candidates for directional selection, were determined for the Anatolian human population.

CHAPTER 3

3.MATERIALS AND METHODS

In this chapter, the methodology of the study is explained, including the collection of samples, compiling the list of phenotype-associated SNPs, SNP calling from DNA sequences, allele frequency comparison between Neolithic and modern Anatolian populations, and selection analysis using F_{ST}-based statistics.

3.1. Genome Data

Published ancient and modern DNA genome sequence data was used in this study. In total, 36 ancient samples from Neolithic Period (8300 – 6200 BCE) from various locations in west and central Anatolia and 16 present-day individuals from the Turkish population were included for analysis. The information about their locations, time periods, median of genome coverages, the number of individuals and source studies are summarized in **Table 3.1**. Neolithic DNA samples were collected from various studies that used different sequencing approaches. Data for 20 individuals from Barcın Höyük and 4 individuals from Menteşe Höyük had been primarily derived using a SNP capture approach that targeted ~1240,000 (1240K) SNPs (Lazaridis, 2016; Mathieson et al., 2015). Two Barcın Höyük individuals (Hofmanová et al., 2016) and one individual from Kumtepe (Omrak et al., 2016), sequenced with whole-genome shotgun sequencing, were also included. Finally, data from 4 individuals from Boncuklu Höyük and 5 individuals from Tepecik-Çiftlik (Kılınç et al.,

2016), generated using shotgun sequencing and the whole-genome capture approach, were included. In the present study, we used BAM files of Neolithic DNA datasets after their FASTQ files had been aligned by CompEvo Lab to NCBI build GRCh37/UCSC hg19 (assembly Feb. 2009, version hs37d5) human reference genome using *bwa* software (v0.7.15) (H. Li & Durbin, 2009) with parameters "*-aln -n 0.01 -o 2*" in single-end mode and with the seed disabled using "*–l 16500*".

Table 3.1: A summary description of samples used in this study, including location, historical period, range and medians of genome coverages, number of individuals (*N*), and data source.

Location	**Period**	**Range of Coverage (X)**	**Median of Coverage (X)**	***N***	**Source**
Boncuklu Höyük	10300-9950 BP	0.03-6.68	0.20	4	Kılınç et al., 2016
Tepecik-Çiftlik Höyük	8750-8212 BP	0.02-0.72	0.47	5	Kılınç et al., 2016
Barcın Höyük	8500-7600 BP	0.03-6.28	0.15	22	Mathieson et al., 2015; Lazaridis et al., 2016; Hofmanová et al., 2016
Menteşe Höyük	8400-7600 BP	0.003-0.03	0.017	4	Mathieson et al., 2015
Kumtepe Höyük	7000-6700 BP	0.1	0.1	1	Omrak et al., 2016
Turkey	Present	32-48	36.45	16	Alkan et al., 2014

For the modern-day dataset, sequences of 16 individuals from different regions of Turkey (Alkan et al., 2014) were used. Modern-day sequences were previously aligned by Seven Bridges Ankara team members to the NCBI build GRCh37/UCSC hg19 (assembly Feb. 2009) human reference genome and genotyped. Here, I used genomic variant calling format files (gVCFs) generated separately for each individual; these files contain genotypes of each detected site in the genome.

3.2. SNP List

3.2.1. Trait-associated SNPs

In this study, I chose 40 phenotypes that I hypothesized to have frequency changes over the last 10,000 years in Anatolia. These phenotypes are mostly composed of the traits that were previously investigated and identified in ancient DNA as explained in Section 2.4 of Literature Review. Therefore, the traits that I hypothesized for changing are mainly associated with metabolism (eg. fatty acid measurement, insulin measurement etc.), immunity (eg. basophil count, serum IgE measurement, interleukin measurement, etc.), externally observable traits (eg. eye color, skin color, hair color, etc.) and complex diseases (eg. obesity, type-1 diabetes, type-2 diabetes, etc.). SNPs of 38 of the phenotypes were downloaded from 3616 phenotype of GWAS Catalog. The list of SNPs associated with two Mendelian phenotypes (lactose intolerance and blood type) were retrieved from the OMIM and SNPedia databases. The genomic locations of those latter SNPs were retrieved from dbSNP. The final number of SNPs for each of the studied phenotypes ranged from 5 to 1379. A detailed summary about the SNP numbers and source databases for each phenotype are provided in **Appendix A**.

To determine genotype and phenotype relations, Genome-Wide Association Studies (GWAS) are being ever commonly used for complex phenotypic traits and clinical conditions, owing to advances in genotyping and sequencing technologies in recent decades (Buniello et al., 2019). The results of these studies are stored and provided to the users via various databases. GWAS Catalog (https://www.ebi.ac.uk/gwas/home) is one of these databases that enables users to access published SNP-trait associations that can be easily searched, visualized and integrated with other resources. GWAS Catalog provides significant SNP-trait associations that were discovered from meta-data of various populations including Africans, South and East Asians, Europeans, Hispanic-Latin Americans (Buniello et al., 2019; Welter et al., 2014). After all eligible GWA studies are identified from the literature and assessed by the curators they can be accessed in GWAS Catalog (Buniello et al., 2019; Welter et al., 2014). Here we

used the GWAS Catalog as a main source to collect information on trait-associated SNPs, including chromosomal location and effect size of alleles. For the phenotypes that were studied here but not found in the GWAS Catalog, SNPedia, dbSNP and OMIM databases were used to obtain the required information of trait-associated SNPs. SNPedia (https://www.snpedia.com/index.php/SNPedia) also stores variant information likewise GWAS Catalog (Cariaso & Lennon, 2012). However, the variants are not clustered based on phenotypes. It only presents the genotypes and their effects retrieved from published studies. Therefore it was used to detect the effect of the trait-associated SNPs that could not be retrieved from GWAS Catalog. Online Mendelian Inheritance in Man (OMIM; https://www.omim.org/), which is a platform where variants among more than 15,000 genes can be searched with respect to their relations to any Mendelian disorders or phenotypes (Amberger et al., 2011; Amberger et al., 2015; Amberger et al., 2019; Amberger & Hamosh , 2017; McKusick, 2007), was used to find SNPs related to the traits absent in GWAS Catalog. Another database that contains variants is dbSNP (https://www.ncbi.nlm.nih.gov/snp/), which is a tool of NCBI, National Center for Biotechnology Information (Sherry et al., 2001). dbSNP has a collection of many variants including SNPs, small-scale multi-base deletions or insertions, retroposable element insertions and short tandem repeats (STR). In this study, dbSNP was only used to retrieve genomic positions of the SNPs associated with the traits that were not found in GWAS Catalog.

From GWAS Catalog, the SNPs were downloaded in TSV file format, including information such as date of the published article, name and link of the published study, disease trait, mapped gene, SNP ID, chromosome number, chromosomal position, risk allele, p-value and effect of the risk allele on phenotype. The downloaded SNPs were first filtered based on association p-values ($p < 10^{-8}$) since this threshold is determined as a standard for identifying common variants after Bonferroni correction (Fadista, Manning, Florez, & Groop, 2016). After that, the chromosome number, position and SNP IDs were stored in separate BED files for each phenotype, to be used during the SNP calling step. Further, information about the SNPs including ID, chromosome number, position, risk allele and effect of the risk allele were saved in other BED files to be used for further directionality analysis.

Chromosome numbers and positions of the SNPs retrieved from GWAS Catalog were based on the NCBI build GRCh38/UCSC hg38 (assembly Dec . 2013) human reference genome. Since the BAM files of DNA sequences were aligned to NCBI build GRCh37/UCSC hg19 (assembly Feb. 2009), all the SNP positions in both BED files were mapped to NCBI build GRCh37/UCSC hg19 (assembly Feb. 2009) reference genome positions using *liftOver* software (https://genome.sph.umich.edu/wiki/LiftOver) (Hinrichs et al., 2006) and the

hg38ToHg19.over.chain.gz chain file, which was downloaded from https://genome.sph.umich.edu/wiki/LiftOver on December 17, 2018 and provides the mapping information between hg38 and hg19 reference genome positions.

The SNPs associated with lactose intolerance and blood type were retrieved from OMIM and SNPedia databases. The chromosome number and positions according to GRCh37 human reference genome were retrieved from dbSNP. The risk alleles and their effects on the phenotypes were taken from SNPedia if applicable and this information was stored in separate BED files, as above.

3.2.2. SNPs in neutral regions

To investigate whether SNPs associated with any trait may have evolved under natural selection, the frequency of trait associated alleles were compared with the frequency of alleles explored in neutral regions. Those regions were determined using the *Neutral Region Explorer (NRE)* tool (http://nre.cb.bscb.cornell.edu/nre/index.html) (Arbiza, Zhong, & Keinan, 2012) The *NRE* tool has been developed to explore neutral regions on the hg19 reference genome by excluding certain genomic regions. In this study, the locations of known genes, repeated elements, gene boundaries, copy number variations (CNVs), spliced expressed sequence tags (ESTs) and self-chains were excluded. I mainly used default parameters, and chose a minimum region size of 200 basepairs (bp), set the recombination rate as 0.9 centiMorgan/Megabase (cM/Mb), used the HapMap genetic map (Belmont et al., 2003), estimating human diversity using the "CEU" (i.e. Utah Residents (CEPH) with Northern and Western European Ancestry) population including all individuals, choosing the "strict" version of Repeat Masker (Smit, Hubley & Green,1996-2010), and setting the minimum background selection coefficient as 0.95. Meanwhile I modified a number of parameters, including the distance to the nearest gene, which was set to 1000 bp, and chromosome numbers, which were set as all autosomes (i.e. chromosomes 1-22). In addition, I selected parameters for the regions of simple repeats that contain a set of repetitive elements, using Repeat Masker v3.2.7 (Smit, Hubley & Green1996-2010) that contains the full set of Repeat Masker Regions, and a multiple alignment of 46 mammals (Blanchette et al., 2004; Karolchik et al., 2003; Siepel et al., 2005) that contains conserved genomic elements across the mammalian phylogeny to overlap with the identified neutral regions.

When a search is performed, the *NRE* tool returns a table that includes minimum, first quartile, median, mean, third quartile and maximum values of length, percentage of simple repeats, percentage of conservation, percentage of Repeat

Masker regions, distance to genes in cM and bp, recombination rate, genetic diversity levels and background selection estimates for the identified neutral regions. Since I aimed to choose neutral positions, which should be independent of any effect from functional regions, I only retained as "neutral" those regions with a conservation rate lower than the mean of the conservation rate of all neutral regions identified. To obtain a list of locations and SNP IDs for neutral SNPs, the chosen "neutral" regions were intersected with 30,761,499 SNP positions using *bedtools intersect* (Quinlan & Hall, 2010) from the 1000 Genome Project Consortium phase3 (Auton et al., 2015), and called based on the NCBI build GRCh37/UCSC hg19 (assembly Feb. 2009) human reference genome. As a result of this intersection 1,192,246 neutral SNPs were determined to be used in downstream analyses.

