University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Johann Koobas

The Pleiotropic Effects of Neanderthal DNA in Present-Day Asian Populations

Mathematical Statistics

Bachelor's thesis (9 ECTS)

Supervisor: Dr. Michael Dannemann

Tartu 2023

THE PLEIOTROPIC EFFECTS OF NEANDERTHAL DNA IN PRESENT-DAY ASIAN POPULATIONS

Bachelor's thesis

Johann Koobas

Abstract

The aim of this thesis is to study the pleiotropic effects of Neanderthal DNA in present-day Asian populations. To address this research question, Neanderthal variants were scanned for pairwise phenotype associations from GWAS summary statistics from a Japanese cohort and the number of Neanderthal associations were compared to those of 1000 background sets of non-Neanderthal variants. The thesis gives a brief overview of genomics research and Neanderthals, then describes the methods, results, and discusses findings. As a result of this thesis, 10 pairs of phenotypes were identified with significantly higher numbers of Neanderthal associations compared to the background sets, suggesting that Neanderthal DNA is potentially affecting these phenotype pairs in a significant and pleiotropic manner. The results provide new insights into the functional role of Neanderthal DNA in people today.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics; B220 Genetics, cytogenetics.

Key words: Neanderthal DNA, genomics, human evolutionary genetics.

NEANDERTALI DNA PLEIOTROOPSED EFEKTID TÄNAPÄEVA AASIA RAHVASTES

Bakalaureusetöö

Johann Koobas

Lühikokkuvõte

Käesoleva bakalaureusetöö eesmärk on uurida neandertali DNA pleiotroopseid efekte tänapäeva Aasia rahvastes. Selleks eraldatakse Jaapani kohordi GWASi kokkuvõtvatest statistikatest neandertallaste päritolu haplotüübid, leitakse variandid, mis näitavad kahe fenotüübiga olulist seost, ning võrreldakse neid 1000 mitte-neandertali päritolu variantide kogumiga. Töös antakse ülevaade genoomikast ja neandertallastest, kirjeldatakse meetodeid, tulemusi ning arutletakse leidude üle. Analüüsi tulemuseks identifitseeriti 10 paari fenotüüpe, kus esines oluliselt rohkem seoseid neandertali DNA-ga kui muude variantidega, vihjates potentsiaalsele olulisele ja pleiotroopsele neandertali DNA mõjule. Tulemused arendavad hetkest arusaamu neandertali DNA rollist tänapäeva inimestes.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika; B220 Geneetika, tsütogeneetika.

Märksõnad: Neandertali DNA, genoomika, inimese evolutsiooniline geneetika.

Contents

Introduction

When Neanderthals were first discovered, they were regarded as primitive brutes with very little connection to modern humans. Now, more than a century later, it is widely accepted that Neanderthals not only resemble modern humans, but have been influencing the evolution of the modern human for thousands of years (Langdon, 2016). This has now made them very appealing to scientists, who are eager to understand their past and present impact in human evolution.

While it has been well established that due to the admixture between modern humans and Neanderthals, present-day non-Africans still carry 2% of Neanderthal DNA in their genomes, the full extent of the biological effects of this archaic DNA in its carriers is still incomplete. Currently, the research of Neanderthal DNA has mostly focussed on populations with European ancestry. However, traces of Neanderthal ancestry have been found across the globe in all non-African populations, which leaves room for many discoveries to still be made. (Green et al., 2010)

To date most phenotypic studies of Neanderthal DNA have focussed on individual phenotypes. The aim of this thesis is to analyze the pleiotropic effects of Neanderthal DNA in modern Asian populations. This is achieved by identifying pairs of phenotypes that display an association with Neanderthal DNA and comparing the number of associations for a given pair to the number of variants that are not of Neanderthal ancestry.

The thesis is divided into five sections. First, we provide some background information on genomics – we briefly cover the advancements made in the field over the last two decades and discuss some of the methods used to study genetic material, as well as their shortcomings. The description of Neanderthals and their genetic impact on modern humans as much as has been studied so far ensues. The motivation behind this thesis is addressed in this chapter as well, in which we also show differences in Neanderthal ancestry between European and Asian populations. The following chapter gives details on the methodology and data used in the analysis.

The analysis's findings are introduced and contextualized in the last two chapters with directions for further research.

The thesis benefited greatly from the contributions of Michael Dannemann, who provided extensive help and guidance throughout this project for which the author is very grateful for.

1 Genomics

In the early 2000s, a new era in genomics had begun, when one of the most remarkable scientific feats was completed – the first complete sequencing of the human genome (Green, Watson, and Collins, 2015). The human genome represents the complete set of genetic information that determines an individual's characteristics, it is comprised of approximately three billion base pairs of DNA organized into 23 pairs of chromosomes (Nurk et al., 2022). The Human Genome Project, which lasted for 13 years, dramatically altered the scientific landscape of genomics, enabling the development of groundbreaking technologies and approaches as well as creating a new way of doing science (Green, Watson, and Collins, 2015).

Nowadays, the cost of sequencing a human genome has dropped exponentially, due to advanced high-throughput technologies, like next-generation sequencing, leading to an increase in the availability of genomes to study genetic variation (McGuire et al., 2020). Over the past decade, the progress in genome sequencing and assembly has positively impacted various areas of human genetics research, such as population genomics, genetic disease mapping and diagnostics, personalized medicine initiatives, cancer research, and prenatal testing (Giani et al., 2019).

1.1 Reconstructing Genomes of Archaic Humans

The benefit of these technologies also expands to exploring ancient modern humans (Homo sapiens) and their extinct counterparts such as Neanderthals and Denisovans. Advancement in DNA extraction, sequencing, and analysis have made it possible to extract ancient DNA from bones and reconstruct the genomes of modern and archaic humans that lived thousands of years ago (Orlando et al., 2021). This task is particularly challenging as the DNA of these ancient individuals is often heavily fragmented, damaged and contaminated with microbial and present-day human DNA. The availability of genome sequences from tens of archaic

humans and thousands of ancient modern humans to date has provided invaluable insights into our evolutionary history, migration patterns, and genetic inheritance (Skoglund and Mathieson, 2018).

1.2 Genome-Wide Association Studies

Understanding the genetic variation among human populations is crucial for understanding the genetic basis of diseases and developing more effective treatments for them as well as determining which kind of medicine is best for a patient. With the availability of genome sequences in large cohorts with phenotype information, genome-wide association studies (GWASs) have established themselves as an important tool for identifying specific genetic variants associated with different diseases, traits or conditions. (Sebastiani et al., 2009)

GWASs aim to uncover genomic knowledge to understand the biology of diseases. These studies compare the genetic makeup of individuals with a particular trait or condition to that of a control group without that trait or condition, allowing researchers to pinpoint the variations that may be a contributing factor. In the last decade, GWASs have significantly improved our understanding of different diseases like type 2 diabetes, auto-immune diseases and schizophrenia. (Visscher et al., 2017)

While GWASs are excellent for identifying these variations, also called single-nucleotide polymorphisms (SNPs), associated with certain diseases, it is important to note that an association does not necessarily refer to causation. Due to the fact that SNPs are generally not independent from one another and are frequently linked to other SNPs, it can be challenging to identify the precise SNP that is responsible for a particular disease or trait (Gibson, 2012). These connected groups of SNPs, which are likely to be inherited together, are often called haplotypes. The concept of SNPs being associated with other SNPs at different loci is referred to as linkage disequilibrium (LD). Nevertheless, insights from GWAS are often a

valuable source to identify genetic variation that can then be subject to functional testing for biological effects.

