COMPARISON OF METHYLATION PATTERNS IN ANCIENT HUNTER-GATHERERS AND FARMERS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SEVIM SEDA ÇOKOĞLU

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

SEPTEMBER 2021

Approval of the thesis:

COMPARISON OF METHYLATION PATTERNS IN ANCIENT HUNTER-GATHERERS AND FARMERS

submitted by **SEVIM SEDA ÇOKOĞLU** in partial fulfillment of the requirements for the degree of **Master of Science in Biology Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Ayşe Gül Gözen
Head of Department, **Biology** _____

Prof. Dr. Mehmet Somel
Supervisor, **Biology, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Pavlos Pavlidis
Institute of Computer Science, FORTH _____

Prof. Dr. Mehmet Somel
Biological Sciences, METU _____

Assist. Prof. Dr. Nihal Terzi Çizmecioğlu
Biological Sciences, METU _____

Date: 06.09.2021

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, Surname: Sevim Seda Çokoğlu

Signature :

ABSTRACT

COMPARISON OF METHYLATION PATTERNS IN ANCIENT HUNTER-GATHERERS AND FARMERS

Çokoğlu, Sevim Seda
M.S., Department of Biology
Supervisor: Prof. Dr. Mehmet Somel

September 2021, 59 pages

As the Neolithic transition dramatically changed human lifestyle and diet, one may expect epigenetic differences between pre-Neolithic Hunter-Gatherers (HGs) and Neolithic farmers (NFs).[1] In this study, we investigate methylation profile differences between the Pre-Neolithic and Neolithic individuals. It is today possible to infer methylation patterns of ancient DNA samples using programs such as epiPALEOMIX, which calculates a methylation score (MS) per CpG position from aDNA data.[2] This approach uses deamination patterns to assign a MS per CpG position in ancient DNA libraries treated by uracil-DNA glycosylase (UDG). Here we compiled published HG and NF genomes produced using UDG, either by shotgun sequencing or by 1240K SNP capture, from bone or tooth remains. After estimating MS values per genome, we applied statistical analyses to find methylation pattern differences between the subsistence types, HGs and NFs. We found 1147 genes showing significantly different MS values between prehistoric HGs and NFs. The subsistence type effect influenced more genes than sex or tissue. Intriguingly, we also observed a significant correlation between HG vs. NF methylation differences we identified in our ancient sample, and modern-day hunter-gatherer (MHG) vs. modern-day farmer

(MF) genome-wide methylation differences identified in African populations. This result raises the possibility of the existence of universal methylation patterns that may have accompanied the Neolithic transition.

Keywords: neolithic, paleolithic, methylation profile, methylation score, CpG positions, epiPaleomix, shotgun sequencing, 1240K SNP capture

ÖZ

ANTİK DÖNEM AVCI-TOPLAYICI VE ÇİFTÇİLERDE METİLASYON YOLAKLARININ KARŞILAŞTIRILMASI

Çokoğlu, Sevim Seda
Yüksek Lisans, Biyoloji Bölümü
Tez Yöneticisi: Prof. Dr. Mehmet Somel

Eylül 2021 , 59 sayfa

Neolitik geçişin insalığın yaşam tarzını ve diyetini önemli ölçüde değiştirmesiyle neolitik öncesi ve sonrası toplumlar arasında epigenetik değişikliklerin olması beklenebilir.[1] Bu çalışmada Neolitik öncesi ve Neolitik bireyler arasındaki metilasyon profili farklılıklarını bulmaya çalışmaktayız. aDNA verilerinden CpG pozisyonu başına bir metilasyon skoru (MS) hesaplayan epiPALEOMIX gibi programları kullanarak aDNA örneklerinin metilasyon modellerini çıkarmak günümüzde mümkün hale gelmiştir.[2] Bu yaklaşım, urasil-DNA glikosilaz (UDG) uygulanan antik DNA kütüphanelerinde CpG başına bir MS atamak için deaminasyon yolaklarını kullanır. Burada, UDG kullanılarak, shotgun dizilimi veya 1240K SNP yakalama yoluyla, kemik ya da diş kalıntılarından üretilen yayınlanmış avcı-toplayıcı ve neolitik çiftçi genomlarını derledik. Genom başına MS değerlerini tahmin ettikten sonra, geçim türleri, avcı-toplayıcılar ve neolitik çiftçiler arasındaki metilasyon modeli farklılıklarını bulmak için istatistiksel analizler uyguladık. Tarih öncesi avcı-toplayıcılar ve neolitik çiftçiler arasında önemli ölçüde farklı MS değerleri gösteren 1147 gen bulduk. Geçim türü etkisinin, cinsiyet veya dokudan daha fazla geni etkilediğini gördük. Şaşırtıcı bir şekilde, antik

örneğimizde tanımladığımız avcı-toplayıcı ve neolitik çiftçi metilasyon farklılıkları ile Afrika popülasyonlarında tanımlanan günümüz avcı-toplayıcı ve günümüz çiftçilerinin genom çapında metilasyon farklılıkları arasında önemli bir korelasyon gözlemledik. Bu sonuç, Neolitik geçişe eşlik etmiş olabilecek evrensel metilasyon yolaklarının mevcudiyeti olasılığını ortaya çıkarmaktadır.

Anahtar Kelimeler: neolitik, paleolitik, metilasyon profili, metilasyon skoru, CpG pozisyonları, epiPaleomix, shotgun dizileme, 1240K TNP yakalama

To my dearest family

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor Prof. Dr. Mehmet Somel for his countless help. He helped me a lot in these three years and I became a different person as I learned. My point of view of life was completely different before I started working with him. Now, I can feel my perspective has expanded in a positive way.

Also, I would like to thank my thesis committee members; Assist. Prof. Dr. Nihal Terzi Çizmecioğlu and Assoc. Prof. Dr. Pavlos Pavlidis for accepting to be a part of the committee and sharing their time.

I would also like to express my appreciation to my Compevo colleagues. All of the members had a positive influence on this study by sharing their thoughts. I would like to especially thank Fatma Rabia Fidan and Dilek Koptekin for their help and patience. I did not hesitate once to ask any question as they were always willing to help.

I also would like to express my sincere gratitude to Scientific and Technological Research Council of Turkey (TÜBİTAK) for funding this research.

My special thanks goes to my family. My mother Tevrat Çokoğlu supports me and believes in me even in the hardest times. My father Mustafa Çokoğlu who passed away a few months ago, was a cheerful man and he was always proud of me and my sisters. My elder sister Sinem Demirci always shared her experiences on academic life and made my problems easy to solve. My beloved little sister always supports me and finds a solution to every problem I face making my anxiety go away. My husband Eren Odacıoğlu is also very supportive of me and he did most of the wedding preparations by himself for me to focus on my thesis.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xiv

LIST OF FIGURES xvi

LIST OF ABBREVIATIONS xix

CHAPTERS

1 INTRODUCTION 1

2 LITERATURE REVIEW 5

2.1 Gene Expression and Epigenetic Marks 5

2.1.1 DNA Methylation 5

2.1.2 Histone Modifications 6

2.1.3 Non-coding RNA Activity 6

2.2 Environmental Factors Affecting Gene Expression 6

2.3 Ancient Genome Analyses 8

2.4 Ancient Epigenomics 9

3 MATERIALS AND METHODS 13

3.1 Presenting the Dataset 13
3.1.1 Data Selection 15
3.1.2 Data Processing for ANOVA 15
3.1.3 Data Processing for other statistical tests 17
3.2 Statistical Tests 18
3.2.1 Analysis of Variance Tests 18
3.2.2 Wilcoxon Rank-Sum Tests 19
3.2.3 Permutation Tests 19
3.2.4 Gene Ontology Enrichment Analysis 21
3.2.5 Subsistence Type-Related Methylation Differences in Ancient Eurasian vs Modern African Datasets 21
3.2.6 Methylation Differences in Shotgun vs 1240K Capture Datasets 22
4 RESULTS 25
4.1 ANOVA Tests Reveal Significant Methylation Differences on Particular Genes 27
4.2 Significant Correlation is Observed between Ancient and Modern Methylation Data Comparisons 30
4.3 Wilcoxon Rank-Sum and Permutation Tests Show No Significant Outcome 37
4.4 Comparison between the Shotgun and 1240K SNP Capture Data Shows No Significant Association 40
5 DISCUSSION 43
REFERENCES 47
APPENDICES
A GO ENRICHMENT ANALYSIS RESULTS 53

B 1240K SNP CAPTURE DATA DESCRIPTION 55

C PMD PLOTS OF SHOTGUN SEQUENCED SAMPLES 57

D MEAN METHYLATION SCORES ON CPG ISLANDS AND ACROSS NON-CPG ISLAND AREAS 59

LIST OF TABLES

TABLES

Table 3.1 Shotgun libraries used in the study. Name of the individual, the subsistence type/group it belongs to, genomic coverage related to the individual, estimated time period, region it is excavated and data source are described. * Individual removed from the analyses. 13

Table 3.2 1240K SNP capture libraries used in the study. Subsistence type of the individuals, N, the number of the individuals in the subsistence type, genomic coverage range, minimum and maximum time period spanning the subsistence type, median coverage and data source are described. . . . 14

Table 3.3 Description of the categories related to shotgun sequenced individuals. Library type, tissue DNA extracted from, genetic sex of the individual, total CpG positions observed in raw data, and total CpG positions observed after filtering for having at least 5 reads per CpG. 16

Table 4.1 Table describing the combinations of the category significance patterns and corresponding number of genes. The categories are Subsistence Type, Tissue, and Genetic Sex. The second column represents the number of genes for shotgun sequenced libraries, the third column defines the number of genes for 1240K sequenced individuals. Significant genes were defined as having BH corrected $P<0.05$ 29

Table 4.2 The count matrix of common genes for each significance level. Rows indicate ancient categories, and columns indicate modern categories. First 2 columns are related to EAD and the last 2 columns are related to WAD. 33

Table 4.3 GO terms from the analysis of genes found common with those published in the original study (EAD). Adjusted $P<0.05$. * Category including recent differentially methylated genes. 35

Table 4.4 Signifcant GO terms from the analysis of common genes found in both our data and WAD. Adjusted $P<0.05$. * Category including recent differentially methylated genes 36

Table A.1 First 5 GO terms from the analysis of group differences for shotgun sequenced samples. 53

Table A.2 First 5 GO terms from the analysis of tissue differences for shotgun sequenced samples. 53

Table A.3 First 5 GO terms from the analysis of genetic sex differences for shotgun sequenced samples. 54

Table B.1 1240K libraries used in the study. Name of the individual, the group it belongs to, genomic coverage related to the individual, total number of CpGs and data source are described. 55

LIST OF FIGURES

FIGURES

Figure 2.1 Illustration of deamination process in nucleotide conversion on methylated and unmethylated CpGs. Methylated CpGs deaminates and become TpGs. Unmethylated CpGs deaminates and converted to UpGs. UDG treatment removes unmethylated cytosines from the DNA. 10

Figure 2.2 An example sample to calculate methylation score. The algortihm decides the state of each CpG (green). The top line indicates the reference sequence, rest are the reads from an ancient DNA data. Counting the number of Ts and Cs, methylation score is attained. The starting TpGs are shown in blue and end TpGs are shown in red, as well as their complement CpAs. The figure is adopted from Hanghøj et al 2016. 11

Figure 3.1 A snippet of ANOVA input file. It contains chromosome, CpG start, CpG end, deamination status of the read (0 for unmethylated CpG reads, 1 for methylated CpG reads), tissue type, individual, genetic sex of the individual and gene the CpG overlaps. 19

Figure 3.2 Illustration of the permutation process. The input file is taken, the randomized subsistence means per CpG is calculated, subsistence differences and 5 kb window means are computed. The results are compared with the observed values and a *P*-value is calculated based on 5000 randomizations. As a final step, multiple testing correction is applied. HG: Hunter-Gatherer, FA: Farmer 20

Figure 4.1 The geographical distribution of the shotgun sequenced ancient individuals. All individuals included in this study are from Eurasia. . . . 26

Figure 4.2 Violin plot distributions of the methylation scores for every individual. The x-axis shows the individuals and the y-axis represents the methylation score. Blue dots represent the mean MS, and black dots represent the median MS per individual. 28

Figure 4.3 Heatmaps showing all combinations of categories *P*-values. Color key on the upper right represents the *P*-values. Side bars on the left show the categories. Red bar shows non-significant subsistence type, pink bar shows non-signifcant tissue and orange bar shows non-significant genetic sex *P*-values. Black lines on the bars indicate BH-corrected $P<0.05$. Left-hand-side heatmap is for shotgun sequenced libraries, and the right-hand-side one is for 1240K sequenced libraries. 31

Figure 4.4 Plots of common genes (the number of genes for EAD comparison=612, for WAD comparison=745) in both modern data which is filtered to have at least one CpG being significant per gene and our MS data filtered by significance observed in ANOVA. The x-axis shows mean difference MS for ancient data, and y-axis shows mean difference logFC on Modern data. Red lines indicate correlation between the variables. The left figure describes the EAD comparison and the right figure shows the WAD comparison. On the top of the figures, Spearman's rank correlation information is also given. 34

Figure 4.5 The correlation diagrams after filtering for the ratio of significant CpGs. First row diagrams show the correlation after (the number of significant CpGs over total number of CpGs per gene) filtering for ratio greater than 0.3, second ones show filtering for ratio greater than 0.5. The left column is for EAD, the right one is for WAD comparison. Gene names are labelled in the plots. 38

Figure 4.6 Histogram of *P*-values. The dashed line shows *P* less than 0.05. The *P*-value distribution is not continuous because of our limited sample size and the test being rank-based. 39

Figure C.1 PMD plots of Hunter-Gatherers. The plots show C to T and G to A transitions. 57

Figure C.2 PMD plots of Neolithic Farmers. The plots show C to T and G to A transitions. 58

Figure D.1 Barplots of mean methylation scores on CpG islands and across non-CpG island areas per shotgun sequenced individual. x-axis shows genomic regions defined; shelf, shore, CpG island,and open sea. y-axis shows mean MS. 59

LIST OF ABBREVIATIONS

aDNA	Ancient DNA
ANOVA	Analysis of Variance
BH	Benjamini-Hochberg
chr	Chromosome
DMS	Differentially Methylated Sites
DNA	Deoxyribonucleic acid
EAD	Eastern African Dataset
GO	Gene Ontology
HG	Hunter-Gatherer
kb	Kilobase
logFC	Logarithm of Fold Change
MF	Modern Farmer
MHG	Modern Hunter-Gatherer
MS	Methylation Score
NF	Neolithic Farmer
PMD	Post-Mortem-Damage
SNP	Single Nucleotide Polymorphism
UDG	Uracil-DNA Glycosylase
WAD	Western African Dataset

CHAPTER 1

INTRODUCTION

The Neolithic transition changed the lifestyle and diet of ancient humans tremendously.[1] Before the Neolithic transition, modern humans were hunting and gathering. Gradually, people started building settlements, farming and eventually herding animals in Southwest Asia approximately 12,000 years ago.[3] It is predicted that Neolithization arose independently around the world in different time periods. Sedentary lifestyle and transition to agriculture was caused by the variable climate conditions due to instability of the climate in early Holocene.[4] Consequently, domesticated animals replaced hunted wild animals in the diet and plant products such as bread from wheat became the base of the diet in the Near East.[4] There is archaeological evidence of grinding and processing the cultivated grains which appear as tools in archaeological remains. Another evidence comes from the bones of the individuals excavated. Signs of extra strains on the bones of the upper vertebrae were observed which imply carrying load, possibly grains and other agricultural products.[5] Also, Neolithic women had changes in the limb bones which were associated with cereal grinding.[5] Another piece of evidence comes from the dental data. According to a study, comparisons between the prehistoric HG and Neolithic human teeth reveal some distinct characteristics. Prehistoric HGs mostly have healthy teeth whereas Neolithic samples have more caries and fractures on their teeth due to the consumption of the carbohydrate-based products more than prehistoric HG humans.[6] Fractures are most probably caused by the consumption of cereals that are not perfectly grinded.[4] Meanwhile, there are other studies which claim a decrease in the dental caries of the individuals in different Neolithic settlements in Southwest Asia and Southeast Europe. [7, 8] Thus, while the topic of dental caries may remain controversial, it is expected that the transition to a Neolithic lifestyle did have various effects on the human body.