3.3. Data Processing

3.3.1. SNP calling from ancient DNA sequencing data

To calculate frequency change of each phenotype-associated SNP, SNPs were called from BAM files of ancient individuals. BAM files are the binary format for storing the information of reads that aligned to a reference genome (Heng Li et al., 2009). Each line in the BAM file composed of read name, chromosome, start position, matched read sequence, quality and alignment tags (The SAM/BAM Format Specification Working Group, 2019). Before SNP calling, each BAM file of ancient individuals were *in silico* trimmed, since the transitions at 3'-ends and 5'-ends are not reliable and real SNPs might be confounded with deamination patterns frequently observed in ancient DNA (Dabney et al., 2013; Lamnidis et al., 2018). To eliminate this problem, 10 positions from the ends of each read were converted to "N" and quality to "!" using the *trimBam* function of *bamUtil* software (v1.0.14) (Jun, Wing, Abecasis, & Kang, 2015). Next, the SNP positions were called using *samtools* (v1.4.1) (Li et al., 2009) *mpileup* from the reads with minimum base quality (-Q) 30 and mapping quality (-q) 30. The output BCF file was converted to VCF format using *bcftools* (v1.4.1) (Li, 2011).

Most ancient genomes we use have genome coverage <1; in other words, the polymorphic sites we study are usually sequenced once (represented by one read), if at all. For such low coverage genomes we cannot detect both alleles of an individual at diploid sites, and the genotypes cannot be called with confidence (Schraiber, 2018). Generally, this problem is overcome in ancient DNA studies using the so-called pseudo-haploidization procedure, which ensures that all analysed individuals' genotypes are inferred with the same confidence, irrespective of the genome coverage. Ancient genotypes are computationally haploidized by randomly sampling a single read from each covered sited or

randomly selecting one of the alleles at heterozygous sites (Allentoft et al., 2015; Fu et al., 2015; Haak et al., 2015; Iosif Lazaridis, 2016; I. Mathieson et al., 2015; Schraiber, 2018; Skoglund et al., 2012). In this analysis I performed pseudo-haploidization for each ancient sample after the genotyping, using a Python (v3.5.2) code. This code finds the heterozygote genotypes for each SNP, chooses one of the alleles randomly and writes the selected allele as a homozygote genotype.

3.3.2. SNP calling from modern data

Modern-day Anatolian genomic polymorphism data was obtained from the Turkish Genome Project, collected and sequenced from 16 volunteer individuals (Alkan et al., 2014). These individuals represent different cities of Turkey including Ankara, Artvin, Erzincan, Erzurum, Hatay, Isparta, Istanbul, Izmir, Kayseri, Konya, Mersin, Muğla, Nevşehir, Ordu, Sinop and Van. I used genome-wide variant call format files (gVCF) that contain genotypes of all positions in each individual's genome. The variants of phenotypes were called from each individual's gVCF file using *GATK* (v3.7-0-gcfedb67) (McKenna et al., 2010) *GenotypeGVCFs* function.

3.4. Population Genetic Analysis

3.4.1. F_{ST} analysis

F_{ST} is a widely used statistic in population genetics to measure genetic differences between two populations. Despite its widespread use, there are different descriptions and estimation ways for F_{ST}. Here, Weir-Cockerham's F_{ST} estimator (Weir & Cockerham, 1984) was used to calculate genetic distances between Neolithic and modern Anatolian individuals. This estimator is based on estimating the ratio of subpopulation variance and total variance, by incorporating the sample sizes in the case of two populations and biallelic SNPs, and subtraction from one with the following formula:

$$\mathrm{F}_{ST}^{WC} = 1 - \frac{2\frac{n_1 n_2}{n_1+n_2}\frac{1}{n_1+n_2-2}[\,n_1 p_1 (1-p_1) + n_2 p_2(1-p_2)]}{\frac{n_1 n_2}{n_1+n_2}(p_1-p_2)^2+\left(2\frac{n_1 n_2}{n_1+n_2}-1\right)\frac{1}{n_1+n_2-2}[\,n_1 p_1(1-p_1)+n_2 p_2(1-p_2)]} \quad (3.1)$$

where n_1 and n_2 indicate the population sizes while p_1 and p_2 indicate allele frequencies for each population. F_{ST} varies between 0 and 1, where the more

similar the populations, the closer will be F_{ST} to 0, and the more genetically distant, the closer will it be to 1. We calculated Weir-Cockerham's F_{ST} using *vcftools* (v0.1.17) (Danecek et al., 2011). *vcftools* calculates F_{ST} from a single VCF file including all individuals from both populations. Thus we first intersected ancient and modern VCF files for each phenotype using *bcftools isec* (v1.4.1) (Li, 2011), which returned two separate outputs including shared positions from each VCF file. Then, those genotypes of common positions were merged into one single VCF file using the *vcf-merge* (vcftools v0.1.17) (Danecek et al., 2011) software. After that, F_{ST} could be calculated using the final merged VCF file.

We calculated F_{ST} in two ways. (a) F_{ST} values for each SNP were calculated separately with *vcftools* (v0.1.17) (Danecek et al., 2011) using the parameters "*--weir-fst-pop ancient.txt --weir-fst-pop modern.txt*" to specify which individuals belong to which population, without giving any window size. (b) We calculated F_{ST} across genomic windows to account for linkage between neighboring SNPs. Specifically, we calculated the mean F_{ST} of the SNPs within 200 kilobase pair (kbp) window sizes by including the "*--window-size 200000*" parameter to the same code before. This window size was chosen since it has been suggested that recombination hotspots occur approximately every 200 kbp in the human genome (McVean et al., 2004). This window size would then ensure the relative independence of windows from each other.

3.4.2. Allele frequencies and directionality

The directions of allele frequency changes of the SNPs were examined to determine if the allele frequency of all alleles that have a significant association and positive effect size on the trait tend to change consistently from the Neolithic Period to the present. The information on risk alleles and their effects were collected from TSV files downloaded from the GWAS Catalog (Welter et al., 2014) into a separate BED file. For the traits that could not be found in the GWAS Catalog, the relevant information was obtained from SNPedia and OMIM databases. If a risk allele had a decreasing effect, the other allele was assumed to have an increasing effect. Allele frequencies of risk alleles from both Neolithic and modern-day populations were calculated using the *vcftools* (v0.1.17) (Danecek et al., 2011) "*--freq*" parameter. After that, those frequencies were compared for each increasing allele in R (v3.5.0, https://www.r-project.org/). To test the significance of allele frequency changes between ancient and modern samples, the Student's paired *t*-test was used and the "Benjamini-Hochberg" multiple testing correction method applied on the 40 p-values for all traits, using the R "stats" package functions (v3.5.0).

3.5. Selection Analysis

3.5.1. Trait-associated vs. neutral SNPs

To test if there are significant changes in allele frequency for SNPs associated with specific traits, the distribution of F_{ST} values of SNPs associated with each trait were compared with SNPs in neutral regions. We performed these comparisons both using individuals SNPs' F_{ST} values and also using 200 kbp window F_{ST} calculations, and the results were described using boxplots in R. Since the sample sizes of the SNPs were not equal and the F_{ST} values were not normally distributed, we used the Mann-Whitney U test (Mann & Whitney, 1947) in R (v3.5.0) as a non-parametric test to compare each trait with neutral regions separately. We then determined significantly changed traits that had *p*-values lower than 0.05 (*p*-value < 0.05) after the application of "Benjamini-Hochberg" multiple test correction method.

3.5.2. Population Branch Statistic

As another way to investigate genetic changes between Neolithic and present-day Anatolian populations, we performed an additional selection analysis. Here the goal was to detect regions that may have differentiated specifically in the modern-day Turkish population, using the population branch statistic (PBS). Basically, this test statistic is a derivative of F_{ST} designed to take advantage of an outgroup population to identify differentiation on a specific population branch (Yi et al., 2010). It is known as a strong method to detect selection among the given populations. The PBS score was calculated following Yi et al. (2010). Specifically, the F_{ST} values calculated between pairwise populations were transformed into *T,* which estimates divergence time (or branch length) scaled by the population size, with the formula:

$$T = -\log\left(1 - F_{ST}\right) \tag{3.2}$$

This value is calculated between the first and second populations (T^{P1P2}), first and third populations (T^{P1P3}), and second and third populations (T^{P2P3}), respectively. Then a PBS score is calculated as:

$$PBS = \frac{T^{P1P2} + T^{P1P3} - T^{P2P3}}{2} \tag{3.3}$$

To detect any region that is highly differentiated from the Neolithic Period to present, and thus could be under selection, a PBS analysis was performed comparing 16 individuals from the modern-day Turkish population (TGP), 36 Neolithic Anatolian individuals (Table 3.1), and 16 Yoruba individuals of

modern-day Nigeria (YRI) from the 1000 Genome Project phase3 (Auton et al., 2015). The 27,586,575 positions from 1000 Genome Project phase3 (Auton et al., 2015) were called from BAM files of Neolithic individuals and TGP gVCF files. Then, the data in separate VCF files (for each population) were merged and F_{ST} was calculated between the following pairs of populations: TGP-Neolithic Anatolia, TGP-YRI, Neolithic Anatolia-YRI, using *vcftools* (v0.1.17) (Danecek et al., 2011) with parameters "*--window-size 50000*" and "*--window-step 10000*" for 27,586,575 SNPs. For each SNP F_{ST} values were converted to T values between modern and Neolithic Anatolia (T^{MN}), between modern Anatolia and YRI (T^{MY}) and between Neolithic Anatolia and YRI (T^{NY}) according to formula (3.2) in R (v3.5.0). Next, PBS scores of each SNP was calculated according to formula (3.3).

To verify our approach, we also performed a positive control, testing for a known case of positive selection signature specific to North Europe (Schlebusch et al., 2013). For this, we used data from Utah Residents (CEPH) of Northern and Western European ancestry (CEU), individuals from Toscani in Italy (TSI) and individuals from Luhya in Webuye, Kenya (LWK) from the 1000 Genome Project phase3 (Auton et al., 2015). From the 1000 Genome Project phase3 (Auton et al., 2015) VCF file, we randomly sampled 16 individuals belonging to these three populations. Any position including at least one missing genotype was discarded using "*--max-missing 1.0*" parameter of *vcftools* (v0.1.17) (Danecek et al., 2011) and separated into VCFs include genotypes of pairwise populations. The F_{ST} and *PBS* scores were calculated with the same parameters as explained above with R (v3.5.0) code using 7,134,475 SNPs.

In both *PBS* analyses, scores higher than 1.0 ($PBS > 1.0$) were listed since those values are found in the 99.9% percentile of empirical distribution. To find which genes overlapping with such high *PBS* score regions, the genomic positions for each high *PBS* score SNP was searched in the UCSC hg19 (assembly Feb. 2009) human reference genome using the UCSC Genome Browser (Kent et al., 2002). Then, gene names retrieved from the browser were searched in GWAS Catalog to find associated functions.

CHAPTER 4

4.RESULTS

4.1. Significant SNP Associations and Effect Type Information

A list of trait-associated SNPs were collected as reported in section 3.2.1 of Material and Methods. The list of 40 traits, chosen to be tested for possible frequency change over the last 10,000 years in Anatolia, is given in Appendix A. From NHGRI-EBI GWAS Catalog, 11,849 trait-associated SNPs were downloaded in total. After p-value filtering ($p<10^{-8}$) and duplicate removal, 6651 SNPs remained. Of these, 6616 of them could be called from Neolithic DNA sequences while 6340 of them could be found from the present-day individuals' genotyped dataset. The information of the effect of risk allele (increasing or decreasing the trait) was available for 5235 of these.

