# Assessing the impact of 20th century internal migrations on the genetic structure of Estonia

Ivan A. Kuznetsov[1,*], Estonian Biobank Research Team[1], Mait Metspalu[1], Uku Vainik[1,2,3], Luca Pagani[1,4], Francesco Montinaro[1,5], Vasili Pankratov[1,*]

[1] Institute of Genomics, University of Tartu, Tartu 51010, Estonia
[2] Institute of Psychology, University of Tartu, Tartu 50409, Estonia
[3] Montreal Neurological Institute, McGill University, Montreal H3A 2B4, Canada
[4] Department of Biology, University of Padova, Padova 35121, Italy
[5] Department of Biosciences, Biotechnology and Environment, University of Bari, Bari 70124, Italy

*Corresponding authors: IK (ivan.kuznetsov@ut.ee), VP (vasili.pankratov@ut.ee)

Estonian Biobank Research Team: Andres Metspalu, Lili Milani, Tõnu Esko, Reedik Mägi, Mari Nelis and Georgi Hudjashov

## Abstract

Spatial genetic structure observed in many human populations is in large part attributed to past demographic events and isolation by distance. However, how intensifying migration affects this structure remains understudied. Here we harness a sample of more than 180 thousand individuals to explore the genetic correlates and consequences of contemporary migrations in Estonia. While we show that migration smoothens the genome-wide genetic structure, it intensifies inter-regional differences in polygenic scores (PGS) for certain traits, derived both from population as well as within-sibship studies. The strongest effect is observed for educational attainment which is consistent with previous observations in the UK and suggests this to be a general pattern. We explore those regional differences in PGS in terms of the driving forces behind them and from a temporal perspective, and suggest urbanisation as a major driver for this pattern in Estonia from at least the first half of the 20th century.

## Introduction

Spatial genetic structure is revealed by differences in allele frequencies across geographic locations[1]. This phenomenon has been observed in human populations from global[2,3] to fine scale[4–8]. It is driven by various demographic phenomena, including prehistoric migrations and admixture as well as isolation due to physical barriers and the relatively low mobility of many human groups[9–13]. However, migration activity, primarily related to urbanisation and political changes, has largely intensified in the past century, blurring such fine-scale population structure[14].

The propensity to migrate is a behavioural trait with potentially some genetic contribution. If so, regions attractive for internal migrations should be enriched for alleles associated with an increased probability of migration. Trait-associated genetic correlates of spatial population structure and migration patterns have been demonstrated in the British population[15]. Since moving individuals change not only their location but sometimes also their environment including lifestyle, migration may generate new genotype-environment correlations[16,17] leading to spurious non-causal genome-wide associations[15]. Capturing such non-causal effects in genetic studies may lead to biased estimates of heritability, genetic correlations, and Mendelian randomisation inferences[18,19].

An essential factor predicting geographic mobility is socio-economic status and, particularly, educational attainment (EA) which refers to the highest level of education completed by an individual. In fact, the level of education has been shown to directly influence migration behaviour in Europe and the US[20–22]. EA is a heritable trait with heritability estimates ranging from 4% to more than 50%, depending on the definition and study design[23–26]. Thus, it is natural to expect that recent migrations can be associated with EA-associated genetic variants and so affect the geographical pattern of allele distribution in a non-random fashion. Indeed, it has been shown that migrants and non-migrants from the same areas in Great Britain differ in their average genetic profiles with the strongest difference in alleles associated with EA[15]. Despite the potential practical implications of such changes in spatial genetic structure due to recent human migrations, little is still known about how widespread and how recent they are. Most of the observations to date come from the UK Biobank, raising the question if those effects are country or cohort specific.

Here we aim at exploring the genetic consequences of recent migrations in Estonia and the genetic associations of migration patterns within the country. We analyse data from the Estonian Biobank (EstBB) which represents a population different from the British one in terms of genetic background, as well as demographic and socio-economic aspects. In particular, during the 20th century, Estonia underwent a series of transitions (Estonia gained independence from the Russian Empire in 1918, was annexed by the Soviet Union in 1940 and re-gained independence in 1991), each of them associated with political, economic and sociological changes. In this regard, Estonia substantially differs from the UK which had more stable social conditions, potentially leading to a long-standing socio-economic structure[27–29]. In addition, Estonia has one of the largest internal migration rates in Europe, with approximately 50% moving at least once in their lives[30]. The recruitment strategy of the EstBB is also different from that of the UK Biobank[31,32]. The EstBB includes data on more than 210,000 participants which represents approximately 20% of the current adult population of all ages and a relatively uniform geographic coverage. Specifically, the variety of birth years of participants allows us to analyse temporal trends in genetic correlates of migration.

In this work, we use the EstBB to explore the genetic correlates of migrations within Estonia, defined as differences between place of birth (POB) and place of current residence (POR). We first analyse changes in the geographical distribution of ancestry captured by genetic principal components. Next, we check if the phenotype-related genetic components captured with polygenic scores (PGS) orthogonal to ancestry are distributed non-randomly across the regions of the country. Then we look at how this distribution changes due to contemporary migrations. As PGS for educational attainment ($PGS_{EA}$) demonstrated the strongest evidence for differential distribution across regions we focus on it in subsequent analyses. We compare mean $PGS_{EA}$ values between different groups of individuals based on their POB and POR to explore how different migration patterns are associated with $PGS_{EA}$. The age-stratified analysis made it possible to give an upper estimate of the time at which differences between regions arose and to describe how these differences have been increasing. Finally, we look into the relationship between migration and EA phenotypes to assess whether the correlation between them can entirely explain the pattern of $PGS_{EA}$ distribution in space and between migration groups.

## Results

### Data overview

We investigated the distribution of genetic ancestry and complex trait variation across different migration groups and geographic areas using genome-wide single-nucleotide polymorphism (SNP) data from 183,576 self-reported Estonian or Russian adult individuals from the Estonian Biobank (EstBB)[31]. Since Estonians are the major relatively homogeneous group in the biobank and the country, we use the cohort of Estonians for all the main analyses. For the sensitivity analyses and comparison between subgroups, we repeated some analyses in partially overlapping subgroups defined based on demography (relatedness, sex and age) and time of the biobank enrolment as enrolment happened in two periods, differing in recruitment strategy (Supplementary Materials, *Supplementary analyses*). We also replicated most of the analyses in the cohort of self-reported Russians - the second largest group in the EstBB (Supplementary Materials, *Supplementary analyses*). Detailed subdivision information and a description of the groups can be found in Supplementary Materials, *Estonian Biobank cohort overview*.

### Effect of recent migrations on regional differences in genome-wide ancestry and polygenic scores

It has been previously reported that the Estonian population shows a geography-correlated genetic structure which can be captured by principal component analysis (PCA)[7]. To explore how this genetic structure is affected by migration of the EstBB participants (defined as a difference between the place of residence (POR) and the place of birth (POB) and referred to as "contemporary migrations") we performed PCA[33] separately for Estonians and Russians and compared the proportion of variance in principal component coordinates (PCs) explained by differences between counties ($Var_{county}$; see Methods) for POB versus POR. Between-county differences explain a significant proportion of variance of all 100 PCs for POB and 98 out of 100 PCs for POR in Estonians (Figure 1). The proportion explained by POR is smaller for all PCs where the difference is significant. This is expected if we assume that contemporary migrations are random with respect to ancestral background. Hence, regional ancestry differences are decreasing over time, blurring the population genetic structure.