The Neolithic transition can also be studied using evidence of diet. In addition to archaeobotanical and zooarchaeological studies, human remains also reveal dietary shifts. For example, evidence for consuming dairy products has been reported approximately 8000 years ago in Çatalhöyük, Turkey.[9] Milk protein was found in both ceramics and the dental calculus of the excavated individuals. Another evidence for consuming dairy was found to be approximately 6000 years ago in Neolithic Britain.[10] Researchers used dental calculus to do proteomics analysis and discovered milk protein on the teeth of these Neolithic individuals.

An open question is whether and how this major transition in the diet and lifestyle may have affected gene expression in ancient humans. Because soft tissue and RNA are generally not preserved, the main source of epigenetic information about ancient organisms is DNA methylation, which can be inferred using the uracil-DNA glycosylase (UDG) enzyme. UDG application is a popular method in aDNA sequencing. This is an enzyme that removes uracil nucleotides which are the result of post-mortem deamination in ancient DNA.[11] Fortunately, we can estimate a methylation score per CpG directly from the sequence data coming from UDG-treated libraries thanks to its working principle. The post-mortem deamination rates are high at both termini of aDNA resulting in C $\rightarrow$ T and G $\rightarrow$ A transitions.[12] Deamination converts unmethylated cytosines into uracil and methylated cytosines into thymine nucleotides.[13] The software epiPALEOMIX uses this knowledge in order to estimate methylated and unmethylated residues in ancient genomes.[2] This is an open-source pipeline that uses post-mortem damage signals to estimate the ancient methylation patterns by using shotgun and/or capture data. One needs to input data having over 2x coverage to obtain confident results and it works efficiently on UDG-treated ancient samples. epiPALEOMIX has several modules but the MethylMap module which outputs the number of deaminated reads was used in our analyses.

Here, I present my work on methylation patterns of prehistoric HGs and NFs estimated using computational tools on published aDNA data.

In Chapter 2, a literature review on the topic is presented. Description of the related topics are given. I describe a compilation of past studies and share inferences from the literature.

In Chapter 3, materials and methods of the study are described. The dataset is defined in detail and processes that the data went through are narrated. Statistical methods executed on the data are reported in detail.

In Chapter 4, the results related to the research on methylation pattern differences in ancient DNA are presented. First, our dataset is introduced generally including some statistical metrics. Then, the results of the statistical methods are reported. Subsequently, comparisons between the datasets are given and all the results are discussed thoroughly.

In Chapter 5, I discuss the overall study and my results. The difficulties and the limits of the area are highlighted. Some points of the study are questioned and possible hypotheses are put forward. Recommendations on future applications are also provided.

CHAPTER 2

LITERATURE REVIEW

2.1 Gene Expression and Epigenetic Marks

Epigenetic changes occur continuously throughout life and they are associated with gene expression. Epigenetic marks change gene expression by switching the gene status to inactive from active and vice versa. The mechanisms work differently at different stages of life. In humans, it starts when the zygote is created and all the epigenetic marks belonging to the parents are removed from the DNA. Then, related mechanisms are activated in order for the zygote to become a fully grown human. For example, X chromosome inactivation occurs around the time of gastrulation in females, and other major epigenetic shifts occur as organs are formed during embryonesis. Epigenetic patterns are specific to tissues, developmental stage, sex and other factors affecting both the mother and the individual, including the environment. There are several mechanisms affecting gene expression including DNA methylation, histone modifications, and non-coding RNA regulations. The most widely studied epigenetic mark among them is DNA methylation.[14] The mentioned epigenetic mechanisms affecting gene expression are described in the following.

2.1.1 DNA Methylation

DNA methylation involves the addition of methyl groups on a cytosine preceding in 5'-3' direction a guanine (CpG). This context-dependent cytosine methylation is known to alter gene expression by usually repressing the gene in concern, but also, more rarely, activating transcription, with the effect depending on the location of the CpG within the gene.[15, 16] Bisulfite genomic sequencing is a popular method in-

troduced in 1992 to identify methylation patterns in modern genomes. It is based on the treatment of sodium bisulfite. The outcome of sodium bisulfite - DNA reaction differs when there is a methylated cytosine or an unmethylated cytosine.[17] According to the reactions the methylated cytosines remain as cytosines and unmethylated cytosines become uracils and are read as thymines following PCR. [18] This way, we can distinguish methylated cytosines from unmethylated ones.

2.1.2 Histone Modifications

Histones are proteins which structurally support the genome and they also play roles in epigenetic changes. Modifications on histones activate or deactivate genes.[19] Histone acetylation, phosphorylation and methylation are the categories of histone modifications and they act on the amino acid residues of histones.

2.1.3 Non-coding RNA Activity

Non-coding RNAs are not translated into proteins but influence gene expression, and thus are a major epigenetic mechanism. Two categories represent non-coding RNAs: short non-coding RNAs and long non-coding RNAs (lncRNAs). Examples of short non-coding RNAs are siRNAs, miRNAs and piRNAs. Studies on siRNAs and miRNAs showed that it has a role in gene silencing.[20, 21] piRNAs, on the other hand, binds to Piwi proteins.[22] Piwi proteins with bound piRNAs are shown to repress homeobox gene and this characteristic make piRNAs an epigenetic regulatory component. [23]

2.2 Environmental Factors Affecting Gene Expression

It was well known that both DNA methylation and gene expression patterns can be affected by environmental factors in a predictable as well as stochastic ways. Diet, physical activity, exposure to toxins, psychological trauma and other factors can influence DNA methylation patterns.[24] For instance, it has been shown that consuming broccoli can protect against several cancer types, as organosulfur compounds

which are found in broccoli regulate DNA methylation levels and activate pathways that suppress tumor growth.[24] Monozygotic twin studies are important in this area as they are genetically identical but experience life differently. Methylation studies on adult monozygotic twins show significant differences on the phenotype of the individuals.[25]

In the context of our topic, it was recently shown that there exist differentially methylated regions in the genomes of modern-day rainforest hunter-gatherers (MHGs) and modern-day farmers (MFs) from Central Africa.[26] Fagny et al. used bisulfite conversion of DNA from blood samples from these groups and obtained reliable methylation patterns using Illumina arrays. They defined two types of categorical methylation differences; "historical" differences and "recent" differences.

The "recent" differentially methylated site (DMS) category was defined by comparing two groups of populations that had the same historical subsistence type and genetic background but different recent habitat. Both populations were agriculturalists in the past; however, while one population still lived in rural, deforested areas (which the authors called w-AGR, abbreviation for Western Central African agriculturalists living in rural areas), the other population, which was genetically closely related, had recently (within the last millennia) moved to the forest and started to practice regular hunting (which the authors called f-AGR, abbreviation for forest agriculturalists). Thus, w-AGR and f-AGR were genetically closely related and shared an agriculturalist past, but recently, the f-AGR had moved to the forest. Thus, any DMS between w-AGR and f-AGR would indicate the effects of this recent environmental change.

The "historical" category, in turn, was defined by comparing two populations today living in the same habitat but had different genetic background and historical lifestyles. Both populations were living in the forest. One was f-AGR, described earlier, and represented agriculturalists who had recently turned forest hunters. The other population was composed of traditional hunter-gatherers, who had probably never become full-scale agriculturalists (which the authors called w-RHG, abbreviation for Western Central African rainforest hunter-gatherers).

Comparing these two pairs of groups, Fagny et al. reported thousands of DMS in both cases. Moreover, historical methylation differences usually involved developmental

genes, while recent differences included immune system-related genes. Age, sex, and tissue (blood cell type) composition differences were also taken into account in the study as they may affect the methylation patterns as well.

The study also included a set of independent populations who lived in Eastern Central Africa (e-AGR: Eastern agriculturalists and e-RHG: Eastern rainforest hunter-gatherers) as a control group and compared them with Western Central African populations in terms of methylation profile differences. They found similar methylation differences in Eastern and Western population subsistence type comparisons. Using these findings, they also found that the historical DMS category was enriched in methylation quantitative trait loci (meQTLs), with 30% overlap with higher association. However, the overlap for the recent category was 9%. This suggests that methylation differences between agriculturalist and hunter-gatherer groups are affected by the genetic background.

This interesting study shows that DNA methylation and other epigenetic mechanisms may be affected by subsistence type significantly.

2.3 Ancient Genome Analyses

Because aDNA is highly fragmented and in low amounts, and it is not easy to obtain and sequence in large amounts. However, the development of sequencing technologies and new laboratory protocols has allowed the study of ancient genomes. Today about 5000 ancient human genomes have been published. One important development here has been the so-called 1240K SNP capture protocol, which uses probes to target DNA fragments that contain about 1 million common human SNPs.[27] The alternative is genome-wide shotgun sequencing in which the reads are sequenced randomly and then mapped to the desired reference genome.[28] Although shotgun sequencing produces more information, it is more expensive and therefore many genomes have been produced using 1240K capture.

aDNA scientists analyse data coming from sequencing devices in order to reveal population history, such as demographic changes, migration patterns, genetic kinship between individuals, etc. Since aDNA is highly degraded, there are metrics to decide

if the sequencing was successful or not. Coverage is one the metrics and it is also used in modern genomes as well. If the sample is of good quality, then the coverage should be high too. The average coverage levels of ancient genomes usually are not as high as modern genomes but if it is sufficient to do statistical analyses for a sample, it is included in studies. Other metrics are also crucial to check, including estimates of modern human contamination. There are tools to estimate these and scientists use them to decide whether the sample is usable or not.

2.4 Ancient Epigenomics

Besides population genetics on ancient genomes, another study area includes epigenetic analyses which mainly include DNA methylation studies. Detecting methylation in aDNA is based on the deamination patterns. Post-mortem damage/deamination (PMD) is an advantage in terms of ancient methylation studies even though it may be a disadvantage for other areas. Biochemistry of methylated and unmethylated cytosine bases with time after death provides a clue on methylated CpGs. After the death of the individual, deamination of cytosines and other chemical processes occur on DNA. Thanks to deamination, we can recognize which cytosines are methylated using an enzyme which removes uracils (deaminated cytosines) on the DNA, uracil DNA glycosylase (UDG), before library preparation.[11] We know that methylated cytosines turn into thymines and unmethylated cytosines are converted to uracils. [11] After removing uracils, preparing a library and sequencing, if a read contains a TpG where we would have expected a CpG, this suggests that the CpG on that DNA molecule was originally methylated, and then got deaminated. This would not happen at an unmethylated CpG, which would be converted to UpG and then be removed from the DNA fragment by UDG. Methylation patterns are attained counting the number of TpG reads for CpGs in concern. The schema of the above explanation is shown in Figure 2.1.

In the last decade, there have been a number of studies on aDNA methylation. The first study included the retrieval of the methylation profile of 6 ancient bisons from fossilized remains using bisulphite allelic sequencing and found similar methylation patterns with modern bisons.[29] The first ancient human to be analysed epigeneti-

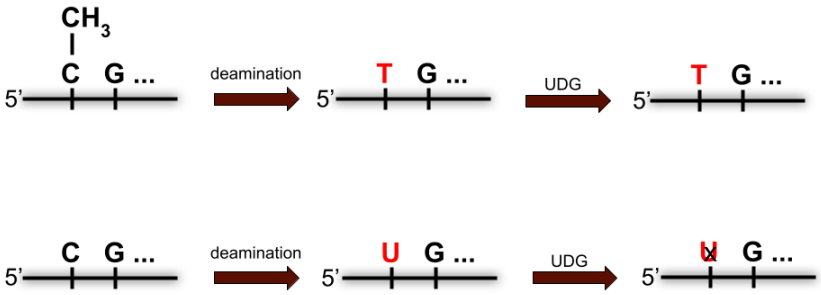

Figure 2.1: Illustration of deamination process in nucleotide conversion on methylated and unmethylated CpGs. Methylated CpGs deaminates and become TpGs. Unmethylated CpGs deaminates and converted to UpGs. UDG treatment removes unmethylated cytosines from the DNA.

cally was a Paleo-Eskimo individual related to Saqqaq culture.[13] The individual's genome had over 20x coverage and had been UDG treated. Pedersen et al. used DNA methylation patterns estimated from this genome using the above-described method, and further predicted the age-of-death of the individual to be between 44-69 years. They also predicted gene expression patterns using epigenetic analyses and generated a nucleosome map.[13] Another study investigated the differentially methylated regions between high coverage Neanderthal, Denisovan individuals, and modern humans.[15] Gokhman et al. reported that only 1% of CpG show methylation differences between archaic hominins and present-day humans. Their method included dividing the number of C to T substitutions by the total number of reads per CpG (read depth). It may be notable that this study used only one genome each per species and therefore had low power to detect differences.

A number of groups had also developed computer programs for detecting ancient methylation patterns. One of these, called epiPALEOMIX[2], utilizes the method introduced by Pedersen et al. to calculate the methylation score (MS) per CpG dinucleotide. The authors suggest to apply it on genomes with a a coverage greater than 2x. The input files consist of a BAM file belonging to the individual, reference genome of the organism and a BED file containing the regions of interest. The output

gives the total number of the deaminated reads per CpG and the total number of reads per CpG. One can use this information to calculate a ratio and carry out statistical analyses. The illustration of the above calculations are shown in Figure 2.2. The first 5' to 3' line shown in the figure is the reference sequence. The following lines are the reads for the region of interest. The last line is the reverse complement. Both the number of TpGs at read starts and the CpA's at read ends are considered while deciding on the methylation state of a CpG. The authors showed that epiPALEOMIX methylation estimates accurately captured known variation in CpG methylation across genomic regions, such as CpG islands, promoters, splice sites, and mtDNA.

