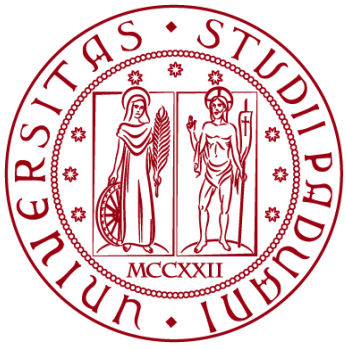
**UNIVERSITÀ DEGLI STUDI DI PADOVA**

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea magistrale in Molecular Biology




**TESI DI LAUREA**

# To the bottom of genetic assortative mating: a genetic investigation of what drives mate choice within two present day European populations

**Relatore: Prof. Luca Pagani**
**Dipartimento di Biologia**

**Laureanda: Konstantina Cheshmedzhieva**

**ANNO ACCADEMICO 2022/2023**

**CONTENTS**

# Abstract

The choice of mating partner is a topic which has been an object of both biological and sociological interest. Although much research has been done on the basis of phenotype data, the age of next generation sequencing together with genome-wide association studies (GWAS) and the establishment of biobanks across Europe have provided researchers with the possibility to gain a better understanding of mate choice in humans from a genetic point of view. This work focuses on detecting the presence of genetic assortative mating in contemporary human populations in Europe and studying its patterns. It is based on a newly developed method for genotype analysis different from the available research, since it builds on single genomes rather than on pedigrees or couples. Using individual genomes, we detected signals of assortative mating in the previous generation, based on a computed score from genetic windows containing single nucleotide polymorphisms (SNPs) associated with complex traits of interest, and compared it to each individual's phenotype data. The outcome of this work represents an initial insight into the genetic perspective of partner selection to have taken place in the previous generation of individuals with homogenous backgrounds.

# 1. Introduction

## 1.1. What is assortative mating?

Mate choice is an integral, yet complicated process that has been taking place in the course of evolution. Opposed to mating under random selection, assortative mating is a form of sexual selection in which individual phenotypic preferences are taken into consideration and thus the random model of mating is no longer valid. [1]

Assortative mating can be represented both by choosing phenotypically and/or genotypically similar or dissimilar partners, respectively known as positive assortative mating (PAM) and negative assortative mating (NAM). Among human populations, there is scientific evidence of PAM for traits defining physical appearance [2], cognitive style and personality [3], physiological features and longevity [4,5]. However, mate choice based on the major histocompatibility complex in humans has been proven to be a representation of NAM [6], explaining the complexity of partner selection regarding the immune system.

## 1.2. How and why does assortative mating take place?

The reasons explaining PAM occurrence in human populations are proximate and ultimate. Proximate explanations focus on how the phenomenon arises on the bases of population stratification and mate choice according to a mate-value preference, and ultimate explanations elaborate on the phenomenon as an evolutionary force for positive selection of given traits. [8]

Researching PAM from the proximate point of view would require one to consider this form of deviation from panmixia as a consequence of population segregation into separate clusters according to phenotype (or genotype). This could be also referred to as ‘social homogamy’, where mates tend to originate from the same cluster of individuals.[8] Therefore, even if phenotypic assortment is absent, a stratified population which strictly follows the social homogamy principles, in the long term has a higher probability of demonstrating some degree of PAM, compared to a well-mixed population. [8]

When researching patterns of assortative mating, the population scale is also of importance. Depending on whether the object of study is a local group, a country's or a continental population, different patterns could be observed. This is due to the probability of coming across clines with distinct mate preferences and culturally mediated social rules. Thus, the deviation from the random pattern of mating caused by social norms and consequently, evolutionary events like genetic drifts or natural selection, leaves a door open for differentiation of groups within a certain spatial continuum regardless of the degree of their shared genetic ancestry. [8]

The differentiation would then lead to a higher phenotypic variation which often goes together with distinct socio-cultural attributes. These attributes may be at the core of another possible explanation for PAM, which is mate-value preference (where "value" has to be intended as subjectively perceived value), or the preference within a population of a mate in possession of a certain complex trait. [8] The preference of a trait, ranging from physiological to behavioural characteristics, may be influenced by a myriad of reasons including population-specific culture, environment, susceptibility to disease etc.[8] These factors could define the mate values within a population, and be the reason for further cluster formation depending on paired mate-values.[8]

## 1.3. How is assortative mating accounted for?

Literature regarding assortative mating is abundant in studies investigating the similarity between spousal couples based on phenotypic studies of traits of interest, reported as numerical values and further processed by statistical methods. [3,4,7] With the advancement of genotype sequencing, the establishment of biobanks and the broad enlargement of genome-wide association studies (GWAS) data, phenotypic data has been collected to provide numerical estimates of genetic correlations for trait-associated loci across the genome, therefore focusing on the more specific phenomenon of genetic assortative mating (GAM). [2, 5-6, 10] As most complex traits associated with assortative mating are polygenic, multiple predictors are taken into consideration when performing a correlation estimate. [10] Polygenic Risk Scores (PRS) are often performed on different configurations of related and/or non-related individuals so as to estimate the putative parents' PRS similarities and to infer whether any type of assortment has taken place from a genetical point of view. Using the latter

approach, it has been found that partners tend to be similar in their PRS for educational attainment, height and depression. [10]

## 1.4. What approaches have been applied so far?

From a genetic viewpoint, evaluation of assortative mating usually tends to happen by either using the above-described strategy of PRS comparison between partners, or by a comparatively new methodology, which seeks traces of genetic assortative mating in the offspring's genome. [12]

The latter focuses on calculating the correlation between trait-increasing alleles (TIAs) on odd versus even chromosomes, as an estimate for gametic phase disequilibrium (GPD). GPD estimate under random mating is expected to be approximately 0, while if assortative mating has taken place, the TIAs will be equally dispersed throughout the genome. Therefore, the correlation between $GPD_{odd}$ and $GPD_{even}$ is assumed to differ from 0 under assortative mating.

Assortative mating investigations are a genetic gateway to understanding the underlying biological bases for mating behaviours in human populations. Despite the need for establishing a global understanding of this process, the genotype data available for analysis is centred mainly around European populations.[2,5,9,14] A recent endeavour has been made by Yamamoto et al. in order to trace the genetic footprints of assortative mating in the Japanese population.[13] The study implements the approach developed by Yengo et al.[12] so as to identify the traits showing most significant levels of GPD, and further compare the results with such obtained from the UK Biobank.

## 1.5. What is the approach applied in this study? How does it stand out from the previous ones?

In the current piece of work, we develop a novel method in order to account for assortative mating via individual genomes as by Yengo *et al.*, however, by tracking the degree of genetic assortative mating (GAM) through an estimated heterozygosity score. The method is based on analysing individual autosomal DNA broken down into windows of set size in kilobases (i.e. 50kb wide). The estimated heterozygosity for each

window containing single nucleotide polymorphisms (SNPs) associated with a trait of interest, when being significantly different from the estimated heterozygosity for the rest of the windows, can be regarded as an indicator of genetic assortative mating for the trait of interest between the parents of the individual. The genome-division strategy has been previously proved to be effective in accounting for ancestral contributions to complex traits in contemporary Europeans [11] .

As stated above, the scores and differences between the windows associated with the trait of interest and the rest of the genome could be interpreted as a marker for a form of sexual selection, which is different from the random one. Thus, it can be both a form of negative assortment and a form of positive assortment, depending on the direction of the difference between the heterozygosity score for genome-wide and trait-associated windows.

This methodology aims to provide a better insight into the patterns of recent assortative mating of homogenous-background individuals from Europe, looking for a different genetic architecture than the one expected under random mating and comparing it with the reported phenotypic characteristics.[2,12] Simultaneously, it represent a novel way for numeric representation and analysis of mating events that took place a generation ago without the usage of couples' data and PRS as sources.

# 2. Aims

The aim of this work is to investigate key factors in mate choice from a genetic perspective in two present day European populations. This goal is set to be fulfilled by the assessment of contrasts between heterozygosity levels in parts of the genome where SNPs relevant to the trait of interest are located against the rest of the genome.

In order to acquire a numerical representation of the biological phenomenon, a calculated heterozygosity score is obtained through computational means designed for the purpose. The results undergo a series of statistical analysis downstream, which aim to detect signals of genetic assortative mating and to assess its extent on a population scale.

The outlook of this study is to provide a new insight on mate choice, investigating in-depth the genetic perspective of the phenomenon via a novel method that has been applied for the first time in this context. Through a targeted search for signals of assortative mating for various complex traits, the goal is to discover whether patterns of sexual selection are uniform or different across modern day Europe. In addition to the thorough genetic scans, the processed genotype data is juxtaposed against the phenotypes in order to control for extreme geno-phenotype patterns.

# 3. Methods

## 3.1. Data Availability

The current work uses data from the UK Biobank project GWAS and the Estonian Biobank via the access of prof. Luca Pagani due to collaborative research with the University of Edinburgh and his research affiliation with the Institute of Genomics of the University of Tartu, Estonia, which hosts the biobank. Each Biobank provided genotypes for around 800K genome-wide SNPs for each individual. Genotype and phenotype data for 50,000 anonymous and unrelated individuals from each biobank were processed and the per window summary statistics obtained by prof. Luca Pagani using the script developed for the purposes of the study. The aggregated and anonymous results were then further processed on a local machine.

## 3.2. Traits of Interest: Definition, Categorisation, Source

The traits of interest that were analysed in this study, given the available genotype and phenotype data, are divided into 4 broad categories:

- Anthropometrics and pigmentation
- Reproductive behaviour
- Educational attainment
- Subjective well-being

The above-listed classes are composed of traits selected from the GWAS Catalog, maintained by EMBL-EBI. The table attached on the next page contains information about the category, trait, EFO IDs and whether phenotype data per each trait was provided by the respective biobank.

| Anthropometrics and Pigmentation | | | | |
|---|---|---|---|---|
| EFO ID in EMBL-EBI GWAS Catalog | Trait | Number of GWAS-significant SNPs included in the current analysis | UK Biobank | Estonian Biobank |
| EFO_0004339 | Body height | 4919 | + | + |
| EFO_0004340 | Body mass index (BMI) | 4520 | + | + |
| EFO_0004342 | Waist circumference | 4880 | + | - |
| EFO_0005093 | Hip circumference | 4769 | + | - |
| EFO_0004343 | Waist-hip ratio | 1605 | + | - |
| EFO_0007788 | BMI-adjusted waist-to-hip ratio | 4895 | - | + |
| EFO_0007789 | BMI-adjusted waist circumference | 4858 | - | + |
| EFO_0007777 | Base metabolic rate measurement | 11 | + | - |
| EFO_0003924 | Hair colour | 463 | + | + |
| EFO_0006336 | Diastolic blood pressure | 3859 | + | + |
| EFO_0006335 | Systolic blood pressure | 4898 | + | + |
| EFO_0007805 | HDL | 72 | - | + |
| EFO_0007804 | LDL | 33 | - | + |
| EFO_0003949 | Eye colour | 102 | - | + |
| EFO_0009902 | Handedness | 73 | + | + |
| **Reproductive Behaviour** | | | | |
| EFO_0004703 | Age at menarche | 629 | + | + |
| EFO_0004704 | Age at menopause | 334 | + | - |

| | | | | |
|---|---|---|---|---|
| **Educational Attainment** | | | | |
| EFO_0011015 | Educational Attainment | 4983 | + | - |
| **Subjective Well-Being** | | | | |
| EFO_0007878 | Alcohol consumption measurement | 3221 | * | + |
| EFO_0005271 | Sleep duration | 754 | + | + |
| EFO_0002009 | Major depressive disorder | 453 | + | + |
| EFO_0008328 | Chronotype | 1194 | + | - |
| EFO_0006781 | Coffee consumption measurement | 325 | + | + |

*"+" signifies availability of phenotype data on the trait of interest*
*"-" signifies lack of phenotype data on the trait of interest; any results reported on these traits are based only on genotype data.*
*"*" taken as a summed up phenotype of multiple alcohol consumption related categories provided in the UK BioBank phenotype data.*

***Table 1.*** *Traits for each major category with provided EFO ID from the EMBL-EBI GWAS catalogue and information on phenotype availability.*

## 3.3. Heterozygosity Score

The foundation for any further calculations and analyses in this work is the estimated heterozygosity score for each of the windows, which we set to be as wide as 50kb in size, along the autosomal DNA of the sampled individuals.

The score ($\gamma$) of for each window is estimated as:

$$\gamma = \alpha/(\alpha + \beta),$$

where $\alpha$ is the heterozygous positions against the overall positions count in the window ratio, and $\beta$ is the estimated population average

heterozygosity for the window. Thus $\gamma$ is the heterozygosity score per window per individual controlled against the population average.

