nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-019-0757-5

Genetic correlates of social stratification in Great Britain

Abdel Abdellaoui[1*], David Hugh-Jones[2], Loic Yengo[3], Kathryn E. Kemper[3], Michel G. Nivard[4], Laura Veul[1], Yan Holtz[3], Brendan P. Zietsch[5], Timothy M. Frayling[6], Naomi R. Wray[3,7], Jian Yang[3,7], Karin J. H. Verweij[1] and Peter M. Visscher[3,7*]

Human DNA polymorphisms vary across geographic regions, with the most commonly observed variation reflecting distant ancestry differences. Here we investigate the geographic clustering of common genetic variants that influence complex traits in a sample of ~450,000 individuals from Great Britain. Of 33 traits analysed, 21 showed significant geographic clustering at the genetic level after controlling for ancestry, probably reflecting migration driven by socioeconomic status (SES). Alleles associated with educational attainment (EA) showed the most clustering, with EA-decreasing alleles clustering in lower SES areas such as coal mining areas. Individuals who leave coal mining areas carry more EA-increasing alleles on average than those in the rest of Great Britain. The level of geographic clustering is correlated with genetic associations between complex traits and regional measures of SES, health and cultural outcomes. Our results are consistent with the hypothesis that social stratification leaves visible marks in geographic arrangements of common allele frequencies and gene–environment correlations.

The first law of geography states that "everything is related to everything else, but near things are more related than distant things"[1]. Humans living near each other tend to share more ancestry with each other than with humans who live further away, which is reflected in genome-wide patterns of genetic variation on a global scale[2] and on finer scales[3–5]. Regional differences in allele frequencies are driven by genetic drift (that is, the random fluctuations of allele frequencies in each generation), natural selection pressures, migrations or admixture (that is, two previously isolated populations interbreeding). Out of these four mechanisms, genetic drift is the only mechanism not expected to disproportionately affect genetic variants that are associated with heritable human traits. Natural selection targets heritable traits over extended periods of time, thereby affecting allele frequencies of the genetic variants that are associated with the traits under selection. Earlier studies have identified natural selection pressures on many trait-associated variants by looking for extreme allele frequency differences between different ancestries[3,6,7]. Migration is a behaviour, and since most behavioural traits have heritable components[8], migration is likely to be associated with genetic variants that influence behaviour. Long-distance migratory events may in turn result in admixture. Internal migrations (that is, migrations within countries) may lead to geographic clustering of trait-associated genetic variants beyond the clustering of ancestry, and may occur for a variety of reasons. They may be driven by the search for specific neighbourhood, housing and inhabitant characteristics, and/or socioeconomic factors[9], such as the mass migrations from rural to industrial areas during industrialization[10]. These geographic movements may coincide with the regional clustering of a range of heritable outcomes such as socioeconomic status (SES), health and cultural outcomes[11–14].

Understanding what drives the geographic distribution of genome-wide complex trait variation is important for a variety of reasons. Studying regional differences of genetic variants associated with education, wealth, growth, health and disease may help explain why these traits are unevenly distributed across a country. Besides the known regional differences in income and SES, significant regional differences have been reported for mental[14] and physical[13] health problems. Regional differences in wealth and health are probably linked to each other[15–17] and have been shown to be partly driven by migration[13,18]. If genome-wide complex trait variation is geographically clustered, this should be taken into account in certain genetically informative study designs. Mendelian randomization, for example, uses genetic variants as instrumental variables to identify causality, under the assumption that the genetic instrument is not associated with confounders that influence the two traits under investigation[19]. Geographic clustering of genetic complex trait variation could introduce gene–environment correlations that violate this assumption[20]. Such gene–environment correlations could also introduce bias in heritability estimates in twin and family studies[21], and could affect signals from genome-wide association studies (GWASs).

Here, we investigate whether genome-wide complex trait variation in Great Britain is geographically clustered after accounting for ancestry differences; if so, this may reflect the genetic consequences of more recent (internal) migration events. We then investigate the role of migration in the geographic clustering of genome-wide complex trait variation. Finally, we examine whether genome-wide complex trait variation and its geographic clustering are associated with regional SES, health and cultural outcomes.

[1]Department of Psychiatry, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands. [2]Department of Economics, University of East Anglia, Norwich, UK. [3]Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia. [4]Department of Biological Psychology, VU University, Amsterdam, The Netherlands. [5]School of Psychology, University of Queensland, Brisbane, Queensland, Australia. [6]Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK. [7]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia. *e-mail: a.abdellaoui@amsterdamumc.nl; peter.visscher@uq.edu.au