4.2. F_{ST} Comparisons Suggest Loci Related to Lipid Metabolism Have Changed Significantly

We first asked whether alleles associated with specific phenotypes may have changed in their frequencies more than alleles in neutral regions, which represent change as a result of genetic drift (Kimura, 1983) or population admixture. Trait-associated and neutral SNPs were called from Neolithic Anatolia (n=36) and modern-day Turkish (n=16) genomes. Note that here we assume some degree of population continuity in Anatolia, and thus Neolithic individuals are assumed as among the ancestors of the modern population, although not necessarily the only ancestors, which is supported by demographic analyses (Damgaard et al., 2018; Feldman et al., 2019).

As a result of genotype calling, the number of individuals with missing genotypes in ancient data ranged from 5 to 36 (13% to 100%, median = 86%), while all 16 individuals were successfully genotyped in the modern population for 6340 SNPs. After genotyping, to detect allele frequency changes, we used Weir-Cockerham's F_{ST} (Weir & Cockerham, 1984), which is a measure of population differentiation also commonly used in studies on recent positive selection. F_{ST} was estimated between Neolithic Anatolian individuals and the modern Turkish population using *vcftools* (v0.1.17) (Danecek et al., 2011), for each SNP. Then, for each phenotype, F_{ST} estimates for trait-associated SNPs were compared with the F_{ST} estimates of neutral SNPs, using a two-sided Mann-Whitney U test (Mann & Whitney, 1947). The results of this non-parametric test were adjusted using the "Benjamini-Hochberg" (BH) multiple test correction method.

F_{ST} of neutral SNPs ranged between 0 and 1, with a median of 0 and a mean of 0.075, shown with a vertical red dashed line in Figure 4.1. F_{ST} distributions of trait-associated SNPs and SNPs in neutral regions are shown in the same figure with boxplots colored blue and gray, respectively. Three of these 40 tested traits showed significantly different F_{ST} distributions than neutral: trans fatty acid measurement (p-value = 6.3E-08), body mass index (p-value = 5.1E-06) and waist-hip ratio (p-value =1.9E-03); with median 0.08, 0.01, 0.03 and mean 0.14, 0.08, 0.09 respectively. These traits are indicated with red color in Figure 4.1. All of the significant traits had higher mean F_{ST} values in comparison to mean F_{ST} of the neutral SNPs. As higher-than-neutral F_{ST} values may suggest that the changes in allele frequency is adaptive (Myles et al., 2007), our results imply that there could be a polygenic adaptation signal on these traits, differentiating Neolithic and modern populations. Trans fatty acid measurement is directly related to the lipid metabolism and regulation of conversion between small and large chain fatty acids, and genes in this pathway have been shown to be subject to selection in human populations (Harris et al., 2019; S. Mathieson & Mathieson, 2018). Body mass index (BMI) is defined as a measure for body fat based on height and weight (WHO, 2008). Similar to BMI, waist-hip ratio is an additional measure for distribution of fat, calculated by dividing circumference of the waist to the hip (WHO, 2008). Both BMI and waist-hip ratio are considered as indicators for risk of many diseases including obesity, cardiovascular diseases, diabetes and stroke (WHO, 2008). Notably, all these three traits have an association with fat and lipid metabolism. Therefore, assuming Neolithic Anatolians were ancestral to the modern Turkish population, this differentiation should have happened during the last 10,000 years, after the adoption of farming. Our result may thus suggest that lipid metabolism in Anatolian populations became adapted to the transitions in the diet.

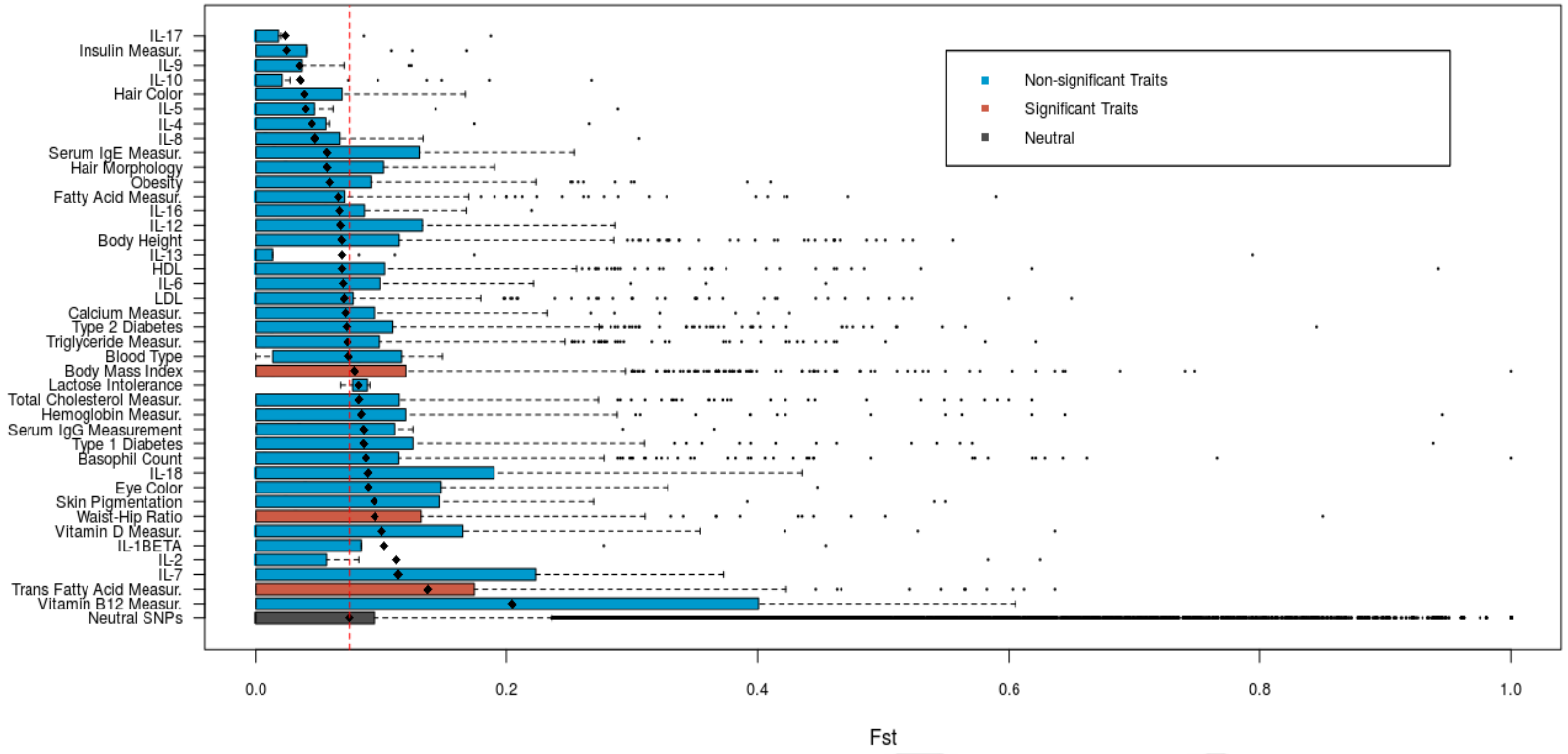

Figure 4.1. F_{ST} distributions of trait-associated and neutral SNPs. The traits were sorted according to the mean F_{ST} shown with diamonds. The red dashed vertical line indicates the mean F_{ST} of neutral SNPs at the bottom of the plot. Trans fatty acid measurement (p-value=6.3E-08), waist-hip ratio (p-value=1.9E-03) and body mass index (p-value=5E-06) were significantly different traits than neutral regions according to Mann-Whitney U Test and are shown with red boxplots. All the p-values were adjusted using Benjamini-Hochberg multiple test correction.

In the analysis presented in Figure 4.1, all trait associated SNPs were treated independently, ignoring linkage. To ensure independence of observations for both trait-associated and neutral SNPs, the same F_{ST} analysis was performed but this time using the mean F_{ST} values of trait-associated or neutral SNPs within non-overlapping 200 kbp windows. The mean F_{ST} of neutral SNPs decreased from 0.075 to 0.047 (Figure 4.2). After comparisons using the Mann-Whitney U test (Mann & Whitney, 1947) and Benjamini-Hochberg multiple test correction, 23 of 40 traits were found more or less differentiated than neutral regions shown with red boxplots in Figure 4.2. Of these 23 significant traits, 6 traits had lower mean F_{ST} values than neutral mean while the remaining had higher mean F_{ST} values. The significant traits were basophil count (p-value = 3.6E-06), body height (p-value = 3.9E-07), body mass index (p-value = 2.3E-08), calcium measurement (p-value = 1.1E-02), fatty acid measurement (p-value = 1E-05), HDL (p-value= 3.1E-03), hemoglobin measurement (p-value = 2.1E-02), IL-2 (p-value = 2E-02), IL-4 (p-value = 1.7E-02), IL-5 (p-value = 2.5E-03), IL-6 (p-value = 1.1E-02), IL-8 (p-value=2.5E-02), IL-9 (p-value = 1.1E-02), IL-10 (p-value = 1.2E-05), IL-13 (p-value = 6.2E-04), IL-17 (p-value = 5.6E-04), insulin measurement (p-value = 2.2E-04), LDL (p-value = 2.2E-04), obesity (p-value = 1.5E-03), total cholesterol measurement (p-value = 8.3E-05), trans fatty acid measurement (p-value = 2.1E-07), triglyceride measurement (p-value = 2.8E-07), type-2 diabetes (p-value = 7.5E-05) and vitamin D measurement (p-value = 2.1E-02). Notably, body mass index and trans fatty acid measurement were again found as significantly differentiated, as in the previous F_{ST} analysis, while waist-hip ratio was not significant (p-value > 0.05) (Figure 4.1).

This time, mean F_{ST} of some significant traits, including insulin measurement and multiple immunity-related traits (e.g. IL-4, IL-5, IL-8, IL-9, IL-10 and IL-17), were lower than the mean of neutral SNPs. If the allele frequencies associated with these traits indeed did not change as much as in neutral regions, this could suggest negative selection on these traits. On the other hand, there were some traits, namely basophil count, body height, body mass index, calcium measurement, fatty acid measurement, hemoglobin measurement, HDL, IL-2, IL-6, LDL, obesity, total cholesterol measurement, trans fatty acid measurement, triglyceride measurement, type-2 diabetes, vitamin D measurement, that displayed significantly higher mean F_{ST} values than neutral mean F_{ST}. This group mostly consists of metabolism- and diet-related traits, except IL-2 and IL-6.

Consequently, taken together with the results depicted in Figure 4.1, these results imply that diet might have been the strong force in adaptively shaping human genomic variation in Anatolia over the last 10,000 years.