Despite the success that GWASs have seen, they are not without their limitations. For example, currently these studies are heavily biased towards populations of European descent. This is a result of most GWASs conducted with populations of European ancestry, leading to an under-representation of genetic variants lacking in European populations as well as not accounting for differences of LD between populations. As a result, findings in GWASs often cannot be accurately translated to populations of non-European ancestry (Gasperini et al., 2019). In this thesis, we seek to address this knowledge gap and analyze GWAS summary statistics from the Asian cohort provided by BioBank Japan (Sakaue et al., 2021).

2 Background of Neanderthals

2.1 Historical Background

Neanderthals (Homo neanderthalensis) are an extinct species that inhabited Europe and parts of Asia between 400 000 and 40 000 years ago (Green et al., 2010). They existed around the same time as Homo sapiens, who first appeared in Africa about 300 000 years ago (Hublin et al., 2017). Neanderthals and modern humans shared a common ancestor between 550 000 and 765 000 years ago, and the two lineages separated after Neanderthals moved into Eurasia (Prüfer et al., 2013).

Neanderthals had several unique physical characteristics, were skilled hunters and scavengers, and had complex social lives. They had a robust and heavily muscled body build, with a stocky torso and shorter limbs relative to body size (Stewart et al., 2018). These physical characteristics were likely an adaptation to living in cold environments. They also had a larger brain size than modern humans, although this is likely due to their larger overall body size (Holloway, 1981). Neanderthals likely lived in small, tribal groups and there is evidence that Neanderthals took care of those who were ill or injured within their group, which suggests that they displayed empathy and compassion (Trinkaus and Villotte, 2017). They may have had language and culture, although the extent of these traits is still debated (Dediu and Levinson, 2018). Moreover, studies have revealed that Neanderthals exhibited cultural behaviors such as burying their dead and possibly even creating art (Pettitt, 2002; Vidal et al., 2014).

The reasons for the disappearance of Neanderthals around 40 000 years ago remain a subject of debate among scientists. Some theories suggest that they could not compete with the more technologically advanced Homo sapiens, who arrived in Europe around the same time (Higham et al., 2014). Others suggest that interbreeding between Neanderthals and Homo sapiens 47 000 to 65 000 years ago led to their eventual integration into the Homo sapiens population (Sankararaman

et al., 2012). Climatic changes, disease, and resource competition might have also contributed to their decline. While the exact cause of their extinction remains uncertain, the legacy of Neanderthals endures through the 1-2% of Neanderthal DNA found in the genomes of non-African modern humans (Green et al., 2010).

2.2 The Effect of Neanderthal DNA on Modern Humans

The sequencing of the Neanderthal genome in 2010 sparked a renewed interest in understanding the biological and cultural impact of these extinct hominids. It is now widely accepted that Neanderthals and anatomically modern humans interbred, resulting in 1-2% of Neanderthal DNA in the genomes of non-African populations (Green et al., 2010). This genetic admixture has been linked to several aspects of human physiology, immunity, and behavior, highlighting the impact of Neanderthal DNA on modern human populations.

2.2.1 Physiological Effects

One of the most significant physiological effects of Neanderthal DNA is its influence on skin and hair characteristics. Researchers have identified some genetic variants, that are of Neanderthal ancestry, associated with hair and skin traits in present-day humans, which are thought to have been advantageous in adapting to the colder environments of Europe and Asia (Sankararaman et al., 2014). These traits include increased keratin filament production, which contributes to thicker hair and tougher skin and provides better insulation and protection against the cooler climate (Dannemann and Kelso, 2017).

2.2.2 Behavioral and Cognitive Effects

Neanderthal-origin DNA may also have had an impact on the psychological side of modern humans, some studies have found evidence of Neanderthal variants affecting behavioral and cognitive abilities. Neanderthal-derived genetic variants have been found in genes associated with mood regulation, sleep patterns, and circadian rhythms (Simonti et al., 2016). These variations may influence how individuals feel about themselves, their quality of sleep, and how much energy they have during the day, although their precise effects on modern human behavior remain unclear.

In addition, genetic variations originating from Neanderthals have been found in genes associated with brain function and development (Vernot et al., 2016). Some of these variants have been associated with increased risk for neurological disorders, such as schizophrenia and autism (Srinivasan et al., 2015). However, as with aforementioned behavioral effects of Neanderthal DNA on modern humans, the exact nature by which Neanderthal genes influence modern human cognition and neural functioning are still not well understood.

2.2.3 Immunity and Infections

The admixture of genetic material between Neanderthals and modern humans has also been associated with the development of the human immune system. For example, Neanderthal genes have been found to play a role in shaping the human leukocyte antigen (HLA) system, which is responsible for immune response regulation (Abi-Rached et al., 2011). This enhanced immune response might have been advantageous in fighting off infections encountered by earlier human populations as they migrated across the globe. However, this heightened immune sensitivity may also increase the likelihood of developing autoimmune diseases and allergic reactions in contemporary populations (Dannemann, Prüfer, and Kelso, 2017; Dannemann, Andrés, and Kelso, 2016). Modern humans may have benefited from the introduction of Neanderthal HLA genes by expanding the range of pathogens that could be

recognized and countered by the immune system.

Genetic variations with Neanderthal ancestry may also increase vulnerability to some viral diseases. For example, research has shown that Neanderthal genes could increase the likelihood of developing severe COVID-19 symptoms (Zeberg and Pääbo, 2020). This finding suggests that while Neanderthal DNA may have transferred some adaptive advantages to our ancestors, it could also have left modern human populations vulnerable to certain diseases and infections.

2.3 Differences of Neanderthal Legacy in Modern Humans

It is estimated that 40% of Neanderthal DNA is found in people across the globe today (Dannemann et al., 2020). However, a large proportion of Neanderthal DNA is either not shared or displays large allele frequency (AF) differences between populations from different continents. As demonstrated in Figure 1, populations within Asia and Europe show high frequency similarity of Neanderthal variants.

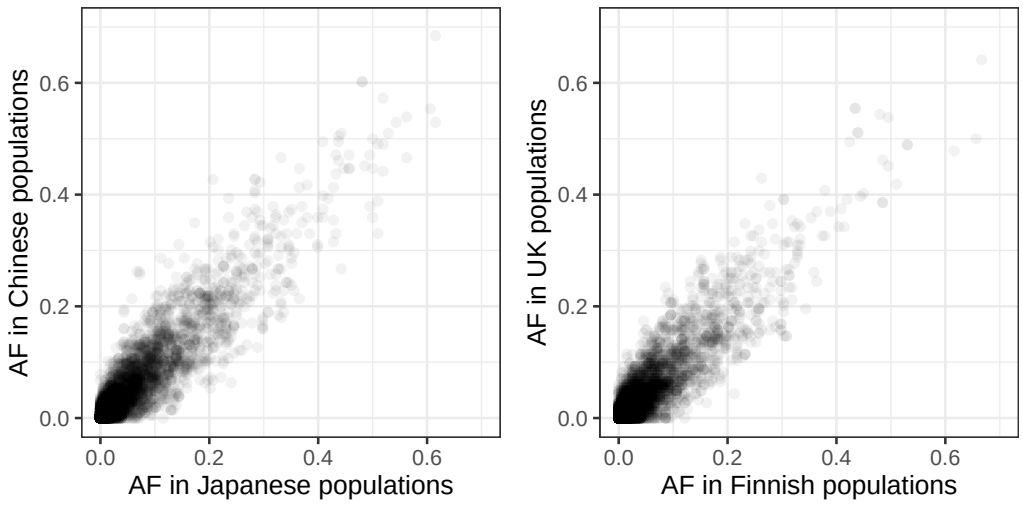

Figure 1: Comparison of allele frequencies between two Asian populations and two European populations.