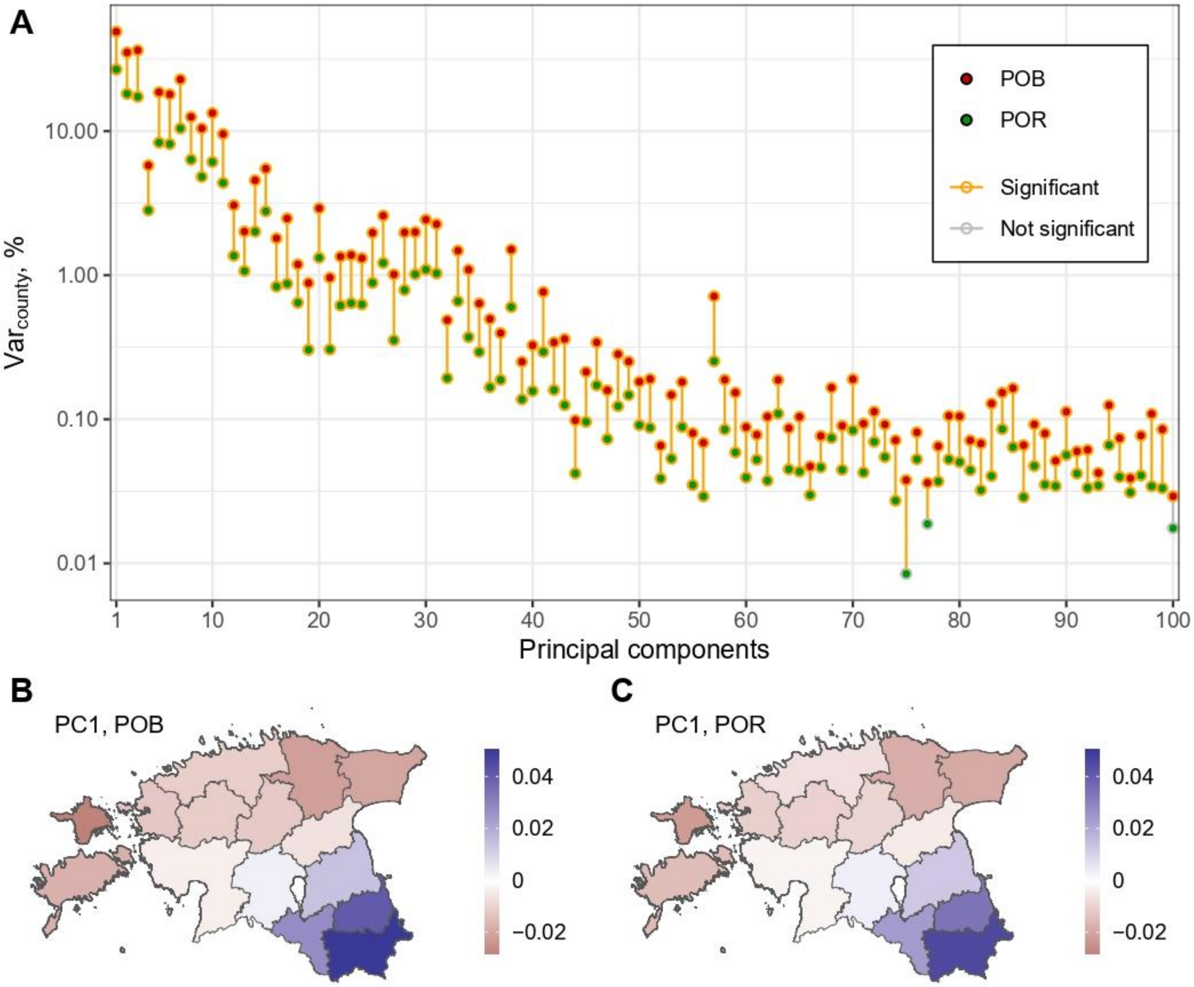

**Figure 1. Inter-individual variance of PCs explained by county of birth (POB) and county of residence (POR)**. (A) Estimates of inter-individual variance of PCs explained by POB and POR. Red and green dots refer to the POB and POR, correspondingly. Estimates significantly different from zero are outlined in yellow. The line connecting the two points is yellow when the variance explained by POB and POR together is significantly larger than the variance explained by only the weaker predictor (which is always POR in this case). The significance level is 0.05, adjusted for 100 tests with Bonferroni correction. (B-C) The map of Estonia with mean PC1 coordinates for individuals' POB (B) and POR (C).

It has been previously shown that migration patterns are associated with heritable phenotypes, particularly related to socio-economic status (SES)[15]. Therefore, we might expect that migration can enhance geographic differences in frequencies of alleles associated with such traits. To check this hypothesis in the EstBB we explored the spatial distribution of PGS for 169 diverse phenotypes, with a particular focus on traits related to behaviour and SES (Supplementary Table 1 and Methods). The population-based polygenic scores (PGS), used in all the analyses, unless stated otherwise, were calculated using summary statistics from genome-wide association studies (GWAS) conducted on the UK Biobank European-ancestry cohort[32,34]. All the polygenic scores were adjusted for demographic covariates (see Methods) and the first 100 PCs for the corresponding ethnic subgroup. For most PGSs regional differences in both POB and POR explain a non-zero fraction of variance, however, unlike the PCs, $\mathrm{Var}_{\mathrm{county}}$ values for POR are higher than for POB (Figure 2A). In other words, most PGSs show a geographic structure orthogonal to the first 100 PCs and this structure is enhanced by contemporary migrations. In agreement with a study conducted on a British population sample[15], the largest $\mathrm{Var}_{\mathrm{county}}$ is observed for PGS for educational attainment ($\mathrm{PGS}_{\mathrm{EA}}$; “College or university degree”). $\mathrm{Var}_{\mathrm{county}}$ for other traits is related approximately linearly to the absolute value of the correlation between the trait’s PGS and $\mathrm{PGS}_{\mathrm{EA}}$ ($r[\mathrm{PGS}_{\mathrm{trait}}, \mathrm{PGS}_{\mathrm{EA}}]$) starting from ~0.15, for both POB and POR (Figure 2A). Correlation between polygenic scores is a good measure of their shared characteristics as it accumulates the effects of true genetic correlation, heritability, demographic confounders and GWAS sample size. This suggests that the pattern for other PGSs is for a big part, if not entirely, driven by their correlation with the $\mathrm{PGS}_{\mathrm{EA}}$.

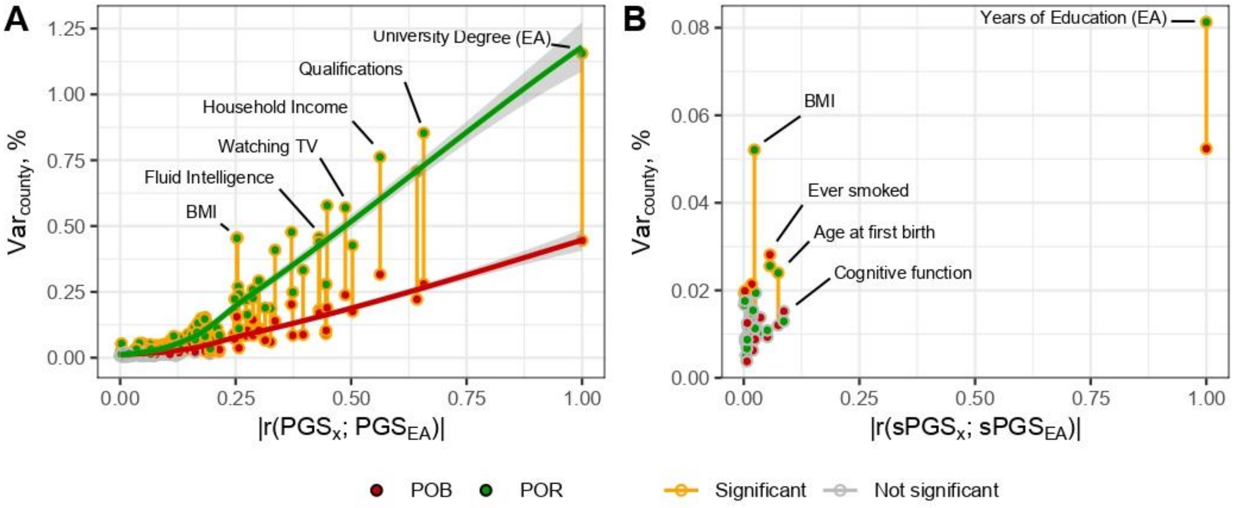

**Figure 2. Estimates of the inter-individual variance of (A) PGSs and (B) within-sibship GWAS PGSs (sPGSs) explained by POB and POR.** (s)PGSs are adjusted for demographic and genetic ancestry covariates. Estimates significantly different from zero are outlined in yellow. The line connecting the two points is yellow when the variance explained by POB and POR together is significantly larger than the

variance explained by only the weaker predictor. The significance level is 0.05, adjusted for the number of (s)PGS tested with Bonferroni correction.