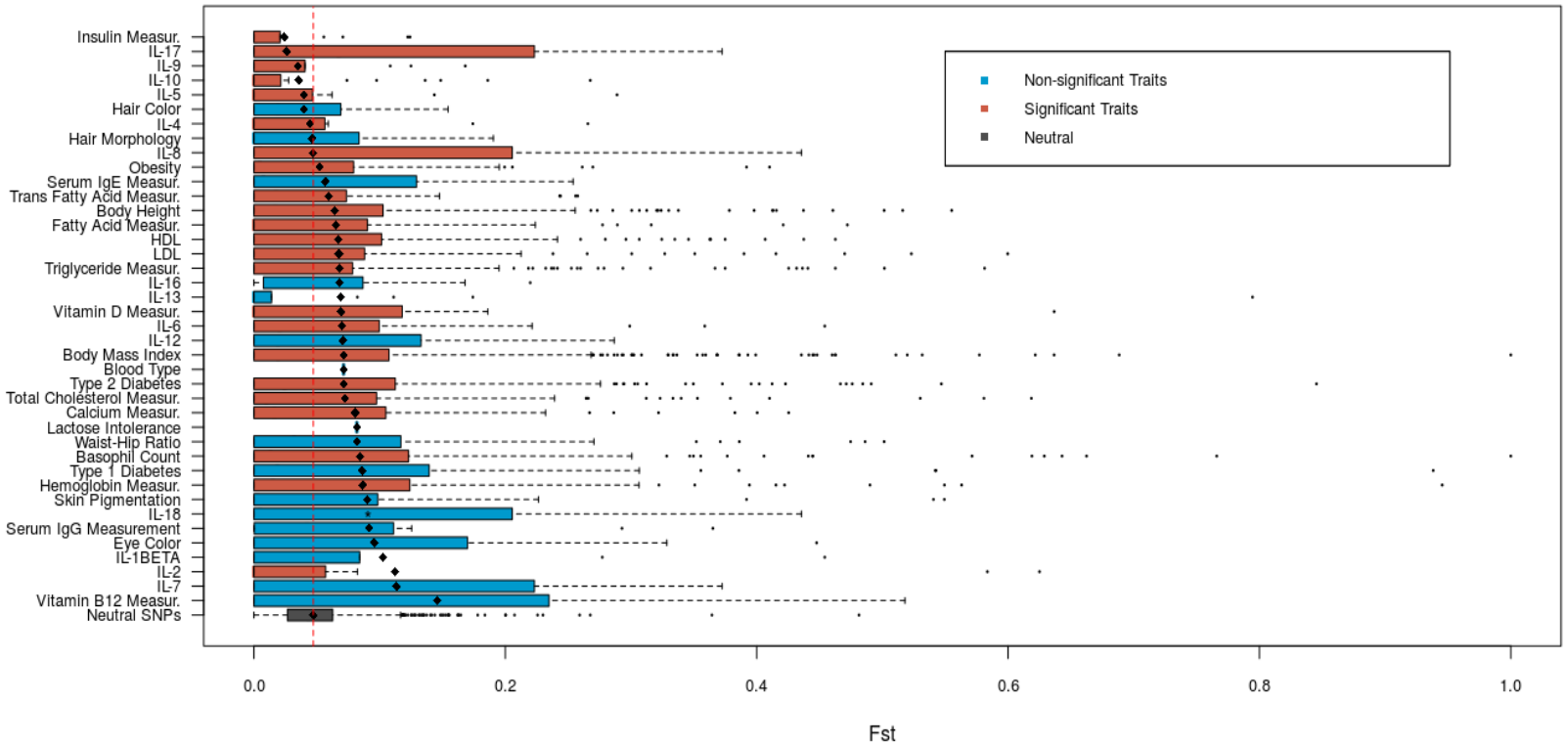

Figure 4.2. Mean F_{ST} distributions of the trait-associated and neutral SNPs within 200 kbp windows. The traits were sorted according to the mean F_{ST} shown as diamonds. The red dashed vertical line indicates the mean F_{ST} of neutral SNPs at the bottom of the plot. The traits indicated with red boxplots were significantly different according to the Mann-Whitney U Test. All the p-values were adjusted using Benjamini-Hochberg multiple test correction.

4.3. Direction Analysis Shows that Alleles Elevating Trans Fatty Acid Levels Have Decreased in Frequency in Anatolia

We next asked whether allele frequency changes between Neolithic and present-day Anatolia could indicate a directional change in phenotypic values, which could be expected under directional selection. For this, we specifically studied frequency changes in alleles associated with each trait that increase the value of the trait, as explained in Section 3.4.2 of Material and Methods. The frequencies of increased alleles were compared between Neolithic Anatolian (n=36) and modern Turkish (n=16) populations using a paired t-test.

Out of 40 tested traits for which we could find 5235 alleles with known effect, we identified three displaying consistent frequency changes for trait-elevating alleles between Neolithic and modern-day Anatolian populations: fatty acid measurement (nominal p-value = 3.5E-05), trans fatty acid measurement (nominal p-value = 7.7E-09) and triglyceride measurement (nominal p-value = 3.3E-03). Figure 4.3 shows the boxplots and strip charts of the comparisons for these significantly changed traits. In all three cases, we found decreased allele frequencies for alleles that elevated trait values. There is only one common SNP for these three traits, which is on chromosome 11 at position 61571477, on the *FADS1* gene.

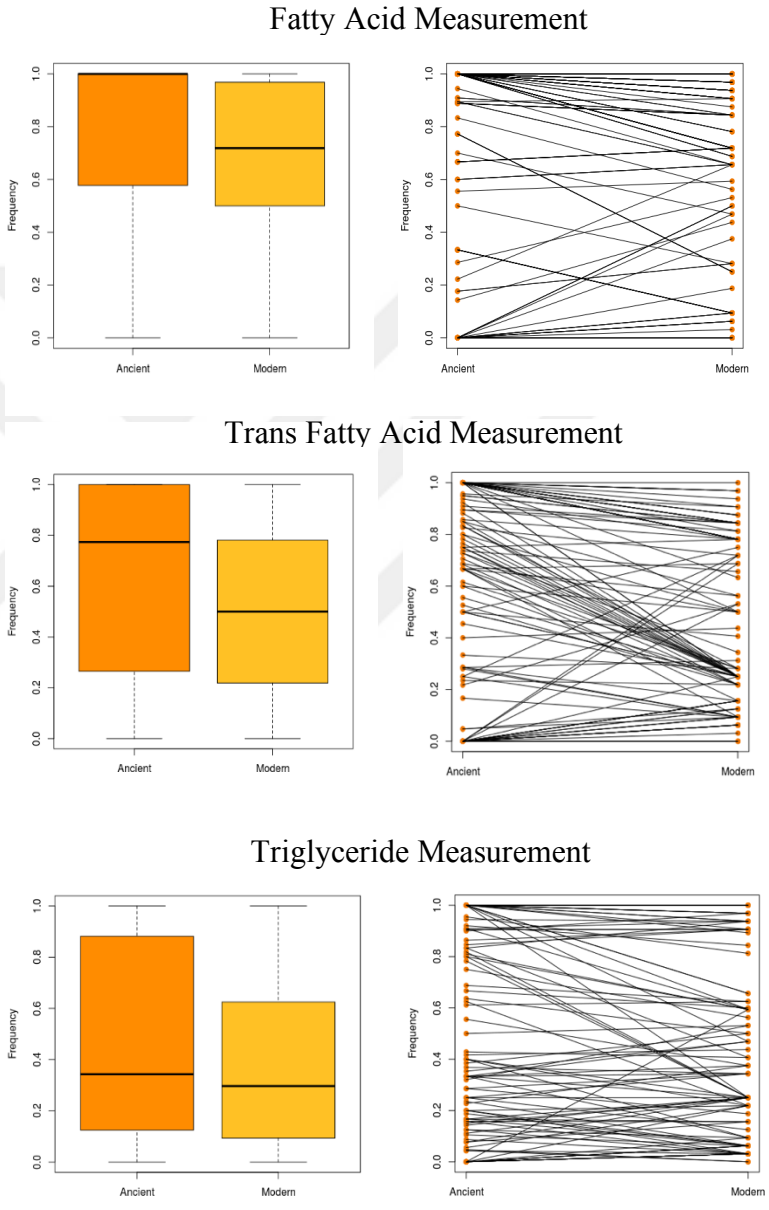

Figure 4.3. Historical allele frequency changes for trait-elevating alleles associated with fatty acid measurement, trans fatty acid measurement and triglyceride measurement. Left panel boxplots show the frequency distributions for Neolithic Anatolia (dark orange) and modern-day Anatolia human populations (yellow). Right panel strip charts show each allele separately, describing the direction of frequency change by lines.

4.4. Population Branch Statistic Can Successfully Detect a Known Selection Signal in Central Europe

In this study, our aim was to find traits that may have evolved under positive selection in Anatolian populations as a result of diet and life-style alterations. In the previous sections, we investigated such candidate traits using F_{ST} analysis, comparing Neolithic and present-day Anatolia. However, F_{ST} only measures differentiation, not change in time. To infer positive selection over time given differentiation between these two populations, we need to additionally assume that Neolithic Anatolians are direct ancestors of modern-day Anatolians, which is not fully accurate (Feldman et al., 2019). Alternatively, we may infer differentiation over time using an outgroup population to both Neolithic and modern-day Anatolians, such as a sub-Saharan African population (López et al., 2015). This selection scan method is known as the population branch statistic (*PBS*) (Yi et al., 2010). *PBS* compares genetic distances among three populations, including two close ones and one distant population, or outgroup. Genetic distances are again measured using F_{ST} calculated pairwise between each population pair. Using these three pairwise comparisons, *PBS* measures how much the branch specific to the population of interest has differentiated. Population-specific differentiation that is higher at a single locus compared to the genome background can be interpreted as a signal of positive selection.

We first sought to confirm the efficacy of the *PBS* method using a known case of positive selection, using limited sample sizes as in our Anatolian datasets. We chose lactase persistence as positive control, where the *LCT* gene has been shown to have undergone strong selection in North and Central European populations. Recently, using 1000 Genomes data, Schlebusch et al., (2013) confirmed that the CEU (Central European descent) population displays a selection signature around the *LCT* and *MCM6* locus, while TSI (Italy) and LWK (Kenya) populations did not. We thus calculated *PBS* values as explained in detail in Section 3.5.2 of Material and Methods, calculating *PBS* for 50,000 kbp windows size and 10,000 kbp window steps across the genome. *PBS* analysis was performed using CEU (n=16), TSI (n=16) and LWK (n=16) populations from the 1000 Genome Project phase3 (Auton et al., 2015) dataset to detect the same selection signal close to the *LCT* and *MCM6* gene region in CEU, which was the focal population.

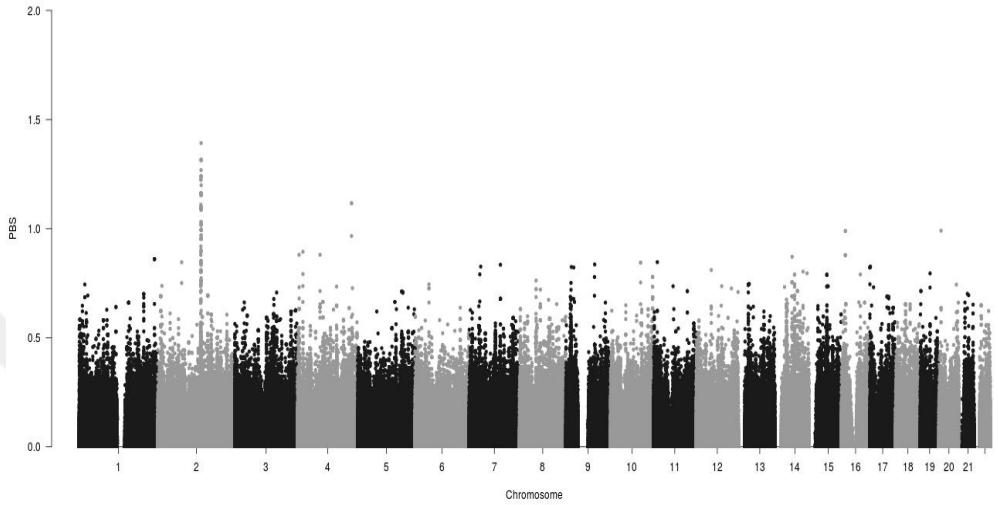

Figure 4.4. Manhattan plot for PBS analysis between CEU-TSI-LWK (n=16 for each) populations, with CEU as the focal population. The peak at the chromosome 2 belonged to *LCT/MCM6* gene region, which is responsible of lactase persistence.