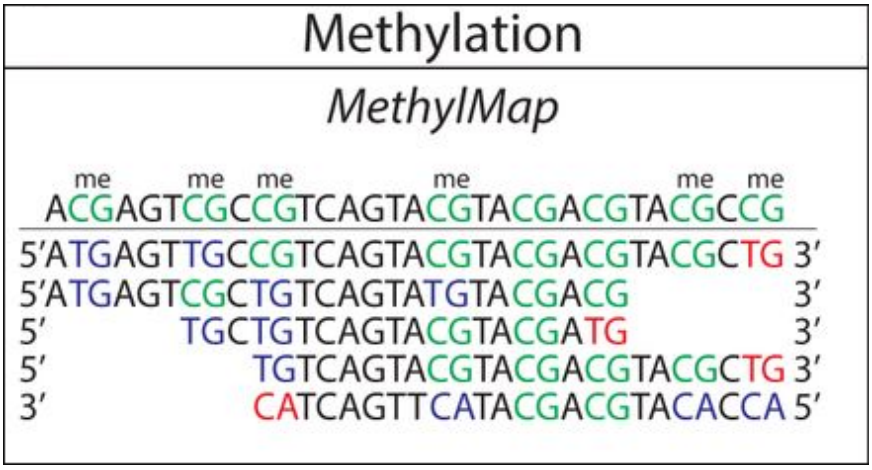

Figure 2.2: An example sample to calculate methylation score. The algortihm decides the state of each CpG (green). The top line indicates the reference sequence, rest are the reads from an ancient DNA data. Counting the number of Ts and Cs, methylation score is attained. The starting TpGs are shown in blue and end TpGs are shown in red, as well as their complement CpAs. The figure is adopted from Hanghøj et al 2016.

A more recently published software is DamMet, which uses maximum likelihood estimation to predict methylation patterns in aDNA.[30] This new method was unfortunately incompatible with our data although it has some advantages over epiPALEOMIX, such as optimizing error rates caused by overlooking some technical issues including true sequence variants in CpGs, errors while mapping and sequencing, unbalanced PMD rates on aDNA. [30] The authors report that it yields accurate results at 20x coverage.

CHAPTER 3

MATERIALS AND METHODS

3.1 Presenting the Dataset

In this study we used shotgun sequencing data of 18 ancient individuals. This contains 11 HG individuals in the context of diet from across Eurasia, and 7 NF individuals from Europe only. Sample related information is given in Table 3.1. At the same time, we used 1240K data from a total of 32 individuals including 14 HGs and 18 NFs, excavated from Eurasia.

BAM files were mapped by my colleague Kıvılcım Başak Vural on *Homo sapiens* genome assembly hs37d5. The genomic coverage range for shotgun samples was 0.86x-57.80x and the median is 4.4x. For 1240K samples coverage was 0.14x-1.53x and the median was 0.22x (note that this is genome-wide coverage, not coverage per 1240K SNP). Information related to 1240K SNP capture samples are shown in Table 3.2.

Table 3.1: Shotgun libraries used in the study. Name of the individual, the subsistence type/group it belongs to, genomic coverage related to the individual, estimated time period, region it is excavated and data source are described. * Individual removed from the analyses.

Individual	**Subsistence Type**	**Coverage**	**Time BP**	**Region**	**Source**
Ust-Ishim	HG	26.39	45,020	Russia	[31]
Motala12	HG	1.93	7,624	Sweden	[32]
Loschbour	HG	15.53	8,050	Luxemburg	[32]

Table 3.1 (continued)

K14	HG	0.86	37,470	Russia	[33]
Sf12	HG	57.80	8,895	Scandinavia	[34]
R15	HG	2.54	9,124	Italy	[35]
R7	HG	2.67	10,681	Italy	[35]
irk025	HG	8.79	4,362	Siberia	[36]
irk034*	HG	14.5	5,533	Siberia	[36]
irk061	HG	8.54	4,336	Siberia	[36]
kra001	HG	13.62	4,185	Siberia	[36]
cta016	HG	4.74	8,266	Siberia	[36]
Primrose2	NF	4.30	5,675	Ireland	[37]
Primrose16	NF	4.50	5,560	Ireland	[37]
Primrose13	NF	1.34	5,450	Ireland	[37]
Primrose9	NF	4.94	5,380	Ireland	[37]
LBK*	NF	14.27	7100	Germany	[32]
R2	NF	3.33	7,984	Italy	[35]
R3	NF	3.70	7,729	Italy	[35]
R9	NF	3.45	7,496	Italy	[35]

Table 3.2: 1240K SNP capture libraries used in the study. Subsistence type of the individuals, N, the number of the individuals in the subsistence type, genomic coverage range, minimum and maximum time period spanning the subsistence type, median coverage and data source are described.

Subsistence Type	***N***	**Coverage Range**	**Median Coverage**	**Time Period BP**	**Source**
Hunter-Gatherer	14	0.14x-1.53x	0.19x	37,470-7,376	[38, 39, 40]
Neolithic Farmer	18	0.18x-1.49x	0.43x	8,300-5,173	[39, 40, 41]

3.1.1 Data Selection

epiPALEOMIX requires genomes (i) that have been UDG-treated, (ii) that have at least 2x genomic coverage. Its input data are (i) the library type, single-stranded or double-stranded, of the sample, (ii) the path of the BAM file related to individual in concern, (iii) the path of the reference fasta file of the organism and (iv) the path of the reference BED file for CpG positions.

Although we had chosen genomes reported to be UDG-treated in their relevant publications, we still decided to check their PMD profiles before further analyses, in order to verify that UDG treatment was indeed effective on the samples. The transition of C $\rightarrow$ T at 5' termini and G $\rightarrow$ A at the 3' termini should be close to 0 if UDG treatment was successful. We calculated the PMD profiles using pmdtools on our shotgun samples, which are provided in Figure C.1 and Figure C.2 in Appendix C. [42] All the samples used meet the criteria defined above.

The CpG reference BED file (hg19) was retrieved by my colleague Dilek Koptekin from the UCSC Genome Browser using the R Bioconductor package BSgenome, and it contains a total of 13,270,411 autosomal CpG positions.

3.1.2 Data Processing for ANOVA

After running epiPALEOMIX, we filtered the output files for having at least 5 reads per CpG position (collecting rows with $>= 5$ on the total read column in the epiPALEOMIX output). This coverage filtering was done to increase the reliability of the methylation estimates. We converted epiPALEOMIX output to a binary vector, with 0's representing non-methylated reads and 1's representing methylated reads, along with the CpG position information.

Gene annotation was carried out using the Ensembl GRCh37 assembly including only promoters and exons to determine which genes each position belonged to. We preferred to use both promoters and exons because promoters-only analyses did not yield any result due to insufficient number of CpGs in the ancient genomes available (data not shown).

The input file for ANOVA testing also contains subsistence type, tissue and genetic sex information of every individual. Subsistence type and the word "group" will be used interchangeably hereafter. Table 3.3 describes the information related to the individuals in terms of their library type (single-stranded or double-stranded), tissue, genetic sex and total CpG positions observed. As seen in Table 3.3, after filtering for having at least 5 reads per CpG, the total number of CpG sites decreases significantly in the majority of samples.

Table 3.3: Description of the categories related to shotgun sequenced individuals. Library type, tissue DNA extracted from, genetic sex of the individual, total CpG positions observed in raw data, and total CpG positions observed after filtering for having at least 5 reads per CpG.

Individual	**Library Type**	**Tissue**	**Genetic Sex**	**Total CpG Position**	**Total CpG after Filtering**
Ust-Ishim	Single-Stranded	Bone	XY	12,164,144	10,546,989
Motala12	Double-Stranded	Tooth	XY	4,033,101	998
Loschbour	Double-Stranded	Tooth	XY	11,902,498	6,407,030
K14	Double-Stranded	Bone	XY	2,849,025	2,338
Sf12	Double-Stranded	Bone	XX	12,427,015	11,395,394
R15	Double-Stranded	Bone	XY	6,174,910	16,959
R7	Double-Stranded	Bone	XY	6,397,309	21,662
irk025	Double-Stranded	Tooth	XY	8,678,201	173,646

Table 3.3 (continued)

irk061	Double-Stranded	Tooth	XY	7,727,457	60,500
kra001	Double-Stranded	Tooth	XY	8,682,548	146,718
cta016	Double-Stranded	Tooth	XX	6,676,778	29,889
Primrose2	Single-Stranded	Tooth	XX	9,797,247	470,696
Primrose16	Double-Stranded	Tooth	XY	7,928,254	86,933
Primrose13	Single-Stranded	Tooth	XY	10,074,830	1,425,413
Primrose9	Single-Stranded	Tooth	XY	9,899,233	626,647
R2	Double-Stranded	Bone	XX	7,229,731	49,501
R3	Double-Stranded	Bone	XX	7,649,881	70,779
R9	Double-Stranded	Bone	XY	7,528,483	67,474

3.1.3 Data Processing for other statistical tests

Besides ANOVA, we employed several different statistical analyses to test our hypothesis. Here we used the epiPALEOMIX output, from which we calculated the proportion deaminated reads, which we term methylation score (MS), $\frac{\text{number of deaminated reads per CpG}}{\text{total number of reads per CpG}}$.

The individual epiPALEOMIX output files were generated and joined per CpG position. The generated file included chromosome, CpG start, CpG end, and MS of the individuals in matrix format.

We applied a minimum 5 read coverage criterion for the MS data. If any of the individuals had a missing value at a particular CpG, the MS value was labelled as NA for that individual.

We used the same process for calculating MS scores from the 1240K SNP capture data.

3.2 Statistical Tests

We employed a number of statistical tests in order to address the question whether there existed a difference in the average methylation patterns between the subsistence types (HG and NF), and the direction of the difference under investigation. We used tests from the R "stats" package as well as random permutations. All tests conducted were two-sided unless otherwise indicated.

3.2.1 Analysis of Variance Tests

The outputs of epiPALEOMIX filtered to have at least 5 reads per CpG for every individual were used for ANOVA testing to increase the reliability of the results. ANOVA was preferred because we wanted to test fixed and random factors affecting the outcome. All the positions are annotated using Ensembl GRCh37 assembly including only promoters and exons. An example snippet of the ANOVA input file is given in Figure 3.1. All the lines indicate reads; one CpG per individual is represented by at least 5 reads; and categorical information is also shown. The *P*-values were calculated per gene.

The R stats package aov() function was employed using the following linear model:

$$deamination \sim subsistence + tissue + genetic\_sex + Error(Individual). \tag{3.1}$$

Here *subsistence*, *tissue* and *genetic_sex* are binary fixed factors, while *Individual* is a random factor, because the individuals selected are random representatives of their

chr	start	end	deam	tissue	individual	genetic_sex	subsistence	gene
1	10000	10001	1	Bone	Ind_1	XX	HG	Gene1
1	10000	10001	1	Bone	Ind_1	XX	HG	Gene1
1	10000	10001	0	Bone	Ind_1	XX	HG	Gene1
1	10000	10001	0	Bone	Ind_1	XX	HG	Gene1
1	10000	10001	0	Bone	Ind_1	XX	HG	Gene1
1	10475	10476	1	Tooth	Ind_2	XY	NF	Gene1
1	10475	10476	1	Tooth	Ind_2	XY	NF	Gene1
.								
.								
.								

Figure 3.1: A snippet of ANOVA input file. It contains chromosome, CpG start, CpG end, deamination status of the read (0 for unmethylated CpG reads, 1 for methylated CpG reads), tissue type, individual, genetic sex of the individual and gene the CpG overlaps.

populations and multiple CpG positions per gene represent an individual, with absent data recorded as NA.

Applying this analysis we could obtain non-NA *P*-values for subsistence, tissue, genetic sex, for a total of 9450 genes. The resulting *P*-values were corrected for multiple testing using the Benjamini-Hochberg method. These *P*-values for subsistence, tissue, genetic sex calculated for the mentioned 9450 genes were plotted using the heatmap() function offered by R the programming language.

3.2.2 Wilcoxon Rank-Sum Tests

To test for MS differences between the two subsistence types, HGs and NFs, at each CpG position, we used Wilcoxon rank-sum tests via the wilcox.test() function offered by the R "stats" package. The same procedure was applied on the 1240K dataset.

3.2.3 Permutation Tests

Permutation tests are carried out for our shotgun samples in order to see whether there is a difference between the subsistence types in terms of their mean methylation scores. Custom scripts are implemented in Python3 for calculating the observed

values and expected values. First, the mean methylation scores per subsistence type group for every single position are calculated and their difference is written to a file. Then, the file is given as input to the program again and the mean differences belonging to subsistence types for 5000 bp non-overlapping windows are computed. This procedure is applied to get observed values for the permutation tests. The experimental part includes the same main steps but instead the individuals are randomized between the subsistence types (with sample sizes fixed) so that we can decide whether the difference seen in the observed data is randomly possible or not. Five thousand repetitions are produced and the output was compared to observed values. The outcome of the comparisons results in empirical *P*-values for every single window. Benjamini-Hochberg correction is applied as a last step. The illustration of the workflow is shown in Figure 3.2.

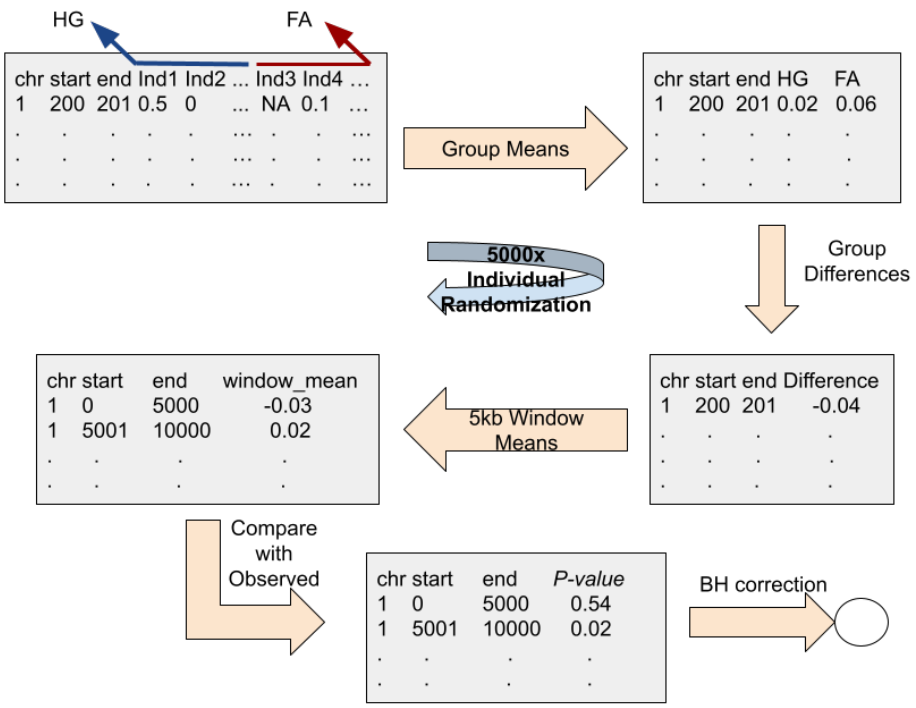

Figure 3.2: Illustration of the permutation process. The input file is taken, the randomized subsistence means per CpG is calculated, subsistence differences and 5 kb window means are computed. The results are compared with the observed values and a *P*-value is calculated based on 5000 randomizations. As a final step, multiple testing correction is applied. HG: Hunter-Gatherer, FA: Farmer

3.2.4 Gene Ontology Enrichment Analysis

Gene Ontology (GO) is a project that provides classification among the genes, gene products and sequences in terms of functional characteristics.[43] GO enrichment analysis is a technique that involves identifying the processes that are over-represented or under-represented in a set of genes by making use of Gene Ontology.[44] The corrected *P*-values from the ANOVA output were further analysed using the topGO package in the R programming environment. GOSlim GOA database was retrieved for our analysis. It contains 163 Gene Ontology terms. Adjusted $P < 0.05$ was chosen as significance threshold. Significant genes were labelled as 1 and non-significant genes as 0. The Fisher's exact test and the "elim" algorithm were chosen for running the analysis. When analysing gene sets, we chose only those passing the significance threshold a single factor (subsistence, tissue and genetic sex), but not for other factors. For example, when analysing genes showing differential methylation due to subsistence type, we selected genes showing a significant subsistence type effect, but removed those genes which also showed significance effects for tissue or genetic sex from this list. The same filters were carried out for other factors as well. We also carried out GO enrichment analysis using g:GOSt provided by g:Profiler with our adjusted significant gene sets for subsistence type, tissue type and genetic sex. The background gene set was chosen to be default.