Results

Data and analysis. We investigated geographic clustering of ancestry and complex trait variation using phenotypic measurements and genome-wide single-nucleotide polymorphism (SNP) data from ~450,000 British individuals of European ancestry from UK Biobank[22]. Ancestry within Great Britain was captured by conducting a principal component analysis (PCA)[23] on genome-wide common SNPs—a method that has been shown to successfully capture ancestry differences within relatively homogeneous populations[3]. Genome-wide complex trait variation was captured by polygenic scores, which are created by weighting an individual's alleles by the estimated allelic effects on the trait of interest and then summing the weights, resulting in predictive scores for each individual. We built polygenic scores for 456,426 individuals from 1,312,100 autosomal SNPs using effect estimates from 33 published GWASs on traits related to psychiatric disease, substance use, personality, body composition, cardiovascular disease, diabetes, reproduction and educational attainment (EA) (Supplementary Table 1). Importantly, the 33 GWASs that produced the effect estimates did not, to our knowledge, include UK Biobank participants[24]. We included five heritable phenotypic outcomes; namely, EA, height, body mass index (BMI), body fat and overall health. Geographic clustering of principal components (PCs), phenotypes and polygenic scores was then investigated in 320,940 unrelated individuals and their birthplace by testing whether their spatial autocorrelation (Moran's I) was significantly greater than zero. The spatial autocorrelation (Moran's I) is the correlation in a measure among nearby locations in space, and its values range between −1 (dispersed) and 1 (spatially clustered), where 0 = spatially random[25]. Supplementary Fig. 1 shows the geographic locations of UK Biobank participants. Furthermore, we tested whether phenotypes and polygenic scores that showed significant geographic clustering were associated with migration into or out of the most economically deprived regions (coal mining areas). Finally, we examined whether the complex traits and their geographic clustering were associated with regional measures of SES, health and cultural outcomes.

Geographic clustering of ancestry and complex trait variation. In line with earlier studies[5], British ancestry showed significant geographic clustering: the first 100 genetic PCs all showed Moran's I values that were >0, with 72 PCs showing an empirical $P < 0.0005$ (the Bonferroni-corrected threshold) and 95 PCs showing an empirical $P < 0.05$, with the top PCs generally having higher Moran's I values (Fig. 1 shows the first five PCs; https://holtzyan.shinyapps.io/UKB_geo/ shows maps of all 100 PCs). The geographic distributions of the ancestry differences captured by the PCs probably reflect historical demographic events[5]. These include old population movements and settlements, followed by generations of relatively isolated (sub) populations that went through genome-wide allele frequency differentiation through genetic drift and, perhaps, differential natural selection pressures.

Both before and after correcting for 100 PCs, the phenotypes all showed significant Moran's I values (0.61 for EA, 0.63 for height, 0.51 for BMI, 0.61 for body fat and 0.52 for overall health (all empirical P values $< 10^{-4}$; Fig. 2 and Supplementary Figs. 2 and 4)). In particular, EA, body fat, BMI and overall health showed considerable clustering in coal mining areas (Supplementary Fig. 2), as further discussed below. Without controlling for ancestry, 30 out of the 33 polygenic scores tested showed significant geographic clustering, with geographic distributions similar to ancestry differences captured by the PCs (Supplementary Fig. 3; see https://holtzyan.shinyapps.io/UKB_geo/ for maps of all of the polygenic scores). After regressing out 100 PCs, 21 polygenic scores remained significantly geographically clustered, with EA polygenic scores showing the highest Moran's I values, and geographic distributions similar to the phenotypic EA outcome. We included two EA polygenic scores based on two similar-sized GWASs: one excluding UK Biobank participants and 23andMe ($n = 217{,}569$), based on the Okbay et al.[26] EA GWAS (EA2); and one excluding all British cohorts and 23andMe ($n = 245{,}621$), based on the larger Lee et al.[27] EA GWAS (EA3). The polygenic score based on the GWAS excluding all British cohorts should be more robust to residual population stratification. The two EA polygenic scores gave very similar results (before PC correction: Moran's $I = \sim 0.6$; empirical P values $< 10^{-4}$; after PC correction: Moran's $I = \sim 0.5$; empirical P values $< 10^{-4}$ (Figs. 2 and 3)).

The polygenic scores of the different traits differed in their predictive power due to variation in the trait heritability and due to different discovery GWAS sample sizes. It is expected that geographic clustering is easier to detect when the polygenic score has more predictive power. We computed the expected r^2 value (i.e., proportion of variance explained) of the polygenic scores with their traits using $h^2_g/(1 + [M/N \times h^2_g])$, where h^2_g is the SNP-based heritability, M is the equivalent number of independent SNPs after accounting for linkage disequilibrium, and N is the discovery GWAS sample size (Fig. 2)[28]. The correlation between Moran's I and the expected r^2 between the polygenic scores and their traits was 0.14 (calculated over 33 traits; $P = 0.45$), indicating that the differences we observed in the estimated geographic clustering of different traits were not driven by differences in the predictive power of the respective polygenic scores.

It has been argued that geographic clustering of genetic variation related to EA in Great Britain is probably due to (subtle) ancestry differences or ascertainment bias[29]. In the Supplementary Information, we discuss in more detail why these are unlikely to drive our observations (see the section 'Population stratification and ascertainment bias'). Instead, we explored a more likely explanation; namely, recent SES-related migrations.

SES-related migration. The geographic clustering of genome-wide trait-associated alleles after correcting for population stratification would be expected if there were migration events that occurred more recently than the pre-modern demographic events that drove the regional ancestry differences captured by the PCs. We investigated whether our observations were consistent with what is expected from relatively recent internal migrations due to SES-related factors, which are known to motivate longer-distance moves especially[30]. SES-related migration would be in line with the rest of the traits being geographically clustered because they share genes with EA: the genetic correlation between EA and the rest of the traits was significantly associated with their Moran's I values ($r = 0.56$; $P = 9.5 \times 10^{-4}$; Supplementary Fig. 5). We hypothesized that two types of migration flows may have affected the geographic clustering of SES-related alleles: (1) labourers and farmers leaving the countryside during the Industrial Revolution to work in the geographically clustered industrial jobs[10]; and (2) more recent migration of higher-educated people, or people seeking a higher education, out of the more economically deprived industrial regions.