The distribution of *PBS* values, displayed in Figure 4.4, revealed a unique peak on chromosome 2, consistent with a selection signal in that region in Central European descendants. The 15 SNPs with the highest *PBS* scores which are found in the 99.9% percentile of empirical distribution and the gene names overlapping those SNPs are shown in Table 4.1. This single region includes *LCT*, consistent with earlier reports (Bersaglieri et al., 2004; Itan, Powell, Beaumont, Burger, & Thomas, 2009; Schlebusch et al., 2013). Therefore, we concluded that the *PBS* analysis could detect a known selection signal successfully, even with highly modest sample sizes.

Table 4.1. The list of top 15 *PBS* values for the CEU branch. The chromosome number (chr), position, overlapping gene names, p-value and related functions are also listed. Genomic locations are given according to NCBI build GRCh 37 human reference genome.

Chr	**Position**	***PBS***	**Gene Name**	**p-Value**	**Function**
2	136658345	1.392	-	1.73E-07	-
2	136633771	1.315	*MCM6*	1.04E-06	Blood protein measurement
2	136670298	1.315	*DARS*	1.04E-06	Blood protein measurement, leukocyte count, neutrophil count
2	136685228	1.315	*DARS*	1.04E-06	Blood protein measurement, leukocyte count, neutrophil count
2	136696138	1.315	*DARS*	1.04E-06	Blood protein measurement, leukocyte count, neutrophil count
2	136740900	1.315	*DARS*	1.04E-06	Blood protein measurement, leukocyte count, neutrophil count
2	136617805	1.268	*MCM6*	1.2E-06	Blood protein measurement
2	136569848	1.239	*LCT*	2.5E-06	Lactose persistence, blood protein measurement
2	136575199	1.239	*LCT*	2.5E-06	Lactose persistence, blood

					protein measurement
2	136576577	1.239	*LCT*	2.5E-06	Lactose persistence, blood protein measurement
2	136578536	1.239	*LCT*	2.5E-06	Lactose persistence, blood protein measurement
2	136580287	1.239	*LCT*	2.5E-06	Lactose persistence, blood protein measurement
2	136583192	1.239	*LCT*	2.5E-06	Lactose persistence, blood protein measurement
2	136586958	1.239	*LCT*	2.5E-06	Lactose persistence, blood protein measurement
2	136588478	1.239	*LCT*	2.5E-06	Lactose persistence, blood protein measurement

4.5. Population Branch Statistic Detects Signals on Cholesterol Related Traits in Anatolia

In the previous section, we demonstrated that *PBS* analysis can be used to detect a known genomic selection signal on the CEU population branch, using modest sample sizes. We now turned to study selection signals on the branch of modern Turkish population (n=16) by comparing this with Neolithic Anatolian individuals (n=36) and using the Sub-Saharan Yoruba (YRI, n=16) from the 1000 Genome Project phase3 (Auton et al., 2015) as outgroup. We used the same method and formula to calculate *PBS* scores as explained in Section 3.5.2 of Material and Methods Chapter.

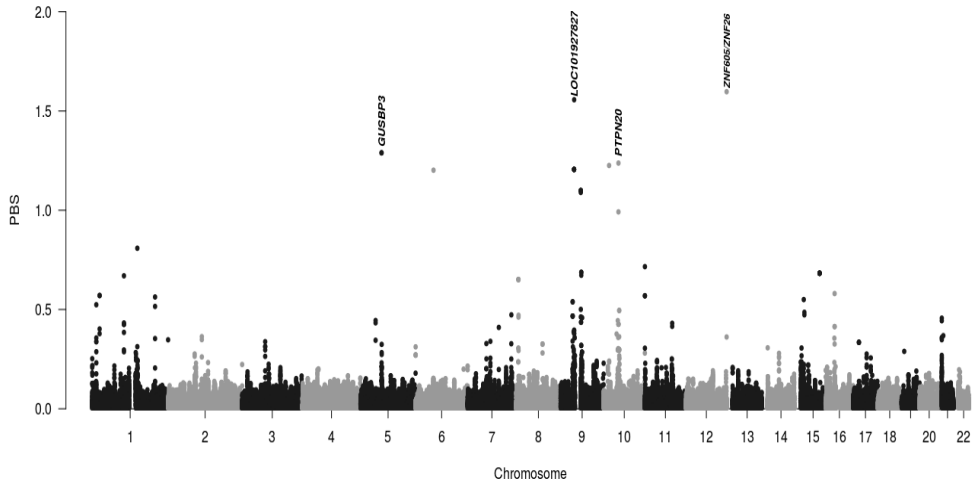

Figure 4.5. Manhattan plot for PBS analysis between modern-day Turkish (n=16)–Neolithic Anatolia (n=36) and YRI (n=16) populations, with the modern-day Turkish as focal population. The genes with the highest four *PBS* values were labeled on the plot.

The *PBS* score distribution, shown in Figure 4.5, revealed several peak points at chromosomes 5, 9, 10 and 12, overlapping with the genes *GUSBP3*, *LOC101927827*, *PTPN20* and *ZNF605/ZNF26*, respectively. The gene names, the related functions and the positions of highest 15 *PBS* values are given in Table 4.2.

Table 4.2. The list of top 15 *PBS* values for the modern-day Turkish population branch. The chromosome number (chr), position, gene names, p-values and functional information obtained from the GWAS Catalog are listed. Genomic locations are given according to NCBI build GRCh 37 human reference genome.

Chr	**Position**	***PBS***	**Gene Name**	**p-Value**	**Function**
12	133520001-133570000	1.597	*ZNF605* *ZNF26*	3.7E-06	Antisaccade response measurement, prostate carcinoma, total cholesterol measurement, triglyceride measurement, LDL, HDL, cholesterol measurement
9	44340001-44390000	1.557	*LOC101927827*	7.4E-06	-
5	68930001-68980000	1.289	*GUSBP3*	1.5E-05	(pseudo gene) Systolic blood pressure, IL-18
5	68940001-68990000	1.289	*GUSBP3*	1.5E-05	(pseudo gene) Systolic blood pressure, IL-18
10	48690001-48740000	1.238	*PTPN20*	1.8E-05	DNA methylation, response to high fat food intake, triglyceride change measurement, mosquito bite reaction, diet measurement

10	17970001-18020000	1.225	*MRC1*	2.2E-05	Blood protein measurement, aspartate aminotransferase measurement, osteoarthiritis biomarker measurement, childhood agressive behaviour measurement
9	43970001-44020000	1.206	*CNN2P4* *CYP4F60P*	3.7E-05	(pseudo gene)
9	43980001-44030000	1.206	*CNN2P4* *CYP4F60P*	3.7E-05	(pseudo gene)
9	43990001-44040000	1.206	*CNN2P4* *CYP4F60P*	3.7E-05	(pseudo gene)
9	44000001-44050000	1.206	*CYP4F60P* *SNX18P5*	3.7E-05	(pseudo gene)
6	58780001-58830000	1.202	-	4.1E-05	-
9	66080001-66130000	1.099	*RP11-93P10.3*	4.1E-05	(pseudo gene)
9	66090001-66140000	1.099	-	4.1E-05	-
9	66100001-66150000	1.099	-	4.1E-05	-
9	66070001-66120000	1.089	*RP11-93P10.3*	5.2E-05	(pseudo gene)

ZNF605 and *ZNF26*, two neighboring genes located on chromosome 12, had the highest *PBS* score. According to GWAS Catalog, *ZNF605* is associated with antisaccade response measurement while *ZNF26* is related to prostate carcinoma, total cholesterol measurement, triglyceride measurement, LDL, HDL, cholesterol measurement as an enhancer according to GWAS Catalog. Both *ZNF605* and *ZNF26* are act as enhancer on a neighbouring allele in rs12423664 that is in the intron of *FBRSL1* (Fibrosin-like 1) gene and this locus is identified as associated with triglyceride level and coronary artery disease in a study by Siewert & Voight (2018). Therefore, *ZNF605* and *ZNF26* may have a relation with the fat metabolism. The second highest *PBS* score belonged to the hypothetical gene *LOC101927827*. The third and fourth ones were located on the *GUSBP3* pseudo-gene, related to systolic pressure and IL-18 measurement according to GWAS Catalog. This gene may not have a direct effect on IL-18 measurement but it is close to *OCLN* gene that is associated with IL-18 measurement directly as identified in the study of Ahola-Olli et al., 2017. The sixth highest *PBS* value overlapped with the gene *PTPN20* that has a role in DNA methylation, response to high fat food intake, triglyceride change measurement, mosquito bite reaction and diet measurement with a near gene *GDF10* (Wojczynski et al., 2015). The next one, *MRC1* is a gene linked with blood protein measurement, aspartate aminotransferase measurement, osteoarthiritis biomarker measurement, childhood aggressive behaviour measurement in East Asian populations (Kamatani et al., 2010). The remaining genes, *CNN2P4*, *CYP4F60P*, *SNX18P5* and *RP11-93P10.3*, listed in Table 4.2, were pseudogenes that do not have relation to any phenotype according to GWAS Catalog.

Although the genes with the highest *PBS* values appeared associated with various traits, we noticed that cholesterol, fat intake and diet were recurrently emerging. While we cannot perform a formal enrichment test using as small sample sizes as here, it is remarkable that these traits also coincide with the previous results based on F_{ST} analysis (Figure 4.1 and 4.2), which also indicated that lipid metabolism may have been be under positive selection in Anatolia. Overall, both F_{ST} and *PBS* results support the notion that dietary, and specifically, lipid-related processes, underwent selection in the Anatolian population, or perhaps across west Eurasia or west Asia, after the Neolithic Period.

CHAPTER 5

5. DISCUSSION

As human populations expanded to a wide range of environments after out-of-Africa migrations, the novel physical and cultural environments increased the potential of positive selection and local adaptations (Jeong & Di Rienzo, 2014; Myles et al., 2007; Vallender & Lahn, 2004; Vatsiou, Bazin, & Gaggiotti, 2016). Similar evolutionary changes are thought to have occurred after Neolithization, the era of agriculture and sedentary life, that started in the Fertile Crescent in the Middle East region about 12,000 years ago and spread to other regions of West Eurasia and North Africa step by step (Bellwood, 2005; Lazaridis, 2016; Mathieson et al., 2015). These changes are assumed to have led to new selective pressures related to different pathogen exposures and dietary transformations (Latham, 2013; Jeong & Di Rienzo, 2014; Vallender & Lahn, 2004; Vatsiou et al., 2016). During or at some point after Neolithic transitions (in west Eurasia, east Asia, or the Americas), some of these farming populations, who started to extensively plant carbohydrate-rich wheat, rice, or maize, are thought to have shifted to more plant-based diets, and also to have started to eat more easily chewable and more cooked foods, on average. As a result, morphological changes in tooth and facial bones and nutritional deficiencies like vitamin A, vitamin B12 and zinc deficiency have been observed until present time compared to hunter-gatherers (Larsen, 2006; Latham, 2013). Besides dietary change, post-Neolithic humans started to maintain more complex communities, which led to large settlements and eventually cities (Larsen, 2006). This caused an environment suitable for pathogens, since pathogens need large and crowded host populations to spread from person to person easily (Armelagos et al., 1991; Latham, 2013). Moreover, increased intimacy with animals during the domestication process introduced new pathogens from animals to humans (Armelagos et al., 1991). The human immune system is thought to have coevolved against these pathogens that evolved after the Neolithic transitions.