3.2.5 Subsistence Type-Related Methylation Differences in Ancient Eurasian vs Modern African Datasets

A study published in recent years has compared the methylation profiles of the Western and Eastern African Rainforest MHGs and MFs using blood samples taken from the individuals.[26] DNA methylation was measured using a bisulphite protocol and the Infinium HumanMethylation450 BeadChip. Fagny et al. had some comparable results with our results although being performed on modern DNA. We downloaded the results of the study as a file which included adjusted *P*-values and difference logarithm of fold change (logFC) of 365,401 CpG positions, which overlap with 19,672 genes. Difference logFC values of the modern data are the fold change differences between the MF subsistence type and MHG subsistence type. This information was

used to calculate correlation between our ancient results and the results of the original study. The Fisher's exact test was used to test whether there is a nonrandom association between the categories: ancient significant, ancient non-significant; modern significant, modern non-significant. When analysing the modern data, we decided to include genes with at least one CpG being adjusted significant per gene, meaning that the adjusted *P*-value of at least one CpG per gene should be less than 0.05. The R utility function fisher.test() was used for testing enrichment.

Further, the co-directionality of farmer vs. HG methylation differences between the two datasets were tested. For this, we annotated our processed MS data and subtracted the mean value per CpG position of HGs from the mean values of the NFs. After that, the mean of the differences were calculated per gene for both datasets. We had significant *P*-values per gene for subsistence type from our ANOVA tests and by using the information provided from ANOVA results, we joined the common genes between our ancient results and the results of the original study with our processed MS data by making use of the bash utility called join. We used the Spearman's rank correlation using the R function cor.test() to the common genes between the modern data and our ancient data. Scatter plots with regression lines were plotted using R.

In further analysis, we performed the same correlation on filtered versions of the modern data, to test the reliability of the correlations we observed. For this, we calculated a proportion on modern data representing the percentage of CpGs observed to be significant. This was calculated as $\frac{\text{Number of Significant CpGs}}{\text{Total Number of CpGs}}$ per gene. The ratio generated for modern data per gene was filtered for having at least 30% and 50% of the CpGs being significant and correlation analyses were also carried out on the new versions.

3.2.6 Methylation Differences in Shotgun vs 1240K Capture Datasets

We performed the same type of Fisher's exact test-based enrichment analysis described in the previous section to compare genes identified as significantly differentially methylated in the shotgun vs in the 1240K SNP capture data. Both shotgun and 1240K processed data had the same format. Therefore, mean difference values per gene were calculated for both data type as described above. Then, significant

common genes were selected in both data types. The significance information were retrieved from the ANOVA results. Fisher's exact test was carried out to attain the association between our shotgun and 1240K categories.

CHAPTER 4

RESULTS

Our shotgun sequenced sample set had 11 HGs from West Eurasia and Siberia belonging to a period spanning 45kya-4kya, and 7 NF from West Eurasia belonging to a period spanning 8kya-5kya (coverage range: 0.86x-57.80x, median coverage: 4.4x; for more information see Table 3.1). The HG subsistence type/group also contains Siberian Bronze Age HG individuals who apparently were not farming and their diet was similar to HGs and included fish products.[36] The geographical distribution of the samples are shown in Figure 4.1. We want to note that we included all available published genomes in this study that fulfilled our criteria. A few years ago, only a few genomes that had been treated with UDG were published and it was not possible to do population analyses. In more recent years, the number of UDG-treated genomes increased, although modestly, allowing this study to be performed.

We also compiled a sample set using 1240K SNP capture data and our sample set had total of 32 individuals including 14 HGs belonging to a period spanning 38kya-7kya and 18 NFs belonging to a period spanning 9kya-5kya, excavated from Eurasia. For 1240K samples coverage was 0.14x-1.53x and the median was 0.22x (note that this is not coverage per 1240K SNP, it is genome-wide coverage). The same criteria to generate a sample set which applied to shotgun samples were applied to 1240K samples as well.

Total CpG positions identified in each genome, both before filtering (raw) and after filtering, are reported in Table 3.2. The mean number of CpGs before filtering was 8,212,258 across the 18 genomes, and after filtering this was 1,755,531, out of more than 13 million CpG positions across the genome. It is important to report the number of CpG positions per genome as it implies how much a genome contributes to the

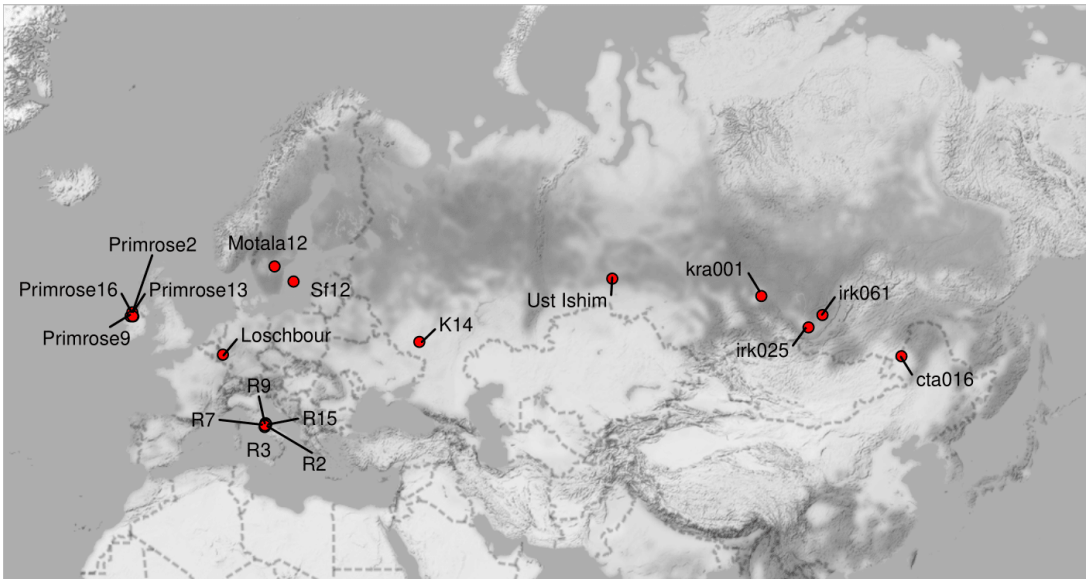

Figure 4.1: The geographical distribution of the shotgun sequenced ancient individuals. All individuals included in this study are from Eurasia.

analyses.

We estimated deaminated (potentially methylated) and non-deaminated reads covering each CpG, as well as methylation scores (MS) per genome using epiPALEOMIX software. We note that we chose to use epiPALEOMIX over DamMet, an alternative ancient methylome mapping software, as the latter requires at least 20x coverage to produce confident results.[30] Thus, we decided to use epiPALEOMIX due to the fact that our sample set includes less than 20x coverage genomes. We then studied the distribution of MSs (i.e. rates) per CpG per individual and plotted the distributions. Violin plots summarizing the filtered MS data per shotgun individual are shown in Figure 4.2. Some of the genomes had very low count of CpG positions and some were good quality genomes and had much more CpGs than the others. At first, we had 20 genomes in the shotgun sequenced sample set instead of 18 individuals. However, some genomes show unexpected behavior in terms of their methylation patterns. The LBK genome was an outlier in terms of mean MSs (Mean MS=0.05) and irk034 (Mean MS=0.006) had very low total CpG count (raw: 1,245,323, filtered: 59) even though it has sufficient coverage (14.6x). For instance, Loschbour genome had a similar coverage to irk034 and it had around 12 million for raw and around 6.5 million CpG positions for filtered category. We decided to remove LBK and irk034 genomes

from our analyses due to the fact that they act as outliers and we do not have any explanation why these genomes show such patterns.

To confirm that the genome-wide MS patterns identified are biologically meaningful we checked their distributions across CpG islands and across non-CpG island areas (e.g. "open sea") of the human genome. Non-CpG Island areas include "shelves" (4 kb from islands), "shores" (2 kb from islands), and "open sea" regions (more than 4 kb from islands). Barplots of mean methylation score on CpG islands and across non-CpG island regions per shotgun sequenced individual are shown in Figure D.1. As expected, CpG islands showed substantially lower MS than "open sea" regions in ancient shotgun genomes.

4.1 ANOVA Tests Reveal Significant Methylation Differences on Particular Genes

We next ran ANOVA on our ancient genome data in order to compare the categories in concern; subsistence type, tissue and genetic sex. These categories were chosen as gene expression depends also on tissue type and sex/gender. These information were readily available for every individual included in this study as they were collected from the literature. Processed files per shotgun sequenced individual were merged and given as input to ANOVA tests. Individual was included in the models as random factor in the ANOVA models.

A total of 9450 genes had sufficient information to perform ANOVA and to calculate *P*-values for subsistence type, tissue and genetic sex. The number of genes that were only significant for subsistence type, at BH-corrected $P<0.05$, was 1147 (12% of tested genes), while 232 genes (3%) were significant for only tissue, and 382 genes (4%) for only genetic sex, after BH correction. All combinations of subsistence type, tissue and genetic sex significance levels are reported as a heat map in Figure 4.3 and the numbers for each combination are reported in Table 4.1. We observe that there were genes that had significant differences in each category and common genes exist in terms of being significant for subsistence type, tissue and genetic sex.

The same analysis was also carried out on 1240K SNP capture dataset, which included a total of only 3730 genes. The significance patterns identified using ANOVA

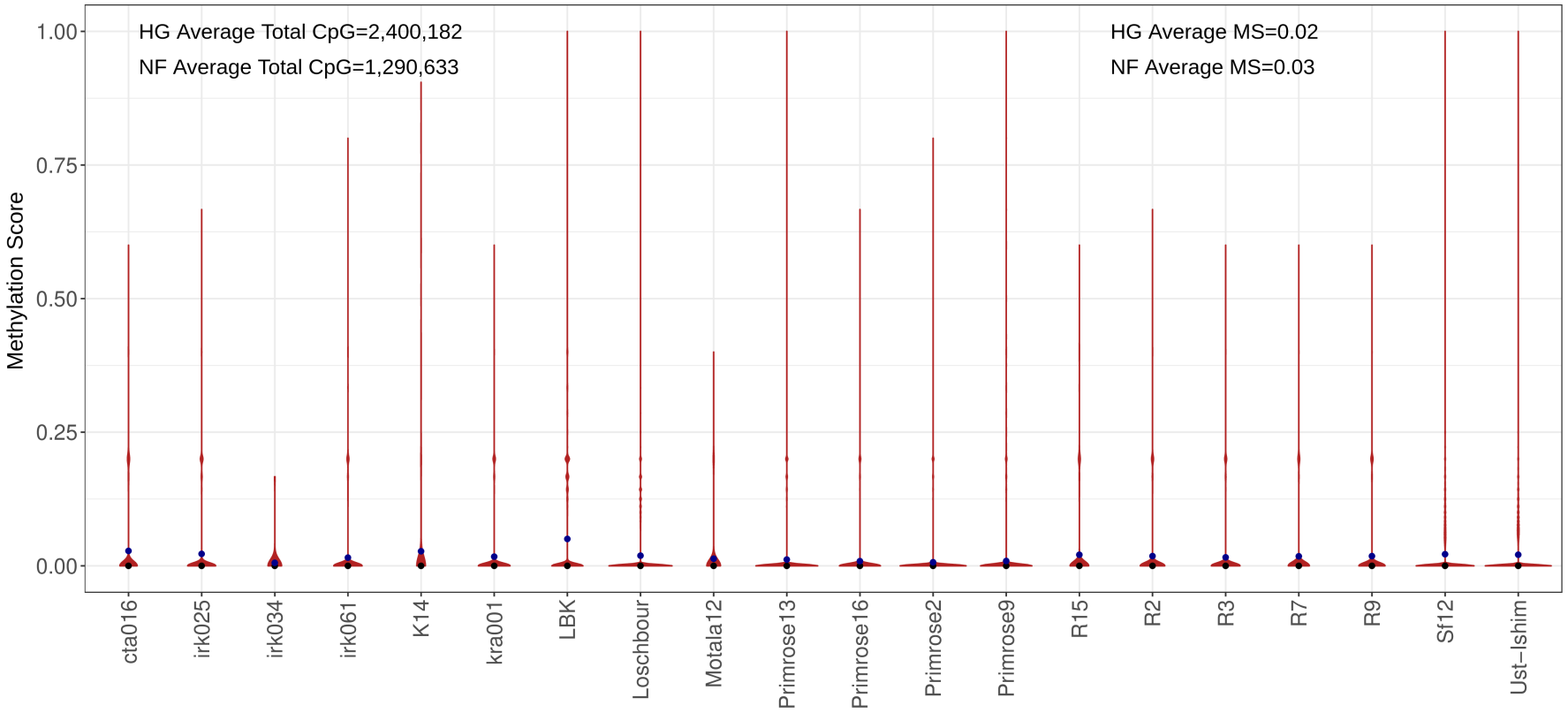

Figure 4.2: Violin plot distributions of the methylation scores for every individual. The x-axis shows the individuals and the y-axis represents the methylation score. Blue dots represent the mean MS, and black dots represent the median MS per individual.

on this dataset are also given in Table 4.1. From the table, we can see that the number of significant genes identified is much lower than those identified using the shotgun dataset. For subsistence type, only 126 genes (3%) were identified.