However, when comparing an European population to an Asian population, there is a significant amount of variants that are either absent from one population while present in the other, or with a remarkably different allele frequency between the cohorts (Figure 2).

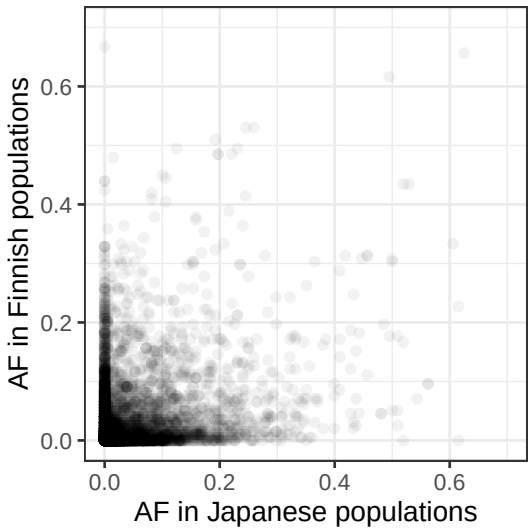

Figure 2: Comparison of allele frequencies between an Asian and European population.

Now, while researching the effects of Neanderthal DNA in modern humans is not a novel topic, the genetic diversity of this field is lacking. As mentioned previously, GWASs tend to be biased towards populations with European ancestry as they make up over 85% of the participants in GWASs (Mills and Rahal, 2019). This bias is reflected in Neanderthal DNA research, where the majority of studies use GWASs to conduct their analysis, therefore covering primarily genomes of European cohorts, despite reports of Europeans having less Neanderthal ancestry than East Asians, for instance (Wall et al., 2013). In addition, studies in this field tend to ignore potential pleiotropic effects, or genes that affect multiple traits, and instead focus on singular associations of phenotypes. In this thesis, we address both issues by studying pleiotropic effects of Neanderthal DNA using GWAS summary statistics in an Asian cohort from BioBank Japan.

3 Methods

3.1 Phenotypic Data

The analysis was conducted using GWAS summary statistics from the BioBank Japan cohort with clinical data from approximately 200 000 individuals and phenotypic information including 159 common diseases, 38 biomarkers, and medication usage data for 23 different drugs. The number of samples ranged from 61 457 to 178 726 for each of the 220 phenotypes. Sakaue et al. (2021) describe the exact procedure of the GWAS.

GWAS summary statistics were divided into data with two different structures – binary traits (disease endpoints, medication) and quantitative traits (biomarkers). Both types of structures included descriptive and analytical data for SNPs, such as the chromosome, position, and name of the marker; allele count and frequency; the effect size, or beta, and its standard error; and the association p-value. (Sakaue et al., 2021)

3.2 Definition of Neanderthal Marker Variants

To identify phenotype associations in the Biobank Japan GWAS data that are linked to Neanderthal DNA, we scanned for the presence of previously annotated variants that match genomic features consistent with Neanderthal admixture (Dannemann, 2020). Those variants have been associated with Neanderthal haplotypes segregating in non-African individuals from the 1000 Genomes project (1000 Genomes Project Consortium, 2015) and followed three criteria: the allele is present in a homozygous state in the Vindija Neanderthal genome, present in at least one non-African individual from the 1000 Genomes Project, and absent from Yoruban populations in the 1000 Genomes Project, which has previously been shown to carry the lowest levels of Neanderthal admixture (Prüfer et al., 2017). In

addition, we filtered out variants with an allele frequency of < 0.01 in the Japanese population.

In order to compare the number of associations of Neanderthal variants to an unbiased set of other, non-archaic variants, we generated datasets of Neanderthal variants and background variants that account for genomic differences between both of them. Due to the relatively recent admixture event, Neanderthal variants typically have an elevated level of linkage disequilibrium and as a result, haplotype structures are generally longer (Dannemann, Prüfer, and Kelso, 2017). To account for that, a single tag SNP was chosen at random to represent each haplotype. Neanderthal haplotypes were defined as sets of Neanderthal variants with a LD of $r^2 > 0.5$. Next, we defined non-archaic background tag SNPs, which were required to show no LD ($r^2 < 0.5$) with a Neanderthal variant or any other non-archaic tag SNP. Ultimately, Neanderthal and all background sets each consisted of 11 000 variants.

Neanderthal variants are still found in present-day people consistent with the initial admixture proportion, which for the vast majority of Neanderthal variants is a frequency below 20%, due to the initial admixture rate of less than 10% (Sankararaman et al., 2014; Vernot and Akey, 2014). Since the minor allele frequency is proportionally associated with the statistical power in GWAS, given the lower allele frequency levels of Neanderthal variants compared to non-archaic variants (Figure 3), we required the background sets of non-Neanderthal variants to be frequency-matched to the set of Neanderthal variants. To be more precise, we created 1000 background sets of non-Neanderthal tag SNPs, which matched the Neanderthal tag SNPs in number and allele frequency (Figure 4).

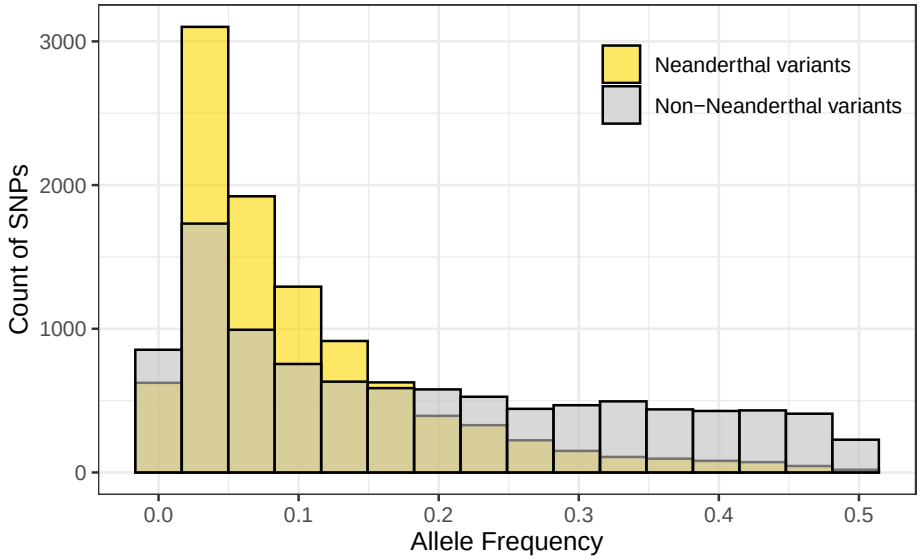

Figure 3: Allele frequency distribution in Atrial flutter GWAS summary statistics. 10 000 Neanderthal variants compared to a sample of 10 000 random variants. Atrial flutter GWAS summary statistics was chosen at random for demonstration purposes.