The purpose of $\gamma$ is to provide a numerical representation of the window heterozygosity per individual, yet normalised against the population average. The implementation of this strategy allows monitoring of heterozygosity levels in different windows across the genome, some of which contain SNPs for the trait of interest, and further comparison in order to detect differences in the distribution of trait-related windows and the rest of the genome for the sample population.

We monitor assortative mating based on heterozygosity scores, as the latter serve as an indicator of the ratio of different haplotypes inherited from the individual's parents. Given the score is lower than the population average for the window, the individual is considered to be more homozygous than the population on average, having inherited more identical haplotypes from their parents. An opposite scenario in which the individual holds a higher heterozygosity score than the population average for this window, would mean the individual is more heterozygous than the population on average, having inherited more different haplotypes from their parents.

Additionally, in order to prevent downstream bias in the analyses, a threshold of available SNPs per window is introduced when the heterozygosity score is computed.

## 3.4. The Delta Test: Primary Screening of Heterozygosity

To provide a generalised view, the delta test was introduced firstly. The delta test essentially is a test which provides the numerical difference between the median genome-wide heterozygosity score for each individual and the median heterozygosity score only for the trait related windows of the same individual, expressed as:

$$\Delta = med(A) - med(B),$$

where *med* signifies median number, *A* - sequence of values of A, containing the genome-wide heterozygosity scores and *B* - sequence of values B, containing the heterozygosity scores only of trait-related windows.

This method is used to provide an initial view of what proportion of individuals in a given population demonstrate extreme levels of heterozygosity, both high and low, at the genetic loci reported to be associated with a given trait. Furthermore, this method facilitates the juxtaposition of the acquired delta scores per individual against the phenotype residuals for the studied complex trait.

The term “residuals” is used to describe the values obtained through multiple linear regression models accounting for the effects of an individual’s sex and age on some numeric phenotype data like BMI, body height, waist circumference etc. Thus the dimensionality of the data was reduced and the juxtaposition of genotype data against the phenotype was possible.

The multiple linear regression models on the UK biobank phenotype data were generated through the Statsmodels Python package by myself controlling for sex and age of the individuals following the formula

$$phenotype = m_1 \times sex + m_2 \times age + age^2 + c,$$

where $m_1$ and $m_2$ denote the slopes of the independent variables and $c$ denotes the residuals. The Estonian biobank residuals were incorporated with the courtesy of Davide Marnetto, University of Turin.

## 3.5. Statistical Analysis: Deciphering Signals of Genetic Assortative Mating

### 3.5.1. Mann-Whitney U-tests

The heterozygosity scores across the genome are further used in comparison of the overall heterozygosity distributions of the trait-related windows and the rest of the genome. In order to correctly identify distribution differences we used the Mann-Whitney U-test, a non-parametric statistical test which aims at identifying whether two independent and unequal in size samples have statistically significant differences in their distributions. The Mann-Whitney test is often called Mann-Whitney-Wilcoxon, as the objective of both the Mann-Whitney and the Wilcoxon test is to compare the distributions between two samples.

However, the Wilcoxon test requires dependent samples of even size, which is practically impossible given the parameters of our analysis. [15]

The U-test is applied to each complex trait, as one of the samples consists of the scores of the SNPs-containing windows, and the other sample contains the scores of the rest of the genome-wide windows. The results we aim for as output are represented by the p-value for each individual in the population sample and if $p < 0.05$, another value, $\boldsymbol{\delta}$, is calculated. $\boldsymbol{\delta}$ estimated as the difference in the distributions between the trait-related windows and the rest of the genome, expressed as:

$$\boldsymbol{\delta} = med(T) - med(R) \quad ,$$

where *T* is the sequence containing all trait-related heterozygosity scores, and *R* is the sequence containing all the rest heterozygosity scores.

Provided there is a statically significant difference in the two distributions, $\boldsymbol{\delta}$ numerically points whether the distribution of the trait-related windows is less heterozygous than the rest of the genome, suggesting PAM, or more heterozygous than the rest of the genome, suggesting NAM.

### 3.5.2. Degree of Assortative Mating Detected

The above-described procedure aims at providing information on what proportion of the sample population demonstrates significant difference in the distributions of the trait and the rest of the genome. This on its own is a good indicator of possible GAM. However, to avoid false discoveries (Type I errors) and to assess the degree of genetic assortative mating for each complex trait, we accounted for multiple testing in two ways : the Bonferroni correction, as a multiple tests correction, and the Benjamini-Hochberg procedure to decrease the false discovery rate (FDR).

The Bonferroni correction aims to filter out any possible false discoveries by setting the threshold p-value as low as the initially set alpha (0.05) divided by the number of individuals forming the population studied. The low p-value serves as a strict threshold, allowing only deep signals of GAM to pass.

Due to the overly conservative nature of the Bonferroni correction [16], another procedure was applied in order to prevent the loss of power. The

Benjamini-Hochberg method of FDR was applied as a milder approach to filter out the significant results on a population scale in parallel. As Waite and Campbell define it, the Benjamini-Hochberg procedure “represents a compromise between the need to correct for multiplicity and the need to conserve power” . [17]

The parallel analyses are introduced in order to ensure maximally that the final results are not excessively conservative, yet to prevent falsely abundant signals of GAM.

## 3.6. Code Availability

The code created for the purposes of this study has been attached as Appendix A.

# 4. Results

## 4.1. Heterozygosity Score Calculation

The heterozygosity data for both biobanks was calculated using the Python script on Appendix A . As each population was equipped with a heterozygosity score, an initial plot of the median heterozygosity on a population-scale was generated for both the UK biobank and the Estonian biobank. As peculiar as it seemed there was an unexpected and sudden decrease in the window-based median heterozygosity score for part of the UKBB individuals, which may indicate inbreeding in a certain individual's family history.

These individuals were excluded from further analyses so as to prevent downstream deviation of the results, due to the lack of information on reasons why this phenomenon was observed. Thus from a total of 50,000 individuals included initially, the downstream data was based on 44,450 individuals, who did not demonstrate visible heterozygosity decrease compared to the rest of the UKBB cohort. The median genome-wide heterozygosity for the UK biobank population is rounded to 0.497. The heterozygosity data of the Estonian biobank, based on 49,646 individuals, showed a consistently lower median genome-wide heterozygosity on a population level with a rounded score of 0.475. This observation is easily explainable by the largely different demographic history of the two populations [26].

## 4.2. Primary Screening

As a preliminary step aimed at evaluating the computed statistics, we performed a primary screening. This term denotes the initial evaluation of individual differences between the genome-wide heterozygosity score and the trait-related scores, also called delta score ($\Delta$). The delta score is bound to have value either higher than 0, marking a lower heterozygosity score for the trait-related genetic windows than the genome-wide median, or lower than 0, marking a higher heterozygosity score for the trait-related genetic windows compared to the genome-wide median. In this context, the results of the primary screening are supposed to serve as an initial assessment of the probability of GAM in both populations.

Additionally, the results of the primary screening are used as reference genetic markers when plotted against the residual phenotypes from both biobanks. This strategy is facilitated by the facts that all SNPs included in the trait analysis are located on autosomes, and all continuous phenotypes are taken as residuals after controlling for sex and age impact.

### 4.2.1. Anthropometrics and pigmentation

The primary screening of the traits assigned to the “Anthropometrics and pigmentation” category, showed a tendency for higher proportion of the Estonian population having $\Delta > 0$, compared to the UK population (Fig.2A) This finding indicates there is a higher percentage of individuals in the Estonian biobank, whose parents are genetically similar for traits in this category. Exceptions from this overall observation are the categories of high-density lipoprotein cholesterol (HDL) and basal metabolic rate (BMR). For HDL, the UK population demonstrated a slightly higher percentage of individuals exhibiting primary similarity for the trait, with 48.03% against the 47.18% demonstrated by the Estonian one. Although the difference is comparatively insignificant, it points at a higher percentage of Estonian individuals whose parents are not genetically similar for HDL genomic windows. Due to a threshold of SNPs available for each window, set at the beginning of the experiment, genotype data for BMR is absent on Fig. 2A for the Estonian biobank sample. Thus the two populations cannot be compared. However, the share of the population demonstrating similarity between mates for BMR in the previous generation is highest in this category with its 55.2%, and third among all traits, surpassed only by chronotype (62.47%) and age at menopause (59.34%).

### 4.2.2. Reproductive behaviour

The scan for autosomal similarities between mates from the previous generation for the age at menarche and menopause showed a different primary pattern. The age at menarche seems to have scored a lower result in the UK population (44.91%) compared to the Estonian one (52.81%). In contrast, the age of menopause demonstrates the second highest similarity in the UK population with 59.34% compared to the Estonian 54.56%. Thus a reversed pattern of similarity for traits in this category can be observed in the primary screening, as depicted in Fig.2B.

### 4.2.3. Educational Attainment

Results show Estonian mates from the previous generation seem to have been more similar in windows related to educational attainment, across the autosomal DNA, with 48.31% of $\Delta > 0$ against the UK's 31.44% (Fig. 2C).

### 4.2.4. Subjective Well-Being

This section aimed to provide an insight on how similar mates from these two different populations tend to be for traits of social relevance. Among the investigated traits, the UK biobank share of individuals holding $\Delta > 0$ was highest for chronotype (Fig. 2D). This is the trait among all the others for which the UK population demonstrates the highest level of similarity in trait-related windows across the genome (62.47%). All other traits demonstrated the traditionally higher proportion of Estonian individuals with more homozygous trait-related windows.

Thus, the primary screening for similarity in the trait-related parts of the genome shows a consistent trend across both populations. Although extreme differences between the two populations are absent, there is a predominantly higher proportion of individuals from the Estonian biobank whose parents have passed on similar genetic characteristics for the traits of interest. This tendency, of course, can be linked with the genome-wide lower heterozygosity score, which was already discussed in section 4.1. Nevertheless, a few significant differences in favour of the UK biobank sample are present, where the share of individuals with $\Delta > 0$ is higher - HDL, age at menopause and chronotype.

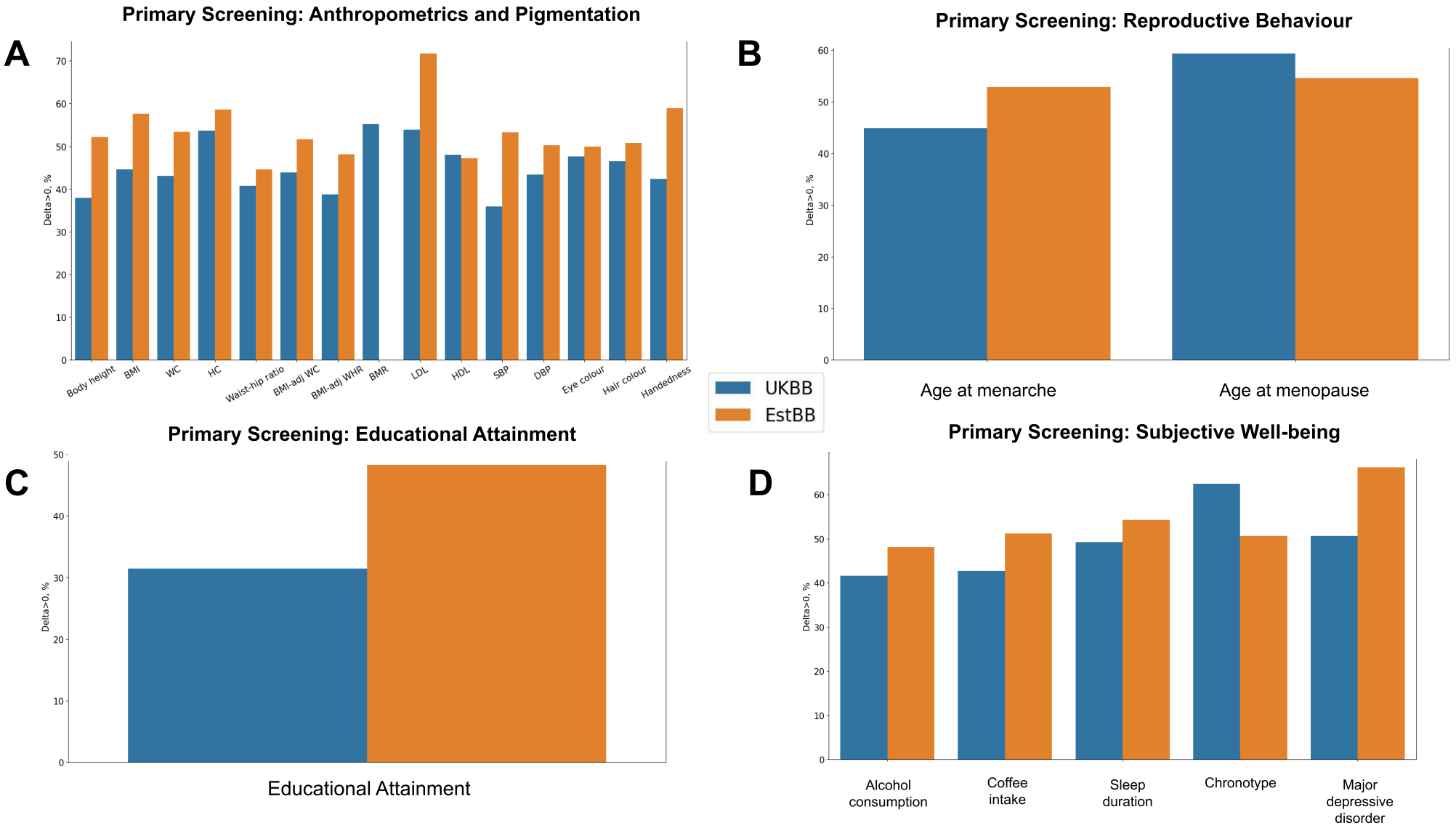



***Figure 2.*** *Comparison between groups exhibiting lower heterozygosity for the trait in the UK Biobank (blue) and the Estonian biobank (orange) for the following categories:* ***A****) Anthropometrics and pigmentation,* ***B****) Reproductive behaviour,* ***C****) Educational attainment and* ***D****) Subjective well-being*

## 4.3. Comparison of Genotype - Phenotype Data

In order to understand better how phenotypes and heterozygosity scores correlate, a juxtaposition of the available phenotypic data against the median heterozygosity scores for the traits and the primary screening results was introduced. Plotting the available data, linear regression models and the densities per phenotype in each trait, a complete scan for correlating deviations in both geno- and phenotype was performed. The information regarding phenotype was used as absolute residuals of multiple regression models accounting for the phenotype dependencies on sex and age of the individuals, as pointed in section 3.4.