EA is known to be influenced by indirect genetic effects of relatives as well as direct genetic effects. On top of that, population GWAS are reported to also capture associations due to demographic factors such as residual population structure and assortative mating. A promising though currently relatively underpowered approach to estimate direct genetic effects on a trait with relatively little bias due to confounders is within-sibship GWAS[35,36]. To test if the effects we observe can be explained solely by GWAS confounders and indirect effects we analysed 24 polygenic scores (Supplementary Table 2) constructed using summary statistics from a recent within-sibship GWAS[25] (sPGS). In accordance with the population-based results, sibship-based $\text{sPGS}_{\text{EA}}$ demonstrates the strongest non-uniform distribution between regions for both POB and POR with $\text{Var}_{\text{county}}$ being larger in the latter case (Figure 2B).

We also observe that PGS and sPGS for BMI demonstrate relatively high $\text{Var}_{\text{county}}$ (Figure 2) that could indicate the relationship between BMI and migration, independent of EA. We leave a full investigation of this hypothesis for future research.

## Geographical distribution of $\text{PGS}_{\text{EA}}$

To explore if the increasing between-county variability of $\text{PGS}_{\text{EA}}$ reported above is driven by some specific regions, we mapped the mean values of $\text{PGS}_{\text{EA}}$ adjusted for demographic and ancestral covariates for every county in Estonia (Figure 3). For both POB and POR, two counties have values significantly higher than the country average: these are Harju (FDR-adjusted p-value 4.1e-77 and 1.1e-168, correspondingly) and Tartu (FDR-adjusted p-value 4.8e-12 and 2.8e-14, correspondingly) Counties, where the two biggest Estonian cities, Tallinn and Tartu City, are located. Most other counties have values significantly lower than the country's average.

To see how the mean $\text{PGS}_{\text{EA}}$ changed due to contemporary migrations, we subtracted the mean values of $\text{PGS}_{\text{EA}}$ individuals born in a corresponding county from the mean values of $\text{PGS}_{\text{EA}}$ of the county's residents. Harju County, which includes the capital Tallinn, is the only county with significantly positive change (FDR-adjusted p-value 1.2e-02). Nine counties demonstrate a significant decrease in their average $\text{PGS}_{\text{EA}}$. Changes in the remaining four counties are insignificant. In three cases this is likely because of insufficient sample size. Still, in Tartu County, where the sample size is the second largest after Harju County (Supplementary Figure 2), this could probably reflect a balance between recent in-migration and out-migration of the county.

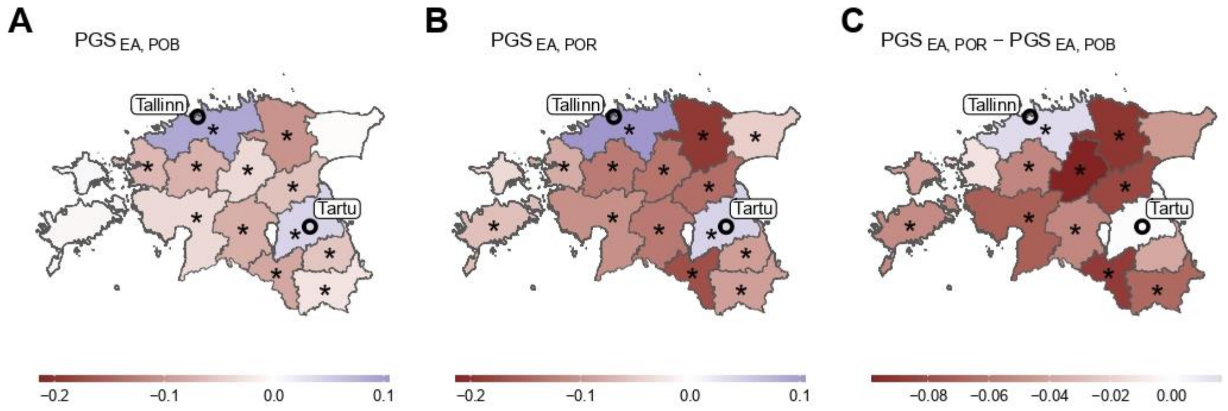

**Figure 3. PGS$_{EA}$ landscape in Estonia.** Mean PGS$_{EA}$ of individuals (A) born or (B) residing in each county. (C) Difference between values in panels "B" and "A". PGS$_{EA}$ is adjusted for demographic and genetic ancestry covariates. Counties with sample mean values significantly different from zero after FDR correction at the 0.05 level are marked with an asterisk (*).

## PGS$_{EA}$ values in groups with different migration profiles

Next, we compared the mean PGS$_{EA}$ between groups with different migration profiles. For this, we divided Estonia into three areas: Harju County (including Tallinn), Tartu County (including Tartu City) and other regions of Estonia (referred to as "ORE" below). All the individuals were classified into 9 groups based on their place of birth and residence. This classification was motivated by the results presented above and by the fact that Harju and Tartu Counties are the most economically developed regions, making them attractive migration destinations[37,38]. In all cases, migration within the defined areas (for instance, between counties defined as the ORE) was ignored.

Individuals who moved to Harju or Tartu Counties from ORE have higher PGS$_{EA}$ in comparison to those who stayed in ORE, explaining the decrease of PGS$_{EA}$ in most counties but Harju and Tartu. We also see that among individuals born in Harju or Tartu Counties, those migrating to ORE show the lowest PGS$_{EA}$ among individuals with non-matching POB and POR while individuals with the highest PGS$_{EA}$ are those who moved between Tartu and Harju Counties.

Tallinn and Tartu are the two biggest cities in Estonia, the main hotspots of urbanisation, centres of education and economic development. Therefore, we questioned if our results are also driven by those cities. To check this, we did the same analysis but keeping only participants born/residing in Tallinn or Tartu City instead of the entire corresponding counties (Figure 4B). The results demonstrate an even larger contrast between those who were born in or moved to Tallinn or Tartu City and those who stayed in ORE. That supports the hypothesis on the driver roles of the cities in the process of the increasing contrast between counties.

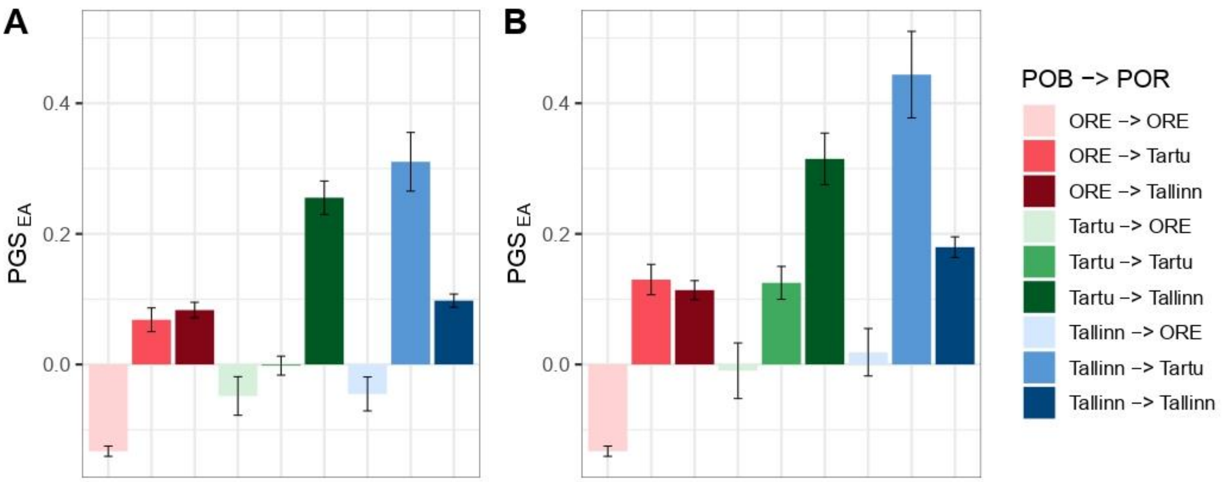

**Figure 4. PGS$_{EA}$ in migration groups by area of birth (POB) and residence (POR).** (A) County-based analysis where POB and POR refer to Tartu County ("Tartu"), Harju County ("Tallinn") and other counties ("ORE"). (B) City-based analysis, where POB and POR refer to Tartu City ("Tartu"), Tallinn ("Tallinn") and other counties ("ORE"). PGS$_{EA}$ is adjusted for demographic and genetic ancestry covariates. Error bars correspond to 95% confidence intervals.