Much of the energy necessary for mass production during the Industrial Revolution came from coal mines, which attracted large numbers of manual labourers. The Industrial Revolution and later deindustrialization had a great impact on the economy of the coal mining areas[31]. The decline of the British coal industry began in the 1920s, and nearly the whole industry has closed since the early 1980s, resulting in major job losses that remained visible in unemployment rates decades later[32]. Economic deprivation is widespread in coal mining areas: 43% of neighbourhoods in coal mining areas fall into the 30% most economically deprived[31]. In our analysis, coal mining areas showed more economic deprivation than the rest of Great Britain from 1971–2011, as measured with the Townsend index[33] (all false discovery rate (FDR)-corrected P values $< 10^{-32}$; Fig. 3 and Supplementary Fig. 6). All regions have become less economically deprived over time, but the difference between coal mining areas and the rest remains highly significant.

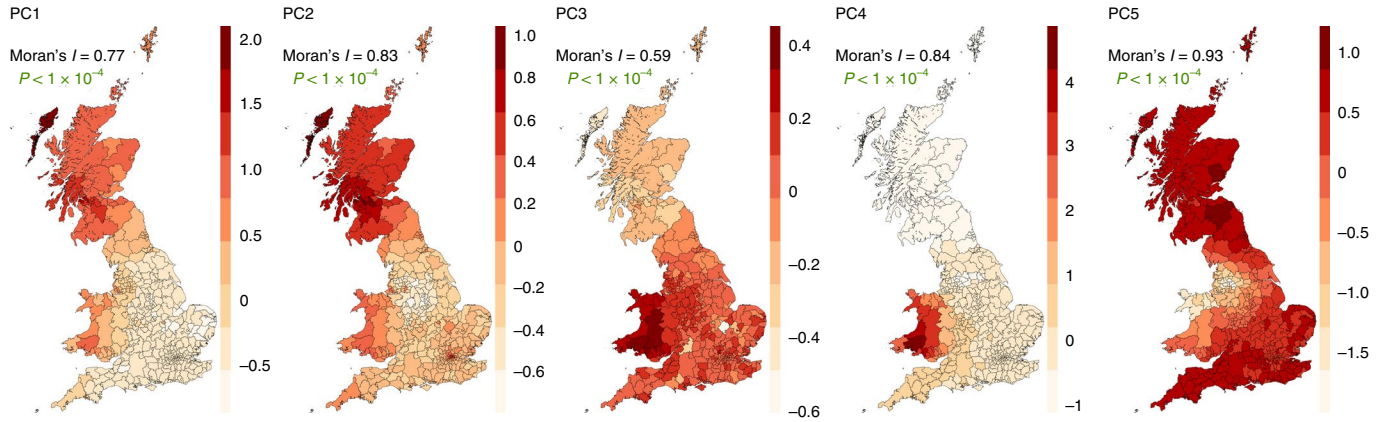


Fig. 1 | Geographic distributions (birthplace) of the first five PCs, Moran's *I* and empirical *P* values for Moran's *I*. *P* values denoted in green are significant after Bonferroni correction ($n = 320{,}940$ unrelated individuals). Maps were adapted from 2011 Census aggregate data (UK Data Service (February 2017 edition). Office for National Statistics; National Records of Scotland; Northern Ireland Statistics and Research Agency (2017); https://doi.org/10.5257/census/aggregate-2011-2).