### 4.3.1. Anthropometrics and pigmentation

Initially, the concept of low heterozygosity scores at the loci associated with a given trait, compared to the rest of the genome, may also point to the expression of an extreme phenotype. To elaborate on this part, the following example may be provided: if the offspring of mates with haplotypes for higher than the average body height give birth to an offspring with a low heterozygosity score for height, one would automatically expect that the offspring would also demonstrate the phenotype by displaying a  higher than the average height. Controversially, the results did not actually correspond to such an expectation. The information gained during the comparison shows individuals holding low trait median heterozygosity scores actually express no deviations or low-to-mild ones, as shown in fig. 3A.

Given that height is a largely polygenic trait with more than 4919 genome-wide significant SNPs involved in the analysis, it could be assumed that the lack of initial filtering may be a premise for vague downstream results. However, traits such as eye colour (100 SNPs), low-density lipoprotein cholesterol (33 SNPs) and BMR (16 SNPs), which hit at least 40-fold less parts of the genome, still exhibit the same behaviour in both their correlations.

The pattern observed above kept repeating throughout both datasets for the anthropometrics and pigmentation category. No significant deviation in phenotype was observed in individuals highly homozygous for the traits neither in trait-median-vs-phenotype correlation nor in the delta-score-vs-phenotype correlation. Individuals who aligned at the

extreme phenotypes demonstrated median heterozygosity for traits close to the population-average.

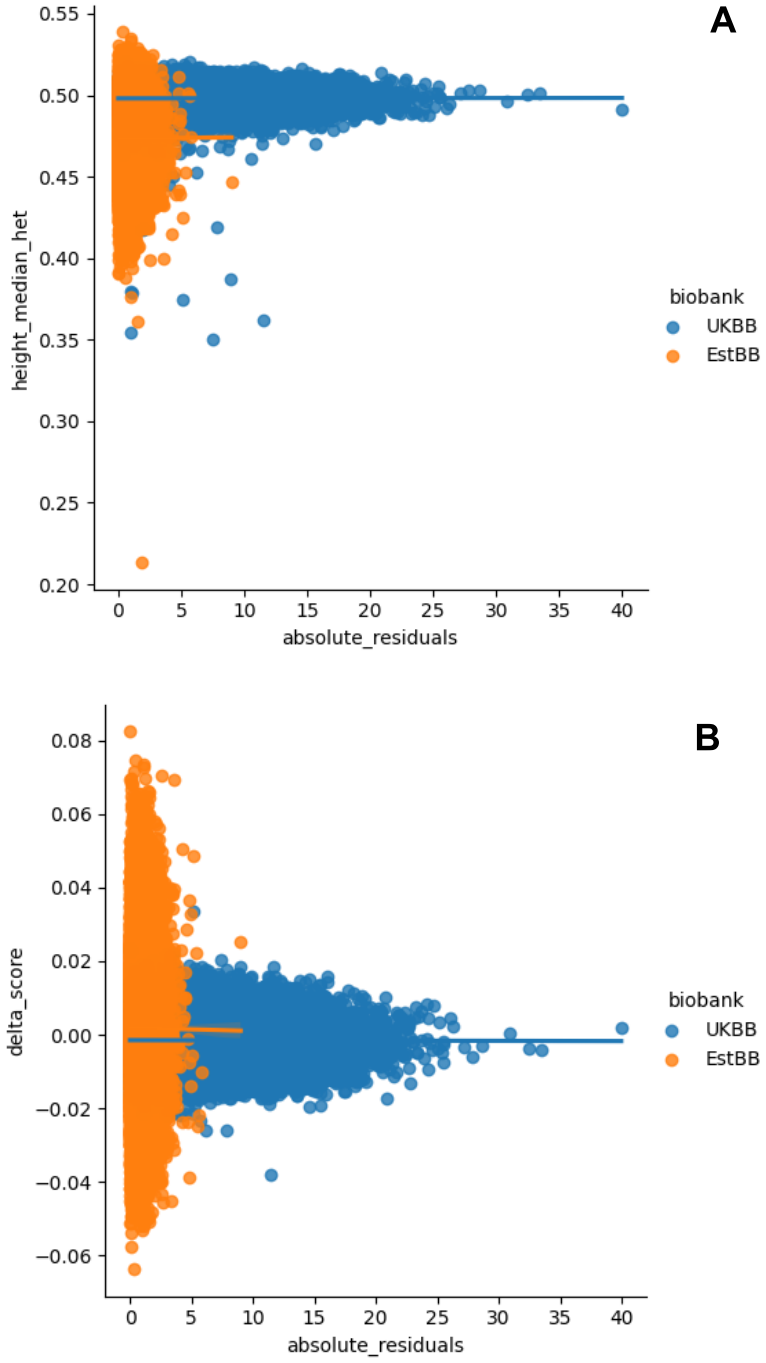



***Figure 3.*** *Linear regression models for **A)** height heterozygosity median score against absolute phenotypic residuals and **B)** primary screening result against absolute phenotypic residuals.*

Due to some differences in the phenotypic data available in the UKBB and the EstBB, marked in Table 1, information on a few categories was only available for one of the two biobanks, and the availability of sex-indicated

individuals in the UKBB allowed for a detailed linear regression models examination. Due to the availability of the sex-indicated individuals from the UKBB, the linear regression models for hair colour showed some unexpected difference in the correlation trends between male and female individuals . As hair colour can be considered a discrete phenotype, correlations between the genome-wide medians and the delta scores were seeked. Among the available phenotypes in the UKBB, the “dark brown” and “other” phenotypes did not show any significant correlation (Fig. 4A, B), whereas the “blonde” and “light brown” phenotypes were positively correlated (Fig. 4C, D).

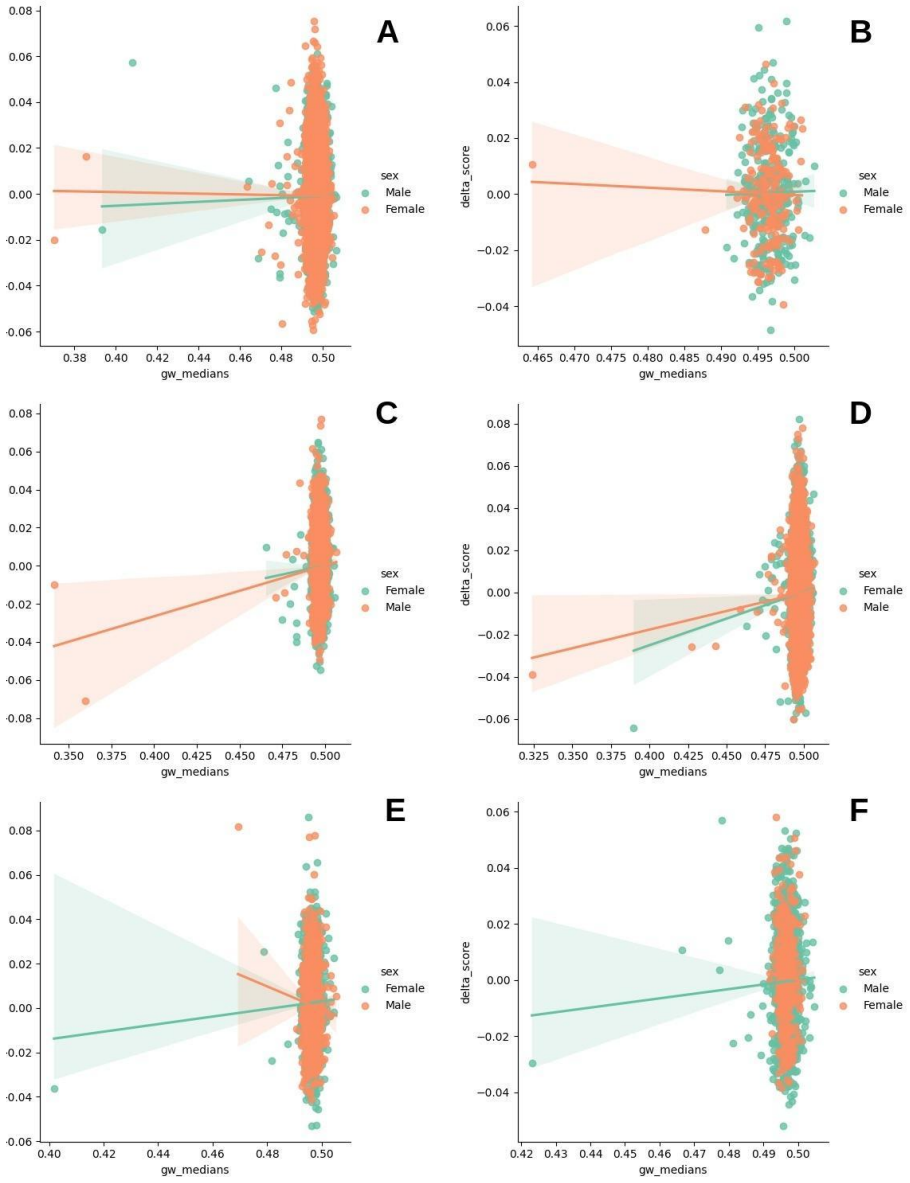



***Figure 4.*** *Hair colour correlations between genome-wide heterozygosity and delta scores for* ***A)*** *dark brown,* ***B)*** *other,* ***C)*** *blonde,* ***D)*** *light brown,* ***E)*** *red and* ***F)*** *black hair colours.*

However, a different trend was spotted between the female and male individuals in the red hair and black hair phenotypes. The delta scores correlation to the genome-wide median is positive in females, while for males the correlation between the two is negative (FIg. 4E). On the contrary, female individuals with black hair demonstrate a short-span negative correlation between the delta scores and their genome-wide medians, while males show a positive one (Fig.4F). This finding looks in detail at the sex-based difference in the correlation between the genome-wide heterozygosity and the hair colour primary assortment score in individuals with red and black hair from the UKBB.

Other traits related to anthropometrics and pigmentation did not demonstrate differences between the correlations as previously described. However, the HDL, LDL and eye colour data in the Estonian biobank cohort showed a certain, but not predominant, proportion of individuals, demonstrating outlined homozygosity for the traits. Those individuals were widely spread across the absolute residuals axis for HDL and LDL (Fig.5), which points at the fact that extreme homozygosity scores for the trait, consequently delta scores ($\Delta HDL_{max}$ = 0.5, $\Delta LDL_{max}$ = 0.5), could be linked both with extreme phenotypes, and with standard, non-deviating ones. No significant correlation trend was found.