## Migration direction and PGS$_{EA}$

Based on the previous results, the cities of Tallinn and Tartu are more attractive to individuals with above-average PGS$_{EA}$. We next asked if the PGS$_{EA}$ of migrants to Tallinn and Tartu City depends on an individual's POB in a city-dependent manner. We calculated differences in mean PGS$_{EA}$ between residents of Tallinn and Tartu City born outside those two cities grouped by their county of birth (Figure 5). It demonstrates that individuals who migrated to Tallinn from counties surrounding Tartu City have on average higher PGS$_{EA}$ compared to individuals born in the same counties and migrated to Tartu City. The opposite is true for counties surrounding Tallinn. This suggests that, in general, shorter-distance movement is less discriminating in terms of PGS$_{EA}$ than longer-distance movement in Estonia. However, the area of "less discriminative attraction" is wider for Tallinn compared to Tartu City, probably reflecting that Tallinn is a more general and stronger migration attracter.

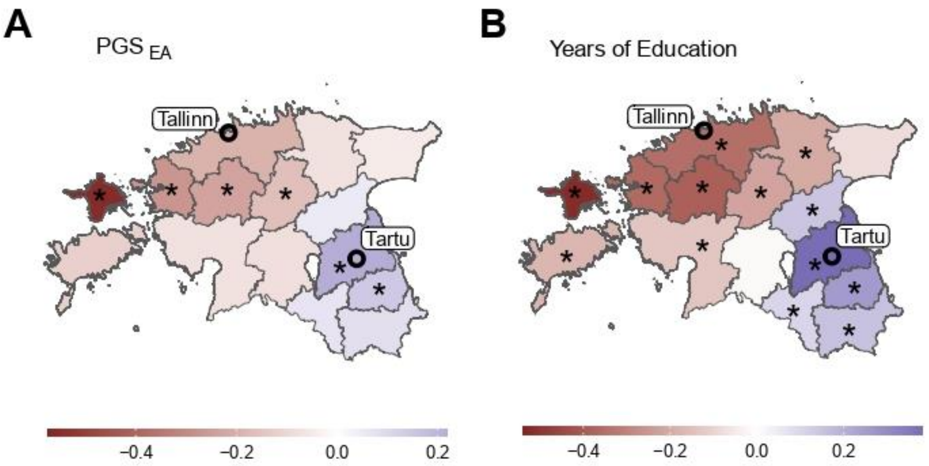

**Figure 5. The contrast in mean $\text{PGS}_{\text{EA}}$ and EA (years of education) between residents of Tallinn and Tartu City by county of birth.** (A) The value for each county corresponds to the mean $\text{PGS}_{\text{EA}}$ of individuals born in that county and living in Tartu City subtracted from the mean $\text{PGS}_{\text{EA}}$ of individuals born in the same county and living in Tallinn. Individuals born in Tallinn or Tartu City are excluded from the analysis. (B) The same but for the "years of education" phenotype. Counties with significant differences between the migrant groups after FDR correction at level 0.05 are marked with an asterisk (*).

## How old is the difference between cities and ORE?

To this end, we showed that contemporary migration increases the $\text{PGS}_{\text{EA}}$ differentiation between Tallinn/Tartu City and ORE. We next set out to explore if this effect accumulated over the last century and if there has been any change in the genetic makeup of migrants over this period of time. We compared mean $\text{PGS}_{\text{EA}}$ in Estonians grouped by place of birth and residence and the birth decade, while the $\text{PGS}_{\text{EA}}$ was adjusted and normalised in the entire Estonian cohort (Figure 6). We used wider birth year bins for the oldest and the youngest participants due to their smaller sample sizes. The comparison between groups of individuals born in Tallinn/Tartu City and ORE shows that individuals born in the cities on average have significantly higher $\text{PGS}_{\text{EA}}$ than those born in ORE starting from the 1940s (p-value 4.2e-3). Furthermore, the contrast between these groups tends to increase over time (Figure 6A). Consistently, $\text{PGS}_{\text{EA}}$ is significantly higher in the group of migrants from ORE to the cities than in the group of participants who stayed in ORE. This difference is significant already in the earliest bin (p-value 1.4e-3) and persists in all subsequent bins (Figure 6B).

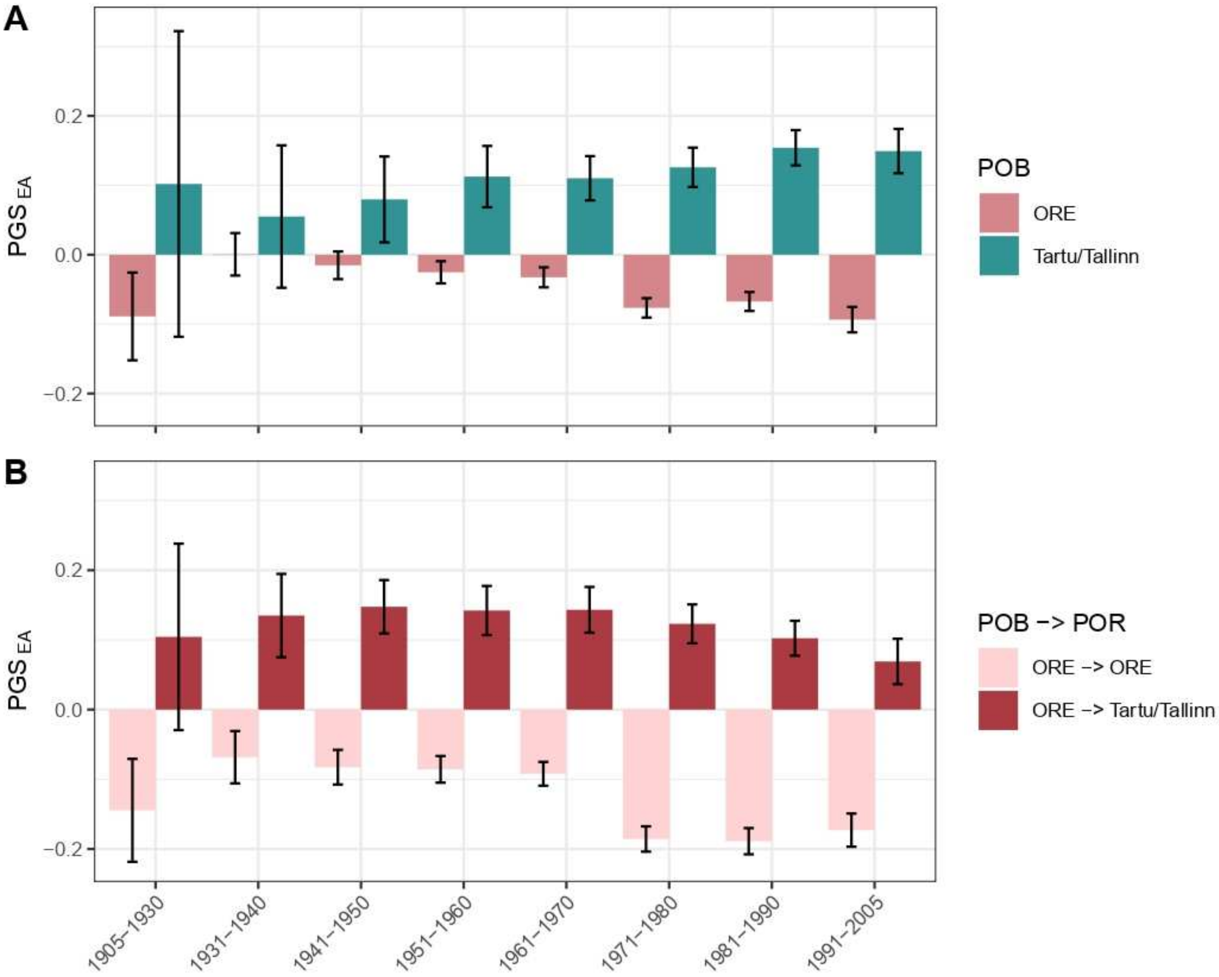

**Figure 6. Difference in average PGS$_{EA}$ between cities (Tallinn or Tartu) and ORE during the 20th century.** (A) Mean PGS$_{EA}$ by area of birth; (B) mean PGS$_{EA}$ of individuals born in ORE by area of residence. PGS$_{EA}$ is adjusted for demographic and genetic ancestry covariates. Error bars correspond to 95% confidence intervals.