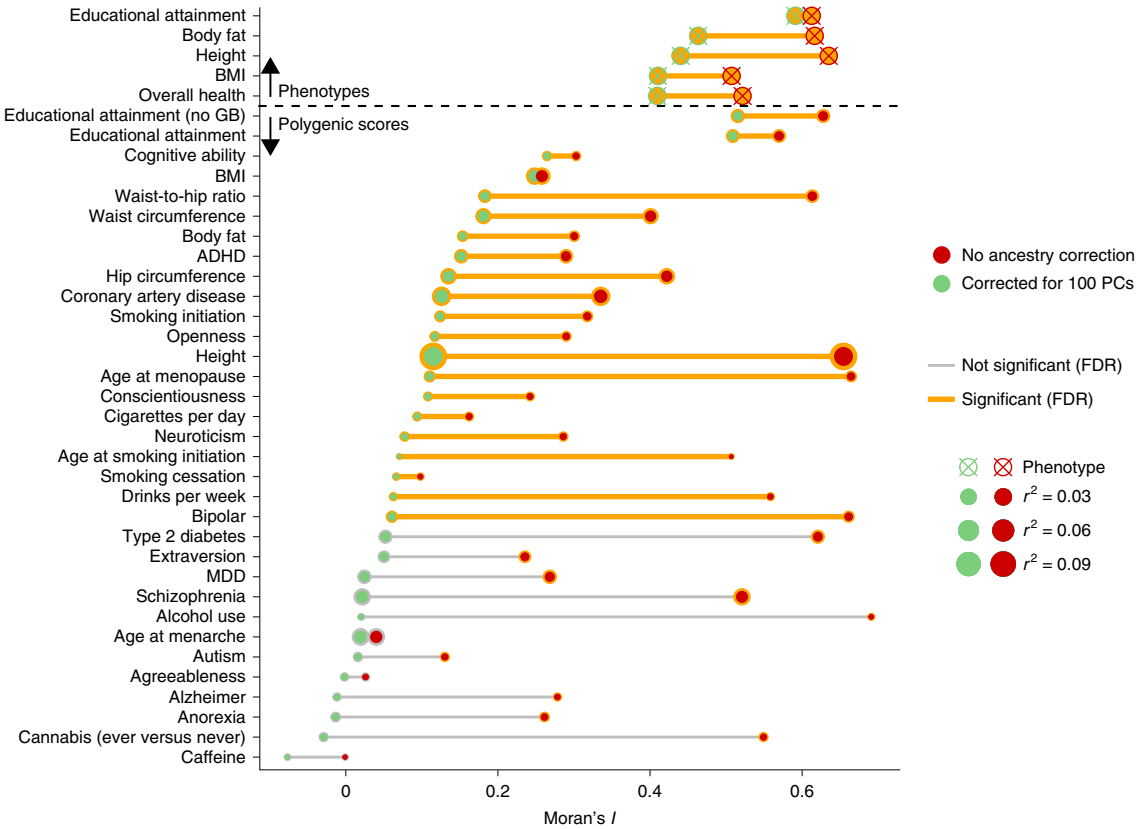


Fig. 2 | Moran's *I* of five phenotypes and 33 SBLUP polygenic scores computed using the average polygenic score per region in 378 local authority regions ($n = 320{,}940$ unrelated individuals). Moran's *I* values of the polygenic scores unadjusted for PCs (red) and adjusted for 100 PCs (green), where orange data points represent a significant FDR-corrected *P* value < 0.05 (corrected for 38 tests). Orange lines between data points represent Moran's *I* values significantly larger than 0, both before and after correcting for 100 PCs. See Supplementary Fig. 4 for the distributions of significant Moran's *I* values from 10,000 permutations that were conducted to obtain the empirical *P* values for the phenotypes and polygenic scores, respectively. ADHD, attention deficit hyperactivity disorder; MDD, major depressive disorder; no GB, excluding British cohorts.

After correcting for ancestry differences, the Townsend index was significantly associated with all five phenotypes and all 21 geographically clustered polygenic scores, with the strongest associations for EA (Supplementary Figs. 7 and 8). All phenotypes and 21 geographically clustered polygenic scores showed significant differences between coal mining areas and the rest of Great Britain, based on both birthplace and current address (Supplementary Fig. 9), with EA showing the strongest differences (FDR-corrected *P* value $< 10^{-200}$). We further compared phenotypes and ancestry-corrected polygenic scores among four groups of unrelated

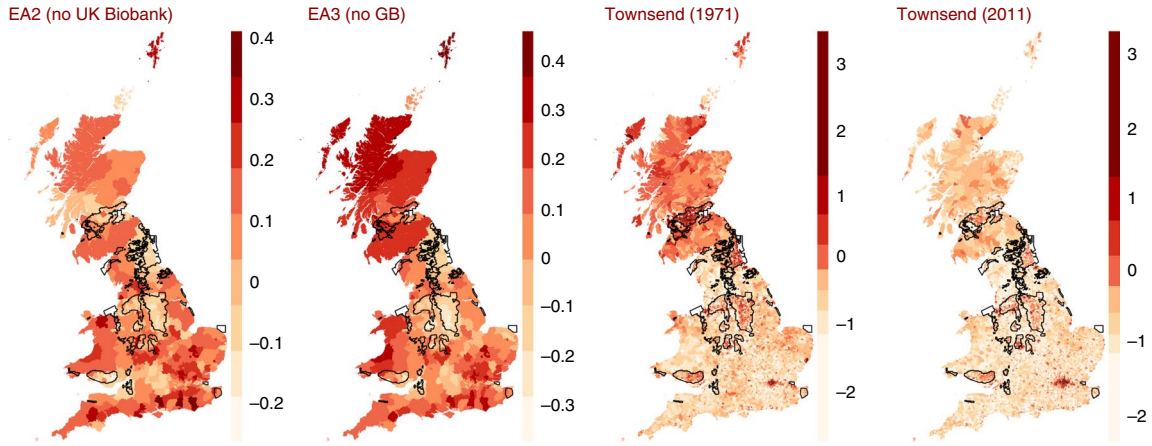


Fig. 3 | Geographic distribution (birthplace) of EA polygenic scores, after regressing out 100 PCs ($n = 320{,}940$ unrelated individuals), and Townsend indices from 1971 and 2011. For EA2 (no UK Biobank), polygenic scores are based on the Okbay et al.[26] EA GWAS, excluding UK Biobank participants ($n = 217{,}569$). For EA3 (no GB), polygenic scores are based on the Lee et al.[26] EA GWAS, excluding all British cohorts ($n = 245{,}621$). The black outlined regions indicate coal mining areas. Maps were adapted from 2011 Census aggregate data (UK Data Service (February 2017 edition). Office for National Statistics; National Records of Scotland; Northern Ireland Statistics and Research Agency (2017); https://doi.org/10.5257/census/aggregate-2011-2). Color bars indicate distribution frequencies, scaled such that each measure has a mean of 0 and s.d of ±1.