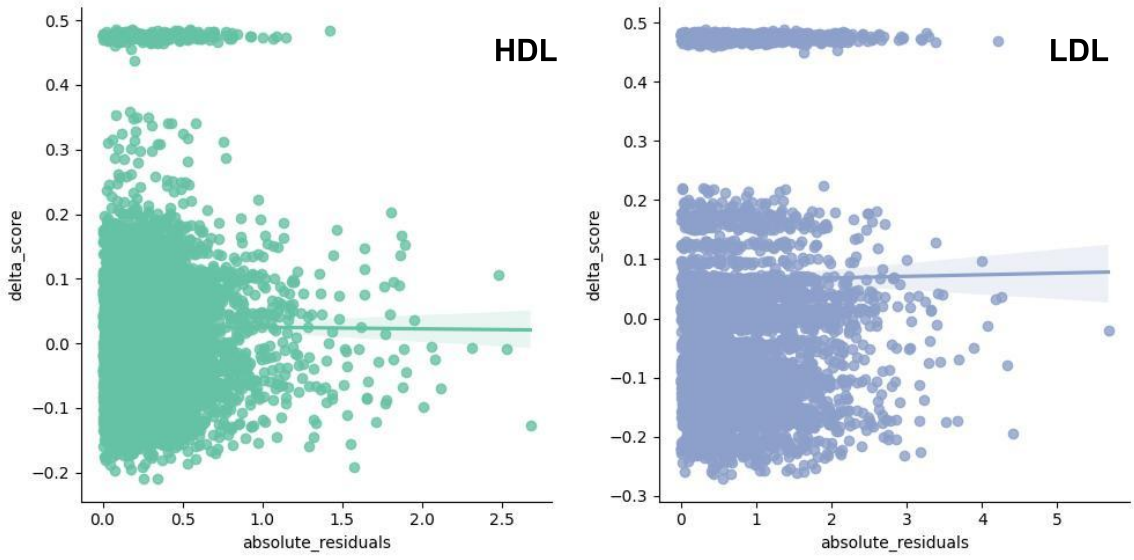



***Figure 5.*** *Linear regression models of HDL and LDL delta scores against the absolute residual values.*

The same strong homozygosity is observed also in a group of people for the eye colour characteristics. Considering eye colour is a discrete phenotype, the generated linear regression models look for any specific correlation between genome-wide medians and delta scores. None of the phenotypes showed a significant correlation. However, a slight gradient of individuals with predominantly light-coloured eyes can be observed in a span of 0.06 score units. The group of individuals at the top of Fig.6, expressing outstandingly high delta scores also span through the whole range of genome-wide medians, and surprisingly, this group is composed of individuals with all the eye colour phenotypes. Therefore, a correlation pattern can be drawn for none of the available eye colour phenotypic data.

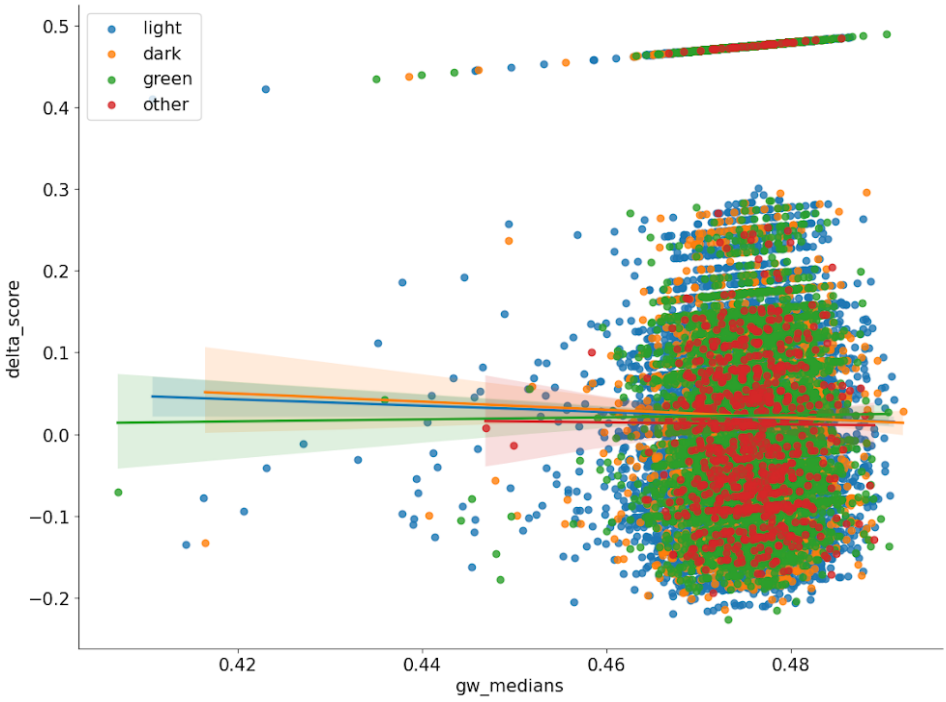



***Figure 6.*** *Linear regression models of eye colour delta scores against genome-wide medians from the Estonian biobank cohort.*

### 4.3.2. Reproductive Behaviour

All of the procedures described in section 4.3.1 were applied to both traits included in the reproductive behaviour category: age at menarche and age at menopause. Phenotypic data on age of menarche was provided by both

cohorts, while age at menopause data was available only for the UKBB cohort. The generated linear regression models did not show any significant trends. In addition, the data distribution was uniform suggesting no distinguished genotype-phenotype patterns.

### 4.3.3. Educational Attainment

The educational attainment phenotype data was only available for the UK population, which limited the scope of the genotype-phenotype comparison for this trait. Looking for patterns linking the age of completed education with heterozygosity scores, it became evident the mean age of completed education for the population was between 15 and 16 years of age. This finding motivated a deeper analysis of the cohort.

According to phenotype data from the UK Biobank, the average age of individuals who have provided information about this trait, is 57 years. The individuals aged 30 to 40 years from the entire dataset are 276. Taking into consideration the weight of the rest of the population against 276 individuals, it is understandable why the mean age of completed education is shifted towards lower values. In a further attempt to link the findings with events of educational relevance, the Raising of School Leaving Age in England and Wales (also known as ROSLA) provides a plausible explanation. This event provides an overview of how the age of obligatory education has changed over less than a century. Reforms on school leaving age were introduced twice in the period which may be linked with the shifted value in the analysis - one in 1947 setting the age of leaving school to 15, and another one in 1972 increasing it to 16. A logical link could be made between the average age of the individuals recruited in the UKBB cohort and the ROSLA reforms throughout the 20th century [25].

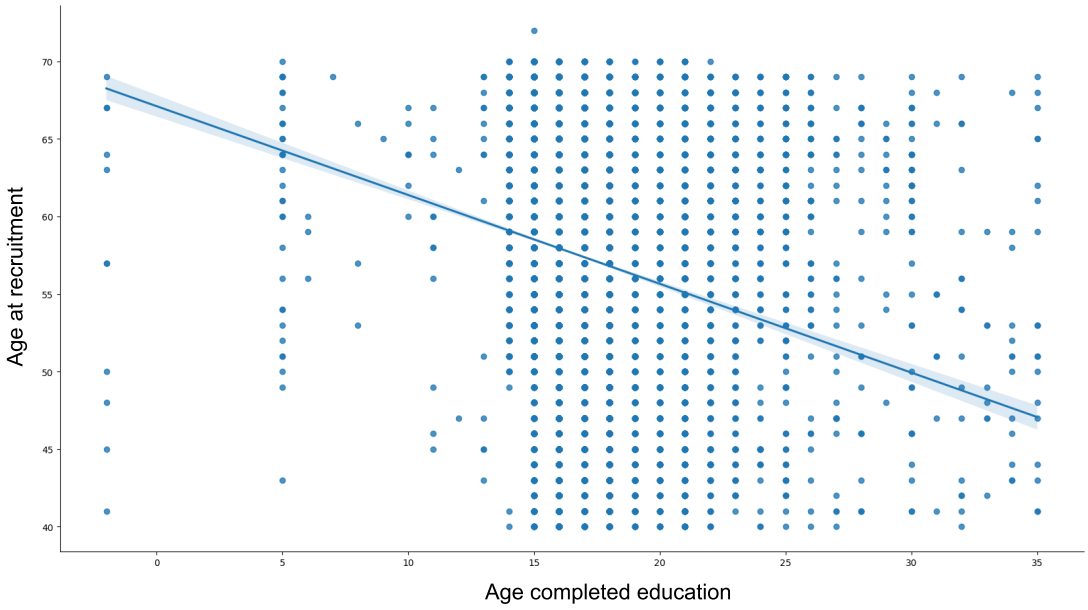



***Figure 7.*** *The ROSLA footprint on the correlation between age of recruitment and educational attainment.*

Thus, the most densely populated groups are the ones of the individuals who concluded their education at the age of 15 and 16. However, there is a moderate gradient towards an increasing age of completed education (Fig.7), corresponding with a well expressed negative correlation between the age of recruitment of the individuals in the population and their school leaving age.

The event, however, did not produce a specific pattern from a genetic point of view, with the exception of a neglectable number of individuals who demonstrated decreased heterozygosity for the trait, but within the 0 to 2.5 years residuals span.

### 4.3.4. Subjective Well-Being

Apart from a general investigation into whether there is a specific correlation between the genotype and phenotype data for any of the traits in this category, a parallel comparative analysis between the two populations was made. No specific correlation between the delta score and the absolute residuals of the phenotypes were identified. However, the Estonian biobank samples were spread across the y-axis in the range $\Delta_{min\ ;\ max}$ = [-0.10; 0.10], while the UK biobank ones were positioned in the range $\Delta_{min\ ;\ max}$ = [-0.03; 0.05]. This would indicate individuals from the Estonian

cohort demonstrated higher levels of both homozygosity and heterozygosity for the trait, compared to the UK one. Comparing the phenotypes in both biobanks associated with highest genetic similarity in the parents of the individuals, the UK one was associated with consumption of alcohol 3 to 4 times a week, and in the Estonian samples - approximately 0.5 units deviation from the mean.

The parts of the genome related with sleep duration showed a similar behaviour as the ranges of the delta score for the two populations were respectively $\Delta_{min\ ;\ max}$ = [ -0.10; 0.15] for the Estonian, and $\Delta_{min\ ;\ max}$ = [-0.04; 0.08] for the UK one. For both populations a higher upper bound was observed, suggesting the presence of individuals who have inherited more similar genetic information for this trait from their parents rather than dissimilar. The observed values were not linked to an extreme phenotype, with the maximum values in both populations exhibiting a deviation from the respective means by not more than 2 hours of sleep.

The tendency of the Estonian cohort to show wider delta score variation was repeated in the coffee consumption group (UK $\Delta_{min\ ;\ max}$ = [-0.05;0.15], Estonian $\Delta_{min\ ;\ max}$ = [-0.2; 0.5]) and the major depressive disorder group (UK $\Delta_{min\ ;\ max}$ = [-0.1; 0.4], Estonian $\Delta_{min\ ;\ max}$ = [-0.2; 0.5]). In terms of phenotypic expression, the coffee consumption in the UK cohort was not linked with a specific behaviour, rather a repetition of the previous traits for which the highest $\Delta_{min\ ;\ max}$ stuck close to neglectable deviations from the mean. The highest delta scores were detected for coffee intake not higher than 15, which is more than 25 units from the maximum extreme. In contrast, the model generated for the Estonian cohort, showed the standard consumption on a population scale is between 4 and 7. Although there was a gradient-like pattern of increasing consumption, the majority of the samples lied in the 4-to-7 range, with the highest $\Delta_{min\ ;\ max}$ lying there as well.

The findings in this category do not differ greatly from the rest of the results. The genotype-phenotype imposition applied to all traits with an available phenotype in at least one of the biobanks, shows the probability for an observed phenotype linked with aver heterozygosity for the trait-related sections of the genome is not limited to the extreme phenotypes, but is often observed

The results in this section point to a tendency of larger absolute differences between the whole genome heterozygosity and the trait-related windows in the Estonian biobank, despite the population-wide lower heterozygosity (0.475). Nevertheless, these results are based on the

primary screening data, which only assessed arithmetically the data from the two biobanks. Further statistical analyses are necessary to reveal to what extent the homozygosity or heterozygosity of the trait-related sections can be considered genetic assortative or disassortative mating.

## 4.4. Statistical Analyses: Proving GAM

The primary screening is able to provide an initial image of the groups per population demonstrating a difference between their genome-wide median heterozygosity score and the trait of interest score. Although these results can be used in the assessment of the phenotype-genotype correlations, they are not enough to prove the existence of genetic assortative mating on population scale. It was necessary to introduce a series of statistical tests to understand the degree to which the differences between genome-wide and trait scores are significant per population.

Applying the Mann-Whitney U-test to compare trait-related with genome-wide windows, as per reasons explained in detail in section 3.5.1, it was possible to acquire per individual p-values at each given trait. However, the individual p-values do not provide sufficient information about the driving factors of mate choice common for the entire population. Therefore a further filter in the form of Bonferroni correction and Benjamini-Hochberg procedure for FDR were applied, as reasoned in section 3.5.2.

The results of both procedures were identical in confirming and rejecting the presence of PAM or NAM for each trait in Table 1. Except for the numerical difference in the results, due to the more restrictive or liberal nature of the methods, the patterns repeated for each trait for both methods. Therefore, the graphs supporting and clarifying the findings are based on the Bonferroni correction results, as it is the stricter form of assessment.