In this study, the effect of the diet and life-style changes on populations in Anatolia was examined by studying the frequency changes of alleles associated with different traits, between the Neolithic Period (c.8500–6000 BCE) and present-day populations. The main purpose of the present study was to detect which type of phenotypes have been genetically changed as a result of human life alterations. In addition, we tried to detect possible selection signals on the studied phenotypes by comparing those traits with neutral regions. For this we used F_{ST} comparisons between Neolithic and modern-day Anatolian samples, assuming that the different time periods represent different population in the same area. In selection studies, F_{ST} statistics are widely used to measure interpopulation or interspecies genetic distances per locus (Berg & Coop, 2014; Myles, Davison, Barrett, Stoneking, & Timpson, 2008; Myles et al., 2007; Vitti et al., 2013). As a result of such comparisons, if certain parts of the genome are found to have much higher F_{ST} values compared to the rest of the genome or to neutral regions, this may indicate that those genes or variants have differentiated through non-random processes between the populations or species (Myles et al., 2007). Such differences can thus be a signal of adaptive changes. Meanwhile, neutral events such as gene flow from another population can also cause high differentiation between populations or over time. Here, we tried to control for such change due to gene flow by comparing trait-associated allele frequencies with the allele frequencies in neutral regions, since gene flow is supposed to affect whole genome roughly in the same way (Futuyma, 2013).

In our study, we focused on 40 pre-chosen traits for their possible roles in metabolism, immunity. Among these we found three traits to be more differentiated than neutral loci using this F_{ST} comparison approach between trait-associated and neutral SNPs: trans fatty acid measurement, body mass index, and waist-hip ratio (Figure 4.1). In a second approach, using mean F_{ST} values calculated within 200 kbp windows, we found 23 traits to be significantly differentiated than neutral (Figure 4.2). Especially, among those traits that were significantly more differentiated than neutral (and candidate targets for past positive selection), we found body mass index, fatty acid measurement, trans fatty acid measurement, HDL, LDL, triglyceride measurement, type-2 diabetes and obesity (Figure 4.2). The general trend indicated selection on fat metabolism-related alleles. Moreover, *PBS* analysis performed to detect selection on the modern-day Anatolia branch revealed that genes with the highest *PBS* scores, *ZNF605* and *ZNF26*, were also associated with fat metabolism (Figure 4.5, Table 4.2). Since a positive control *PBS* analysis was able to detect a known selection signal (Figure 4.4) with the same type of small sample sizes, the *PBS* between Neolithic Anatolia, modern-day Anatolia and YRI, with the given modest sample sizes, was considered reliable. In addition, a directionality analysis that examined the direction of frequency changes of alleles with common effects on each trait showed that the frequencies of alleles

that increase fat metabolism-related trait values were significantly and consistently shifted toward lower values between ancient and modern-day Anatolian populations (Figure 4.3), and again, the only traits that attained significance were these lipid metabolism-related groups.

Consequently, the overall results in the present study suggested that variants regulating fat metabolism might be under positive selection due to dietary shifts after Neolithization. In the literature, this type of selection on lipid metabolism was also reported, especially on the *FADS1/FADS2* genes responsible for the production of the enzyme named fatty acid desaturase (FADS). This enzyme catalyzes short-chained polyunsaturated fatty acids (PUFA) to long-chained PUFAs, which is an essential step in human physiology since LC-PUFAs and their metabolites are crucial for many biological processes, including brain development, innate immunity and energy regulation (Buckley et al., 2017; Harris et al., 2019; S. Mathieson & Mathieson, 2018). However, Mathieson & Mathieson (2018) argue that the selection on the lipid metabolism does not coincide with the Neolithic transition, but they suggest this change in *FADS1/FADS2* genes have been mostly caused by more recent changes in diet due to industrialization or efficiency of the selection due to increased population size during the Bronze Age (Buckley et al., 2017; S. Mathieson & Mathieson, 2018). In future analysis, more samples from different time periods (e.g. Bronze Age and Iron Age) can be added to examine if the observed change in lipid metabolism has started during the Neolithic Period or more recently. Also, the same analysis can be performed for neighboring regions of Anatolia including East Mediterranean and West Eurasian populations to show if this is an adaptation that belongs to a wider region of human populations.

Previous studies suggest that the selection on lipid metabolism or some diseases like cardiovascular diseases, type-2 diabetes and obesity can be explained with the "Thrifty Genotype Hypothesis" (Berg & Coop, 2014; Neel, 1962; Vatsiou et al., 2016), which assumes that the genes associated with those diseases have evolved under positive selection to store fat and carbohydrates in periods of food scarcity (Berg & Coop, 2014; Myles et al., 2008; Neel, 1962; Vatsiou et al., 2016). This hypothesis can be an explanation for the frequency changes in the traits related to lipid metabolism, cardiovascular and metabolic diseases in this study. On the other hand, we showed that alleles elevating lipid-associated traits have decreased in frequency in time; thus, lipid storage capacities might also be reduced. However, the decreasing and elevating effect of the risk-alleles should be investigated in detail to understand what kind of decreasing or increasing effect is observed at the phenotype level. Alternative to this analysis, the directionality of derived allele frequencies can be examined rather than effect of the risk-alleles.

F_{ST} comparisons also indicated that there were significant differences in some immunity-related and diet-related traits relative to neutral loci, including IL-2, IL-6, basophil count, calcium measurement, hemoglobin measurement, body height, and vitamin D measurement (Figure 4.2). Moreover, IL-18 was identified as a possible candidate of positive selection along the modern-day Anatolia branch in the *PBS* analysis (Figure 4.5), even though the joint set of SNPs associated with IL-18 were not found to significantly differentiate in time more than neutral SNPs (Figure 4.2). In fact, many of the immunity traits including IL-4, IL-5, IL-8, IL-9, IL-17 (Figure 4.2) were less differentiated in time than neutral SNPs. In this case, the effect of negative or balancing selection can be considered. This appears surprising as large changes in the immune system have been hypothesized in response to the Neolithic shift in lifestyles (Reher, Key, Andrés, & Kelso, 2019; Vatsiou et al., 2016). Among the 17 immunity-related phenotypes, only three of them were significantly more differentiated and one of them had a selection signal according to *PBS* analysis. This could be a result of not including more diverse types of immunity-related traits into the analysis. For instance, to study immunity changes after the Neolithic Period, MHC-related variants can also be included, since MHC genes are known as the most diverse gene sets in the genome, influenced by adaptation to different pathogens (Key, Teixeira, de Filippo, & Andrés, 2014; Meyer & Thomson, 2001; Reher et al., 2019). In this study, we could not include MHC-associated alleles due to limited information on MHC variants in GWAS Catalog, OMIM or SNPedia. In further analysis, MHC-associated SNPs can be determined and included through more extensive literature search.

Despite their potential, current-day ancient DNA studies suffer from certain limitations involving the use of sequence data from highly degraded and damaged DNA. First of all, it is difficult to obtain endogenous DNA from ancient especially from the Anatolian Neolithic, since it is highly fragmented and degraded as a result of DNA decay after the death of the organism, which happens particularly rapidly in temperate regions (Prüfer et al., 2010). Therefore, the endogenous DNA can be isolated and sequenced at low amounts and at low coverages with high rates of missing nucleotides. Moreover, the signatures of post-mortem damage like deamination that causes C to T and G to A conversions affect the accuracy of the sequencing (Sawyer, Krause, Guschanski, Savolainen, & Pääbo, 2012). Most of the time those transitions are excluded from the analysis, especially if they are seen at the end of the reads, which limits information (Prüfer et al., 2010). Low-coverage data cannot enable diploid genotype calling and genotypes can be skewed to reference alleles (Prüfer et al., 2010; Schubert et al., 2012). To overcome these problems in low coverage aDNA analyses, sampling a random read to obtain a haploid genotype call is used as a widespread solution (Allentoft et al., 2015; Fu et al., 2016; Haak et al., 2015; Lazaridis, 2016; Mathieson et al., 2015; Skoglund et al., 2012)

Alternatively, genotype likelihood-based population allele frequencies can also be calculated from insufficient diploid calls using counts of sequences of each site (I. Mathieson et al., 2015; Racimo, Renaud, & Slatkin, 2016; Schraiber, 2018). Furthermore, if certain genotypes cannot be called, then those regions can be predicted by the imputation method, which uses modern haplotype data and linkage information between the sites (Berg & Coop, 2014; Buckley et al., 2017; Gamba et al., 2014; Gelabert et al., 2017; Martiniano et al., 2017).

In this study, we used diploid genotype calling, and then subjected the data to pseudo-haploidization, and the observed allele frequencies from pseudo-haploid samples were used to compare with modern samples, instead of using genotype likelihood. However, pseudo-haploidization unavoidably leads to information loss (Schraiber, 2018). In addition, many target SNPs could not be called from ancient samples due to low coverage. Therefore, the calculated frequencies were not the best possible representation of Neolithic Anatolian populations. In further analysis, genotype likelihood and imputation methods can be performed on ancient samples to increase the power of comparisons between modern and ancient samples.

The main part of this study was the collection of trait-associated SNPs from different databases that store variant-phenotype associations. SNPs associated with 38 of 40 phenotypes were downloaded from GWAS Catalog, which provides accession to various SNP-trait associations found in the literature (Buniello et al., 2019; Welter et al., 2014). However, there are more than 3000 traits in GWAS Catalog, which makes difficult to analyze all of them. Therefore a new system can be followed to group GWAS traits based on the ontology. Then, the frequencies of GWAS alleles can be considered to select neutral alleles having equal frequency in the modern population. Moreover, GWAS were determined from different populations including Africans, Europeans, East Asians or Latin/Hispanic American, etc. Previous studies have shown that SNP-trait associations or risk-alleles show high heterogeneity and differ from population to population (Ioannidis, Ntzani, & Trikalinos, 2004; Myles et al., 2008). Also, most of the SNP-trait associations are based on samples of European ancestry (Wojcik et al., 2019). As a result of this, GWAS results may contain false positives when the same associations are used for different populations (Berg & Coop, 2014; Wojcik et al., 2019). In this study, association of all downloaded SNPs were found in different populations, but not in Anatolian populations. Thus, the allele frequency and directionality comparisons using those SNPs may include false positives. To eliminate those false positives, downloaded SNPs can be filtered according to the population by selecting ones closer to Anatolia.

In summary, this study showed that selection signals involving lipid metabolism genes can be detected in Anatolia, using either polygenic selection signatures with trait-associated SNPs, or genome-wide *PBS* analysis Our results suggest that after Neolithization, dietary shifts may have had a greater impact on the human genome in Anatolia rather than immunity. However, in future work, more phenotypes should be included not to miss out other regions that could have evolved under recent selective pressures. In the further studies, candidate genes under positive selection should be identified with the examination of other candidate regions. This may provide a more extensive perspective on how life-style shifts shaped the human genome.