## Relation between genetic factors of educational attainment and migration

It has been previously shown that a higher EA level is associated with higher migration activity[20,21,39]. Hence, the patterns we report above for PGS$_{EA}$ can merely reflect migration patterns of individuals with various levels of EA. This is supported by the observation that EA shows similar geographic distribution as well as similar distribution between different migration-profile groups (Supplementary Figures 20-25, 38-49).

To test if the results for PGS$_{EA}$ can be entirely explained by the trait itself we first regressed EA out of PGS$_{EA}$. With either binary and continuous measures of EA (university degree and years of education, correspondingly), regressed out of (s)PGS$_{EA}$, the differences between the migration groups become less pronounced but are not eliminated completely (Supplementary Figures 50-63).

Next, we defined a migration phenotype for individuals born in ORE by distinguishing between those who moved to Tallinn or Tartu City (cases, N = 24,827) versus those who stayed in ORE (controls, N = 61,373). We a) used logistic regression to test if EA and $PGS_{EA}$ predict migration in joint effect models and b) estimated the genetic correlation of migration with EA (Table 1). $PGS_{EA}$ is a significant predictor for migration (p-value 4.9e-258). Years of education attenuates the regression coefficients of $PGS_{EA}$ but keeps it significant (p-value 5.4e-64), which is in agreement with recent study results[40]. Note, however, that converting EA categories to years of education has an empirical rather than theoretical background and can be suboptimal in reflecting the reality in any particular country. Moreover, the "years of education" measure does not follow a normal distribution which can cause statistical artefacts. Thus, we also used reported EA as a categorical covariate. In this model, the effect of $PGS_{EA}$ on migration is close to that with years of education as a covariate and still significant (p-value 2.0e-57). GREML-GCTA analysis shows that migration is a heritable trait ($h^2$ = 0.13, $CI_{95}$: 0.10 - 0.16) and demonstrates a genetic correlation of 0.8 ($CI_{95}$: 0.7 - 0.9) with having versus not having a university degree. This suggests the two traits have largely but not fully overlapping genetic backgrounds.

**Table 1. Genetic aspects of migration phenotypes.** The migration phenotype corresponds to individuals born in ORE and residing in either Tallinn or Tartu City (cases) or in ORE (controls). The logistic regression section provides the odds ratio for $PGS_{EA}$ as a migration predictor in a model without or with EA. Two models with EA as a covariate were tested: years of education translated from the reported categories of EA (Supplementary Table 3) and the reported categorical EA. GREML-GCTA section tabulates heritability estimates for binary educational attainment - university degree ($h^2_{EA}$) and migration ($h^2_{Migr}$) as well as the genetic correlation between them in the corresponding cohort.

**Logistic regression, $OR_{PGS(EA)}$**

| | Estimate, $CI_{95}$ | P-value |
|---|---|---|
| $PGS_{EA}$ | 1.31 [1.29; 1.33] | 4.9e-258 |
| $PGS_{EA}$ + Years of education | 1.15 [1.13; 1.17] | 5.4e-64 |
| $PGS_{EA}$ + EA (categories) | 1.14 [1.12; 1.16] | 2.0e-57 |

**GREML-GCTA**

| | Estimate, $CI_{95}$ | P-value |
|---|---|---|
| $h^2_{EA}$, % | 25.9 [23.1; 28.6] | 8.8e-77 |
| $h^2_{Migr}$, % | 12.9 [10.2; 15.6] | 2.7e-21 |
| $r_g$, % | 79.9 [69.9; 89.9] | 2.3e-55 |

## Replication of the analyses in Estonian subgroups, Russian cohort and using $sPGS_{EA}$

We repeated most of our analyses in subgroups of the Estonian cohort and in the Russian cohort as well as on the entire Estonian cohort using $sPGS_{EA}$ (Supplementary Materials, *Supplementary analyses*). The results in the subgroups are consistent with the observations made in the entire sample, although the statistical power is diminished due to the smaller subgroup sample sizes. This indicates that the observed patterns of inter-regional variance, geographical distribution, and group differences based on migration destination are not driven by the presence of related individuals, variations in sex, age, recruitment strategy, or self-reported ethnicity. Specifically, the results obtained from the Russian cohort largely corroborate the overarching patterns for

Estonians, despite disparities in population structure and geographical distribution between them. Also, while all the trends for $sPGS_{EA}$ are less pronounced than for $PGS_{EA}$, they generally align with the trends observed in $PGS_{EA}$. Collectively, these results underscore the robustness of our key findings with regard to demographic factors and ethnicity, as well as the characteristics of the polygenic score.

## Discussion

In this study, we demonstrated that although contemporary migrations smoothen spatial genetic structure in Estonia, described by genome-wide PCs, such migration enhance inter-regional differences in PGSs, with $PGS_{EA}$ showing the strongest differentiation. Hence, similar patterns described by Abdellaoui et al.[15] are unique neither to the UK Biobank cohort nor to the UK population in general. Importantly, in the 20th century, Estonia went through a series of political transitions related to drastic changes in economic and social organisation. It first gained independence in 1918 and lost it during the Soviet period from 1940 to 1991, which was interrupted by German occupation from 1941 to 1944. This turbulence makes long-term SES inheritance in Estonia less likely than in the UK[27]. Those differences between the UK and Estonia make us suggest that the effect of recent migrations on PGS distribution is a more general phenomenon for urbanised societies, largely independent of political and economic aspects and probably shared with other countries, at least within Europe. We also replicated the patterns of $PGS_{EA}$ distribution in sex-, age- and recruitment strategy-based subcohorts and in self-reported Russians further supporting these patterns to be genuine and general.

Next, we extended our work beyond replicating the study of Abdellaoui et al.[15] in several directions. First, as Estonia is a small country with only two major urbanisation centres (Tallinn and Tartu City) we could show that the non-uniform distribution of $PGS_{EA}$ is driven mostly by the difference between these two cities and the rest of the country and can be related to urbanisation-driven migrations.

Second, due to the wide age range of the Estonian Biobank participants, we were able to add a chronological perspective to the effects of migrations on $PGS_{EA}$ distribution. We showed that differences in average $PGS_{EA}$ between cities and other regions existed already in the first half of the 20th century and consistently increased during and after the Soviet period.

Third, we recapitulated our findings using within-sibship GWAS PGS ($sPGS_{EA}$) instead of population-based GWAS PGS. The within-sibship GWAS provides considerably lower heritability estimates for EA compared to population-based ones[25] which suggests population-derived estimates of effect sizes to incorporate confounders and/or parental effects. Nevertheless, as we recovered qualitatively the same patterns using $sPGS_{EA}$ we can hypothesise that they are at least partially driven by direct genetic effects. Note, however, that a recent study suggested that within-sibship GWAS estimates can still carry some residual confounding[41]

Fourth, we demonstrated that migrants to the cities of Tallinn or Tartu differ in their $PGS_{EA}$ depending on their county of birth which roughly reflects the migration distance and that individuals who moved between Tallinn and Tartu City have on average higher $PGS_{EA}$ than individuals staying in the city of birth. Both observations suggest that $PGS_{EA}$ is not just associated with migration to the cities in general but with more intricate migration patterns, probably linked to search for very specific jobs or educational opportunities, not always present in the closest city. This is in line with previous reports that educational and job opportunities are more often the reasons for long-distance movements than for short-distance in Sweden[42] and that the average EA is higher in longer-distance migrants in the UK[43]. A similar pattern has already been observed phenotypically in the early 20th century in Estonia, where students from farther away from Tartu City had on average higher scores on an intelligence test than students born closer to the city[44]. Although the test used in that study is considered outdated, factors affecting the result are in line with those currently affecting EA[45].

Finally, we explored if the association between migration behaviour and $PGS_{EA}$ can be entirely explained through the EA phenotype. In agreement with a study of mobility in Sweden[40], our results demonstrate that EA only partially explains the relationship between migration and $PGS_{EA}$. While differences in the underlying genetic architecture of EA and migration behaviour ($r_g < 1$) can play some role here, there are other mechanisms potentially contributing to this observation. In fact, the same genetic variants can affect EA and migration behaviour through different pathways (horizontal pleiotropy). Furthermore, a reverse causal relationship between EA and migration may be observed, for example, when migration is a required condition for gaining a certain education level. Third, some individuals might migrate with their parents or partners, whose migration could be related to job or education opportunities. As $PGS_{EA}$ is naturally correlated between parents and offsprings and has been shown to be correlated between partners[24], accompanying family members will have on average higher $PGS_{EA}$ than non-migrants, regardless of their EA. Such “accompanying” migration results in genotype-environment correlations. These correlations can be seen as passive when children move with their parents, and mostly active when spouses move together[46].