### 4.4.1. Anthropometrics and pigmentation

The statistical analyses showed clear signals of positive assortment (PAM) for several traits - waist circumference, hip circumference, BMI-adjusted waist circumference, BMI-adjusted waist-hip ratio. A comparison between the strength of the signals from the UK cohort and the Estonian one reveals a 2-fold to 10-fold difference in favour of the

Estonian population. A deeper look at Fig.8 also presents a very slight signal for positive assortative mating from the Estonian population for systolic blood pressure (1 individual from the entire cohort).

On the other hand, there were also sound negative assortment (NAM) results for this category. The UK biobank population demonstrated signals for NAM for more traits than it did for PAM. As visible on Figure 9, the strongest signal for NAM comes from the Estonian biobank population and is for the BMI-adjusted waist-hip ratio. However, for all the other traits the UK biobank cohort shows persistently a stronger NAM signal. Traits that were found to be under NAM from the UK, are body height, BMI, waist circumference, hip circumference, waist-hip ratio, BMI-adjusted waist circumference, BMI-adjusted waist-hip ratio, systolic blood pressure, diastolic blood pressure and hair colour. All of those traits are predominantly under negative assortative mating in the UK biobank individuals, according to the numbers acquired during the statistical tests.

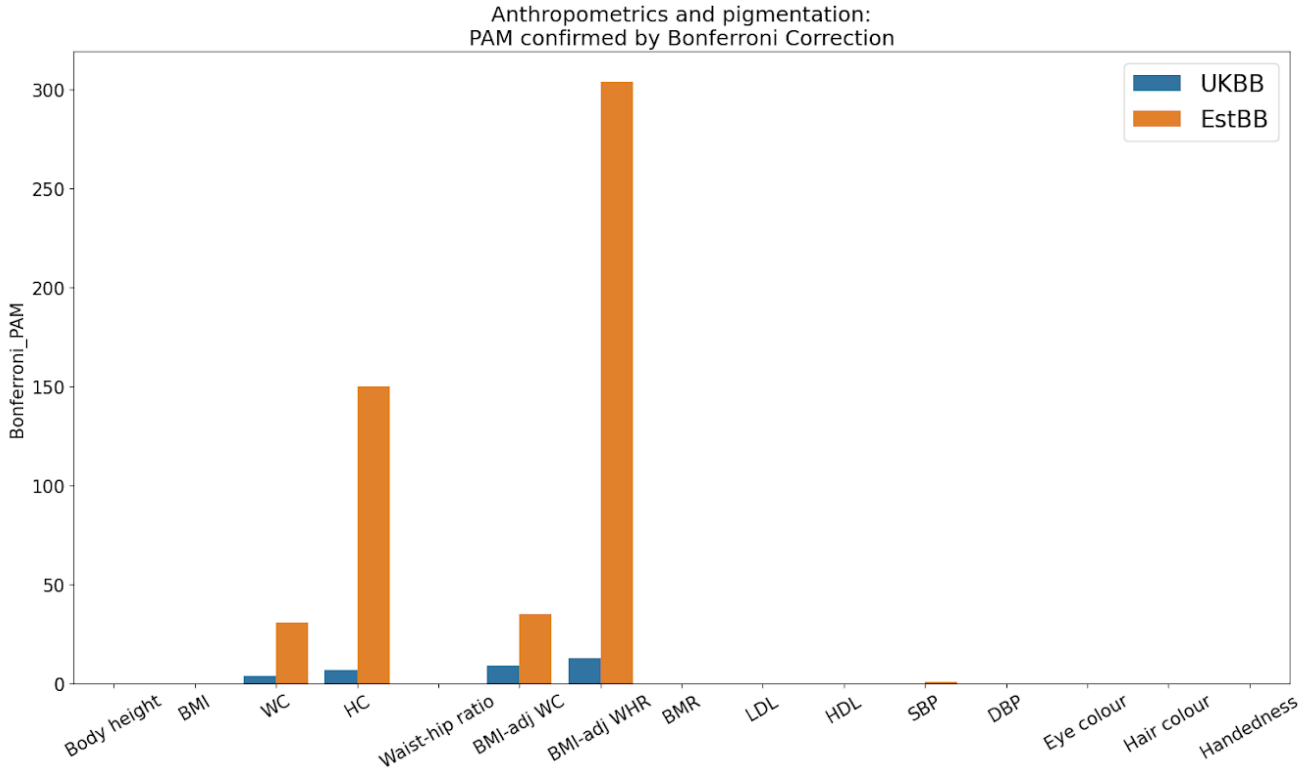



***Figure 8.*** *Number of PAM individuals confirmed by Bonferroni correction for traits in Anthropometrics and pigmentation category (WC standing for waist circumference, HC for hip circumference, WHR for waist-hip ratio).*

The predominant proportion of individuals showing clear signs of disassortative mating in the UK biobank would essentially signify that their

biological parents did not share similar haplotypes for these traits, thus the “opposites attract each other” hypothesis would be rather accepted, than rejected in this case. This statement could easily be defended as the results from the Estonian biobank cohort do not indicate such a behaviour. On the contrary, they deeply resonate with a population where mates tend to be much more genetically similar than dissimilar for traits, concerning pigmentation, metabolism and body characteristics.

The statistical tests applied in this work cannot provide an explanatory evidence of *why* there is such a difference in the mating patterns concerning anthropometrics between the two populations studied. They are designed so as to determine the patterns already present in the studied individuals, rather than seek the reasons behind it.

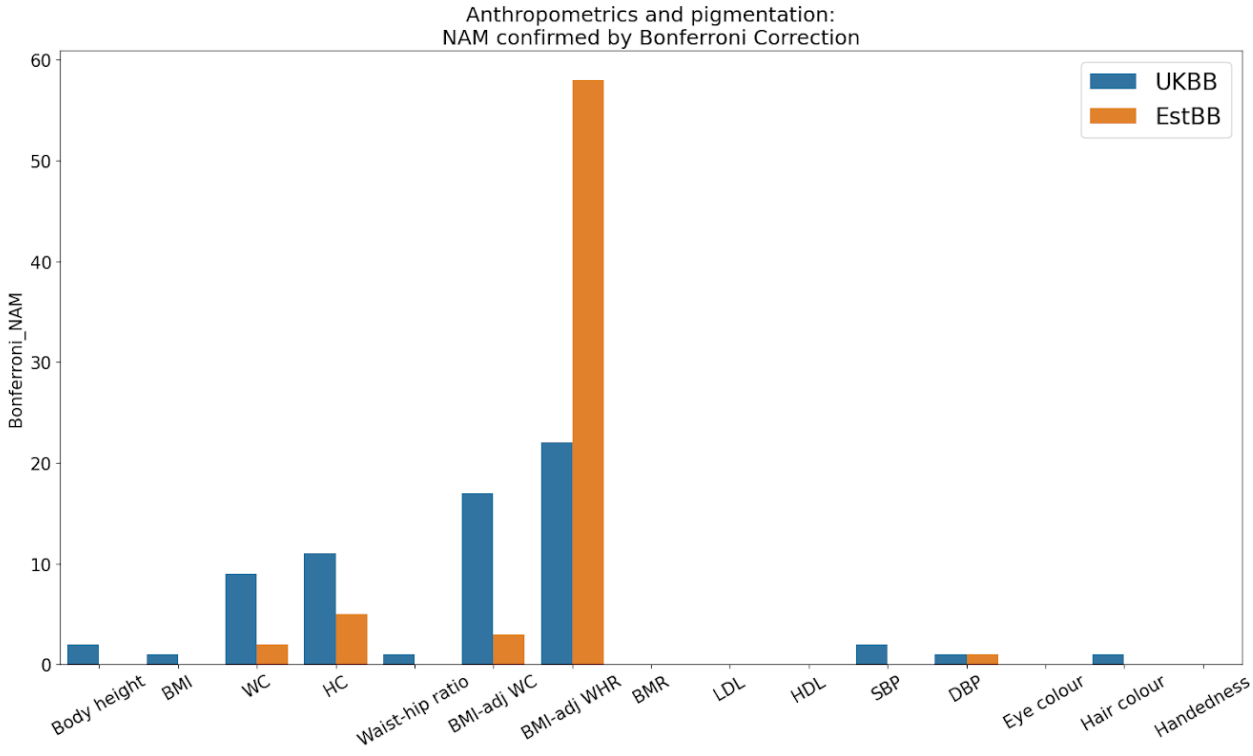



***Figure 9.*** *Number of NAM individuals confirmed by Bonferroni correction for traits in Anthropometrics and pigmentation category (WC standing for waist circumference, HC for hip circumference, WHR for waist-hip ratio).*

### 4.4.2. Reproductive Behaviour

The genetic windows associated with the ages of menarche and menopause did not show outstandingresults in terms of neither PAM nor NAM. The only results who showed significance after the performed statistical analyses come from the UK biobank cohort and are very slight, which in numbers translates as 1 individual out of the entire cohort per trait.

Due to the low signals for both traits in the UK biobank and the lack of any signal from the Estonian biobank, a comparison of the polygenicity of the traits was made against some of the anthropometric features which showed significance, either in PAM or NAM. The age of menopause analysis included the scan of genetic windows across the genome containing a total of 334 SNPs, both trait increasing and trait decreasing. Identical procedure was followed for the age of menarche, containing a total of 629 SNPs. In comparison, traits from the anthropometrics and pigmentation category contained 5 to 10-fold more SNPs included in the analyses: body height (4,919), BMI-adjusted waist-hip ratio (4,895), waist circumference (4,880), BMI-adjusted waist circumference (4,858), hip circumference (4,769) and waist-hip ratio (1,606). Thus, the hypothesis that the weak signal for traits in reproductive behaviour originates from their polygenicity does not seem to be credible.

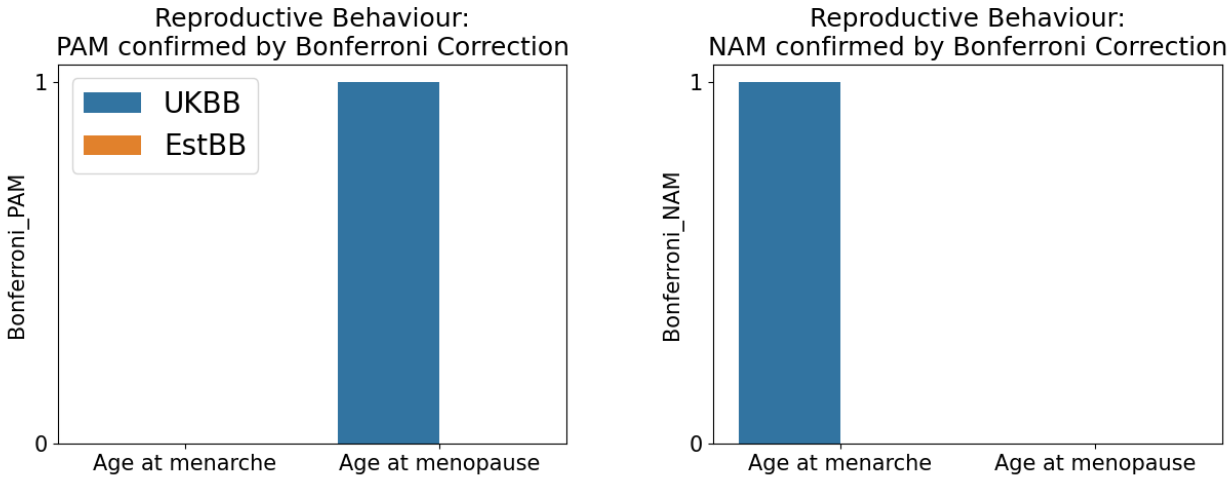



***Figure 10.*** *Individuals affected by PAM and NAM in traits from the Reproductive behaviour category from both cohorts.*

Despite the scarce results, a check-up for positive and negative assortative mating was done. The signal for PAM comes from the age at menopause, whereas the NAM signal comes from the age at menarche trait. The results, however, are not as explicit as the ones acquired for the anthropometric traits. Therefore, the weak to non-existing signals from both cohorts make it difficult to outline a specific genetic signature of these traits as assortative mating factors.

### 4.4.3. Educational Attainment

Educational attainment is one of the categories which PRS research outlines as a strong factor in positive assortative mating (*citations here)*. The statistical filtering applied to the Mann-Whitney test results aimed at confirming whether the findings would replicate using the novel method. Despite the predominance of Estonian biobank individuals having positive delta scores as described in section 4.2.3, the degree of assortative mating was found to be equal in both cohorts.

Although the strength of the signals from both biobanks was equal, the patterns in terms of positive and negative assortative mating differed. It was found that 75% of the total signal from the Estonian population showed mates tend to be genetically alike for genetic windows associated with educational attainment, while only 25% tend to be under NAM for the trait (fig. A). In contrast, the UK biobank cohort demonstrated solid 100% of NAM for the trait (fig. B).