Table 4.1: Table describing the combinations of the category significance patterns and corresponding number of genes. The categories are Subsistence Type, Tissue, and Genetic Sex. The second column represents the number of genes for shotgun sequenced libraries, the third column defines the number of genes for 1240K sequenced individuals. Significant genes were defined as having BH corrected $P<0.05$

Significance Pattern	**Number of Genes (Shotgun)**	**Number of Genes (1240K)**
Subsistence type Non-Significant & Tissue Non-Significant & Genetic Sex Non-Significant	6492	3502
Subsistence type Non-Significant & Tissue Non-Significant & Genetic Sex Significant	382	8
Subsistence type Non-Significant & Tissue Significant & Genetic Sex Non-Significant	232	14
Subsistence type Non-Significant & Tissue Significant & Genetic Sex Significant	54	0
Subsistence type Significant & Tissue Non-Significant & Genetic Sex Non-Significant	1147	126
Subsistence type Significant & Tissue Non-Significant & Genetic Sex Significant	543	12
Subsistence type Significant & Tissue Significant & Genetic Sex Non-Significant	338	21

Table 4.1 (continued)

Subsistence type Significant & Tissue Significant & Genetic Sex Significant	262	47

Since we identified genes showing statistically significant differential methylation patterns, we thought it would be useful as a next step to perform GO Enrichment Analysis and to investigate the functional roles of the genes we found significant. The background gene pool included 9450 genes tested in ANOVA. GOSlim GOA database was used for this purpose. The analysis was tested over 1147 (subsistence type), 232 (tissue), and 382 (genetic sex) BH-corrected significant genes. No significantly enriched GO term was found for genes showing tissue and genetic sex effects, after multiple testing correction on Fisher's exact test *P*-values. There were two GO terms that passed 0.05 FDR for genes showing only subsistence type effects, and these were "biological process" and "circulatory system process" (adjusted $P<0.02$; see Appendix A for first 5 terms for all categories); "cell-cell signalling" was marginally significant at adjusted $P=0.051$. Because these GO terms were too general, we considered that they might not carry much biological meaning. We also tried GO enrichment analysis using g:GOSt provided by g:Profiler on our adjusted significant genes for subsistence type including 1147 genes, tissue type including 232 genes and genetic sex including 382 genes. We found results similar to those found in our previous GO enrichment analysis approach.

4.2 Significant Correlation is Observed between Ancient and Modern Methylation Data Comparisons

We next asked whether the putative subsistence-related methylation differences identified in ancient genomes could at least partly overlap with similar methylation differences identified in modern-day populations. For this we took advantage of a study that published blood methylation patterns in MHGs and MFs, that is, the two subsistence types that we are interested in Africa.[26]. The samples were collected from Eastern Central Africa and Western Central Africa, and each contained MHGs and MFs

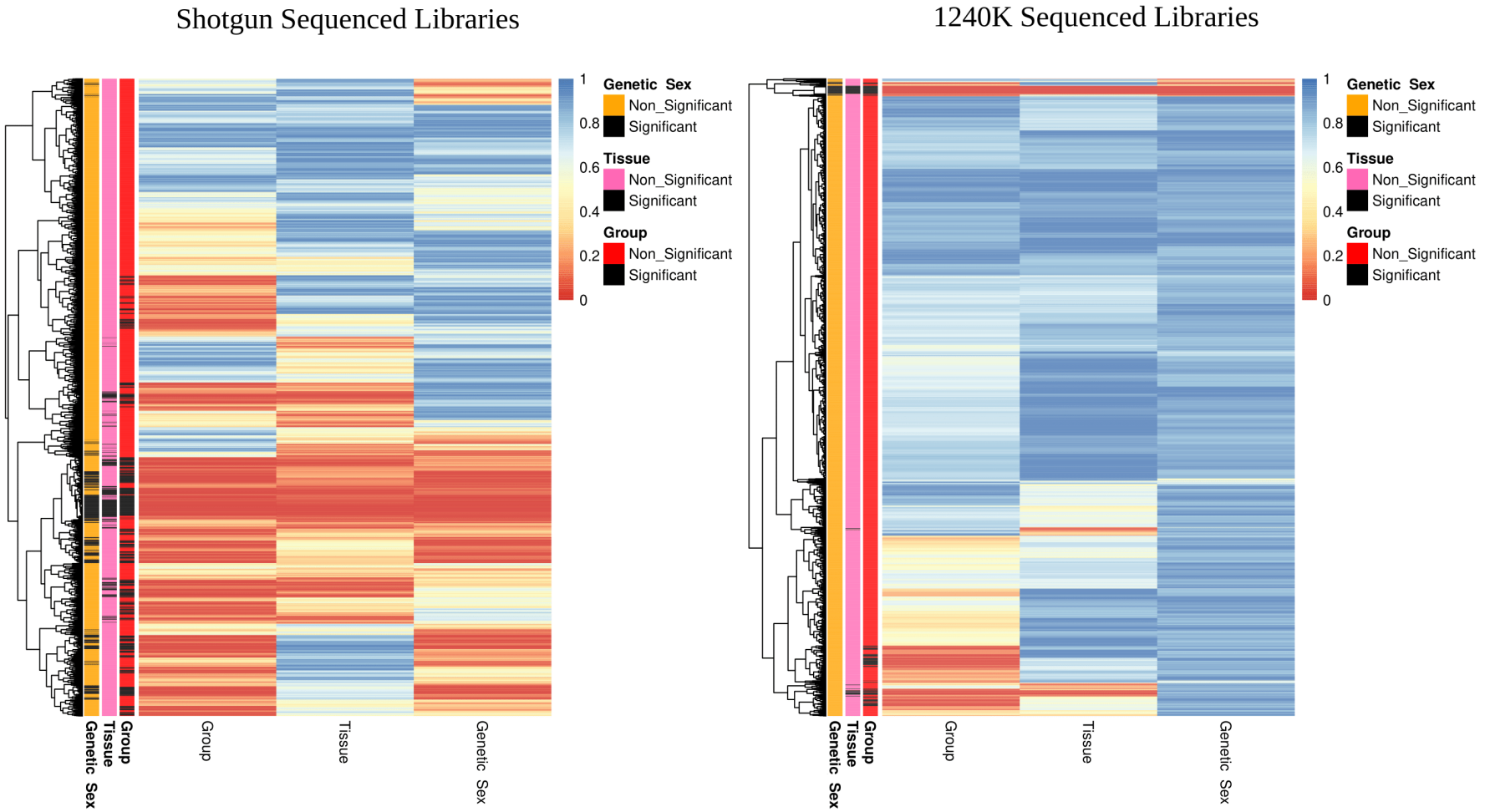

Figure 4.3: Heatmaps showing all combinations of categories *P*-values. Color key on the upper right represents the *P*-values. Side bars on the left show the categories. Red bar shows non-significant subsistence type, pink bar shows non-signifcant tissue and orange bar shows non-significant genetic sex *P*-values. Black lines on the bars indicate BH-corrected $P<0.05$. Left-hand-side heatmap is for shotgun sequenced libraries, and the right-hand-side one is for 1240K sequenced libraries.

(with sample sizes ranging from 29 to 68 per group); these datasets will be named as Eastern African Dataset (EAD) and Western African Dataset (WAD) hereafter. The authors used a linear model to estimate differential methylation.

We first compared the genes showing subsistence type effect in ANOVA in our ancient dataset, and the genes showing the same effect in the modern dataset. Specifically, we tested whether genes being significantly different between subsistence types in both datasets occur more than randomly expected. For this, we created a contingency matrix based on the subsistence type significance (tissue and genetic sex, being non-significant) of BH-corrected ANOVA results for shotgun sequenced ancient genomes, and the significant genes reported from the modern dataset. We filtered the modern data genes by having at least 1 significant CpG position per gene. The matrices are shown in Table 4.2. The number of common significantly differentially methylated genes between our ancient dataset and those in the EAD was 612. The same number for WAD was 745. To evaluate these overlaps we applied Fisher's Exact Test. This suggested a nonrandom overlap in both cases (Fisher's exact test for EAD comparison, odds ratio: 1.33, $P<10^{-4}$; for WAD, odds ratio: 1.51, $P<10^{-6}$).

This result was interesting but also surprising, as we did not expect such strong association. The reason we were sceptical was due to the fact that aDNA is highly fragmented and the estimation of methylation in ancient datasets is highly noisy, depending on PMD occurrance, and also only observable at DNA molecule ends (i.e. ends of the reads) but not in the middle nucleotides.

We next reasoned that, if the overlap signal is indeed real, we should expect the directions of the methylation differences (i.e. differences between farmers vs. hunter-gatherers in different settings) to be correlated as well. Therefore, as the next step, it would be appropriate to calculate a correlation between our ancient mean difference MS data and mean difference logFC values per gene provided by the modern data. We calculated the correlation on processed MS data filtered by the subsistence type significant genes from ANOVA results and genes which have at least one adjusted significant CpG in modern data. Spearman's rank correlation between our results and the results of the original study yields significant positive correlation (EAD comparison: ρ: 0.22, $P<10^{-7}$, n = 612 genes; WAD comparison: ρ: 0.14, $P<10^{-4}$, n = 745

genes). We also fitted the data to a linear model in order to visualize the relationship, which is shown in Figure 4.4. These results suggest a positive correlation between the ancient and modern data in terms of subsistence type methylation differences per common gene.

Table 4.2: The count matrix of common genes for each significance level. Rows indicate ancient categories, and columns indicate modern categories. First 2 columns are related to EAD and the last 2 columns are related to WAD.

Number of Genes	**Modern Significant (EAD)**	**Modern Non-Significant (EAD)**	**Modern Significant (WAD)**	**Modern Non-Significant (WAD)**
Subsistence type Significant	**612**	366	**745**	233
Subsistence type Non-Significant	2989	2381	3645	1725

We were still suspicious about the results. Hence, we examined the correlation plots and decided to remove the left hand side (MS difference $<$ -0.035) as the majority of the correlation might be resulting from these genes. Leftmost 14 genes from the EAD and 10 genes from the WAD comparison were removed from the analysis. The correlation is decreased but it does not have a major effect (Spearman's rank correlation for EAD comparison: ρ=0.19, P<10^{-5}; for WAD comparison: ρ=0.12, P<0.01). It turned out that the leftmost genes did not make up most of the correlation we observed.

As previous analyses were on modern data having at least one CpG significantly differentially methylated per gene, and one CpG might not have an impact on gene expression, we filtered the modern data to have greater than 30% of the corresponding CpGs of a gene to be significant between the MF subsistence type and the MHG subsistence type. A total of 89 and 107 genes were overlapping with our ancient data for EAD and WAD comparisons, respectively. We observed a stronger correlation in WAD comparison, the correlation remained the same for EAD and the outcome

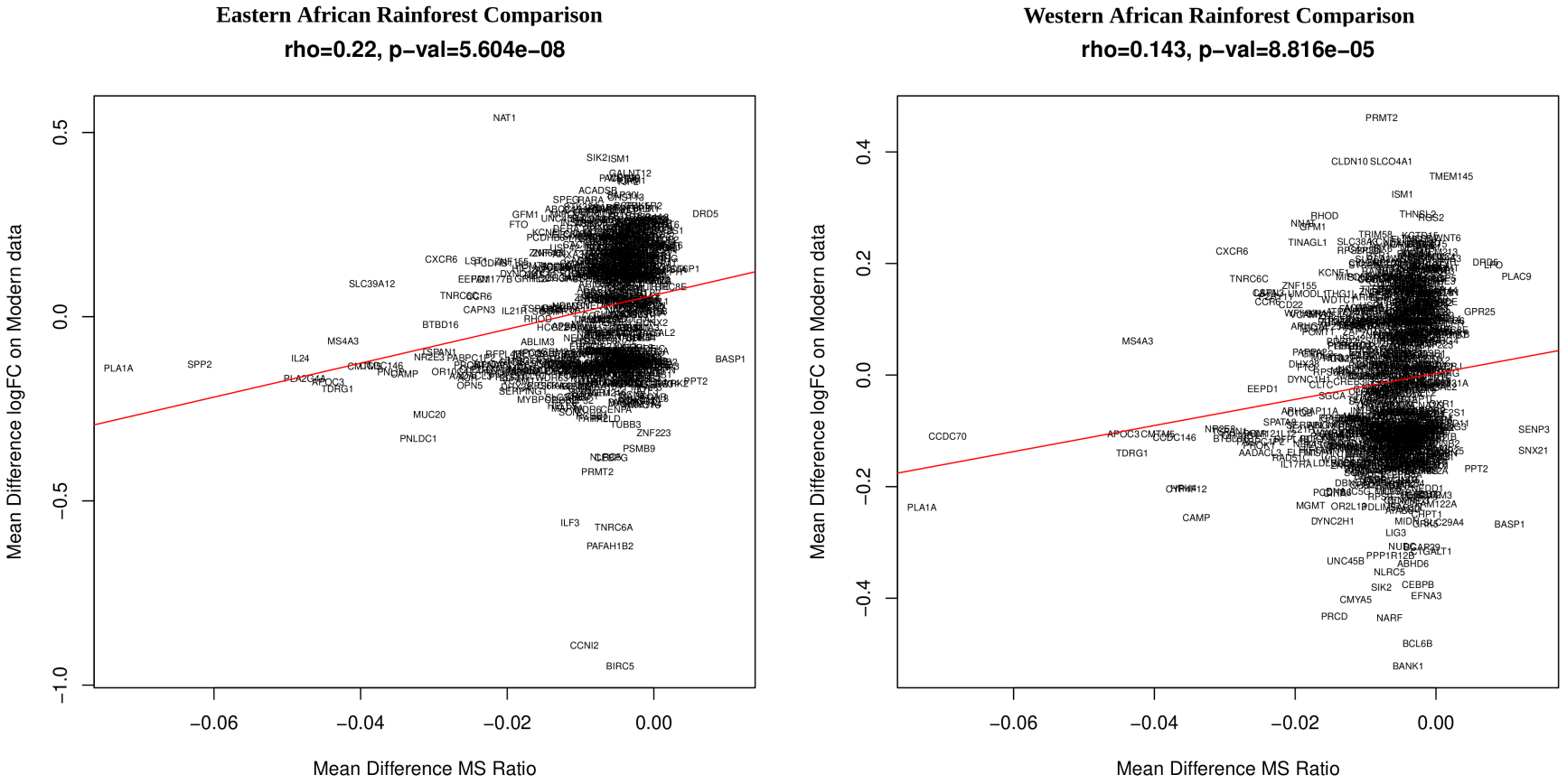

Figure 4.4: Plots of common genes (the number of genes for EAD comparison=612, for WAD comparison=745) in both modern data which is filtered to have at least one CpG being significant per gene and our MS data filtered by significance observed in ANOVA. The x-axis shows mean difference MS for ancient data, and y-axis shows mean difference logFC on Modern data. Red lines indicate correlation between the variables. The left figure describes the EAD comparison and the right figure shows the WAD comparison. On the top of the figures, Spearman's rank correlation information is also given.

was significant for both (Spearman's rank correlation for EAD comparison: ρ=0.22, P=0.04; for WAD comparison: ρ=0.25, P=0.01).