The non-random distribution of PGS between regions and migration groups even after a thorough correction for population structure not only provides interesting insights into the interplay between recent social dynamics and genetics but also poses challenges for genetic studies[47]. Regardless of the causality, it generates genotype-environment (G-E) correlations. In the case of active G-E correlations, the environment may be considered dependent on the individual’s genotype thus intermediating the phenotype manifestation. However, interpretability may especially face limitations due to passive G-E correlations, as in this case, the environment depends on the genotypes of parents or even more distant ancestors and not the individual’s genotype (like in the case of “accompanying” migrations). In this study, we showed that alleles associated with higher EA are also associated with staying in or moving to the cities, where the conditions of living are different from those in towns and rural areas. The urban population, for instance, has been shown to be healthier in general[48–50]. This is most probably due to differences in environment rather than genetics[51–53]. Moreover, the environment is being inherited not just because of geography but also due to cultural transmission of lifestyle. So, this issue might be even more complex than we show here and would be present even in the situation when all the population would settle in a single location without any spatial segregation. Such G-E correlations, especially passive ones, may lead to inflated estimates of heritability, and genetic correlations and affect GWAS and other genetic analyses such as Mendelian randomisation[15,47]. Given the patterns we report here, it is reasonamle to assume that the EA phenotype can be especially prone to such confounders[25,47]. Thus, most estimates of direct genetic effects on EA are likely to be inflated.

This study has several limitations. First, although the biobank data includes information on approximately 20% of the adult population in Estonia it has been shown not to be a completely representative population cohort[31] (Supplementary Materials, *Estonian Biobank cohort overview*). We expect the EstBB to be more representative than the UK Biobank because of the fraction of the population covered and the diversity of participants but not to be bias-free. Replication of the results in the subcohorts and in the Russian cohort reduces the risk of artefacts due to systematic participation bias. However, it should be taken into account when interpreting the results. Second, there is a minor uncertainty in the EA phenotype. The information is received from the population register as well as from the questionnaire. In some cases, this information can be outdated or inaccurate. If these errors are not random, it can lead to a systematic bias in the results[54]. Converting EA from a categorical to a continuous scale probably is not an ideal strategy as it includes, although commonly used, a partially arbitrary rescaling procedure that leads to the loss of information[55]. Third, the information on the places of birth or residence may not be perfect as well. The reported place of birth may in some cases correspond to the settlement where the maternity hospital was located and not to the actual place where the family lived at the time of birth. For this reason, it is safer to consider counties than individual cities and our main conclusions are not sensitive to this issue. The information on the place of residence is updated regularly, synchronising with the population register. However, people do not always report their movements to the register. Fourth, reporting the results for separate age groups we consider the age of the dead individuals to be fixed at the time of death. Given that migration behaviour can change both with an individual's age and over historical periods, the desynchronisation of year of birth and age can smoothen some patterns conditioned on one of these factors. Still, this effect is expected to be negligible (Supplementary Materials, *Estonian Biobank cohort overview*).

Finally, we would like to make a caution note about interpreting our results within a broader sociological framework. In most analyses, we used the polygenic score based on population-based GWAS for EA. It has been shown by many studies that it is influenced by lots of diverse confounders[15,24,25,47,56–61]. Thus, $\text{PGS}_{\text{EA}}$ cannot be interpreted as a cumulative genetic factor directly affecting EA outcome. It is rather a correlate of EA, likely only modestly determined by direct genetic effects. Though we reproduced the main pattern of increased inter-county difference with $sPGS_{EA}$ as well, it cannot be perceived as final proof of a genetic-driven mechanism (Supplementary Materials, *General Summary & Frequently Asked Questions*). One should also note that the differences in mean $PGS_{EA}$ between migration groups are subtle despite being statistically significant. Moreover, the corresponding distributions strongly overlap for all the migration groups considered (Supplementary Materials, *General Summary & Frequently Asked Questions*).

Our findings demonstrate that people's geographic mobility, particularly related to urbanisation, is accompanied by changes in the genetic structure of a population. The comparison of Estonia and the UK shows this phenomenon can manifest in countries with different socio-economic systems as well as population sizes. Such migrations, non-random with respect to genetics, generate genotype-environment correlations which are not only a technical issue for genetic studies but also a potential burden for society. In this context, it implies that potentially tiny differences in genetic factors affecting EA translate into environmental differences not linearly, but being amplified by the environment. Consequently, individuals with a lower genetic predisposition to EA have fewer opportunities to fulfil their potential. We speculate that the same pattern can be observed on a finer within-settlement scale and even without spatial segregation due to social stratification. Thus, active measures might be needed if a society aims at truly equal opportunities in education and related aspects for all its members.

## Methods

### *Participants*

The participants of this study were sourced from the Estonian Biobank (EstBB), which is a volunteer-based cohort of the Estonian resident adult population[31]. It includes (as of 2022) genetic and diverse phenotype data on 210,438 individuals (72,384 men and 137,180 women) corresponding to ~20% (~14% men and ~24% women) of the contemporary adult population of Estonia[62]. Participants' age ranges from 18 to 107, determined as of 2022 for alive participants or at the year of death. The EstBB is linked with the Estonian national register so the information on education level and place of residence is being constantly updated. The participants were recruited over two decades from 2001 to 2021 across the country, covering all the regions and a variety of different settings providing socio-economic and ethnic heterogeneity. Besides genetic and demographic data, participants provided health data, blood samples and lifestyle information.

### *Ethics statement*

The activities of the EstBB are regulated by the Human Genes Research Act, which was adopted in 2000 specifically for the operations of the EstBB. Individual level data analysis in the EstBB was carried out under ethical approval "1.1-12/3593" from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs), using data according to release application "4-1.6/GI/79" from the Estonian Biobank.

*Genotypes and quality control*

Samples were genotyped on the Infinium Global Screening Array (GSA) of different versions (depending on the time of recruitment) with approximately 550,000 overlapping positions. Samples with <95% call rate or mismatch between genetic and self-reported sex were excluded. Before the imputation step all non-SNP polymorphisms and strand ambiguous SNPs were filtered out. The final number of SNPs before the imputation step was 309,258. The genotypes were imputed with Beagle 5.4[63] using the Estonian Reference panel as a reference set[64]. To create polygenic scores, we extracted a set of 1,075,599 autosomal HapMap 3 SNPs with a minor allele count >5, and info score >0.7. Unrelated individuals were defined as having less than 2nd-degree relationship inferred with KING[65].

For GREML analysis, the non-imputed genotyping data were used after keeping SNPs with minor allele frequency >0.01, Hardy–Weinberg equilibrium (HWE) p-value $>10^{-5}$ and missingness <0.015. Related individuals with a 2nd-degree relationship and closer were excluded. Relationships were inferred with KING[65].

*Ancestry and PCA*

Ancestry grouping was estimated with bigsnpr[66]. For ancestry inference, genotypes were imputed using 1000 Genomes Project phase 3 samples[3]. Individuals from "Europe (East)", "Europe (North West)" and "Finland" inferred ancestry groups were kept for further analysis. Next, individuals with no self-reported Estonian or Russian ethnicity were excluded from the participants who passed the ancestry filter.

A principal component analysis (PCA) was conducted separately on individuals of Estonian (182,252 individuals) and Russian (17,954 individuals) self-reported ethnicity to capture ancestry differences within the corresponding populations. Before the analysis genotypes were filtered for minor allele frequency >0.01, Hardy–Weinberg equilibrium (HWE) p-value $>10^{-5}$ and missingness <0.05. Long-range linkage disequilibrium regions were removed[67]. Genotypes were pruned for linkage disequilibrium with PLINK2[68,69] with window size 50kb, step 5kb and $r^2$ threshold 0.1. The PCA to construct PCs on Estonian and Russian individuals was conducted on this SNP set using flashPCA version 2[70].