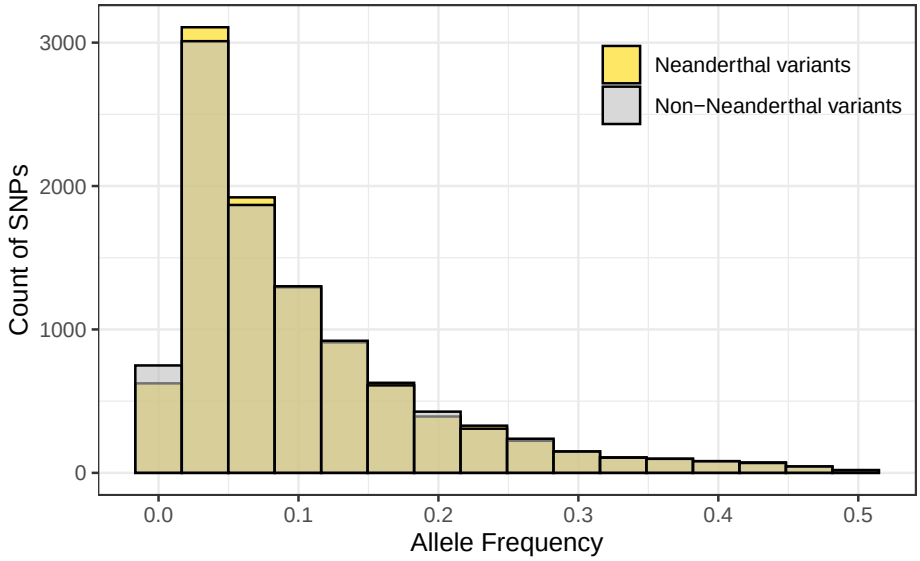

Figure 4: Allele frequency distribution of SNPs from Atrial flutter GWAS summary statistics. 10 000 Neanderthal variants compared to a sample of 10 000 variants from a random frequency matched non-Neanderthal set.

3.3 Testing for Significant Pleiotropic Effects of Neanderthal DNA

For the analysis, the Neanderthal tag SNPs and 1000 non-Neanderthal sets of tag SNPs were extracted from 220 GWAS summary statistics. Only SNPs with a minor allele frequency of at least 0.01 were included. The extracted Neanderthal tag SNPs were then used to find all pairs of phenotypes with an overlapping tag SNP that displayed a significant association with both phenotypes. The relative number of associations for each such pair was calculated by dividing the number of significant overlapping tag SNPs for both phenotypes by the total number of tag SNPs present in both phenotype datasets. This relative number of associations was calculated for the Neanderthal set, as well as for each of the 1000 background sets. Using this approach, we also take into account the various degrees of polygenicity for the phenotype pairs scanned. Polygenicity describes the level of associated variants with a given phenotype.

This distribution of relative numbers for the non-archaic sets, could then be used to decide whether the Neanderthal DNA has elevated levels of associations for a given pair compared to non-Neanderthal variants, suggesting a potential pleiotropic effect of Neanderthal DNA. The p-value for a certain pair of phenotypes is equal to the quantity of non-Neanderthal sets that have a higher number of relative associations than the Neanderthal set for that pair, divided by the amount of non-Neanderthal sets.

In this process we are using a one-sided test, which means we are only testing for instances where Neanderthal variants show a larger number of pair-wise phenotype associations compared to their non-archaic counterparts. To account for multiple testing, we used the Benjamini-Hochberg correction to calculate adjusted p-values; for pairs where the Neanderthal set had the highest relative number of significant associations, the original p-value of < 0.001 was replaced with 0.001 for more accurate adjusted p-values. It is important to note, the Benjamini-Hochberg correction

was only applied to pairs of phenotypes, which showed at least one overlapping significant tag SNP association, in order to limit computational efforts. This restriction likely skews the p-values towards more significance and possibly increases the rate of false positives.

For each pair with a significant overlapping tag SNP, we calculated odds ratios (OR) by dividing the relative number of significant associations of the Neanderthal set by the mean relative number of significant associations in the background variants. The analysis was conducted two times with different SNP association significance thresholds, 5×10^{-8} and 10^{-5}.

4 Results

4.1 Scanning GWASs from BioBank Japan for Neanderthal Pleiotropic Effects

In order to identify phenotype pairs for which Neanderthal DNA shows elevated levels of associations compared to non-archaic variants, we analyzed 220 phenotypes. We tested each phenotype pair that had at least one overlapping significant Neanderthal tag SNP association, to determine whether the number of significant tag SNPs is higher than one might expect when comparing with non-Neanderthal variants, using GWAS summary statistics made available by BioBank Japan. We conducted this analysis with two significance cutoff points: the first analysis was performed with a threshold of 5×10^{-8}, and the second analysis used a threshold of 10^{-5}.

4.1.1 Results for a threshold of 5×10^{-8}

Out of 48 180 possible pairs, we found 294 pairs of phenotypes that had at least one significant overlapping haplotype of Neanderthal origin with a tag SNP that had association p-values with both phenotypes below the significance threshold of 5×10^{-8}. After comparing the number of Neanderthal tag SNPs with pairwise phenotype associations to the numbers from 1000 sets of non-Neanderthal variants, we found that for 73 out of 294 phenotype pairs, the Neanderthal tag SNPs set had a higher relative number of significant associations than 95% of non-archaic tag SNP sets. After correcting for multiple testing using the Benjamini-Hochberg method, 10 pairs of phenotypes were deemed statistically significant using a significance level of $\alpha = 0.05$ (Table 1).

Table 1: The pairs of phenotypes that were deemed to be significant with an association p-value of $< 5 \times 10^{-8}$.

Phenotypes	Significant tag SNPs	Adjusted p-value	Odds ratio (OR)
Myocardial infarction - Total protein	4	0.0294	6.42
Uterine fibroid - Height	3	0.0294	84.01
Chronic sinusitis - Salicylic acid and derivatives	3	0.0294	17.31
Salicylic acid and derivatives - Total protein	3	0.0294	13.44
Antithrombotic agents - Chronic Sinusitis	2	0.0294	245.06
Antithrombotic agents - Total protein	2	0.0294	45.63
Antithrombotic agents - Lymphocyte count	1	0.0294	Inf
Chronic sinusitis - Gastric cancer	1	0.0294	Inf
Gastric cancer - Total protein	1	0.0294	Inf
Mean corpuscular hemoglobin concentration (MCHC) - Basophil count	1	0.0294	Inf

For all pairs in the results in Table 1, the Neanderthal set had the highest relative number of significant associations when comparing with the non-archaic sets; as a consequence, these pairs share the same p-value. There are four pairs of phenotypes where it was impossible to calculate OR precisely, as no significant overlapping non-Neanderthal variants were identified for that pair.

Total protein level was the most prevalent phenotype in these pairs, which was associated with four other phenotypes. It was followed by chronic sinusitis, and usage of antithrombotic agents, with three occurrences each.