We further filtered the modern data in the same way described above, but this time the threshold was 50%. A total of 25 genes were common between EAD and our results and 29 genes were overlapping with WAD. We observed stronger correlation with EAD and the correlation turned out to be significant (Spearman's rank correlation ρ=0.5, P=0.01). However, the same result was not achieved with WAD comparison (Spearman's rank correlation ρ=0.19, P=0.37). The correlation graphs are shown in Figure 4.5 for both filters. Although we could not achieve a significant P-value for WAD filtered for having 50% significant CpGs per gene, we observed that the correlations generally increase with filtering.

The sign, positivity or negativity, of the difference values of our results were odd as ancient results shown in x-axis in the correlation graphs had mostly negative difference values. We considered this situation and had an hypothesis. The differences are calculated by subtracting the mean MS of HG subsistence type from NF subsistence type. Our hypothesis was that the HG subsistence type (mean coverage=13.03x) had greater mean coverage than the NF subsistence type (mean coverage=3.65x) in our ancient dataset. We tested this assessment with all of the individuals and found that there is no systematic difference between the coverages of the subsistence types in our ancient data (Mann-Whitney U test, P=0.14).

We also conducted GO enrichment analysis using GOSlim GOA database on significant common genes between the modern data and our ancient data. We found significant GO categories. The significant GO categories for both EAD and WAD comparisons are given in Tables 4.3 and 4.4. We compared the GO categories and found common GO terms with those published in the original study. The common categories with those published in the original study are shown in bold.

Table 4.3: GO terms from the analysis of genes found common with those published in the original study (EAD). Adjusted $P<0.05$.

* Category including recent differentially methylated genes.

Table 4.3 (continued)

GO ID	Term	Significant	Expected	Adjusted *P*-value
GO:0007267	**cell-cell signaling**	104	60.64	1.61×10^{-6}
GO:0003013	circulatory system process	50	22.03	2.5×10^{-8}
GO:0007165	signal transduction	241	201.94	0.013583
GO:0006810	transport	204	167.48	0.014263
GO:0030198	extracellular matrix organization	27	13.73	0.016572
GO:0000902	**cell morphogenesis**	56	36.26	0.016572
GO:0006464	**cellular protein modification process***	176	143.52	0.019094
GO:0030154	**cell differentiation**	173	142.16	0.00135
GO:0034330	cell junction organization	40	24.6	0.025537
GO:0048646	**anatomical structure formation involved in morphogenesis**	54	36.9	0.040306
GO:0040007	growth	49	32.82	0.040306
GO:0040011	**locomotion**	83	62.57	0.0490254
GO:0008283	cell proliferation	85	64.5	0.0490254

Table 4.4: Signifcant GO terms from the analysis of common genes found in both our data and WAD. Adjusted $P<0.05$.

* Category including recent differentially methylated genes

GO ID	Term	Significant	Expected	Adjusted *P*-value
GO:0003013	circulatory system process	53	26.67	7.2×10^{-7}
GO:0007267	**cell-cell signaling**	108	73.44	0.001141

Table 4.4 (continued)

GO:0040007	growth	64	39.75	0.003858
GO:0030154	**cell differentiation**	209	172.16	0.019968
GO:0006810	transport	238	202.81	0.043684
GO:0006464	**cellular protein modification process***	206	173.8	0.0489
GO:0030198	extracellular matrix organization	29	16.63	0.0489

Approximately half of the categories we found are common in both studies. According to the Fagny et al. study, there are genes which are differentially methylated historically or recently. The historically differentially methylated gene set is associated with the historical lifestyle differences between pairs of populations, and the recent differentially methylated gene set is related to the lifestyle changes occurred recently on the timeline of the populations.[26] We found only one term associated with recently differentially methylated GO terms. The rest (38%) are in the historical category, including cell-cell signaling, anatomical structure formation involved in morphogenesis.

4.3 Wilcoxon Rank-Sum and Permutation Tests Show No Significant Outcome

As an alternative to gene-based ANOVA, we also carried out Wilcoxon rank-sum test on each position to find out whether we could identify differentially methylated singular CpG positions using this sample. The tests on each CpG position for HG and NF subsistence types were carried out on a total of 336,149 CpG positions on filtered shotgun MS data. We found 47 positions that had $P<0.05$. No CpG position with $P<0.05$ were left after BH correction. The distribution of P-values are given in Figure 4.6. Since we could not find any significantly different position using Wilcoxon rank-sum tests, it might indicate that the differences may not be detectable at the single CpG level with our current sample size.

We further carried out custom random permutation tests for 5kb non-overlapping win-

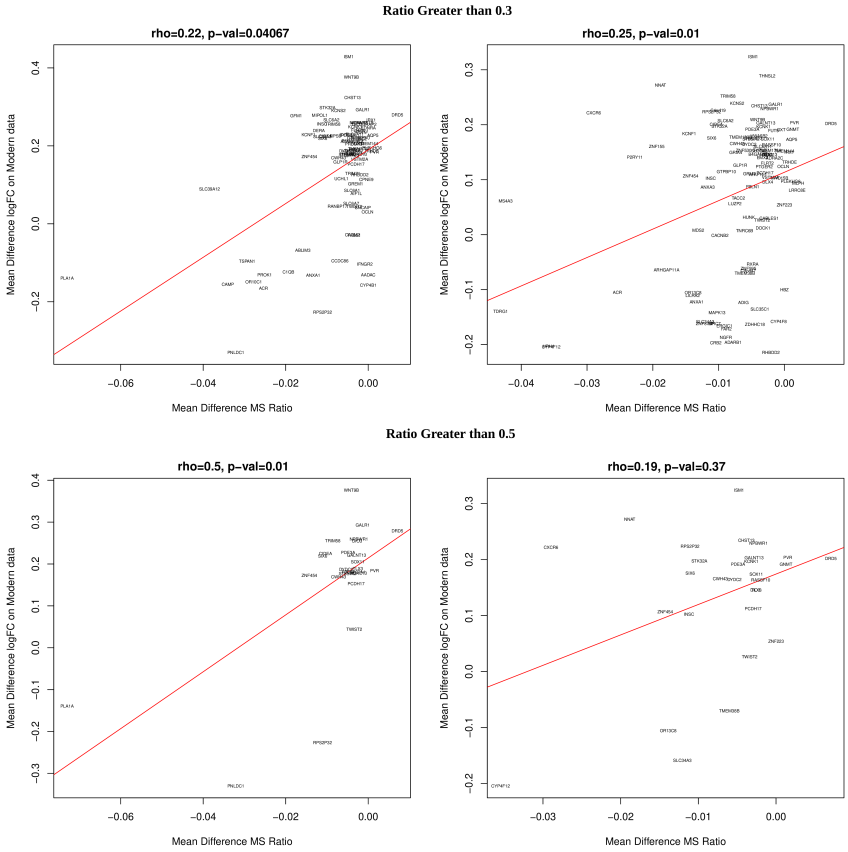

Figure 4.5: The correlation diagrams after filtering for the ratio of significant CpGs. First row diagrams show the correlation after (the number of significant CpGs over total number of CpGs per gene) filtering for ratio greater than 0.3, second ones show filtering for ratio greater than 0.5. The left column is for EAD, the right one is for WAD comparison. Gene names are labelled in the plots.

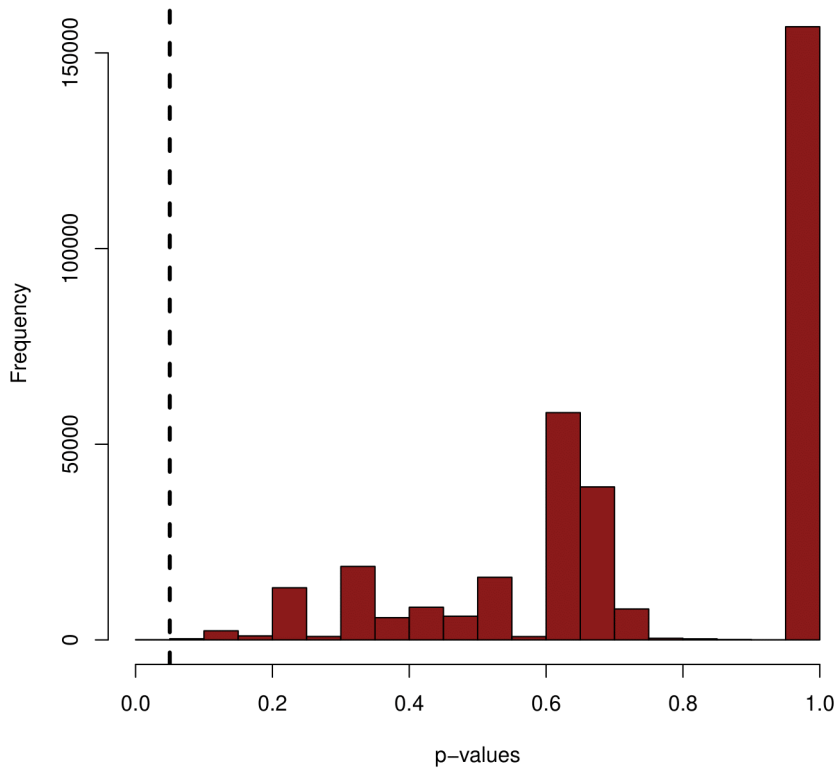

Figure 4.6: Histogram of *P*-values. The dashed line shows *P* less than 0.05. The *P*-value distribution is not continuous because of our limited sample size and the test being rank-based.

dows across the genome (n=525,112) to identify any loci showing differential methylation signals between HGs and NFs, using our shotgun ancient genome dataset. As in the previous analysis, there remained no significant 5kb window after BH correction. The reason we do not detect any significant windows in the permutation test, although we do detect significant gene loci in ANOVA, could be related to various reasons, including the sensitivity of the tests, the loci being studied (genome-wide vs. gene-focused), the total number of tests performed and the multiple testing correction effect, etc.

4.4 Comparison between the Shotgun and 1240K SNP Capture Data Shows No Significant Association

As it was mentioned, we estimated methylation levels using n=NNN 1240K SNP capture genomes, and conducted most of the analyses carried out on shotgun sequenced data with 1240K data as well. See Appendix B for more information on 1240K SNP capture dataset. We compared our shotgun data with 1240K data in the same way we compared modern data with our shotgun data. ANOVA results of both shotgun sequenced and 1240K sequenced data were compared. We observed total of 3716 common genes between the shotgun sequenced and 1240K sequenced data. Together with the previous analyses, we thought that it would be beneficial to calculate the association between the numbers of common genes per significance pattern per category between the 1240K SNP capture and shotgun sequenced results. We found a non-significant association for subsistence type significant, other categories non-significant combination (Fisher's exact test, odds ratio=0.83, P=0.32). Besides subsistence type differences, we also tested tissue and genetic sex differences. It also did not support any association between the two in those categories as well. This result was probably due to the total number of examined CpGs being much lower in 1240K sequenced data and insufficient information provided per gene. As 1240K data targets specific positions in the genome, we can only have information about the CpGs around the SNPs. This also means that their genomic coverages are generally lower than shotgun sequenced ones (Overall mean genomic coverage for 1240K data is 0.42x). Therefore, filtering for having at least 5 reads per CpG results in few CpG

positions. See Appendix B for full genomic coverage table of 1240K SNP capture data.

CHAPTER 5

DISCUSSION

Studying epigenetics has several problems even in modern genomes. Distinguishing true positive results from false positives due to sampling error and experimental artifacts is a problem in the field. DNA extraction and working principles of the chemicals used in the processes can create problems while obtaining the information needed.[45] These are general problems for most of the assays and epigenetics is one of the fields affected. Epigenetics is a complicated area of study as it includes dynamism, tissue-specificity and involves many mechanisms.[46] Furthermore, there is not a threshold in the mechanisms of epigenetics. For example, a gene can be downregulated with only 2 CpGs being methylated or all of the CpGs in the promoter region of a gene methylated. This results in a difficulty interpreting the findings in the studies. Moreover, ethical principles limit the area as it affects many other areas involving tissue collection. For instance, aging is one of the study topic in epigenetics and studying it with elderly people can involve ethical problems. If natural causes results in the death of the participants after the sample collection, the study will most likely to be blamed by the relatives of the participants.

Ancient DNA studies have their own difficulties. Generally, low quantity of DNA is extracted due to fragmentation over time.[47] Human and microbe contamination is a major problem in aDNA studies. Therefore, the facilities for aDNA extraction should be suitable to prevent contamination. Furthermore, post-mortem damage causes another difficulty as it may cause nucleotide conversion. This, in turn, provides information (about the genotype) that is not valid in actuality. Last but not least, efficient sequencing technologies are expensive and this limits ancient genome studies.[48] Studying epigenetics on aDNA is further complicated due to the fact that aDNA is

degraded and the only epigenetic marker that can efficiently be studied is methylation. Both the difficulties of epigenetics and aDNA studies together make it a hard task to study epigenetics on aDNA.

Although bisulfite conversion is the main method for studying DNA methylation in modern-day material, it cannot be used on aDNA as it requires active methyl groups on cytosines, which is not the case for aDNA due to deamination and fragmentation. We can only obtain MSs of CpGs at the molecular ends of the reads on aDNA because CpGs in the middle of the reads are rarely deaminated. As it was mentioned, deamination of the cytosines in CpGs are required in order UDG to remove unmethylated cytosines (uracils). For this reason, MS data usually have noise which makes it difficult to do statistical analyses and obtain reliable results.

Here we attempted to determine possible epigenetic differences between ancient hunter-gatherer and farmer genomes that could be attributable to subsistence type changes associated with the Neolithic transition. Although we could not find any relation using Wilcoxon rank-sum tests and permutation tests on single CpGs, we could find various significantly different genes between the HGs and NFs using ANOVA. The reason why we could not find significant results in the former tests might be due to the fact that we used more information (tissue, genetic sex etc.) while conducting ANOVA per gene. However, we should also treat the ANOVA results with caution as we conducted ANOVA on data which violates certain assumptions of ANOVA. First of all, our ANOVA input data was discrete and this violates the assumption of normal distribution of the response. Secondly, the variances between the subsistence types were not equal, which violates the assumption of equal variances. Thirdly, we did not use equal sample size, as the HG subsistence type included 11 individuals and NF subsistence type included 7 individuals. Lastly, some individuals were sequenced in the same laboratory, this might violate the assumption of independent samples.

We adopted two approaches to confirm the ANOVA results and increase the reliability. One approach is to conduct random permutation tests on the ANOVA results, but I could not finish this permutation tests within the time frame of this thesis work. In another approach, we compared our results (genes showing subsistence type effects) with another methylation dataset involving the same subsistence types. Interestingly,

we could find correlation between the modern and our ancient data. However, it was not very obvious why and how farmers and hunter-gatherers living thousands years ago in Eurasia would have similar methylation patterns to modern-day African farmers and hunter-gatherers. Indeed, the genetic background, flora and fauna should be different in Eurasia than Africa. Nevertheless, this result, if true, suggests that some dietary similarities might explain this observed correlation. It is known that Europeans arrived to Africa in the 15th century and they introduced many agricultural products such as cereals, crops etc. and domesticated animals.[49] Alternatively, common patterns between ancient Eurasian and modern-day African farmers and hunter-gatherers in the consumption of proteins and carbohydrates (e.g. higher carbohydrate amounts in farmers) might affect the outcome of these results, since the digestion of the foods results in basic biochemical products after all. As another hypothesis, the genes we found common between our results and the results of the original study might be over-expressed worldwide in farmers or hunter-gatherers. If this is the case, it might be the reason why we saw such positive correlation. Still, these are just hypotheses, and the question may be resolved with more information from anthropology, archaeobotany and other related sciences.