*Polygenic score calculations*

Polygenic scores were computed for 169 phenotypes using population-based GWAS summary statistics from the UK Biobank (PGSs)[32] and 24 phenotypes using within-sibship GWAS summary statistics (sPGSs)[25]. The PGSs were calculated using summary statistics from GWAS in the European ancestry cohort of the UK Biobank conducted by the Pan-UKBB team[34]. The Pan-UKBB project particularly presents an analysis of 7,228 phenotypes, spanning 16,131 studies. The list of traits selected for the analysis included the maximally independent set of 146 phenotypes (with correlation between them <0.1) for which GWAS results passed the quality control. Additionally, 23 phenotypes related to education, mental health, fluid intelligence, height and body mass index (BMI) were added. The complete list of the phenotypes and the numbers of individuals included in the study is presented in Supplementary Table 1.

The sPGSs were calculated for all phenotypes analysed in the original study presenting a set of within-sibship GWAS results estimating direct genetic effects. Supplementary Table 2 lists 24 traits with corresponding sample sizes.

Polygenic scores with both sets of summary statistics were calculated using SBayesR with default parameters including LD matrix built using data on 50,000 UK Biobank participants[71]. To remove the effect of the ancestral genetic structure on polygenic scores, the top 100 ancestry-informative principal components (PCs) specific to Estonian or Russian ancestry were regressed out. Sex, age, sex×age and age$^2$ were also regressed out of the PGSs to mitigate the influence of potential sex and age bias reported for population volunteer cohorts[57,72]. In analyses of PGS adjusted for educational attainment, binary or continuous EA (see section "*Educational attainment phenotypes*") was also regressed out.

*Sources of education and geographic information*

Initial information on the highest level of education, place of birth and place of residence was obtained from the questionnaire completed by participants when enrolled in the biobank. The EstBB regularly synchronises its information with the Estonian Population Register on the highest level of education and municipality of residence. The data used in this study was last updated in 2022. Participants without information on the counties of birth and residence in Estonia or born outside the country were excluded from the analysis. Participants born or residing in Harju or Tartu Counties and lacking information on the municipality were excluded from the analyses where it was necessary to distinguish Tallinn/Tartu City from other municipalities of the corresponding counties. After filtering, the analysed sample included 172,376 individuals of self-reported Estonian ethnicity and 11,200 individuals of self-reported Russian ethnicity.

*Educational attainment phenotypes*

Continuous and binary traits corresponding to educational attainment were considered. The continuous "years of education" phenotype was derived according to the ISCED 2011 methodology. The link table for the reported level of education, ISCED 2011 and "years of education" is presented in Supplementary Table 3. Alternatively, attainment of a Bachelor's degree or higher was used as a binary phenotype. The quantitative EA phenotype was adjusted to mitigate possible sampling bias in the corresponding analyses. Sex, age, sex×age and age$^2$ and

100 genetic PCs were regressed out of the quantitative EA using linear regression.

*Geographic variability of ancestry and polygenic score variation*

The measure of geographic variability was the proportion of variance explained by county differences:

$$Var_{county} = SSB\,/\,(SSB + SSW)$$

where *SSB* is the sum of squares between counties, *SSW* is the sum of squares within counties. P-values were calculated from the ANOVA test. The chi-square test was implemented to test whether the difference of variance explained by county of birth and county of residence together is significantly larger than by exclusively one of them. The base model to compare with was a less powerful model with either county of birth or county of residence as an independent variable. Statistical significance was determined using a level of 0.05 after the Bonferroni correction for the number of tests (100 for PCs, 169 for PGSs and 24 for sPGS).

*Logistic regression*

Logistic regression with migration phenotype as a dependent variable was performed with $PGS_{EA}$ or $PGS_{EA}$ and EA (years of education or categories) as independent variables. Sex, age, $age^2$, sex×age, sex×$age^2$ and 100 genetic PCs were included in the models as covariates.

*Heritability and genetic correlation calculations*

Bivariate GREML analysis implemented in GCTA software[73,74] was used to estimate heritabilities and genetic correlations. Sex, age, $age^2$, sex×age, sex×$age^2$ and 10 genetic PCs were included as covariates in the models.

*Geographic data visualisation*

Shapefiles used to plot maps of Estonia with county borders were retrieved from the Estonian Land Board website (Administrative and Settlement Division, 2023.02.01)[75]. Geographic data were visualized in R[76] with the aid of the following packages: “sf”[77,78], “geos”[79] and “ggplot2”[80].

## Data availability

Access to the Estonian Biobank data (https://genomics.ut.ee/en/content/estonian-biobank) is restricted to approved researchers and can be requested.

## Code availability

Custom R code used for statistical analyses is available from the corresponding authors on request.

## Author contributions

IK, LP, FM and VP conceived and designed the study. IK performed all the analyses. IK and VP wrote the initial draft of the manuscript. All co-authors contributed to the interpretation of the results, reviewed and approved the submitted version of the manuscript.

## Acknowledgements

We want to acknowledge the participants of the Estonian Biobank. This study was funded by the European Union through the European Regional Development Fund Project No. 2014-2020.4.01.15-0012 GENTRANSMED. Data analysis was carried out in part in the High-Performance Computing Center of the University of Tartu. IK was a student of and supported by the Center of Life Sciences, Skolkovo Institute of Science and Technology, 'Moscow, Russia during the initial phase of the project. VP and MM were supported by the European Union through Horizon 2020 research and innovation programme under grant no 810645 and through the European Regional Development Fund project no. MOBEC008. MM was also supported by the Estonian Research Council grant PUT (PRG1899). UV was supported by the Estonian Research Council grant PUT (PSG759). Luca Pagani was supported by the Italian Ministry of University and Research (2022B27XYM). FM was supported by Fondazione con il Sud (2018-PDR-01136) and by the Italian Ministry of University and Research (2022P2ZESR). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

## Ethics declarations

*Competing interests*

The authors declare no competing interests.

## References

1. Charlesworth, B. & Charlesworth, D. *Elements of Evolutionary Genetics*. (W. H. Freeman, 2010).

2. Cavalli-Sforza, L. L. Genes, peoples, and languages. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 7719–7724 (1997).

3. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

4. O’Dushlaine, C. *et al.* Genes predict village of origin in rural Europe. *Eur. J. Hum. Genet.* **18**, 1269–1270 (2010).

5. Abdellaoui, A. *et al.* Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet.* **21**, 1277–1285 (2013).

6. Kerminen, S. *et al.* Fine-scale genetic structure in Finland. G3 (Bethesda) 7: 3459--3468. Preprint at (2017).

7. Pankratov, V. *et al.* Differences in local population history at the finest level: the case of the Estonian population. *Eur. J. Hum. Genet.* **28**, 1580–1591 (2020).

8. Nait Saada, J. *et al.* Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nat. Commun.* **11**, 6130 (2020).

9. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).

10. Kivisild, T. *et al.* Patterns of genetic connectedness between modern and medieval Estonian genomes reveal the origins of a major ancestry component of the Finnish population. *Am. J. Hum. Genet.* **108**, 1792–1806 (2021).

11. Rodríguez Díaz, R. & Blanco Villegas, M. J. Genetic structure of a rural region in Spain: distribution of surnames and gene flow. *Hum. Biol.* **82**, 301–314 (2010).

12. Gilbert, E. *et al.* The Irish DNA Atlas: Revealing Fine-Scale Population Structure and History within Ireland. *Sci. Rep.* **7**, 17199 (2017).

13. Byrne, R. P. *et al.* Dutch population structure across space, time and GWAS design. *Nat. Commun.* **11**, 4556 (2020).

14. Kerminen, S. *et al.* Changes in the fine-scale genetic structure of Finland through the 20th

century. *PLoS Genet.* **17**, e1009347 (2021).