REFERENCES

Adler, C. J., Dobney, K., Weyrich, L. S., Kaidonis, J., Walker, A. W., Haak, W., … Cooper, A. (2013). Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nature Genetics*, *45*(4), 450–455. https://doi.org/10.1038/ng.2536

Ahola-Olli, A. V., Würtz, P., Havulinna, A. S., Aalto, K., Pitkänen, N., Lehtimäki, T., … Raitakari, O. T. (2017). Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2016.11.007

Alkan, C., Kavak, P., Somel, M., Gokcumen, O., Ugurlu, S., Saygi, C., … Bekpen, C. (2014). Whole genome sequencing of Turkish genomes reveals functional private alleles and impact of genetic interactions with Europe, Asia and Africa. *BMC Genomics*. https://doi.org/10.1186/1471-2164-15-963

Alkorta-Aranburu, G., Beall, C. M., Witonsky, D. B., Gebremedhin, A., Pritchard, J. K., & Di Rienzo, A. (2012). The Genetic Architecture of Adaptations to High Altitude in Ethiopia. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1003110

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., … Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature*, *522*(7555), 167–172. https://doi.org/10.1038/nature14507

Amberger, J., Bocchini, C., & Hamosh, A. (2011). A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Human Mutation*. https://doi.org/10.1002/humu.21466

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gku1205

Amberger, J. S., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2019). OMIM.org: Leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky1151

Amberger, J. S., & Hamosh, A. (2017). Searching online mendelian inheritance in man (OMIM): A knowledgebase of human genes and genetic phenotypes. *Current Protocols in Bioinformatics*. https://doi.org/10.1002/cpbi.27

Andrews, C. A. (2010). Natural Selection, Genetic Drift, and Gene Flow Do Not Act in

Isolation in Natural Populations. *Nature Education Knowledge*, *3(19):5*.

Arbiza, L., Zhong, E., & Keinan, A. (2012). NRE: a tool for exploring neutral loci in the human genome. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-13-301

Armelagos, G. J., Goodman, A. H., & Jacobs, K. H. (1991). The origins of agriculture: Population growth during a period of declining health. *Population and Environment*. https://doi.org/10.1007/BF01256568

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*. https://doi.org/10.1038/nature15393

Bellwood, P. S. (2005). *First farmers : the origins of agricultural societies*. Blackwell Pub.

Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'Ang, L. Y., … Tanaka, T. (2003). The international HapMap project. *Nature*. https://doi.org/10.1038/nature02168

Berg, J. J., & Coop, G. (2014). A Population Genetic Signal of Polygenic Adaptation. *PLoS Genet*, *10*(8), 1004412. https://doi.org/10.1371/journal.pgen.1004412

Bergstrom, C. T., & Dugatkin, L. A. (2012). *Evolution*. Norton.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., … Hirschhorn, J. N. (2004). Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *The American Journal of Human Genetics*. https://doi.org/10.1086/421051

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*. https://doi.org/10.1101/gr.154831.113

Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J. M., Mei, R., … Shriver, M. D. (2010). Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1001116

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smith, A. F. A., Roskin, K. M., … Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*. https://doi.org/10.1101/gr.1933104

Booker, T. R., Jackson, B. C., & Keightley, P. D. (2017). Detecting positive selection in the genome. *BMC Biology*. https://doi.org/10.1186/s12915-017-0434-y

Buckley, M. T., Racimo, F., Allentoft, M. E., Jensen, M. K., Jonsson, A., Huang, H., … Nielsen, R. (2017). Selection in Europeans on Fatty Acid Desaturases Associated with Dietary Changes. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msx103

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky1120

Cariaso, M., & Lennon, G. (2012). SNPedia: A wiki supporting personal genome annotation, interpretationand analysis. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkr798

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., … Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1314445110

Damgaard, P. de B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., … Willerslev, E. (2018). 137 ancient human genomes from across the Eurasian steppes. *Nature*, *557*(7705), 369–374. https://doi.org/10.1038/s41586-018-0094-2

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btr330

Demenocal, P. B., & Stringer, C. (2016). Human migration: Climate and the peopling of the world. *Nature*. https://doi.org/10.1038/nature19471

Deng, L., & Xu, S. (2018). Adaptation of human skin color in various populations. *Hereditas*. https://doi.org/10.1186/s41065-017-0036-2

Egea, R., Casillas, S., & Barbadilla, A. (2008). Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkn337

Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*. https://doi.org/10.1038/ejhg.2015.269

Feldman, M., Fernández-Domínguez, E., Reynolds, L., Baird, D., Pearson, J., Hershkovitz, I., … Krause, J. (2019). Late Pleistocene human genome suggests a local origin for the first farmers of central Anatolia. *Nature Communications*. https://doi.org/10.1038/s41467-019-09209-7

Fu, Q., Hajdinjak, M., Moldovan, O. T., Constantin, S., Mallick, S., Skoglund, P., … Pääbo, S. (2015). An early modern human from Romania with a recent Neanderthal ancestor. *Nature*. https://doi.org/10.1038/nature14558

Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., … Reich, D. (2016). The genetic history of Ice Age Europe. *Nature*. https://doi.org/10.1038/nature17993

Fumagalli, M., Moltke, I., Grarup, N., Racimo, F., Bjerregaard, P., Jørgensen, M. E., … Nielsen, R. (2015). Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science*. https://doi.org/10.1126/science.aab2319

Fumagalli, M., Vieira, F. G., Linderoth, T., & Nielsen, R. (2014). NgsTools: Methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu041

Futuyma, D. (2013). *Evolution* (3rd Editio). Massachussets: Sinauer Associates.

Gamba, C., Jones, E. R., Teasdale, M. D., McLaughlin, R. L., Gonzalez-Fortes, G., Mattiangeli, V., … Pinhasi, R. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, *5*, 5257. https://doi.org/10.1038/ncomms6257

Gelabert, P., Olalde, I., De-Dios, T., & Civit, S. (2017). Malaria was a weak selective force in ancient Europeans. *Scientific Reports*. https://doi.org/10.1038/s41598-017-01534-5

Gulsah, Merve Kılınç, G., Omrak, A., Özer Füsün, Günther, T., Metin Büyükkarakaya, A., Bıçakçı, E., … Gö therström, A. (2016). The Demographic Development of the First Farmers in Anatolia. *Current Biology*, *26*, 2659–2666. https://doi.org/10.1016/j.cub.2016.07.057

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., … Alt, K. W. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*. https://doi.org/10.1038/nature14317

Hancock, A. M., & Di Rienzo, A. (2008). Detecting the Genetic Signature of Natural Selection in Human Populations: Models, Methods, and Data. *Annual Review of Anthropology*. https://doi.org/10.1146/annurev.anthro.37.081407.085141

Harris, D. N., Ruczinski, I., Yanek, L. R., Becker, L. C., Becker, D. M., Guio, H., … O'connor, T. D. (2019). Evolution of Hominin Polyunsaturated Fatty Acid Metabolism: From Africa to the New World. *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evz071

Hinrichs, A. S., Nam, S., Cho, E., Seong, I., Limb, J.-K., Lee, S., … Hsu, F. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*.

https://doi.org/10.1093/nar/gkj144

Hofmanová, Z., Kreutzer, S., Hellenthal, G., Sell, C., Diekmann, Y., Díez-Del-Molino, D., … Burger, J. (2016). Early farmers from across Europe directly descended from Neolithic Aegeans. *PNAS*. https://doi.org/10.1073/pnas.1523951113

Hudson, R. R., Kreitman, M., & Aguadé, M. (1987). A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics*, *116*(1).

Ioannidis, J. P. A., Ntzani, E. E., & Trikalinos, T. A. (2004). "Racial" differences in genetic effects for complex diseases. *Nature Genetics*. https://doi.org/10.1038/ng1474

Iosif Lazaridis, et al. (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature*. https://doi.org/10.1038/nature19310

Itan, Y., Powell, A., Beaumont, M. A., Burger, J., & Thomas, M. G. (2009). The origins of lactase persistence in Europe. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1000491

Jeong, C., & Di Rienzo, A. (2014). Adaptations to local environments in modern human populations. *Current Opinion in Genetics and Development*. https://doi.org/10.1016/j.gde.2014.06.011

Jobling, M. A., Hurles, M., & Tyler-Smith, C. (2004). *Human evolutionary genetics : origins, peoples & disease* (2nd Editio). Garland Science.

Jun, G., Wing, M. K., Abecasis, G. R., & Kang, H. M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. https://doi.org/10.1101/gr.176552.114

Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., … Kamatani, N. (2010). Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nature Genetics*. https://doi.org/10.1038/ng.531

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M, Sugnet, C. W., Haussler, D., Kent, W. J. (2003). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkh103

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102

Key, F. M., Teixeira, J. C., de Filippo, C., & Andrés, A. M. (2014). Advantageous diversity maintained by balancing selection in humans. *Current Opinion in Genetics and Development*. https://doi.org/10.1016/j.gde.2014.08.001

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-14-289

Kryazhimskiy, S., & Plotkin, J. B. (2008). The population genetics of dN/dS. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1000304

Lamason, R. L. (2005). SLC24A5, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans. *Science*, *310*(5755), 1782–1786. https://doi.org/10.1126/science.1116238

Lamnidis, T. C., Majander, K., Jeong, C., Salmela, E., Wessman, A., Moiseyev, V., … Schiffels, S. (2018). Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nature Communications*, *9*(1), 5018. https://doi.org/10.1038/s41467-018-07483-5

Larsen, C. S. (2006). The agricultural revolution as environmental catastrophe: Implications for health and lifestyle in the Holocene. *Quaternary International*. https://doi.org/10.1016/j.quaint.2006.01.004

Latham, K. J. (2013). *Human Health and the Neolithic Revolution: an Overview ofImpacts of the Agricultural Transition on Oral Health, Epidemiology, and the Human Body*. Retrieved from http://digitalcommons.unl.edu/nebanthro/187

Laurence D, H. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in Genetics*. https://doi.org/10.1016/s0168-9525(02)02722-1

Le Corre, V., Roux, F., & Reboud, X. (2002). DNA polymorphism at the FRIGIDA gene in Arabidopsis thaliana: Extensive nonsynonymous variation is consistent with local selection for flowering time. *Molecular Biology and Evolution*. https://doi.org/10.1093/oxfordjournals.molbev.a004187

Lefterova, M. I., Suarez, C. J., Banaei, N., & Pinsky, B. A. (2015). Next-Generation Sequencing for Infectious Disease Diagnosis and Management A Report of the Association for Molecular Pathology. https://doi.org/10.1016/j.jmoldx.2015.07.004

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

Li, Heng. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btr509

Li, Heng, Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp352

López, S., van Dorp, L., & Hellenthal, G. (2015). Human dispersal out of Africa: A lasting debate. *Evolutionary Bioinformatics*. https://doi.org/10.4137/EBo.s33489

Luo, L., Boerwinkle, E., & Xiong, M. (2011). Association studies for next-generation sequencing. *Genome Research*. https://doi.org/10.1101/gr.115998.110

Makova, K., & Norton, H. (2005). Worldwide polymorphism at the MC1R locus and normal pigmentation variation in humans. *Peptides*, *26*(10), 1901–1908. https://doi.org/10.1016/j.peptides.2004.12.032

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*. https://doi.org/10.1214/aoms/1177730491

Marciniak, S., & Perry, G. H. (2017). Harnessing ancient genomes to study the history of human adaptation. *Nature Genetics*. https://doi.org/10.1038/nrg.2017.65

Martiniano, R., Cassidy, L. M., In, M., McLaughlin, R., Silva, N. M., Manco, L., … Bradley, D. G. (2017). The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1006852

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., … Reich, D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, *528*. https://doi.org/10.1038/nature16152

Mathieson, S., & Mathieson, I. (2018). FADS1 and the Timing of Human Adaptation to Agriculture. *Molecular Biology and Evolution*, *35*(12), 2957–2970. https://doi.org/10.1093/molbev/msy180

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*. https://doi.org/10.1038/351652a0

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*. https://doi.org/10.1101/gr.107524.110

McKusick, V. A. (2007). Mendelian Inheritance in Man and its online version, OMIM. *American Journal of Human Genetics*. https://doi.org/10.1086/514346

McVean, G. A. T., SR, M., S, H., P, D., DR, B., & P, D. (2004). The fine-scale structure

of recombination rate variation in the human genome. *Science*. https://doi.org/10.1126/science.1092500

Meyer, D., & Thomson, G. (2001). How selection shapes variation of the human major histocompatibility complex: a review. *Annals of Human Genetics*, *65*(Pt 1), 1–26. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11415519

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

Myles, S., Hradetzky, E., Engelken, J., Lao, O., Nürnberg, P., Trent, R. J., … Stoneking, M. (2007). Identification of a candidate genetic variant for the high prevalence of type II diabetes in Polynesians. *European Journal of Human Genetics : EJHG*, *15*(5), 584–589. https://doi.org/10.1038/sj.ejhg.5201793

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*.