On the other hand, our ancient data included mostly negative difference values between HGs and NFs besides the correlation. Since the difference was taken as NFs-HGs, we thought that the negativity might be due to HGs having individuals with excessively high coverages than NFs in our dataset, which may be translated into higher MS values. It turned out that the coverages are not different between the groups. Another reason might be due to differences in technical procedures used among the laboratories producing the data, as our HG individuals were from various laboratories and NF subsistence type were derived from only 2 laboratories. More research and more data are required to address this question.

In this thesis work we also explored the possible utility of 1240K genomes, as there are plenty of UDG-treated 1240K genomes published. However, despite the larger sample size, we could not find a strong signal supporting our initial hypothesis about farmer vs. hunter-gatherer differences, as we had using the shotgun ancient genomes. The genomic coverages of the 1240K individuals are generally lower than 2x (because the procedure concentrates on only a fraction of the genome); while >2x coverage is

required for epiPALEOMIX to work properly. Possibly because of this restriction, we could not obtain any significant results other than a few genes showing differences between the subsistence types, which were most probably false-positives, and also these few genes did not overlap with the genes observed significantly different between the subsistence types in our shotgun results. We therefore believe there is not enough resolution to discover any such signal in the currently available 1240K ancient genomes. With the current pace of re-sequencing studies, it might be expected that some of these genomes will later be shotgun sequenced to reach much higher coverages, which would allow us to repeat this study with DamMet in a couple of years.

In the future, we may perform differential methylation analyses on specific genes, e.g. developmental genes or immune system related genes, that are known to be differentially expressed between the modern subsistence types, instead of using all autosomal CpGs. This would increase the stringency and reveal the true significance of the genes eliminated by the multiple testing corrections. The study can be extended more for differences in tissue-specific or genetic sex-specific genes if more UDG-treated high coverage ancient genomes are published. Also, we may try to find the genetic basis of methylation differences and compare them with adaptation signals as it was carried out by Fagny et al.

In conclusion, we report -to our knowledge- the first indication of genes differentially methylated between ancient hunter-gatherer and agriculturalist populations. These differences do not seem to be strongly enriched in a specific functional category. However, there is a relatively strong correlation between these methylation patterns identified in ancient groups from Eurasia and those reported between modern-day hunter-gatherer and agriculturalist groups from Africa, which raises the possibility of a universal subsistence type effect on the bone and blood methylome. Still, more research is needed to verify this in the future. The scarce number of shotgun genomes treated with UDG limits the study area. We are excited about the future of the epigenetic studies in aDNA and wish there will be more genomes treated with UDG in the future so that researchers can find more interesting and solid results.

REFERENCES

[1] V. G. Childe, A. Wolf, H. Pledge, G. Perazich, P. M. Field, and J. Bernal, "Man makes himself," *Science and Society*, vol. 4, no. 4, 1940.

[2] K. Hanghøj, A. Seguin-Orlando, M. Schubert, T. Madsen, J. S. Pedersen, E. Willerslev, and L. Orlando, "Fast, accurate and automatic ancient nucleosome and methylation maps with epipaleomix," *Molecular biology and evolution*, vol. 33, no. 12, pp. 3284–3298, 2016.

[3] O. Bar-Yosef and A. Belfer-Cohen, "From foraging to farming in the mediterranean levant," *Transitions to agriculture in prehistory*, pp. 21–48, 1992.

[4] M. P. Richards, "A brief review of the archaeological evidence for palaeolithic and neolithic subsistence," *European journal of clinical nutrition*, vol. 56, no. 12, pp. 1270–1278, 2002.

[5] T. Molleson, "The eloquent bones of abu hureyra," *Scientific American*, vol. 271, no. 2, pp. 70–75, 1994.

[6] T. Molleson, K. Jones, and S. Jones, "Dietary change and the effects of food preparation on microwear patterns in the late neolithic of abu hureyra, northern syria," *Journal of Human Evolution*, vol. 24, no. 6, pp. 455–468, 1993.

[7] V. Eshed, A. Gopher, R. Pinhasi, and I. Hershkovitz, "Paleopathology and the origin of agriculture in the levant," *American Journal of Physical Anthropology*, vol. 143, no. 1, pp. 121–133, 2010.

[8] A. Papathanasiou, "Health status of the neolithic population of alepotrypa cave, greece," *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, vol. 126, no. 4, pp. 377–390, 2005.

[9] J. Hendy, A. C. Colonese, I. Franz, R. Fernandes, R. Fischer, D. Orton, A. Lucquin, L. Spindler, J. Anvari, E. Stroud, *et al.*, "Ancient proteins from ceramic

vessels at çatalhöyük west reveal the hidden cuisine of early farmers," *Nature communications*, vol. 9, no. 1, pp. 1–10, 2018.

[10] S. Charlton, A. Ramsøe, M. Collins, O. E. Craig, R. Fischer, M. Alexander, and C. F. Speller, "New insights into neolithic milk consumption through proteomic analysis of dental calculus," *Archaeological and Anthropological Sciences*, vol. 11, no. 11, pp. 6183–6196, 2019.

[11] A. W. Briggs, U. Stenzel, M. Meyer, J. Krause, M. Kircher, and S. Pääbo, "Removal of deaminated cytosines and detection of in vivo methylation in ancient dna," *Nucleic acids research*, vol. 38, no. 6, pp. e87–e87, 2010.

[12] A. W. Briggs, U. Stenzel, P. L. Johnson, R. E. Green, J. Kelso, K. Prüfer, M. Meyer, J. Krause, M. T. Ronan, M. Lachmann, *et al.*, "Patterns of damage in genomic dna sequences from a neandertal," *Proceedings of the National Academy of Sciences*, vol. 104, no. 37, pp. 14616–14621, 2007.

[13] J. S. Pedersen, E. Valen, A. M. V. Velazquez, B. J. Parker, M. Rasmussen, S. Lindgreen, B. Lilje, D. J. Tobin, T. K. Kelly, S. Vang, *et al.*, "Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome," *Genome research*, vol. 24, no. 3, pp. 454–466, 2014.

[14] H.-S. Lee and Z. Herceg, "Chapter 30 - nutritional epigenome and metabolic syndrome," in *Handbook of Epigenetics (Second Edition)* (T. O. Tollefsbol, ed.), pp. 465–475, Academic Press, second edition ed., 2017.

[15] D. Gokhman, E. Lavi, K. Prüfer, M. F. Fraga, J. A. Riancho, J. Kelso, S. Pääbo, E. Meshorer, and L. Carmel, "Reconstructing the dna methylation maps of the neandertal and the denisovan," *science*, vol. 344, no. 6183, pp. 523–527, 2014.

[16] S. Mao, "Dna methylation promotes transcription," 2018.

[17] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul, "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual dna strands.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 5, pp. 1827–1831, 1992.

[18] Y. Li and T. O. Tollefsbol, "Dna methylation detection: bisulfite genomic sequencing analysis," in *Epigenetics Protocols*, pp. 11–21, Springer, 2011.

[19] L. Mariño-Ramírez, M. G. Kann, B. A. Shoemaker, and D. Landsman, "Histone structure and nucleosome stability," *Expert review of proteomics*, vol. 2, no. 5, pp. 719–729, 2005.

[20] K. V. Morris, S. W.-L. Chan, S. E. Jacobsen, and D. J. Looney, "Small interfering rna-induced transcriptional gene silencing in human cells," *Science*, vol. 305, no. 5688, pp. 1289–1292, 2004.

[21] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets," *cell*, vol. 120, no. 1, pp. 15–20, 2005.

[22] N. C. Lau, A. G. Seto, J. Kim, S. Kuramochi-Miyagawa, T. Nakano, D. P. Bartel, and R. E. Kingston, "Characterization of the pirna complex from rat testes," *Science*, vol. 313, no. 5785, pp. 363–367, 2006.

[23] H. Lin, "pirnas in the germ line," *science*, vol. 316, no. 5823, pp. 397–397, 2007.

[24] Q. A. Abdul, B. P. Yu, H. Y. Chung, H. A. Jung, and J. S. Choi, "Epigenetic modifications of gene expression by lifestyle and environment," *Archives of pharmacal research*, vol. 40, no. 11, pp. 1219–1237, 2017.

[25] M. F. Fraga, E. Ballestar, M. F. Paz, S. Ropero, F. Setien, M. L. Ballestar, D. Heine-Suñer, J. C. Cigudosa, M. Urioste, J. Benitez, *et al.*, "Epigenetic differences arise during the lifetime of monozygotic twins," *Proceedings of the National Academy of Sciences*, vol. 102, no. 30, pp. 10604–10609, 2005.

[26] M. Fagny, E. Patin, J. L. MacIsaac, M. Rotival, T. Flutre, M. J. Jones, K. J. Siddle, H. Quach, C. Harmant, L. M. McEwen, *et al.*, "The epigenomic landscape of african rainforest hunter-gatherers and farmers," *Nature communications*, vol. 6, no. 1, pp. 1–11, 2015.

[27] M. Meyer and M. Kircher, "Illumina sequencing library preparation for highly multiplexed target capture and sequencing," *Cold Spring Harbor Protocols*, vol. 2010, no. 6, pp. pdb–prot5448, 2010.

[28] R. Staden, "A strategy of dna sequencing employing computer programs," *Nucleic acids research*, vol. 6, no. 7, pp. 2601–2610, 1979.

[29] B. Llamas, M. L. Holland, K. Chen, J. E. Cropley, A. Cooper, and C. M. Suter, "High-resolution analysis of cytosine methylation in ancient dna," *PLoS One*, vol. 7, no. 1, p. e30226, 2012.

[30] K. Hanghøj, G. Renaud, A. Albrechtsen, and L. Orlando, "Dammet: ancient methylome mapping accounting for errors, true variants, and post-mortem dna damage," *GigaScience*, vol. 8, no. 4, p. giz025, 2019.

[31] Q. Fu, H. Li, P. Moorjani, F. Jay, S. M. Slepchenko, A. A. Bondarev, P. L. Johnson, A. Aximu-Petri, K. Prüfer, C. de Filippo, *et al.*, "Genome sequence of a 45,000-year-old modern human from western siberia," *Nature*, vol. 514, no. 7523, pp. 445–449, 2014.

[32] I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, *et al.*, "Ancient human genomes suggest three ancestral populations for present-day europeans," *Nature*, vol. 513, no. 7518, pp. 409–413, 2014.

[33] A. Seguin-Orlando, T. S. Korneliussen, M. Sikora, A.-S. Malaspinas, A. Manica, I. Moltke, A. Albrechtsen, A. Ko, A. Margaryan, V. Moiseyev, *et al.*, "Genomic structure in europeans dating back at least 36,200 years," *Science*, vol. 346, no. 6213, pp. 1113–1118, 2014.

[34] T. Günther, H. Malmström, E. M. Svensson, A. Omrak, F. Sánchez-Quinto, G. M. Kılınç, M. Krzewińska, G. Eriksson, M. Fraser, H. Edlund, *et al.*, "Population genomics of mesolithic scandinavia: Investigating early postglacial migration routes and high-latitude adaptation," *PLoS biology*, vol. 16, no. 1, p. e2003703, 2018.

[35] M. L. Antonio, Z. Gao, H. M. Moots, M. Lucci, F. Candilio, S. Sawyer, V. Oberreiter, D. Calderon, K. Devitofranceschi, R. C. Aikens, *et al.*, "Ancient rome: A genetic crossroads of europe and the mediterranean," *Science*, vol. 366, no. 6466, pp. 708–714, 2019.

[36] G. M. Kılınç, N. Kashuba, D. Koptekin, N. Bergfeldt, H. M. Dönertaş, R. Rodríguez-Varela, D. Shergin, G. Ivanov, D. Kichigin, K. Pestereva, *et al.*, "Human population dynamics and yersinia pestis in ancient northeast asia," *Science advances*, vol. 7, no. 2, p. eabc4587, 2021.

[37] F. Sánchez-Quinto, H. Malmström, M. Fraser, L. Girdland-Flink, E. M. Svensson, L. G. Simões, R. George, N. Hollfelder, G. Burenhult, G. Noble, *et al.*, "Megalithic tombs in western and northern neolithic europe were linked to a kindred society," *Proceedings of the National Academy of Sciences*, vol. 116, no. 19, pp. 9469–9474, 2019.

[38] Q. Fu, C. Posth, M. Hajdinjak, M. Petr, S. Mallick, D. Fernandes, A. Furtwängler, W. Haak, M. Meyer, A. Mittnik, *et al.*, "The genetic history of ice age europe," *Nature*, vol. 534, no. 7606, pp. 200–205, 2016.

[39] I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, S. A. Roodenberg, E. Harney, K. Stewardson, D. Fernandes, M. Novak, *et al.*, "Genome-wide patterns of selection in 230 ancient eurasians," *Nature*, vol. 528, no. 7583, pp. 499–503, 2015.

[40] I. Mathieson, S. Alpaslan-Roodenberg, C. Posth, A. Szécsényi-Nagy, N. Rohland, S. Mallick, I. Olalde, N. Broomandkhoshbacht, F. Candilio, O. Cheronet, *et al.*, "The genomic history of southeastern europe," *Nature*, vol. 555, no. 7695, pp. 197–203, 2018.

[41] M. Lipson, A. Szécsényi-Nagy, S. Mallick, A. Pósa, B. Stégmár, V. Keerl, N. Rohland, K. Stewardson, M. Ferry, M. Michel, *et al.*, "Parallel palaeogenomic transects reveal complex genetic history of early european farmers," *Nature*, vol. 551, no. 7680, pp. 368–372, 2017.

[42] P. Skoglund, B. H. Northoff, M. V. Shunkov, A. P. Derevianko, S. Pääbo, J. Krause, and M. Jakobsson, "Separating endogenous ancient dna from modern day contamination in a siberian neandertal," *Proceedings of the National Academy of Sciences*, vol. 111, no. 6, pp. 2229–2234, 2014.

[43] G. O. Consortium, "The gene ontology project in 2008," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D440–D444, 2008.

[44] "Go enrichment analysis."

[45] A. T. Lorincz, "The promise and the problems of epigenetic biomarkers in cancer," *Expert opinion on medical diagnostics*, vol. 5, no. 5, pp. 375–379, 2011.

[46] A. C. Carter, H. Y. Chang, G. Church, A. Dombkowski, J. R. Ecker, E. Gil, P. G. Giresi, H. Greely, W. J. Greenleaf, N. Hacohen, *et al.*, "Challenges and recommendations for epigenomics in precision health," *Nature biotechnology*, vol. 35, no. 12, pp. 1128–1132, 2017.