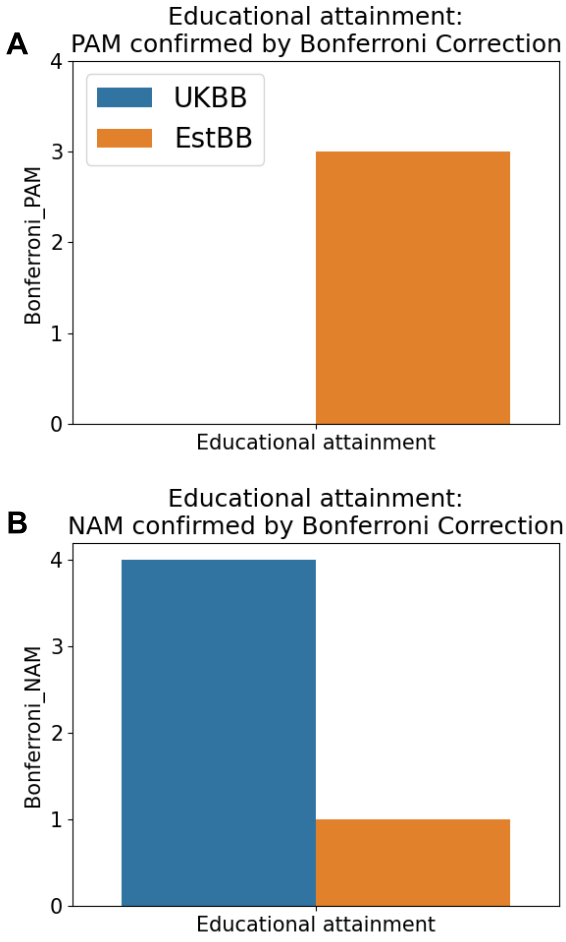



***Figure 11.*** *Individuals affected by PAM and NAM for educational attainment from both cohorts.*

### 4.4.4. Subjective Well-Being

One of the best expressed signals for assortative mating was found to belong to this category, namely the major depressive disorder. The confirmed cases of GAM for depression status in the UK reached a total count of 68, out of which 46 for PAM and 22 for NAM. At the same time, the Estonian biobank reached a count of 10 cases, out of which 7 for PAM and 3 for NAM.

Although a stronger signal for GAM could be expected in regard to anthropometric measurements, the highest number of individuals under GAM in the UK cohort is linked to the depression status. A glance at Table 2 clarifies that the major depressive disorder trait passed the Bonferroni correction test with approximately 2-fold higher count (68) that one of the most significant body characteristics - the BMI-adjusted waist hip ratio (35). This finding means the most important trait under GAM in the UK biobank samples is the depression status.

For the Estonian biobank samples, the depression status was found to be the only trait from the subjective-well being category to pass the Bonferroni correction. Taking into consideration the numbers in Table 2, it is possible to state that the Estonian biobank cohort shows stronger signals of GAM for depression status rather than educational attainment or reproductive behaviour.

In addition to the major GAM factor discovery, the UK cohort demonstrated PAM for chronotype (3) and coffee intake (1), and NAM for alcohol consumption (1). The results for positive assortative mating for chronotype match with the previous findings from the primary screening, where the UK biobank showed a greater proportion of individuals whose biological parents tend to carry similar haplotypes.

| Trait | UKBB | | | | | | EstBB | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bonferroni correction | | | BH FDR | | | Bonferroni correction | | | BH FDR | | |
| | total | PAM | NAM | total | PAM | NAM | total | PAM | NAM | total | PAM | NAM |
| Body height | 2 | 0 | 2 | 11 | 1 | 10 | 0 | 0 | 0 | 0 | 0 | 0 |
| BMI | 1 | 0 | 1 | 7 | 1 | 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| Waist circumference | 13 | 4 | 9 | 790 | 215 | 575 | 33 | 31 | 2 | 2167 | 1526 | 640 |
| Hip circumference | 18 | 7 | 11 | 920 | 281 | 639 | 155 | 150 | 5 | 5493 | 4083 | 1407 |
| Waist-hip ratio | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| BMI-adjusted waist circumference | 26 | 9 | 17 | 1149 | 399 | 750 | 38 | 35 | 3 | 2825 | 1878 | 946 |
| BMI-adjusted waist-hip ratio | 35 | 13 | 22 | 2067 | 501 | 156 | 362 | 304 | 58 | 9098 | 4966 | 4129 |
| BMR | 0 | 0 | 0 | 0 | 0 | 0 | - | - | - | - | - | - |
| LDL | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| HDL | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| SBP | 2 | 0 | 2 | 24 | 3 | 21 | 1 | 1 | 0 | 1 | 1 | 0 |
| DBP | 1 | 0 | 1 | 48 | 7 | 41 | 1 | 0 | 1 | 8 | 3 | 5 |
| Eye colour | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Hair colour | 1 | 0 | 1 | 5 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 |

| Trait | UKBB | | | | | | EstBB | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bonferroni correction | | | BH FDR | | | Bonferroni correction | | | BH FDR | | |
| | total | PAM | NAM | total | PAM | NAM | total | PAM | NAM | total | PAM | NAM |
| Handedness | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Educational attainment | 4 | 0 | 4 | 208 | 0 | 208 | 4 | 3 | 1 | 70 | 43 | 27 |
| Age at menarche | 1 | 0 | 1 | 3 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| Age at menopause | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Alcohol consumption | 1 | 0 | 1 | 9 | 1 | 8 | 0 | 0 | 0 | 0 | 0 | 0 |
| Coffee intake | 1 | 1 | 0 | 3 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sleep duration | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Chronotype | 3 | 3 | 0 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Major depressive disorder | 68 | 46 | 22 | 6573 | 2569 | 4004 | 10 | 7 | 3 | 3692 | 2791 | 899 |

***Table 2.*** *Summary count of post-MCT and post-FDR signals of genetic assortative mating in the studied populations.*

### 4.4.5. Common Tendencies

#### 4.4.5.1. The UK cohort

Having obtained a full-scale picture of the genetic assortative mating processes taking place in the UK cohort, it can be concluded that the traits demonstrating strongest tendency to be under GAM in the 44,450 individuals surveyed here are major depressive disorder (68), BMI-adjusted waist-hip ratio (35), BMI-adjusted waist circumference (26), hip circumference (18) and waist circumference (13).

The major trend among the traits regarding physical appearance converges towards NAM, implying a tendency for attraction between individuals who are genetically dissimilar for anthropometric characteristics On the other hand, deciphering the mating preferences regarding depression status, reveals the opposite pattern. The majority of the individuals prefer genetically similar mating partners in regard to their depression status supported by a 68% fraction of individuals under PAM, whilst the rest of the individuals demonstrate an inclination for partners with different haplotypes, resulting in the NAM counts in table 2.

#### 4.4.5.2. The Estonian cohort

The top five complex traits for which the 49,646 individuals of the Estonian cohort tends to be under GAM are BMI-adjusted waist-hip ratio (362), hip circumference (155), BMI-adjusted waist circumference (38), waist circumference (33) and major depressive disorder (10).

In contrast to the data from the UK, the GAM signal in the Estonian biobank is mainly composed of PAM, with the lowest percentage of PAM being 83.9% for anthropometric traits, and 70% for major depressive disorder, based on the data in table 2.

These findings mark a significant difference between the patterns of genetic assortative mating in the two populations, as although the complex traits in the spotlight are the same, but inclination towards shared similar or dissimilar haplotypes in the genetic windows associated with these traits is different.

#### 4.4.5.3. Detection resolution

Two consecutive procedures were followed in this study in order to evaluate the preliminary possibilities of assortative mating (section 3.4) and then sort out the actual traces of assortative mating in both populations (section 3.5). The results of the primary screening suggested higher chances for GAM, especially PAM, in the Estonian cohort, as the percentage of individuals demonstrating $\Delta > 0$ was generally higher. However, these expectations were not met by the final results. Firstly, the same complex traits took the main role in GAM processes in both populations. Secondly, the primary screening suggested very good chances of PAM for LDL, educational attainment or even major depressive disorder. These chances, however, were eliminated by the statistical procedures downstream.

This comparison proves the primary screening procedure as a useful tool for phenotype-genotype juxtaposition. However, the primary screening and the ranges of the delta scores failed to predict the presence of GAM and its patterns, acquired further.

Thus, evaluating the necessary resolution in order to detect GAM and identify its type, a statistical pipeline relying on MTC and FDR procedures seems to be the most reliable solution applied so far.

# 5. Discussion

## 5.1. Shared patterns

To review assortative mating in full, one needs to consider both the reasons behind its occurrence and the genetic footprint it leaves. Some of the variables modulating the choice of partner include phenotypic preferences, phenotype convergence over time and the extent to which social and cultural factors influence mate choice[8]. Those variables are not necessarily ubiquitous, as populations tend to have different backgrounds in terms of history, culture, religion and other values forming their perception on sexual selection, or in a more socially applicable context, marriage. Depending on the available data, research through the last decades have made a transition from assessing purely phenotypic self-reported characteristics - like the Eysenck's Big Three (extraversion, neuroticism and psychoticism) [3, 18] or blood pressure records [4], to looking for genetic footprints in genotype data from different biobanks and projects globally, like the UK Biobank [2, 12, 13], MoBa (Norway)[9] , BioBank Japan [13] , 23&Me[2] and 1000 Genomes Project [2].

The existence of assortative mating in couples has been proved on the bases of phenotype data [3,4,7], polygenic risk scores [10] and gametic phase disequilibrium (GDP) [12,13]. A novel method has been developed and utilised in this study, giving the possibility to detect genetic assortative mating in the previous generation. An advantageous feature of this method is the fact it focuses on an individual's genome as a source of information, rather than requiring data on siblings [9] or spousal couples [2-5]. This approach has already been undertaken by Yengo *et al.* [12] (2018) Yamamoto *et al.* [13] (2022) for GPD estimation on odd and even chromosomes. Apart from maximising the data availability, the single-genome approach has been proved as a reliable detector of genetic assortative mating footprints in the offspring of mates, without the need for analysing their own genotype data.

In this work, the single-genome approach was applied in the context of heterozygosity score calculation based on the available genotype data. In contrast to the GPD estimation, which accounts for the distribution of trait-increasing alleles only [12], the current method provides a genome-wide database of individual scores controlled against the studied population, which can be downstream filtered according to the parameters of the research. The results in this work are produced by taking into

consideration all of the available SNPs for each trait, having genome-wide significant p-value. According to the aims of the study, the latter can be modified by preselective criteria set by the researcher. Furthermore, the detailed dataset acquired after the heterozygosity score calculation allows one to compare the genome-wide heterozygosity score to the windows containing SNPs of interest. As was the case with some of the samples from both the UK biobank and the Estonian biobank, the genome-wide low heterozygosity score not mandatorily matched with low heterozygosity scores for the studied trait. This provides the possibility to track whether the low genome-wide heterozygosity scores of some individuals match with the hypothesis of assortative mating for a given trait, or whether it is a mere consequence of genealogically reasoned genetic similarity between the parents.

## 5.2. The UK

Several traits have been repeatedly identified as GAM drivers in previous research on UKBB cohorts, including BMI [13,14], systolic blood pressure [2], waist-hip ratio [2], height [2, 12, 13, 14] and educational attainment [2, 12, 14]. The results from this work suggest a different combination of key factors driving genetic assortative mating in the UK population, as the one with the highest resonance being the depression status. Previous research done in Europe have both accepted [9, 19] the hypothesis of correlation between partners' mental health in Sweden and Finland. Here it is confirmed that major depressive disorder is a key-point from the perspective of genetic assortative mating. A recent study by Horwitz *et al.* [14] outlined demographic factors (year of birth, place of birth, age of completed education etc.) as much more resonant in assortative mating, and proved by meta-analysis depression to be one of the lowest ranking traits in terms of partners' correlations. In contrast, current results reveal mental health plays an important role in mate choice from the genetic viewpoint. The results obtained here point to both positive and negative assortment, albeit the latter seems to spread to a much lower extent. These results offer a novel point of view on the question by implying the existence of an unnoticed factor for partner choice in a long span of time, given the studied individuals have been aged between 40 and 73.

Going further, one of the most cited traits under assortative mating in the UK biobank, namely educational attainment, did not provide significant results. Genetic windows associated with SNPs of educational relevance showed no significant difference compared to the rest of the genome. This

implies the lack of genetically based assortment for the trait, in contrast to GPD estimates suggesting strong genetic correlation between partners for the trait [12, 13]. Nevertheless, the different concept of the previous works reviewing GAM for the trait and the current study, may be a root cause for the alternative results. As a matter of fact, the very few individuals demonstrating assortative mating for educational attainment, showed 100% negative assortment. The translated version of this finding implies that individuals with distinct architecture of the relevant genetic windows stand for very dissimilar partners (for the genotype of the trait).