individuals: (1) people born in coal mining areas who moved away from coal mining areas ($n = 35{,}024$); (2) people born outside of coal mining areas and still living outside of coal mining areas ($n = 129{,}298$); (3) people born outside of coal mining areas who moved into coal mining areas ($n = 47{,}505$); and (4) people born in coal mining areas who still live in coal mining areas ($n = 111{,}838$). Analyses of variance (ANOVAs) for all phenotypes and 21 polygenic scores showed significant differences between the four groups, with EA (EA3 (ref. [27]); no GB) showing the largest and most significant differences ($F_{3,323,661} = 705.9$; $P < 10^{-200}$; Fig. 4). The largest differences were between people born in coal mining areas who moved away and those who remained in the coal mining areas. The people who moved away had significantly higher EA polygenic scores than all of the other groups combined ($t_{44,000} = 19.7$; $P = 1 \times 10^{-85}$), while those who remained had significantly lower EA polygenic scores than all of the other groups combined ($t_{230,660} = 45.2$; $P < 10^{-200}$). Similar differences were observed for all of the other geographically clustered polygenic scores, with more favourable outcomes for people leaving coal mining regions (except for bipolar disorder, which had the opposite effect, consistent with its positive genetic correlation with EA[26]). We observed that the degree of geographic clustering of polygenic scores of the different traits was significantly correlated with the strength of their associations with Townsend, coal mining areas and migration groups; the strongest correlations were between Moran's *I* values and the *F* statistics of the migration group differences ($r = 0.95$ and $P = 5 \times 10^{-11}$ including the two outlier EA polygenic scores; and $r = 0.78$ and $P = 9 \times 10^{-5}$ excluding the two EA polygenic scores (Supplementary Fig. 10)).

The significant difference between coal mining regions and the rest of the country in the EA polygenic score was present across the entire birth year range of the UK Biobank participants (1936–1970; Fig. 5b). Both coal mining regions and the rest of Great Britain showed a significantly decreasing average EA polygenic score over time, with a steeper decrease in coal mining regions, especially after 1945. The interaction between birth year and coal region was significant ($P = 2 \times 10^{-7}$), with a slope three times as large within coal regions than outside of coal regions. The standardized effect size of birth year outside of coal mining regions was −0.001 (that is, 0.001 s.d. of the polygenic score decrease per birth year; $P = 6.8 \times 10^{-6}$). Within coal mining regions, the effect size was −0.003 ($P < 2 \times 10^{-16}$) across the entire birth year range and −0.004 after 1945 ($P < 2 \times 10^{-16}$). Across the entire country, the effect of birth year was −0.002 ($P < 2 \times 10^{-16}$; Fig. 5a). A decrease in the average polygenic score is consistent with the EA polygenic score being negatively associated with fertility rate[34]. It is also consistent with year-of-birth-associated ascertainment bias, such that the oldest participants in the UK Biobank were more selected on traits associated with EA (for example, health and longevity) than the youngest participants. The steeper decrease within the coal mining regions could be due to higher fertility among individuals with lower EA polygenic scores who were over-represented in those regions and/or the migration of individuals with higher polygenic scores out of coal mining regions (Fig. 5c). The strength of the interaction effect (that is, whether the slope was different within versus outside of coal regions) was strongly correlated with Moran's *I* (including the EA scores: $r = 0.84$; $P = 8.3 \times 10^{-10}$; excluding the EA scores: $r = 0.59$; $P = 5 \times 10^{-4}$; Supplementary Fig. 11), which may indicate that geographic clustering due to SES-related migration played a role in the selection pressure, especially on EA being stronger within than outside of coal regions. Phenotypic EA also showed a consistent significant difference between coal mining regions and the rest of the country over time, but unlike the polygenic scores, phenotypic EA (years of education) consistently increased over time (Fig. 5d–f). This increased significantly faster in the coal mining regions than in the rest of the country, which is the opposite one would expect based on the change of polygenic scores over time (within coal regions, the increase was 0.15 years of education per year, whereas outside of coal regions, the increase was 0.13 years of education per year; *P* value of interaction effect $< 2 \times 10^{-16}$), indicating that environmental improvements may have had a stronger effect on the longitudinal change in educational outcome than the change in genetic values.

To get a better sense of the scale of regional differences in phenotypes, polygenic scores and PCs, and of how these changed due to migration, we computed how much of their variation was explained by regional differences for both birthplace and current address (Extended Data Figs. 1 and 2). Without correcting for PCs, there was much variation in how much the regional differences explained per polygenic score (Extended Data Fig. 1). After removing the systematic ancestry differences that were captured by PCs, the regional differences explained a minimum amount of variation in all polygenic scores except for the EA polygenic scores, which were much