4.1.2 Results for a threshold of 10^{-5}

In order to test for the robustness of our inferences, we repeated the analysis using a lower significance threshold of 10^{-5}. At that level, 520 different pairs had at least one significant Neanderthal tag SNP. Out of those 520 pairs, 159 pairs of phenotypes showed a higher relative number of Neanderthal tag SNP associations than 95% of non-Neanderthal sets. After adjusting p-values with the Benjamini-

Hochberg method, 54 pairs had a corrected p-value below the 0.05 threshold. Out of these pairs, 17 had at least one tag SNP significant at the lower level of 5×10^{-8}, and were therefore also included in the first analysis (Table 2).

Table 2: The pairs deemed to be significant at an association p-value level of $< 10^{-5}$ which were also examined in previous analysis.

Phenotypes	**Significant tag SNPs**	**Adjusted p-value**	**Odds ratio (OR)**
Drugs used in diabetes - Type 2 diabetes	16	0.015	2.95
Monocyte count - White blood cell count	9	0.015	3.17
Antithrombotic agents - Total protein	4	0.015	8.98
Antithrombotic agents - Grave's disease	4	0.015	8.98
Antithrombotic agents - Serum creatinine	4	0.015	7.39
Antithrombotic agents - HMG CoA reductase inhibitors	4	0.015	7.25
Uterine fibroid - Height	3	0.015	17.2
Gastric cancer - Total protein	3	0.015	11.1
Gastric cancer - Salicylic acid and derivatives	3	0.015	9.28
Antithrombotic agents - Myocardial infarction	4	0.035	5.51
Antithrombotic agents - Salicylic acid and derivatives	4	0.035	4.99
Antithrombotic agents - Chronic hepatitis B	3	0.035	8.68
Gastric cancer - Myocardial infarction	3	0.035	7.11
Antithrombotic agents - RAS-acting agents	3	0.035	7.01
Antithrombotic agents - Chronic Sinusitis	3	0.046	6.69
Chronic Sinusitis - Total protein	5	0.048	4.23
HMG CoA reductase inhibitors - Gastric cancer	3	0.048	6.78

The most prevalent phenotype out of the pairs, which was found significant when testing at an association p-value threshold of 10^{-5} and included in the former analysis, was the use of antithrombotic agents, appearing in nine of such pairs. Gastric cancer and the total protein level followed with four and three instances respectively.

Some of the identified phenotype pairs are likely correlated genetically. It is therefore not surprising to see some of the pairs with higher numbers of significant tag

SNPs, for example the consumption of drugs used in diabetes and type 2 diabetes, and monocyte and white blood cell count. These relationships are more likely to occur as people with diabetes often take medication for diabetes, and monocytes are a type of white blood cell. Other similarly obvious associations include all combinations of phenotypes that are related to cardiovascular complications: myocardial infarction, usage of antithrombotic agents and agents acting on the renin-angiotensin system (cholesterol-lowering medications). Follow-up analysis will need to disentangle which of these enriched phenotype pairs are genetically correlated and interpret these results taking this feature into account.

In the second analysis with an association p-value threshold of 10^{-5}, out of the 54 pairs of phenotypes that showed an elevated level of associations, the phenotype with the most occurrences in pairs was osteoporosis, which was present in 17 different pairs. It is worth mentioning that, osteoporosis was absent from all pairs at a significance level of $< 5 \times 10^{-8}$. However, upon closer examination, no SNP in the osteoporosis GWAS summary statistics has an association p-value lower than 5×10^{-8}. Other more common phenotypes include intake of antithrombotic agents with 13 different pairs, usage of calcium channel blockers (medication for lowering blood pressure) with nine and gastric cancer with five. The data for all 54 significant pairs is presented in Appendix 1.

4.2 Comparison Between Significance Thresholds

After conducting the analyses, we identified four pairs of phenotypes, which show elevated levels of associations at both significance levels. There are six other pairs that were significant at 5×10^{-8}, and 50 pairs that were significant at 10^{-5}. Out of all pairs, which were included in both analyses, 23 pairs were deemed significant in at least one of the analyses (Figure 5).

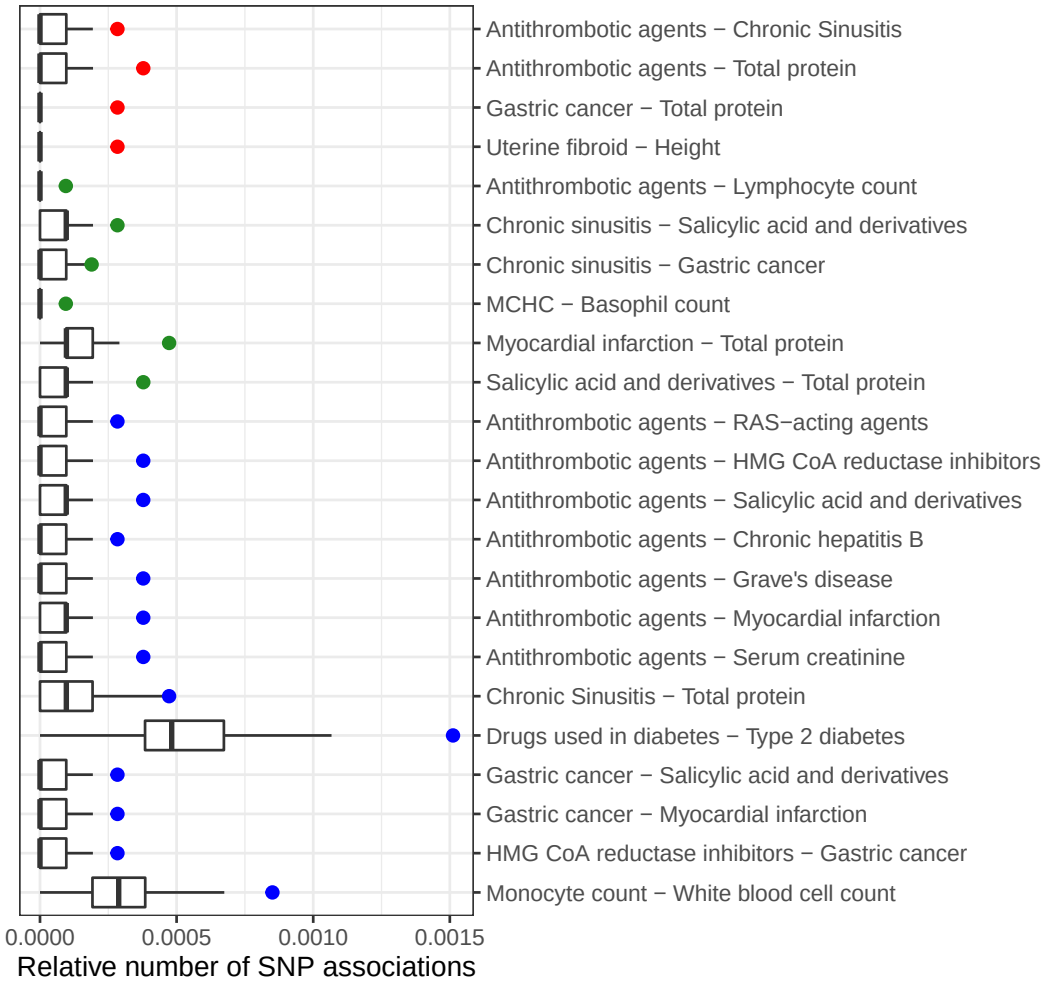

Figure 5: All pairs that were included in both analyses and deemed significant for at least one significance level, 10^{-5} or 5×10^{-8}. Relative number of significant SNP associations from the Neanderthal set are represented by points, box plots represent the distribution of non-Neanderthal relative SNP associations, both at a significance level of $< 10^{-5}$, Red points represent pairs, which were found significant at both thresholds, green points denote pairs that were found significant only at 5×10^{-8}, and blue points denote pairs that were significant at 10^{-5}.