Neel, J. V. (1962). Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *American Journal of Human Genetics*.

Nei, M. (1986). Definition and Estimation of Fixation Indices. *Evolution*, *40*(3), 643–645. https://doi.org/10.1111/j.1558-5646.1986.tb00516.x

Nielsen, R. (2005). Molecular Signatures of Natural Selection SNP: single nucleotide polymorphism. https://doi.org/10.1146/annurev.genet.39.073003.112420

Nielsen, R., & Slatkin, M. (2013). *An introduction to population genetics : theory and applications*. Sinauer Associates.

Olalde, I., Schroeder, H., Sandoval-Velasco, M., Vinner, L., Lobón, I., Ramirez, O., … Lalueza-Fox, C. (2015). A common genetic origin for early farmers from mediterranean cardial and central european LBK cultures. *Molecular Biology and Evolution*, *32*(12), 3132–3142. https://doi.org/10.1093/molbev/msv181

Omrak, A., Günther, T., Valdiosera, C., Svensson, E. M., Malmström, H., Kiesewetter, H., … Götherström, A. (2016). Genomic Evidence Establishes Anatolia as the Source of the European Neolithic Gene Pool. *Current Biology*. https://doi.org/10.1016/j.cub.2015.12.019

Organization, W. H. (2008). *Waist Circumference and Waist-Hip Ratio: Report of a WHO Expert Consultation*. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/44583/9789241501491_eng.pdf;jsessionid=E904ACBA81BAB10A18D9FB6B2FC34B80?sequence=1

Prüfer, K., Stenzel, U., Hofreiter, M., Pääbo, S., Kelso, J., & Green, R. E. (2010). Computational challenges in the analysis of ancient DNA. *Genome Biology*. https://doi.org/10.1186/gb-2010-11-5-r47

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq033

Racimo, F., Renaud, G., & Slatkin, M. (2016). Joint Estimation of Contamination, Error and Demography for Nuclear DNA from Ancient Humans. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1005972

Rana, B. K., Hewett-Emmett, D., Jin, L., Chang, B. H., Sambuughin, N., Lin, M., … Li, W. H. (1999). High polymorphism at the human melanocortin 1 receptor locus. *Genetics*, *151*(4), 1547–1557. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10101176

Reher, D., Key, F. M., Andrés, A. M., & Kelso, J. (2019). Immune Gene Diversity in Archaic and Present-day Humans. *Genome Biology and Evolution*, *11*(1), 232–241. https://doi.org/10.1093/gbe/evy271

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., & Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0034131

Schlebusch, C. M., Sjodin, P., Skoglund, P., & Jakobsson, M. (2013). Stronger signal of recent selection for lactase persistence in maasai than in europeans. *European Journal of Human Genetics*. https://doi.org/10.1038/ejhg.2012.199

Schraiber, J. G. (2018). Assessing the Relationship of Ancient and Modern Populations. *Genetics*. https://doi.org/10.1534/genetics.117.300448

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., Al-Rasheid, K. A. S., Willerslev, E., … Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, *13*, 178. https://doi.org/10.1186/1471-2164-13-178

Sherry, S. T., Bumpstead, S., Weyden, L. Van Der, Reinholdt, L. G., Wilming, L. G., Adams, D. J., … Deloukas, P. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*. https://doi.org/10.1093/nar/29.1.308

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., … Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*. https://doi.org/10.1101/gr.3715005

Siewert, K. M., & Voight, B. F. (2018). Bivariate Genome-Wide Association Scan Identifies 6 Novel Loci Associated With Lipid Levels and Coronary Artery Disease.

Circulation. Genomic and Precision Medicine. https://doi.org/10.1161/CIRCGEN.118.002239

Skoglund, P., Malmström, H., Raghavan, M., Storå, J., Hall, P., Willerslev, E., … Jakobsson, M. (2012). Origins and genetic legacy of neolithic farmers and hunter-gatherers in Europe. *Science*. https://doi.org/10.1126/science.1216304

Slatkin, M., Racimo, F., By, E., & Klein, R. G. (2016). *Ancient DNA and human history*. *PNAS*. https://doi.org/10.1073/pnas.1524306113

Smit AFA, Hubley R, G. P. (n.d.). RepeatMasker Open-3.0. Retrieved from http://www.repeatmasker.org

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3).

The SAM/BAM Format Specification Working Group. (2019). *Sequence Alignment/Map Format Specification*. Retrieved from https://github.com/samtools/hts-specs.

Thiltgen, G., dos Reis, M., & Goldstein, R. A. (2017). Finding Direction in the Search for Selection. *Journal of Molecular Evolution*. https://doi.org/10.1007/s00239-016-9765-5

Vallender, E. J., & Lahn, B. T. (2004). Positive selection on the human genome. *Human Molecular Genetics*, *13*(suppl_2), R245–R254. https://doi.org/10.1093/hmg/ddh253

Vatsiou, A. I., Bazin, E., & Gaggiotti, O. E. (2016). Changes in selective pressures associated with human population expansion may explain metabolic and immune related pathways enriched for signatures of positive selection. *BMC Genomics*. https://doi.org/10.1186/s12864-016-2783-2

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013a). Detecting Natural Selection in Genomic Data. *Annu. Rev. Genet*, *47*, 97–120. https://doi.org/10.1146/annurev-genet-111212-133526

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013b). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, *47*(1), 97–120. https://doi.org/10.1146/annurev-genet-111212-133526

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*. https://doi.org/10.1371/journal.pbio.0040072

Walsh, B., & Lynch, M. (2018). *Evolution and selection of quantitative traits*. Oxford University Press.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*. https://doi.org/10.2307/2408641

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt1229

Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterländer, M., … Burger, J. (2014). Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1316513111

Willing, E. M., Dreyer, C., & van Oosterhout, C. (2012). Estimates of genetic differentiation measured by fst do not necessarily require large sample sizes when using many snp markers. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0042649

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., … Carlson, C. S. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 1. https://doi.org/10.1038/s41586-019-1310-4

Wojczynski, M. K., Parnell, L. D., Pollin, T. I., Lai, C. Q., Feitosa, M. F., O'Connell, J. R., … Borecki, I. B. (2015). Genome-wide association study of triglyceride response to a high-fat meal among participants of the NHLBI Genetics of Lipid Lowering Drugs and Diet Network (GOLDN). *Metabolism: Clinical and Experimental*. https://doi.org/10.1016/j.metabol.2015.07.001

Wright, S. (1965). The Interpretation of Population Structure by F-statistics with Special Regards to System of Mating. *Evolution*, *19*(3), 395–420. https://doi.org/10.1111/j.1558-5646.1965.tb01731.x

Xu, S., Li, S., Yang, Y., Tan, J., Lou, H., Jin, W., … Jin, L. (2011). A genome-wide search for signals of high-altitude adaptation in tibetans. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msq277

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Xi Ping Cuo, Z., Pool, J. E., … Wang, J. (2010). Sequencing of Fifty Human Exomes Reveals Adaptation to High Altitude. *Science*, *329*(5987), 75–78. https://doi.org/10.1126/science.1190371

Zhang, G., Muglia, L. J., Chakraborty, R., Akey, J. M., & Williams, S. M. (2013). Signatures of natural selection on genetic variants affecting complex human traits ☆. *Applied&Translational Genomics*. https://doi.org/10.1016/j.atg.2013.10.002

Zhong, M., Lange, K., Papp, J. C., & Fan, R. (2010). A powerful score test to detect positive selection in genome-wide scans. *European Journal of Human Genetics*. https://doi.org/10.1038/ejhg.2010.60

Zhu, L., & Bustamante, C. D. (2005). A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics*. https://doi.org/10.1534/genetics.104.035097

APPENDICES

APPENDIX A

TRAIT-ASSOCIATED SNPS

Table A: The studied traits and the source database to download them with the number of SNPs that are associated with the traits

Traits	Source	SNP Count
Basophil count	GWAS	484
Blood type	SNPedia, OMIM, dbSNP	10
Body height	GWAS	770
Body mass index	GWAS	1379
Calcium measurement	GWAS	54
Eye color	GWAS	45
Fatty acid measurement	GWAS	150
Hair color	GWAS	28
Hair morphology	GWAS	18
Hemoglobin measurement	GWAS	207
High density lipoprotein (HDL)	GWAS	440

Insuline measurement	GWAS	20
Interleukin 1-beta measurement (IL-1BETA)	GWAS	24
Interleukin 2 measurement (IL-2)	GWAS	26
Interleukin 4 measurement (IL-4)	GWAS	24
Interleukin 5 measurement (IL-5)	GWAS	18
Interleukin 6 measurement (IL-6)	GWAS	68
Interleukin 7 measurement (IL-7)	GWAS	23
Interleukin 8 measurement (IL-8)	GWAS	23
Interleukin 9 measurement (IL-9)	GWAS	25
Interleukin 10 measurement (IL-10)	GWAS	46
Interleukin 12 measurement (IL-12)	GWAS	37
Interleukin 13 measurement (IL-13)	GWAS	33
Interleukin 16 measurement (IL-16)	GWAS	23
Interleukin 17 measurement (IL-17)	GWAS	34
Interleukin 18 measurement (IL-18)	GWAS	48
Lactose intolerance	SNPedia, OMIM, dbSNP	5
Low density lipoprotein (LDL)	GWAS	277
Obesity	GWAS	141
Serum IgE measurement	GWAS	27
Serum IgG measurement	GWAS	13
Skin pigmentation	GWAS	53
Total cholesterol measurement	GWAS	338

Trans fatty acid measurement	GWAS	127
Triglyceride measurement	GWAS	383
Type-1 diabetes mellitus	GWAS	232
Type-2 diabetes mellitus	GWAS	797
Vitamin B12 measurement	GWAS	24
Vitamin D measurement	GWAS	53
Waist-hip ratio	GWAS	139