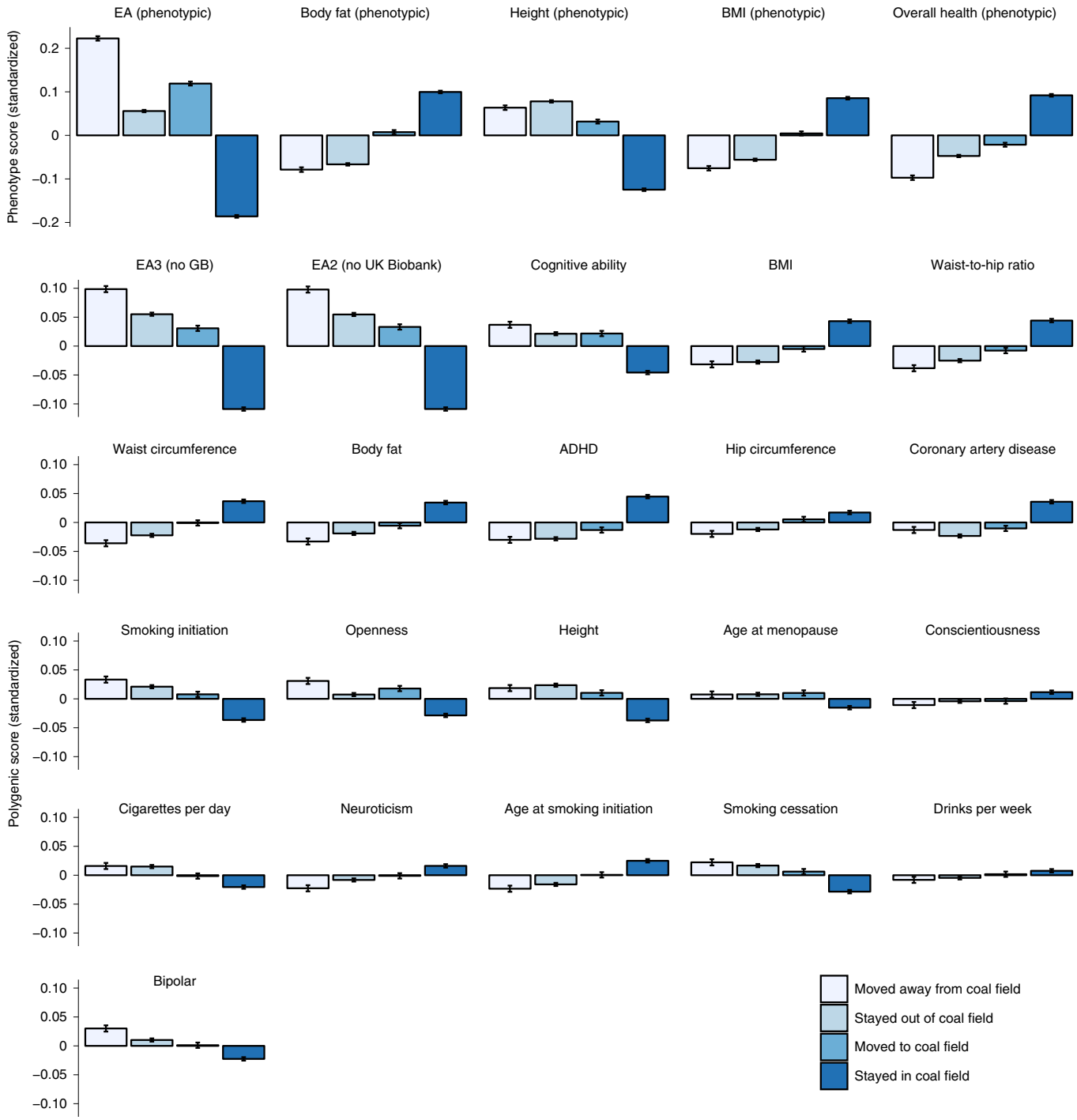


Fig. 4 | Geographically clustered polygenic scores ($n = 16$; ordered by Moran's *I*) for the four migration groups. All polygenic scores are standardized residuals after regressing out 100 ancestry-informative PCs. ANOVA was conducted for each polygenic score to test the presence of group differences, which were all significant, with the least significant FDR-corrected *P* value of 1×10^{-4} for conscientiousness. Bar heights represent average values, with s.e. indicated by the error bars. Migration group *n* values: $n = 35{,}024$ (born in coal field area and moved away); $n = 129{,}298$ (born outside of coal mining area and stayed out); $n = 47{,}505$ (born outside of coal mining area and moved to coal mining area); and $n = 111{,}838$ (born in coal mining area and stayed).

less affected by controlling for PCs. The regional differences were greatest for the EA polygenic scores, with ~0.6–2.6% of individual differences being explained by regional differences, depending on how fine the regional scale was (the finer the scale, the more individual differences explained) and by whether the calculations were based on the birthplace or the current address (Extended Data Fig. 2). For the EA polygenic scores, the regional differences were ~38–72% greater for the current address than for birthplace. The increase in variation explained by regional differences for the ancestry-corrected EA polygenic scores (that is, the difference between birthplace and current address in percentage variance explained by region) was greater than the total variance explained by region for any other ancestry-corrected polygenic score. As would be expected from recent migration events, ancestry showed the opposite effect: comparing birthplace with current address, the variance explained by region on average decreased by 37–73% for the first 30 PCs