We compared the odds ratios of all pairs that were included in both analyses. Between the two significance thresholds, we found a significant correlation Spearman's $\rho = 0.52$ ($p < 2.2 \times 10^{-16}$), indicating a moderate correlation, suggesting

that our results are robust towards the selection of the significance level. This is also illustrated on Figure 6.

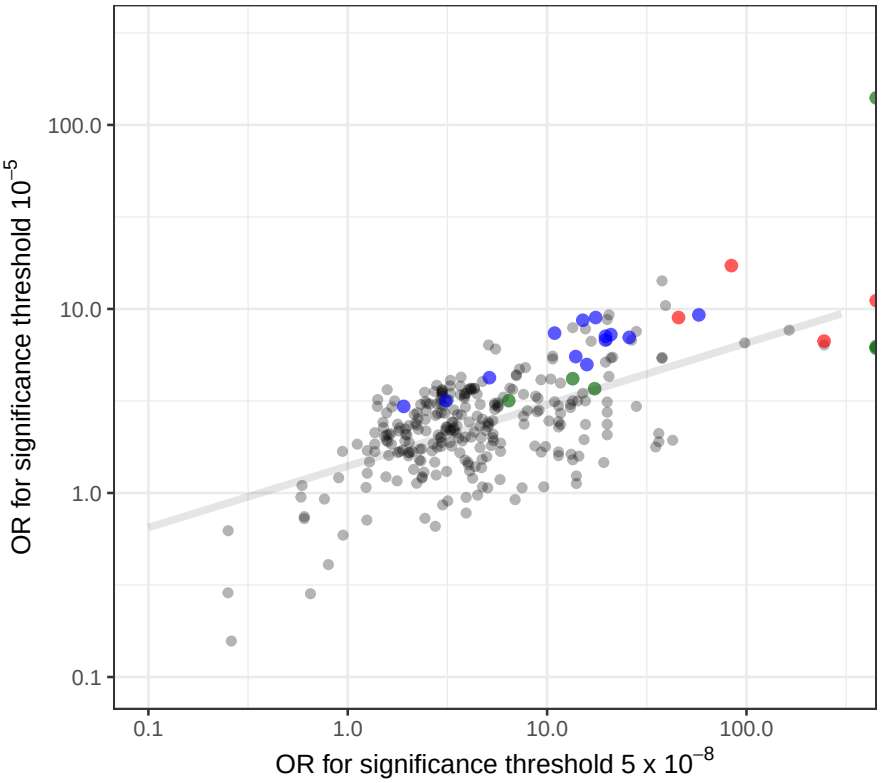

Figure 6: Odds ratios for pairs of phenotypes included in both analyses at significance levels of $< 5 \times 10^{-8}$ and $< 10^{-5}$ on a logarithmic scale. Pairs that were determined to be statistically significant at both levels are denoted in red, while green points denote pairs that were significant only at 5×10^{-8}, and blue points denote pairs that were significant at 10^{-5}. Odds ratios that could not be precisely calculated were marked as infinite, and are positioned on the edge of the right side of the diagram.

5 Discussion

In this study, we investigated pleiotropic effects of Neanderthal DNA in Asian populations by testing for Neanderthal variants with pairwise phenotype associations in GWAS from Biobank Japan. We identified four such phenotype pairs that were significant at both of the applied significance thresholds (5×10^{-8} and 10^{-5}): usage of antithrombotic agents and chronic sinusitis; usage of antithrombotic agents and total protein levels; gastric cancer and total protein levels; and uterine fibroid and height. In the initial analysis, three of these four pairs had the highest ORs of pairwise Neanderthal variant associations compared to sets of frequency-matched non-archaic variants.

Other pairs that we found were: myocardial infarction and total protein level; salicylic acid and derivatives and total protein level; chronic sinusitis with salicylic acid and derivatives and gastric cancer; antithrombotic agents and lymphocyte count; mean corpuscular hemoglobin concentration and basophil count. These phenotype pairs do not display any obvious biological correlation, which likely potentially indicates them as targets of pleiotropic effects of Neanderthal DNA.

While the usage of antithrombotic agents has yet to be shown to be associated with chronic sinusitis and total protein levels, studies point to an existing correlation between gastric cancer and total protein levels. Serum albumin, which is measured in the total protein test, is associated with poor prognosis of gastric cancer (Isik et al., 2014), while a study conducted on a Korean population found a low albumin to globulin ratio to increase risk of cancer incidence and mortality (Suh et al., 2014). The relationship between uterine fibroid and height has been examined by some studies, however, other factors, such as weight and obesity, are thought to have a bigger impact (Yang et al., 2014). Nevertheless, follow-up studies that evaluate these results in light of genetic correlation between these phenotypes are needed to aid the interpretation of these results.

A noteworthy finding from the analysis is the fact that osteoporosis was included

in a third of significant pairs when using a lowered significance threshold. It has been shown before, that higher bone density is associated with some variants with Neanderthal ancestry (McArthur, Rinker, and Capra, 2021), which could indicate, with further research, existence of more pleiotropic effects of Neanderthal DNA associated with osteoporosis.

Despite reports of Neanderthal-derived variants increasing risk of type 2 diabetes (Williams et al., 2014), we were unable to confirm any meaningful pleiotropic effects associated with the disease. The only significant result of our analysis concerning type 2 diabetes came from the second analysis with a lowered threshold, where we found the evident association between drugs used against diabetes and type 2 diabetes itself. While we identified significant SNPs of Neanderthal ancestry between type 2 diabetes and other phenotypes, like blood sugar levels and platelet count, they failed to show elevated levels of SNP associations compared to non-Neanderthal variants.

For future studies, pleiotropic effects covered in this thesis should be analyzed to assess how they affect modern humans, to understand impact of Neanderthal haplotypes and uncover additional information about these diseases. In addition, a comparison of these results to other populations would help us to evaluate whether these results are specific for Asian populations or shared across different ancestries.

Conclusions

In order to enhance the genetic diversity in Neanderthal research, and find pleiotropic effects of Neanderthal DNA, we analyzed GWAS summary statistics from the Asian cohort provided by BioBank Japan. We included 220 phenotypes, which consisted of disease endpoints, biomarkers, and medicine usage. We extracted tag SNPs with Neanderthal ancestry that showed a pairwise significant association with phenotypes, and compared the number of those associations against 1000 sets of non-archaic tag SNPs. The analysis was conducted with SNP association significance thresholds of 5×10^{-8} and 10^{-5}.

Pairs of phenotypes that displayed an elevated level of significant Neanderthal DNA associations compared to non-Neanderthal variants at both significance thresholds were antithrombotic medicine and chronic sinusitis; antithrombotic medicine and total protein levels; gastric cancer and total protein levels; and uterine fibroid and height. In addition, six more pairs were found significant at a threshold of 5×10^{-8} and 50 more pairs were found significant at a threshold of 10^{-5}.

Our results suggest that some biological and medical phenotypes are targets of pleiotropic effects in Asians today. In order to put these results in context, further research conducted using more populations of diverse ancestries is needed.