A body feature which tends to be put in the centre of the already available GAM research is height [2, 12, 13, 14]. Despite this fact, the results acquired via this method do not confirm the ubiquitous GAM for height. Rather than the stature, other anthropometric traits were identified as drivers of GAM - BMI-adjusted waist-hip ratio, BMI-adjusted waist circumference, hip and waist circumferences. These are the traits which showed strong signals of genetic assortative mating in the UK cohort. Another significant point to mention is all the anthropometric traits which were proved to be under GAM, consisted of proportionally larger groups of negatively assorted individuals. This finding reveals a tendency for individuals, who are dissimilar in the relevant genetic windows, to mate and give birth to offspring. Although the statistical procedures filtering for GAM are different in their essence, the juxtaposition of phenotype against genotype revealed individuals with highest delta scores (both suspected to be under PAM and NAM) were positioned very close to the 0 residual unit. This would mean that the individuals suspected of PAM or NAM did not necessarily exhibit physically distinct phenotypes. Characteristics like BMI [13, 14] and systolic blood pressure [2] produced very low signals of GAM, which reduces their significance value compared to the previously mentioned traits.

In summary, the pattern of genetic assortative mating in the UK population analysed in this work points to similarities for depression status and dissimilarities for anthropometrics traits from the genetic point of view. However, phenotype-genotype juxtaposition did not define those similarities and differences as ones demonstrating extreme phenotypes.

## 5.3. Estonia

While the UK biobank datasets have been in the focus of a myriad of assortative mating studies, little is known for Estonia precisely. Previous research on countries from the Nordic-Baltic Eight provided insights on

assortative mating mainly for traits like BMI [22, 23], psychopathologies (including depression) [9, 20], height [9, 23] and educational attainment [9].

Current results for the Estonian population demonstrated how putative culturally mediated behaviour may yield contradictory patterns of human sexual selection as opposed to each other. Although the traits under genetic assortative mating tend to be the same as the ones in the UK cohort, there are several significant differences.

To begin with, the key GAM factor in the UK was found to be depression status. However, in the Estonian population this factor takes the last place among the top 5 traits under GAM. Nevertheless, the major trend is the same as in the UK biobank cohort where the GAM signals were predominantly indicating PAM. Taking into consideration previous research on assortative mating in the Baltic region, more specifically in Finland, suggesting depression status as a mating factor between partners [9], this work confirmed it to be a trait under a significant degree of GAM in the Estonian population as well.

On the other hand, traits demonstrating the strongest degree of GAM among Estonians tend to be related to anthropometrics: BMI-adjusted waist-hip ratio, hip circumference, BMI-adjusted waist circumference and waist circumference. These 4 traits ranked as the most significant factors in the process of mate choice, followed by depression status. In contrast to the results from the UK cohort, the GAM signals indicated mostly PAM. The high positive assortment signal corresponds to a significant genetic similarity between mating partners for the relevant parts of the genome.

It should be noted that educational attainment and height, highly cited traits from previous research, failed to classify as major factors in the current search for genetic assortative mating. While educational attainment achieved a modest signal for PAM, body height was found to pass neither the Bonferroni correction nor the Benjamini-Hochberg procedure. This finding presents a peculiar case of GAM detection, which could be explained either by the different analytical approach the method applies or by the SNPs selection in the current work against previous works. As an additional reference to the latter, Border *et al*. released their work confirming that when only selected SNPs, and not all of the trait-related ones, are included in the heritability analysis, the estimates are inflated. [24] They focus in particular on the height heritability inflation in the UK biobank dataset, which turns out to be expanded by 14% to 23%. [24] Thus, when not all causal variants are taken into account, a possible assortative-mating-induced bias could persist.

## 5.4. What drives mate choice in the UK and Estonia? And how?

As similar as the results between the two populations seem to be at first hand, the patterns of preference they reveal are much different. De facto, the complex traits under GAM are the same for both the UK and Estonia. However, the order of significance is different with depression being the most important factor in mate choice in the UK, while BMI-adjusted waist-hip ratio tends to be the more serious genetic factor in Estonia. Apart from the mutual tendency for PAM expression in terms of psychopathologies like depression, the GAM pattern in anthropometrics is completely different for the studied populations. While a distinctly inclined pattern for NAM is observed in the UK, a much more conservative approach is being followed in Estonia where close to 100% of the detected GAM signals indicated PAM. Thus although it could be stated that the same features play key roles in GAM among these two present-day European populations, the current work proved the selection pattern is largely different.

It does come as a surprise that some of the so much cited pillars of assortative mating in Europe, like height and educational attainment, failed to pass the statistical analyses for significance. Polygenicity is unlikely to be the major reason for this finding due to the relatively equal number of SNPs included for complex traits which passed the statistical filters and the ones which did not. Thus, further investigations might be necessary to fully explore the resolution of the method and the statistical significance of the results.

## 5.5. Future perspectives and caveats

In order to gain a better understanding of GAM via the method used here, future improvements in terms of computation and theoretical basis are necessary.

To begin with, SNPs selection might be a primary and foundational improvement on the way to gain comparative statistics. Until now Yengo [12] and Yamamoto [13] have provided convincing results on the basis of GPD on trait-increasing alleles only. However, from a theoretical stand,

trait-decreasing alleles could also be subject to natural selection and play important roles in adaptation. On the other hand, a last year study revealed not including all causal variants might lead to AN-induced bias [24]. Thus, a comparative characteristic of results obtained through this method on different theoretical bases, might present a plausible prospect for investigation on the topic of genetic assortative mating.

Secondly, a promising direction of research is the inclusion of other European countries so as to gain a more comprehensive picture of the genetic assortative mating bases across Europe. Previous research based on phenotypic markers stated that “there is some tendency towards higher marital associations for physical characteristics among populations in the Mediterranean region than among those in Northern and Western Europe” [21] This hypothesis could be considered and looked through via the method developed here as it requires genotype data for individuals only.

Thirdly, having in mind the migrations across the globe in the last several generations, a crucial optimisation of the method would be to develop it mathematically and computationally further in order to apply it to individuals with heterogeneous backgrounds as well, who have reported ancestry admixture in their family trees dating up to a few generations back. This prospective direction of development would not only expand the available dataset for analysis, but would also represent an investigation of which traits fall under genetic assortative mating despite and because of different genetic ancestries.

Lastly, a computational improvement of the method is required in order to minimise the hardware resources necessary for each cycle. Additionally, the current Python scripts available in Appendix A could be merged into a common pipeline which would reduce the error rate, the runtime and the human resource necessary for processing the data.

# 6. Appendix A

## 6.1. Heterozygosity Calculation Script

```
#!/usr/bin/env python3

import sys
import io
import os
import pandas as pd
import numpy as np
import statistics
import gzip
import csv
import allel
import re

if len(sys.argv) == 1:
 print('Syntax:' + sys.argv[0] + 'path to vcf' + 'included list'
+ 'chrom' + 'csv heterozygosity values' + 'csv with frame' +
'window size' + 'inclusion threshold for snps in a window')

if len(sys.argv) == 8:
 print('\n Process initiated.')
else:
 print('\n Maybe you forgot an argument? Check command line.')
 sys.exit()

file = str(sys.argv[1])
chrom = int(sys.argv[3])
het = str(sys.argv[4])
frame = str(sys.argv[5])
windowsize = int(sys.argv[6])
threshold = float(sys.argv[7])
output1 = str('./{}_{}_{}.csv'.format(chrom,het,windowsize))
output2 = str('./{}_{}_{}.csv'.format(chrom,frame,windowsize))
if sys.argv[2] == 'SKIP':
 included_list = []
else:
 included = open(sys.argv[2], 'r')
 pre_included_list = included.readlines()
 included_list = []
 for sample in pre_included_list:
  sample = sample.replace('\n','')
  included_list.append(sample)

chromoptions = [0, 249000000, 243000000,
199000000,191000000,182000000,171000000,160000000,146000000,139000000
                  ,134000000,136000000,134000000,1

chromlength = int(chromoptions[chrom])
windowranges = np.arange(0,chromlength, windowsize)
windows = []
for i in range(0,(len(windowranges)-1)):
 start = windowranges[i]
```

```
 end = windowranges[i+1]
 coordinate = str('{}:{}-{}'.format(chrom, start, end))
 windows.append(coordinate)

if len(included_list) > 0:
 all_samples = included_list
else:
 header = allel.read_vcf_headers(file)
 all_samples = header[4]
header_line = []
for i in all_samples:
 header_line.append(i)
header_line.append('pop_het')

def estimate_het(file, windows):
 print('\n Calculation of heterozygosity started...')
 windows_new = []
 with open(output1, 'w') as out1:
  out1wr = csv.writer(out1, delimiter = ',')
  out1wr.writerow(header_line)
  for reg in windows:
   data = allel.read_vcf(file,
   region = reg, samples =
   all_samples) if data is None:
    continue
   else:
    datarr = allel.GenotypeArray(data['calldata/GT'])
    datalst = datarr.tolist()
    if len(datalst) < threshold:
     continue
    else:
     windows_new.append(reg)
     df = pd.DataFrame(datalst)
     inds_het = []
     inds_het_updated = []
     for col in df.columns:
      ind_het = []
      slice = df[col].tolist()
      w = []
      for i in slice:
       w.append(str(i))
      hom = w.count('[0, 0]' or '[1, 1]')
      het = w.count('[0, 1]' or '[1, 0]')
      if het == 0:
       ind_het = 0
      else:
       ind_het = het/(het+hom)
      inds_het.append(ind_het)
   pop_het = np.mean(inds_het)
   for i in inds_het:
    k = i / (i + pop_het)
    inds_het_updated.append(k)
   inds_het_updated.append(pop_het)
   out1wr.writerow(inds_het_updated)
   data = []
   datarr = []
   datalst = []
   df = []
 with open(output2, 'w') as out2:
```

```
  out2wr = csv.writer(out2)
  out2wr.writerow(['CHROM','START','END'])
  for i in windows_new:
   out2wr.writerow(re.split(':|-',i))
 print('\n Completed successfully.')
 return output1, output2


almost_there = estimate_het(file,windows)
print('Done')
```

## 6.2. Primary Screening Script

```
#!/usr/bin/env python3

import sys
import numpy as np
import pandas as pd
import csv



if len(sys.argv) == 1:
 print('Syntax:' + sys.argv[0] + 'frame csv path' + 'het csv path' +
'csv with snps of interest path' + 'txt output')

if len(sys.argv) == 5:
 print('\n Process initiated.')
else:
 print('\n Maybe you forgot an argument? Check command line.')
 sys.exit()

framefile = str(sys.argv[1])
hetfile = str(sys.argv[2])
snpslist = str(sys.argv[3])
output = str(sys.argv[4])

print('\n Loading frame and SNPs data.')
frame = pd.read_csv(framefile, compression = 'gzip')
frame = frame.reset_index()
frame = frame.drop_duplicates(keep = False)
for col in frame.columns:
 frame[col] = frame[col].apply(pd.to_numeric)
snps = pd.read_csv(snpslist, compression = 'gzip')
snps = snps.reset_index()
snps = snps.drop(['index', 'Unnamed: 0'], axis = 1)

print('\n Estimating coordinates.')
chrom = snps['CHR'].tolist()
pos = snps['POS'].tolist()
coordinates = []
for i in range(len(chrom)):
 coordinates.append([chrom[i],pos[i]])

firstivs = []
for i in coordinates:
 iv = frame[(frame['CHROM'] == i[0]) & (frame['START'] <= i[1]) &
(frame['END'] >= i[1])].index
 if len(iv) > 0:
  firstivs.append(iv)

print('\n Processing coordinates.')
```

```
secondivs = []
for i in firstivs:
 if i not in secondivs:
  secondivs.append(i)
firsivs = []
thirdivs = list(secondivs)
secondivs = []
allindexes = np.arange(0, len(frame.index))

traitindexes = []
for i in thirdivs:
 traitindexes.append(i[0])

snps = []

print('\n Loading score data.')
het = pd.read_csv(hetfile, compression = 'gzip')
het = het.reset_index()
for col in het.columns:
 het[col] = het[col].apply(pd.to_numeric,
errors = 'coerce') het = het.drop('index',
axis = 1)
ind_id = list(het.columns)
print(het.memory_usage(deep=True).sum())

alltrait = []

print('\n Data extraction per individual.')
for col in het.columns:
 trait = []
 df = het[col].tolist()
 for i in traitindexes:
  trait.append(df[i])
 alltrait.append(trait)

median_windows_of_interest = []
for i in alltrait:
 value = np.median(i)
 median_windows_of_interest.append(value)

print('\n Writing output.')
f = open('{}'.format(output), 'w')
for i in median_windows_of_interest:
 f.write(str(i))
 f.write('\n')
f.close()

print('\n Completed successfully.')
```