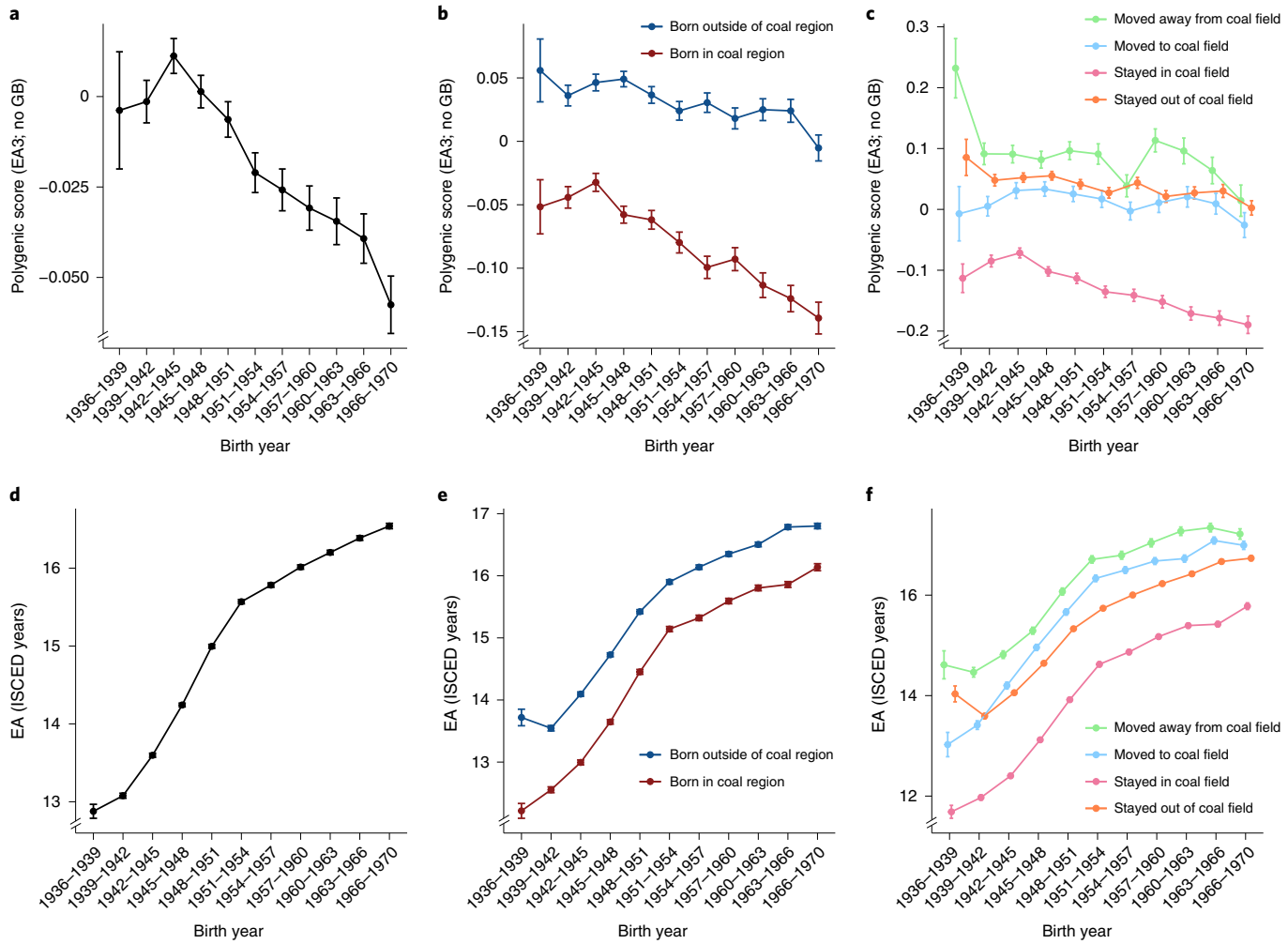


Fig. 5 | Polygenic scores and EA outcomes over time. a–f, Average polygenic scores based on EA3 (Lee et al.[27]), excluding British cohorts (**a**–**c**), and average phenotypic EA outcomes (**d**–**f**) for 11 3-year bins of the entire range of birth years of the UK Biobank participants ($n = 320{,}940$ unrelated individuals), for all unrelated individuals (**a** and **d**), split up into whether the participants were born inside or outside of coal mining regions (**b** and **e**), and for the four migration groups (**c** and **f**). The polygenic scores are standardized residuals after regressing out 100 ancestry-informative PCs. The phenotypic EA outcomes are years of education, as analysed in the EA2[26] and EA3[27] GWASs. ISCE, International Standard Classification of Education.

(Extended Data Fig. 3). A similar decrease was observed for many polygenic scores before correcting for the PCs (Extended Data Fig. 1), indicating that the geographic clustering of those polygenic scores was largely due to them capturing older ancestry differences.

Regional health, economic and cultural outcomes. The geographic clustering of socioeconomic resources and associated genetic variants may coincide with a range of regional economic, behavioural and health outcomes, as well as collective views and attitudes. We tested the association between all 33 ancestry-corrected polygenic scores and 33 publicly available regional measures of economic, health and cultural outcomes that were likely to correlate with SES. The 33 regional measures included measures related to economic outcomes (for example, EA, employment, income and electricity consumption), nutrition and health (for example, obesity and diabetes rates, children's BMI and height, physical exercise, fast food outlets and fruit/vegetable consumption) and major ideologies (political preference and religion). There were many significant associations between polygenic scores and regional health, economic and cultural outcomes, with the EA polygenic scores consistently outperforming the other polygenic scores in the strength of their association (Extended Data Figs. 4–7). If the geographic clustering of polygenic scores is due to SES-related migration, the strength of the association between the polygenic scores and the regional SES-related outcomes should be in line with the geographic clustering, which was largely what we observed. For nearly all of the regional measures, the absolute standardized betas of regression analyses of polygenic scores and regional outcomes were highly correlated with the Moran's *I* values of the polygenic scores (Supplementary Fig. 12).