References

1000 Genomes Project Consortium (2015). "A global reference for human genetic variation". *Nature* 526.7571, p. 68.

Abi-Rached, L., M. Jobin, S. Kulkarni, A. Mcwhinnie, L. Gragert, et al. (2011). "The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans". *Science (New York, N.Y.)* 334, pp. 89–94. DOI: 10.1126/science.1209202.

Dannemann, M. (2020). "The Population-Specific Impact of Neandertal Introgression on Human Disease". *Genome Biology and Evolution* 13. DOI: 10.1093/gbe/evaa250.

Dannemann, M., A. Andrés, and J. Kelso (2016). "Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors". *American journal of human genetics* 98, pp. 22–33. DOI: 10.1016/j.ajhg.2015.11.015.

Dannemann, M., Z. He, J. Heide, B. Vernot, L. Sidow, et al. (2020). "Human Stem Cell Resources Are an Inroad to Neandertal DNA Functions". *Stem Cell Reports* 15, pp. 214–225. DOI: 10.1016/j.stemcr.2020.05.018.

Dannemann, M. and J. Kelso (2017). "The Contribution of Neanderthals to Phenotypic Variation in Modern Humans". *The American Journal of Human Genetics* 101, pp. 578–589. DOI: 10.1016/j.ajhg.2017.09.010.

Dannemann, M., K. Prüfer, and J. Kelso (2017). "Functional implications of Neandertal introgression in modern humans". *Genome Biology* 18, pp. 1–11. DOI: 10.1186/s13059-017-1181-7.

Dediu, D. and S. Levinson (2018). "Neanderthal language revisited: not only us". *Current Opinion in Behavioral Sciences* 21, pp. 49–55. DOI: 10.1016/j.cobeha.2018.01.001.

Gasperini, M., A. Hill, J. McFaline-Figueroa, B. Martin, S. Kim, et al. (2019). "A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens". *Cell* 176, pp. 377–390. DOI: 10.1016/j.cell.2018.11.029.

Giani, A., G. Gallo, L. Gianfranceschi, and G. Formenti (2019). "Long walk to genomics: History and current approaches to genome sequencing and assembly". *Computational and Structural Biotechnology Journal* 18. DOI: 10.1016/j.csbj.2019.11.002.

Gibson, G. (2012). "Rare and common variants: Twenty arguments. Nat". *Nature Reviews Genetics* 13, pp. 135–145.

Green, E., J. Watson, and F. Collins (2015). "Human Genome Project: Twenty-five years of big biology". *Nature* 526, pp. 29–31. DOI: 10.1038/526029a.

Green, R., J. Krause, A. Briggs, T. Maricic, U. Stenzel, et al. (2010). "A Draft Sequence of the Neandertal Genome". *Science (New York, N.Y.)* 328, pp. 710–722. DOI: 10.1126/science.1188021.

Higham, T., K. Douka, R. Wood, C. Ramsey, F. Brock, et al. (2014). "The timing and spatiotemporal patterning of Neanderthal disappearance". *Nature* 512, pp. 306–309. DOI: 10.1038/nature13621.

Holloway, R. (1981). "Volumetric and asymmetry determinations on recent hominid endocasts: Spy I and II, Djebel Ihroud I, and the Sale Homo erectus specimens, with some notes on Neandertal brain size". *American Journal of Physical Anthropology* 55, pp. 385–393. DOI: 10.1002/ajpa.1330550312.

Hublin, J.-J., A. Ben-Ncer, S. Bailey, S. Freidline, S. Neubauer, et al. (2017). "New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens". *Nature* 546, pp. 289–292. DOI: 10.1038/nature22336.

Isik, A., I. Okan, Y. D. Firat, B. Yilmaz, A. Akcakaya, et al. (2014). "A new prognostic strategy for gastric carcinoma: Albumin level and metastatic lymph node ratio". *Minerva chirurgica* 69, pp. 147–153.

Langdon, H. (2016). "Case Study 18. Neanderthals in the Mirror: Imagining our Relatives". In: pp. 141–149. ISBN: 978-3-319-41584-0. DOI: 10.1007/978-3-319-41585-7_18.

McArthur, E., D. Rinker, and J. Capra (2021). "Quantifying the contribution of Neanderthal introgression to the heritability of complex traits". *Nature Communications* 12. DOI: 10.1038/s41467-021-24582-y.

McGuire, A., S. Gabriel, S. Tishkoff, A. Wonkam, A. Chakravarti, et al. (2020). "The road ahead in genetics and genomics". *Nature Reviews Genetics* 21, pp. 1–16. DOI: 10.1038/s41576-020-0272-6.

Mills, M. and C. Rahal (2019). "A scientometric review of genome-wide association studies". *Communications Biology* 2. DOI: 10.1038/s42003-018-0261-x.

Nurk, S., S. Koren, A. Rhie, M. Rautiainen, A. Bzikadze, et al. (2022). "The complete sequence of a human genome". *Science (New York, N.Y.)* 376, pp. 44–53. DOI: 10.1126/science.abj6987.

Orlando, L., R. Allaby, P. Skoglund, C. Sarkissian, P. Stockhammer, et al. (2021). "Ancient DNA analysis". *Nature Reviews Methods Primers* 1. DOI: 10.1038/s43586-020-00011-0.

Pettitt, P. (2002). "The Neanderthal dead: exploring mortuary variability in Middle Palaeolithic Eurasia". *Before Farming* 2002, pp. 1–19.

Prüfer, K., C. Filippo, S. Grote, F. Mafessoni, P. Korlević, et al. (2017). "A high-coverage Neandertal genome from Vindija Cave in Croatia". *Science* 358, pp. 655–658. DOI: 10.1126/science.aao1887.

Prüfer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman, et al. (2013). "The complete genome sequence of a Neandertal from the Altai Mountains". *Nature* 505, pp. 43–49. DOI: 10.1038/nature12886.

Sakaue, S., M. Kanai, Y. Tanigawa, J. Karjalainen, M. Kurki, et al. (2021). "A cross-population atlas of genetic associations for 220 human phenotypes". *Nature Genetics* 53, 1415–1424. DOI: 10.1038/s41588-021-00931-x.

Sankararaman, S., S. Mallick, M. Dannemann, K. Prüfer, J. Kelso, et al. (2014). "The landscape of Neandertal ancestry in present-day humans". *Nature* 507, 354–357. DOI: 10.1038/nature12961.

Sankararaman, S., N. Patterson, H. Li, S. Pääbo, and D. Reich (2012). "The Date of Interbreeding between Neandertals and Modern Humans". *PLoS genetics* 8, e1002947. DOI: 10.1371/journal.pgen.1002947.

Sebastiani, P., N. Timofeev, D. Dworkis, T. Perls, and M. Steinberg (2009). "Genome-wide association studies and the genetic dissection of complex traits". *American journal of hematology* 84, pp. 504–515. DOI: 10.1002/ajh.21440.

Simonti, C., B. Vernot, L. Bastarache, E. Bottinger, D. Carrell, et al. (2016). "The phenotypic legacy of admixture between modern humans and Neandertals". *Science* 351, pp. 737–741. DOI: 10.1126/science.aad2149.

Skoglund, P. and I. Mathieson (2018). "Ancient Human Genomics: The First Decade". *Annual Review of Genomics and Human Genetics* 19, 381–404. DOI: 10.1146/annurev-genom-083117-021749.