[47] O. Handt, M. Höss, M. Krings, and S. Pääbo, "Ancient dna: methodological challenges," *Experientia*, vol. 50, no. 6, pp. 524–529, 1994.

[48] A. Grigorenko, S. Borinskaya, N. Yankovsky, and E. Rogaev, "Achievements and peculiarities in studies of ancient dna and dna from complicated forensic specimens," *Acta Naturae*, vol. 1, no. 3 (3), 2009.

[49] J. R. McNeill, "Europe's place in the global history of biological exchange," *Landscape Research*, vol. 28, no. 1, pp. 33–39, 2003.

APPENDIX A

GO ENRICHMENT ANALYSIS RESULTS

Table A.1: First 5 GO terms from the analysis of group differences for shotgun sequenced samples.

GO ID	Term	Significant	Expected	adjusted *P*-value
GO:0008150	biological_process	1000	990.2	0.002771
GO:0003013	circulatory system process	61	38.67	0.012225
GO:0007267	cell-cell signaling	136	106.47	0.051073
GO:0005975	carbohydrate metabolic process	48	36.03	0.492124
GO:0008152	metabolic process	649	620.87	0.492124

Table A.2: First 5 GO terms from the analysis of tissue differences for shotgun sequenced samples.

GO ID	Term	Significant	Expected	adjusted *P*-value
GO:0065007	biological regulation	93	80.97	0.64
GO:0051169	nuclear transport	8	4.1	0.64
GO:0006913	nucleocytoplasmic transport	8	4.1	0.64
GO:0009987	cellular process	179	171.4	0.64

Table A.2 (continued)

GO:0050794	regulation of cellular process	82	71.55	0.64

Table A.3: First 5 GO terms from the analysis of genetic sex differences for shotgun sequenced samples.

GO ID	**Term**	**Significant**	**Expected**	**adjusted *P*-value**
GO:0008150	biological_process	315	311.6	0.796447
GO:0006091	generation of precursor metabolites and energy	15	9.32	0.796447
GO:0003013	circulatory system process	18	12.17	0.796447
GO:0043933	protein-containing complex subunit organization	39	31.01	0.796447
GO:0065003	protein-containing complex assembly	39	31.01	0.796447

APPENDIX B

1240K SNP CAPTURE DATA DESCRIPTION

Table B.1: 1240K libraries used in the study. Name of the individual, the group it belongs to, genomic coverage related to the individual, total number of CpGs and data source are described.

Individual	Group	Genomic Coverage	Total CpG	Source
I0061	Hunter-Gatherer	0.34	694,524	[39]
I0585	Hunter-Gatherer	1.02	2,024,105	[39]
I1507	Hunter-Gatherer	0.20	484,114	[39]
I4550	Hunter-Gatherer	0.18	538,597	[40]
I4551	Hunter-Gatherer	0.14	472,172	[40]
I4552	Hunter-Gatherer	0.16	460,983	[40]
I4595	Hunter-Gatherer	0.17	504,476	[40]
I4596	Hunter-Gatherer	0.19	535,831	[40]
I4873	Hunter-Gatherer	0.18	487,658	[40]
I4875	Hunter-Gatherer	0.20	523,943	[40]
I4877	Hunter-Gatherer	0.19	491,431	[40]
I4878	Hunter-Gatherer	0.20	550,262	[40]
I4880	Hunter-Gatherer	0.16	457,62	[40]
Kostenki14	Hunter-Gatherer	1.53	3,757,073	[38]
I0054	Farmer	0.98	1,476,557	[39]
I0100	Farmer	0.43	942,704	[41]
I0172	Farmer	1.49	2,260,473	[39]
I0406	Farmer	0.18	462,591	[39]
I0408	Farmer	0.67	1,575,371	[39]

Table B.1 (continued)

I0412	Farmer	0.86	2,214,574	[39]
I0707	Farmer	0.49	1,023,088	[39]
I0708	Farmer	0.37	913,824	[39]
I0709	Farmer	0.50	1,067,211	[39]
I0745	Farmer	0.47	1,248,822	[39]
I0746	Farmer	0.48	1,119,047	[39]
I1495	Farmer	0.20	498,144	[39]
I1496	Farmer	0.21	541,709	[39]
I1583	Farmer	0.43	1,294,211	[39]
I5204	Farmer	0.20	443,446	[40]
I5205	Farmer	0.19	412,994	[40]
I5207	Farmer	0.23	590,713	[40]
I5208	Farmer	0.22	580,019	[40]

APPENDIX C

PMD PLOTS OF SHOTGUN SEQUENCED SAMPLES

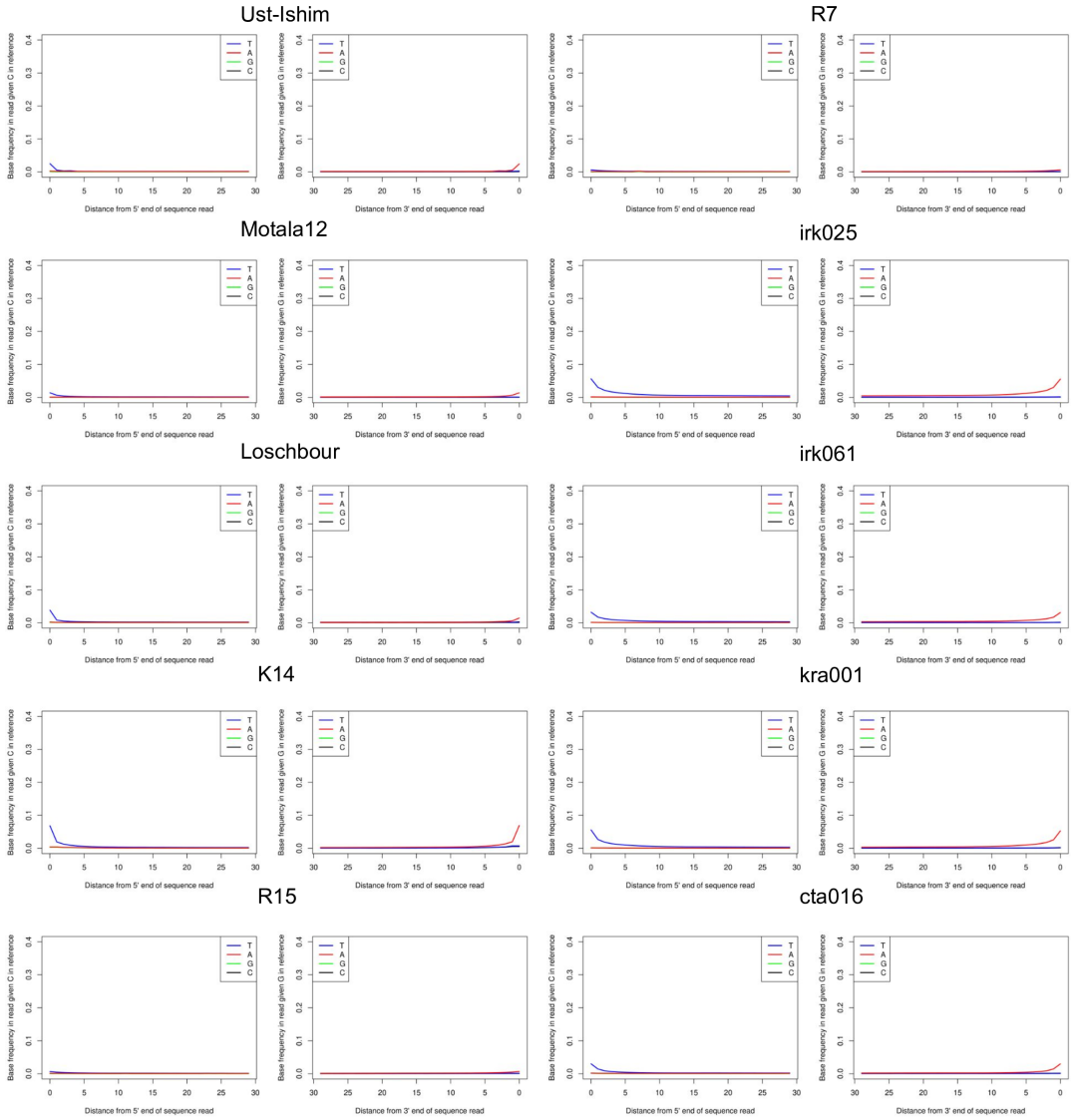

Figure C.1: PMD plots of Hunter-Gatherers. The plots show C to T and G to A transitions.

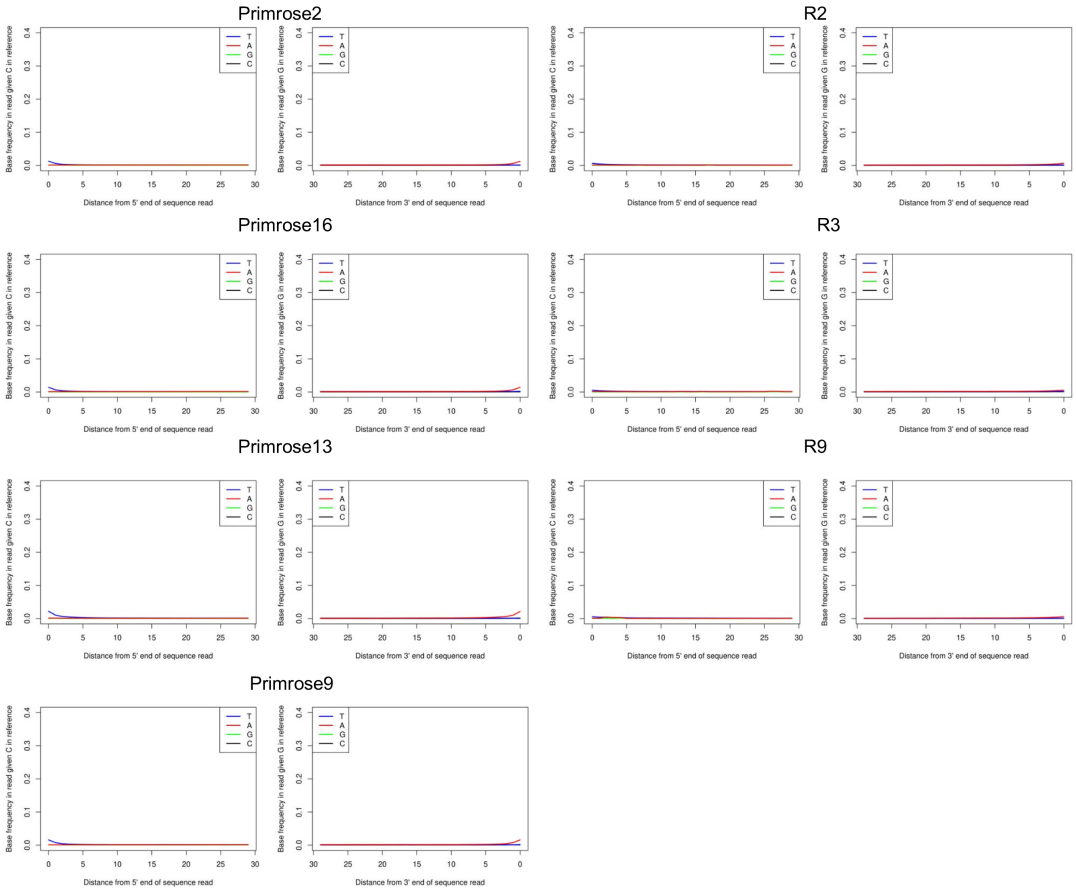

Figure C.2: PMD plots of Neolithic Farmers. The plots show C to T and G to A transitions.

APPENDIX D

MEAN METHYLATION SCORES ON CPG ISLANDS AND ACROSS NON-CPG ISLAND AREAS

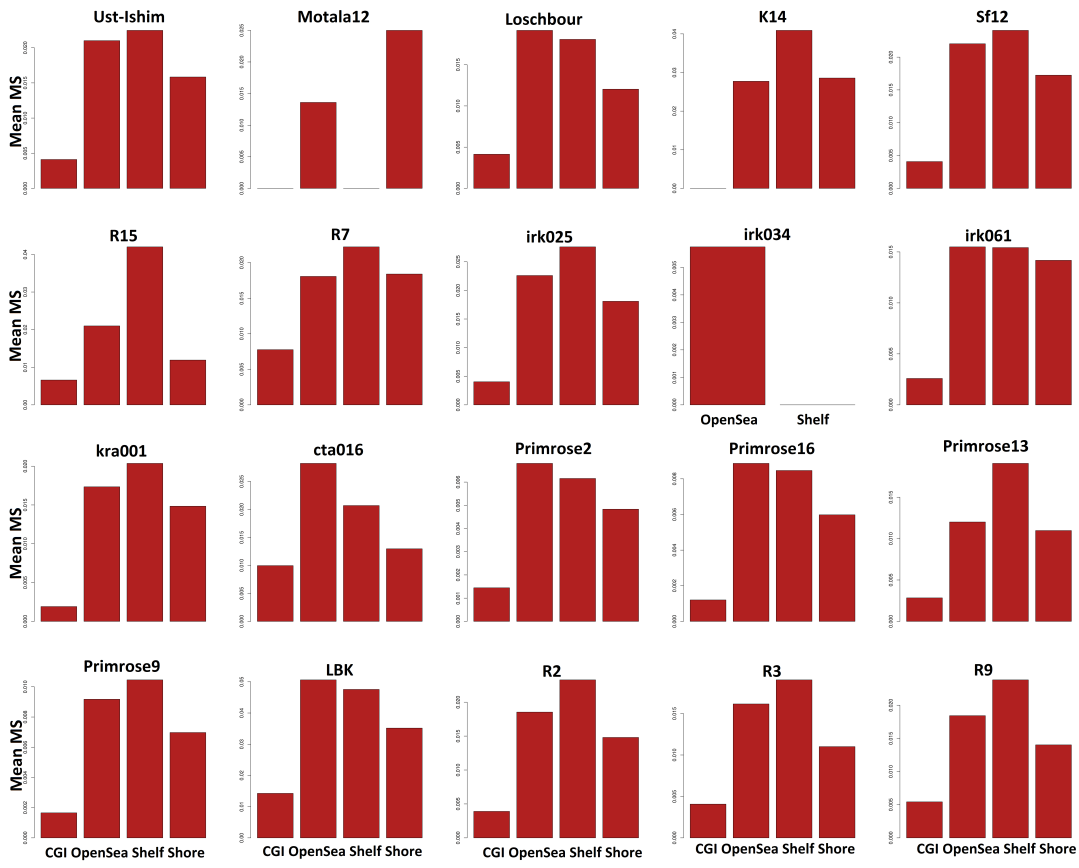

Figure D.1: Barplots of mean methylation scores on CpG islands and across non-CpG island areas per shotgun sequenced individual. x-axis shows genomic regions defined; shelf, shore, CpG island,and open sea. y-axis shows mean MS.