15. Abdellaoui, A. *et al.* Genetic correlates of social stratification in Great Britain. *Nat Hum Behav* **3**, 1332–1342 (2019).
16. Jaffee, S. R. & Price, T. S. Genotype-environment correlations: implications for determining the relationship between environmental exposures and psychiatric illness. *Psychiatry* **7**, 496–499 (2008).
17. Saltz, J. B. Gene-Environment Correlation in Humans: Lessons from Psychology for Quantitative Genetics. *J. Hered.* **110**, 455–466 (2019).
18. Verweij, K. J. H., Mosing, M. A., Zietsch, B. P. & Medland, S. E. Estimating Heritability from Twin Studies. in *Statistical Human Genetics: Methods and Protocols* (eds. Elston, R. C., Satagopan, J. M. & Sun, S.) 151–170 (Humana Press, 2012).
19. Richards, J. B. & Evans, D. M. Back to school to protect against coronary heart disease? *BMJ* vol. 358 j3849 (2017).
20. Malamud, O. & Wozniak, A. The Impact of College on Migration: Evidence from the Vietnam Generation. *J. Hum. Resour.* **47**, 913–950 (2012).
21. Haapanen, M. & Böckerman, P. More educated, more mobile? Evidence from post-secondary education reform. *Spatial Economic Analysis* **12**, 8–26 (2017).
22. Britton, J., van der Erve, L., Xu, X. & Waltmann, B. *London calling? Higher education, geographical mobility and early-career earnings*. https://ifs.org.uk/publications/london-calling-higher-education-geographical-mobility-and-early-career-earnings (2021).
23. Silventoinen, K. *et al.* Genetic and environmental variation in educational attainment: an individual-based analysis of 28 twin cohorts. *Sci. Rep.* **10**, 12681 (2020).
24. Okbay, A. *et al.* Polygenic prediction of educational attainment within and between families

from genome-wide association analyses in 3 million individuals. *Nat. Genet.* **54**, 437–449 (2022).

25. Howe, L. J. *et al.* Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nat. Genet.* **54**, 581–592 (2022).

26. Wolfram, T. & Morris, D. M. Conventional twin studies overestimate the environmental differences between families relevant to educational attainment. (2022) doi:10.31234/osf.io/m4eqv.

27. Meissner, B. The change in the social structure of Estonia. *J. Balt. Stud.* **18**, 301–322 (1987).

28. Clark, G. & Cummins, N. Surnames and social mobility in England, 1170-2012. *Hum. Nat.* **25**, 517–537 (2014).

29. Clark, G. The inheritance of social status: England, 1600 to 2022. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2300926120 (2023).

30. Bernard, A. Does the association between internal migration and personality traits hold in different countries? *J. Res. Pers.* **101**, 104300 (2022).

31. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).

32. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

33. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

34. Pan UKBB. https://pan.ukbb.broadinstitute.org.

35. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association

study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

36. Brumpton, B. *et al.* Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nat. Commun.* **11**, 3519 (2020).

37. Sjoberg, O. & Tammaru, T. Transitional statistics: internal migration and urban growth in post-Soviet Estonia. *Eur. Asia. Stud.* **51**, 821–842 (1999).

38. Lang, T. *et al.* Socio-spatial polarisation and policy response: Perspectives for regional development in the Baltic States. *Eur. Urban Reg. Stud.* **29**, 21–44 (2022).

39. Xu, X., Waltmann, B., van der Erve, L. & Britton, J. *London calling? Higher education, geographical mobility and early-career earnings*. https://ifs.org.uk/publications/15622 (2021) doi:10.1920/re.ifs.2021.0198.

40. Ojalehto, E., Finkel, D., Russ, T. C., Karlsson, I. K. & Ericsson, M. Influences of genetically predicted and attained education on geographic mobility and their association with mortality. *Soc. Sci. Med.* **324**, 115882 (2023).

41. Veller, C. & Coop, G. Interpreting population and family-based genome-wide association studies in the presence of confounding. *bioRxiv* (2023) doi:10.1101/2023.02.26.530052.

42. Niedomysl, T. How Migration Motives Change over Migration Distance: Evidence on Variation across Socio-economic and Demographic Groups. *Reg. Stud.* **45**, 843–855 (2011).

43. Lawson, D. J. *et al.* Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Hum. Genet.* **139**, 23–41 (2020).

44. Tork, J. *Eesti laste intelligents [The intelligence of Estonian children.]*. (Estonia, Tartu: Koolivara, 1940).

45. Must, O., te Nijenhuis, J., Must, A. & van Vianen, A. E. M. Comparability of IQ scores over time. *Intelligence* **37**, 25–33 (2009).

46. Plomin, R., DeFries, J. C. & Loehlin, J. C. Genotype-environment interaction and correlation in the analysis of human behavior. *Psychol. Bull.* **84**, 309–322 (1977).

47. Abdellaoui, A., Dolan, C. V., Verweij, K. J. H. & Nivard, M. G. Gene-environment correlations across geographic regions affect genome-wide association studies. *Nat. Genet.* **54**, 1345–1354 (2022).

48. Leinsalu, M. Social variation in self-rated health in Estonia: a cross-sectional study. *Soc. Sci. Med.* **55**, 847–861 (2002).

49. Galal, S. B. & Al-Gamal, N. Health problems and the health care provider choices: a comparative study of urban and rural households in Egypt. *J. Epidemiol. Glob. Health* **4**, 141–149 (2014).

50. Singh, G. K. *et al.* Social Determinants of Health in the United States: Addressing Major Health Inequality Trends for the Nation, 1935-2016. *Int J MCH AIDS* **6**, 139–164 (2017).

51. Spasojevic, N., Vasilj, I., Hrabac, B. & Celik, D. RURAL - URBAN DIFFERENCES IN HEALTH CARE QUALITY ASSESSMENT. *Mater Sociomed* **27**, 409–411 (2015).

52. Lusmägi, P., Einasto, M. & Roosmaa, E.-L. Leisure-time Physical Activity Among Different Social Groups of Estonia: Results of the National Physical Activity Survey. *Physical Culture and Sport. Studies and Research* **69**, 43–52 (2016).

53. Ma, C., Devoti, A. & O'Connor, M. Rural and urban disparities in quality of home health care: A longitudinal cohort study (2014-2018). *J. Rural Health* **38**, 705–712 (2022).

54. Schoeler, T. *et al.* Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat Hum Behav* **7**, 1216–1227 (2023).

55. Schneider, S. L. Nominal comparability is not enough: (In-)equivalence of construct validity of cross-national measures of educational attainment in the European Social Survey. *Res. Soc. Stratif. Mobil.* **28**, 343–357 (2010).

56. Domingue, B. W., Fletcher, J., Conley, D. & Boardman, J. D. Genetic and educational assortative mating among US adults. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 7996–8000 (2014).

57. Pirastu, N. *et al.* Genetic analyses identify widespread sex-differential participation bias. *Nat. Genet.* **53**, 663–671 (2021).

58. Kemper, K. E. *et al.* Phenotypic covariance across the entire spectrum of relatedness for 86 billion pairs of individuals. *Nat. Commun.* **12**, 1050 (2021).

59. Howe, L. J., Evans, D. M., Hemani, G., Davey Smith, G. & Davies, N. M. Evaluating indirect genetic effects of siblings using singletons. *PLoS Genet.* **18**, e1010247 (2022).

60. Young, A. I. *et al.* Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nat. Genet.* **54**, 897–905 (2022).

61. Young, A. S. Estimation of indirect genetic effects and heritability under assortative mating. *bioRxiv* (2023) doi:10.1101/2023.07.10.548458.

62. Select table. https://andmed.stat.ee/en/stat.

63. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).

64. Mitt, M. *et al.* Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur. J. Hum. Genet.* **25**, 869–876 (2017).

65. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

66. Privé, F. Using the UK Biobank as a global reference of worldwide populations: application to measuring ancestry diversity from GWAS summary statistics. *Bioinformatics* **38**, 3477–3480 (2022).

67. Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *American journal of human genetics* vol. 83 132–5; author reply 135–9 (2008).

68. PLINK 2.0. http://www.cog-genomics.org/plink/2.0/.

69. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

70. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).

71. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).

72. Schoeler, T. *et al.* Correction for participation bias in the UK Biobank reveals non-negligible impact on genetic associations and downstream analyses. *bioRxiv* 2022.09.28.509845 (2022) doi:10.1101/2022.09.28.509845.

73. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

74. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).

75. Maa-amet. Administrative and settlement division. https://geoportaal.maaamet.ee/eng/Spatial-Data/Administrative-and-Settlement-Division-

p312.html.

76. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* (2020).

77. Pebesma, E. J. & Bivand, R. *Spatial Data Science: With Applications in R*. (CRC Press, Taylor & Francis Group, 2023).

78. Pebesma, E. Simple features for R: Standardized support for spatial vector data. *R J.* **10**, 439 (2018).

79. Open Source Geometry Engine ('GEOS') R API [R package geos version 0.2.3]. (2023).

80. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Preprint at https://ggplot2.tidyverse.org (2016).