Srinivasan, S., F. Bettella, M. Mattingsdal, Y. Wang, A. Witoelar, et al. (2015). "Genetic Markers of Human Evolution Are Enriched in Schizophrenia". *Biological Psychiatry* 80, pp. 284–292. DOI: 10.1016/j.biopsych.2015.10.009.

Stewart, J., O. García Rodríguez, M. Knul, L. Sewell, H Montgomery, et al. (2018). "Palaeoecological and genetic evidence for Neanderthal power locomotion as an adaptation to a woodland environment". *Quaternary Science Reviews* 217. DOI: 10.1016/j.quascirev.2018.12.023.

Suh, B., S. Park, D Shin, J.-M. Yun, B Keam, et al. (2014). "Low Albumin to Globulin Ratio Associated with Cancer Incidence and Mortality in Generally Healthy Adults." *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO* 25, 2260–2266. DOI: 10.1093/annonc/mdu274.

Trinkaus, E. and S. Villotte (2017). "External auditory exostoses and hearing loss in the Shanidar 1 Neandertal". *PLOS ONE* 12, e0186684. DOI: 10.1371/journal.pone.0186684.

Vernot, B. and J. Akey (2014). "Resurrecting Surviving Neandertal Lineages from Modern Human Genomes". *Science (New York, N.Y.)* 343, pp. 1017–1021. DOI: 10.1126/science.1245938.

Vernot, B., S. Tucci, J. Kelso, J. Schraiber, A. Wolf, et al. (2016). "Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals". *Science* 352, pp. 235–239. DOI: 10.1126/science.aad9416.

Vidal, J., F. d'Errico, F. Pacheco, R. Blasco, J. Rosell Ardèvol, et al. (2014). "A rock engraving made by Neanderthals in Gibraltar". *Proceedings of the National Academy of Sciences* 111, pp. 13301–13306. DOI: 10.1073/pnas.1411529111.

Visscher, P., N. Wray, Q. Zhang, P. Sklar, M. McCarthy, et al. (2017). "10 Years of GWAS Discovery: Biology, Function, and Translation". *The American Journal of Human Genetics* 101, pp. 5–22. DOI: 10.1016/j.ajhg.2017.06.005.

Wall, J., M. Yang, F. Jay, S. Kim, E. Durand, et al. (2013). "Higher Levels of Neanderthal Ancestry in East Asians Than in Europeans." *Genetics* 194, 199–209. DOI: 10.1534/genetics.112.148213.

Williams, A. L., S. B. Jacobs, H. Moreno-Macías, A. Huerta-Chagoya, C. Churchhouse, et al. (2014). "Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico". *Nature* 506, 97–101. DOI: 10.1038/nature12828.

Yang, Y., Y. He, Q. Zeng, and S. Li (2014). "Association of Body Size and Body Fat Distribution with Uterine Fibroids Among Chinese Women". *Journal of women's health (2002)* 23, pp. 619–626. DOI: 10.1089/jwh.2013.4690.

Zeberg, H. and S. Pääbo (2020). "The major genetic risk factor for severe COVID-19 is inherited from Neanderthals". *Nature* 587, pp. 1–3. DOI: 10.1038/s41586-020-2818-3.

Appendix 1. Results at a significance level of 10^{-5}

Table 3: The pairs deemed to be significant at an association p-value level of $< 10^{-5}$.

Phenotypes	Significant tag SNPs	Adjusted p-value	Odds ratio (OR)
Drugs used in diabetes - Type 2 diabetes	16	0.015	2.95
Monocyte count - White blood cell count	9	0.015	3.17
Antithrombotic agents - Calcium channel blockers	4	0.015	11.31
Antithrombotic agents - Grave's disease	4	0.015	8.98
Antithrombotic agents - Total protein	4	0.015	8.98
Antithrombotic agents - Serum creatinine	4	0.015	7.39
Antithrombotic agents - HMG CoA reductase inhibitors	4	0.015	7.25
Antithrombotic agents - Gastric cancer	3	0.015	18.05
Uterine fibroid - Height	3	0.015	17.20
Chronic sinusitis - Calcium channel blockers	3	0.015	11.72
Gastric cancer - Total protein	3	0.015	11.10
Beta blocking agents - Calcium channel blockers	3	0.015	9.84
Gastric cancer - Salicylic acid and derivatives	3	0.015	9.28
Total Bilirubin - Total cholesterol	2	0.015	280.14
Monocyte count - Total cholesterol	2	0.015	103.28
Beta blocking agents - Gastric cancer	2	0.015	40.03
Angina - Osteoporosis	1	0.015	Inf
Antithrombotic agents - Alanine aminotransferase	1	0.015	Inf
Antithrombotic agents - Osteoporosis	1	0.015	Inf
Aspartate transaminase - Osteoporosis	1	0.015	Inf
Basophil count - Glucose	1	0.015	Inf
Beta blocking agents - Osteoporosis	1	0.015	Inf
Calcium channel blockers - Osteoporosis	1	0.015	Inf
Chronic sinusitis - Osteoporosis	1	0.015	Inf

Phenotypes	Significant tag SNPs	Adjusted p-value	Odds ratio (OR)
Drugs used in diabetes - Atrial flutter	1	0.015	Inf
Gastric cancer - Alanine aminotransferase	1	0.015	Inf
Height - Osteoporosis	1	0.015	Inf
Lymphocyte count - Osteoporosis	1	0.015	Inf
RAS-acting agents - Osteoporosis	1	0.015	Inf
Salicylic acid and derivatives - Osteoporosis	1	0.015	Inf
Stable angina pectoris - Osteoporosis	1	0.015	Inf
Total protein - Osteoporosis	1	0.015	Inf
Unstable angina pectoris - Lymphocyte count	1	0.015	Inf
Unstable angina pectoris - Osteoporosis	1	0.015	Inf
Uterine fibroid - Monocyte count	1	0.015	Inf
Calcium channel blockers - Total protein	3	0.029	9.37
Antithrombotic agents - Myocardial infarction	4	0.035	5.51
Antithrombotic agents - Salicylic acid and derivatives	4	0.035	4.99
Calcium channel blockers - Immunosuppressants	3	0.035	9.78
Antithrombotic agents - Chronic hepatitis B	3	0.035	8.68
Calcium channel blockers - Grave's disease	3	0.035	7.74
Gastric cancer - Myocardial infarction	3	0.035	7.11
White blood cell count - Total cholesterol	2	0.035	23.07
Gastric cancer - Height	2	0.035	14.64
Antithrombotic agents - RAS-acting agents	3	0.046	7.01
Chronic sinusitis - Total protein	5	0.048	4.23
Calcium channel blockers - Rheumatoid arthritis	3	0.048	7.60
Calcium channel blockers - Eosinophil count	3	0.048	7.14
HMG CoA reductase inhibitors - Gastric cancer	3	0.048	6.78
Antithrombotic agents - Chronic sinusitis	3	0.048	6.69
Myocardial infarction - Osteoporosis	1	0.048	196.34
White blood cell count - Osteoporosis	1	0.048	196.34
Uric acids - Osteoporosis	1	0.048	196.33
Monocyte count - Osteoporosis	1	0.048	196.32

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Johann Koobas,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis, The Pleiotropic Effects of Neanderthal DNA in Present-Day Asian Populations, supervised by Dr. Michael Dannemann.

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Johann Koobas
09/05/2023