## 6.3. Mann-Whitney U-test Script

```
#!/usr/bin/env python3

import sys
import time
import numpy as np
import pandas as pd
import scipy
from scipy import stats
import csv

if len(sys.argv) == 1:
 print('Syntax:' + sys.argv[0] + 'frame csv path' + 'het csv path' +
'csv with snps of interest path' + 'csv with p-values and delta
median path' + 'trait')

if len(sys.argv) == 6:
 print('\n Process initiated.')
else:
 print('\n Maybe you forgot an argument? Check command line.')
 sys.exit()

start = time.time()
framefile = str(sys.argv[1])
hetfile = str(sys.argv[2])
snpslist = str(sys.argv[3])
output = str(sys.argv[4])
traitname = str(sys.argv[5])

print('\n Loading frame and SNPs data.')
frame = pd.read_csv(framefile, compression = 'gzip')
frame = frame.reset_index()
frame = frame.drop_duplicates(keep = False)
for col in frame.columns:
 frame[col] = frame[col].apply(pd.to_numeric)
frame = frame.drop(['index', 'Unnamed: 0'], axis = 1)
snps = pd.read_csv(snpslist, compression = 'gzip')
snps = snps.reset_index()
snps = snps.drop(['index', 'Unnamed: 0'], axis = 1)

print('\n Estimating coordinates.')
chrom = snps['CHR'].tolist()
pos = snps['POS'].tolist()
coordinates = []
for i in range(len(chrom)):
 coordinates.append([chrom[i],pos[i]])

firstivs = []
for i in coordinates:
 iv = frame[(frame['CHROM'] == i[0]) & (frame['START'] <= i[1]) &
(frame['END'] >= i[1])].index
```

```
 if len(iv) > 0:
  firstivs.append(iv)

print('\n Processing coordinates.')
secondivs = []
for i in firstivs:
 if i not in secondivs:
  secondivs.append(i)
firsivs = []
thirdivs = list(secondivs)
secondivs = []
allindexes = np.arange(0, len(frame.index))

traitindexes = []
for i in thirdivs:
 traitindexes.append(i[0])

backgroundindexes = []
for i in allindexes:
 if i not in traitindexes:
  backgroundindexes.append(i)
snps = []

print('\n Loading score data.')
het = pd.read_csv(hetfile, compression = 'gzip')
het = het.reset_index()
for col in het.columns:
 het[col] = het[col].apply(pd.to_numeric,
errors = 'coerce') het = het.drop('index',
axis = 1)
ind_id = list(het.columns)
print(het.memory_usage(deep=True).sum())

alltrait = []
allbackground = []

print('\n Data extraction per individual.')
for col in het.columns:
 trait = []
 background = []
 df = het[col].tolist()
 for i in traitindexes:
  trait.append(df[i])
 for i in backgroundindexes:
  background.append(df[i])
 alltrait.append(trait)
 allbackground.append(background)

allresult = []
all_delta = [0] * len(alltrait)

print('\n Mann-Whitney.')
for i in range(len(alltrait)):
 a =
 scipy.stats.mannwhitneyu(alltrait[i],all
 background[i]) allresult.append(a[1])

print('\n Calculating median delta.')
for i in allresult:
```

```
 if i < 0.05:
  index = allresult.index(i)
  delta = np.median(alltrait[index]) -
  np.median(allbackground[index]) all_delta[index] =
  delta

print('\n Writing output.')
datafortherun = [ind_id, allresult, all_delta]
datafortherun = pd.DataFrame(datafortherun)
datafortherun = datafortherun.transpose()
datafortherun.columns = ['ind_id',
traitname, 'delta_med']
datafortherun.to_csv(output)


print('\n Completed successfully.')
```

# 7. Bibliography


1. Assortative Mating in Man: A Cooperative Study. (1903). *Biometrika*, *2*(4), 481. https://doi.org/10.2307/2331510

2. Robinson, M. R., Kleinman, A., Graff, M., Vinkhuyzen, A. a. E., Couper, D., Miller, M. I., Peyrot, W. J., Abdellaoui, A., Zietsch, B. P., Nolte, I. M., Van Vliet-Ostaptchouk, J. V., Snieder, H., McIntosh, A. M., Martin, N. G., Magnusson, P. K. E., Iacono, W. G., McGue, M., North, K. E., Yang, J., & Visscher, P. M. (2017). Genetic evidence of assortative mating in humans. *Nature Human Behaviour*, *1*(1). https://doi.org/10.1038/s41562-016-0016

3. Glicksohn, J., & Golan, H. (2001). Personality, cognitive style and assortative mating. *Personality and Individual Differences*, *30*(7), 1199–1209. https://doi.org/10.1016/s0191-8869(00)00103-3

4. Speers, M. A., Kasl, S. V., Freeman, D., & Ostfeld, A. M. (1986). BLOOD PRESSURE CONCORDANCE BETWEEN SPOUSES. *American Journal of Epidemiology*, *123*(5), 818–829. https://doi.org/10.1093/oxfordjournals.aje.a114311

5. Rawlik, K., Canela-Xandri, O., & Tenesa, A. (2019). Indirect assortative mating for human disease and longevity. *Heredity*, *123*(2), 106–116. https://doi.org/10.1038/s41437-019-0185-3

6. Havlíček, J., Winternitz, J., & Roberts, S. (2020). Major histocompatibility complex-associated odour preferences and human mate choice: near and far horizons. *Philosophical Transactions of the Royal Society B*, *375*(1800), 20190260.

https://doi.org/10.1098/rstb.2019.0260

7. Zietsch, B. P., Verweij, K. J. H., Heath, A. C., & Martin, N. G. (2011). Variation in Human Mate Choice: Simultaneously Investigating Heritability, Parental Influence, Sexual Imprinting, and Assortative Mating. *The American Naturalist*, *177*(5), 605–616. https://doi.org/10.1086/659629

8. Versluys, T. M. M., Flintham, E. O., Mas-Sandoval, A., & Savolainen, V. (2021). Why do we pick similar mates, or do we? *Biology Letters*, *17*(11). https://doi.org/10.1098/rsbl.2021.0463

9. Torvik, F. A., Eilertsen, E. M., Hannigan, L. J., Cheesman, R., Howe, L. D., Magnus, P., Reichborn-Kjennerud, T., Andreassen, O. A., Njølstad, P. R., Havdahl, A., & Ystrom, E. (2022). Modeling assortative mating and genetic similarities between partners, siblings, and in-laws. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-28774-y

10. Abdellaoui, A., Yengo, L., Verweij, K. J. H., & Visscher, P. M. (2023). 15 years of GWAS discovery: Realizing the promise. *American Journal of Human Genetics*, *110*(2), 179–194. https://doi.org/10.1016/j.ajhg.2022.12.011

11. Marnetto, D., Pankratov, V., Mondal, M., Montinaro, F., Pärna, K., Vallini, L., Molinaro, L., Saag, L., Loog, L., Montagnese, S., Montagnese, S., Metspalu, M., Eriksson, A., & Pagani, L. (2022). Ancestral genomic contributions to complex traits in contemporary Europeans. *Current Biology*, *32*(6), 1412-1419.e3. https://doi.org/10.1016/j.cub.2022.01.046

12. Yengo, L., Robinson, M. R., Keller, M. C., Kemper, K. E., Yang, Y., Trzaskowski, M., Gratten, J., Turley, P., Cesarini, D., Benjamin, D.

K., Wray, N. R., Goddard, M. E., Yang, J., & Visscher, P. M. (2018). Imprint of assortative mating on the human genome. *Nature Human Behaviour*, *2*(12), 948–954. https://doi.org/10.1038/s41562-018-0476-3

13. Yamamoto, K., Sonehara, K., Namba, S., Konuma, T., Masuko, H., Miyawaki, S., Kamatani, Y., Hizawa, N., Ozono, K., Yengo, L., & Okada, Y. (2022). Genetic footprints of assortative mating in the Japanese population. *Nature Human Behaviour*, *7*(1), 65–73. https://doi.org/10.1038/s41562-022-01438-z

14. Horwitz, T. B., & Keller, M. C. (2022). Correlations between human mating partners: a comprehensive meta-analysis of 22 traits and raw data analysis of 133 traits in the UK Biobank. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2022.03.19.484997

15. Hart, A. (2001). Mann-Whitney test is not just a test of medians: differences in spread can be important. *BMJ*, *323*(7309), 391–393. https://doi.org/10.1136/bmj.323.7309.391

16. Hollestein, L. M., Lo, S., Leonardi-Bee, J., Rosset, S., Shomron, N., Couturier, D., & Ratib, S. (2021). MULTIPLE ways to correct for MULTIPLE comparisons in MULTIPLE types of studies. *British Journal of Dermatology*, *185*(6), 1081–1083. https://doi.org/10.1111/bjd.20600

17. Waite, T. A., & Campbell, L. G. (2006). Controlling the false rate discovery and increasing statistical power in ecological studies. *Ecoscience*, *13*(4), 439–442. https://doi.org/10.2980/1195-6860

18. McCroskey, J. C., Heisel, A. D., & Richmond, V. P. (2001). Eysenck's BIG THREE and communication traits: three

correlational studies. *Communication Monographs*, *68*(4), 360–366. https://doi.org/10.1080/03637750128068

19. Nordsletten, A. E., Larsson, H., Crowley, J. L., Almqvist, C., Lichtenstein, P., & Mataix-Cols, D. (2016). Patterns of Nonrandom Mating Within and Across 11 Major Psychiatric Disorders. *JAMA Psychiatry*, *73*(4), 354. https://doi.org/10.1001/jamapsychiatry.2015.3192

20. Jarvis, B. F., Mare, R. D., & Nordvik, M. K. (2023). Assortative Mating, Residential Choice, and Ethnic Segregation. *Research in Social Stratification and Mobility*, 100809. https://doi.org/10.1016/j.rssm.2023.100809

21. Spuhler, J. N. (1982). Assortative Mating. *Biodemography and Social Biology*, *29*(1–2), 53–66. https://doi.org/10.1080/19485565.1982.9988478

22. Ajslev, T. A., Ängquist, L., Silventoinen, K., Gamborg, M., Allison, D. B., Baker, J. L., & Sørensen, T. I. A. (2012). Assortative marriages by body mass index have increased simultaneously with the obesity epidemic. *Frontiers in Genetics*, *3*. https://doi.org/10.3389/fgene.2012.00125

23. Silventoinen, K., Kaprio, J., Lahelma, E., Viken, R. J., & Rose, R. (2003). Assortative mating by body height and BMI: Finnish Twins and their spouses. *American Journal of Human Biology*, *15*(5), 620–627. https://doi.org/10.1002/ajhb.10183

24. Border, R., O'Rourke, S., De Candia, T. R., Goddard, M. E., Visscher, P. M., Yengo, L., Jones, M., & Keller, M. C. (2022). Assortative mating biases marker-based heritability estimators. *Nature Communications*, *13*(1).

https://doi.org/10.1038/s41467-022-28294-9

25. Woodin, T., McCulloch, G., & Cowan, S. (2013). Secondary Education and the Raising of the School-Leaving Age. In *Palgrave Macmillan US eBooks*. https://doi.org/10.1057/9781137065216

26. Pankratov, V., Montinaro, F., Kushniarevich, A., Hudjashov, G., Jay, F., Saag, L., Flores, R., Marnetto, D., Seppel, M., Kals, M., Võsa, U., Taccioli, C., Möls, M., Milani, L., Aasa, A., Lawson, D., Esko, T., Mägi, R., Pagani, L., . . . Metspalu, M. (2020). Differences in local population history at the finest level: the case of the Estonian population. *European Journal of Human Genetics*, *28*(11), 1580–1591. https://doi.org/10.1038/s41431-020-0699-4

# 8. Acknowledgements

There are a few people whom I would like to explicitly thank for being by my side and supporting me throughout my personal and academic life so far.

First of all, it is unavoidable to be overwhelmingly emotional when expressing how thankful I am to Sharlota Prodanova, my mother. Mom, I would like to express my gratitude to you with a simple, but huge “Thank you!”. Everything else I would like to say will sound too much like the Oscar-winning speech, which I am sure you can fully imagine! To my family members and friends - thank you for being there!

Then, I would like to sincerely thank my supervisor - prof. Luca Pagani, for always being immensely supportive, patient and considerate! Thank you for sparking my interest in the field of molecular anthropology, for everything you have taught me during this time and always tuning in with the figurative references!

Finally, I would like to thank prof. Massimo Mezzavilla and dott. Leonardo Vallini for all their help, support and warm welcome from the very first day of my internship! Thank you for being so kind and for always making time to answer my questions!

I genuinely hope to be able to work with you in the future again!
Thank you!