The significant associations between polygenic scores and regional outcomes must be interpreted with great care as they do not necessarily imply direct causal links between genetic variants and regional outcomes. They do imply that there are regional outcomes that are directionally associated with trait-associated alleles, and these associations could in principle be detected by GWAS. We therefore conducted GWASs on the 33 regional SES, health and cultural outcomes by assigning all participants from the same region the same regional value as a phenotype (from hereon, referred to as regional GWAS (RGWAS)). An RGWAS is expected to detect a genetic signal (that is, SNP-based heritability) if the regional measure is correlated with systematic regional differences in geographically clustered alleles that are associated with a complex trait beyond ancestry. A similar approach has been previously used to successfully capture a genetic signal for household income (household

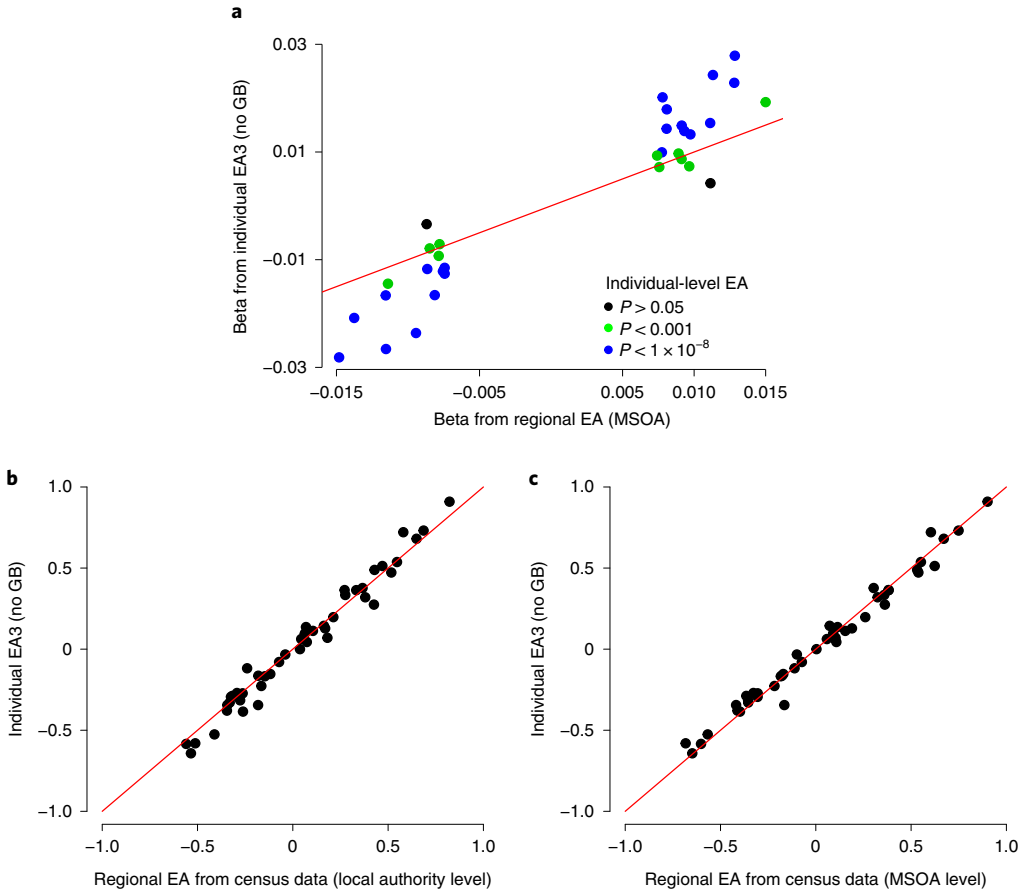


Fig. 6 | Comparisons between the results of the RGWASs on EA from census data and from an individual-level EA GWAS that excluded British participants. a, Effect sizes of 33 independent SNPs ($r^2 < 0.1$ and at least 1 megabase apart) that reached $P < 5 \times 10^{-8}$ in the RGWAS on census EA (MSOA; $n = 416{,}061$) plotted against their effect sizes from the individual-level EA3 GWAS without British cohorts ($n = 245{,}621$). The correlation between the beta values is 0.94 ($P < 2 \times 10^{-16}$). **b**,**c**, Scatterplots of the genetic correlations (r_g) with individual-level GWASs of 45 complex traits. Each dot represents a complex trait. In **b**, the correlation between the r_g values from the RGWAS on census EA at the local authority level ($n = 402{,}552$) and the individual EA3 GWAS without British cohorts ($n = 245{,}621$) is 0.99. In **c**, the correlation between the r_g values from the RGWAS on census EA at the MSOA level ($n = 416{,}061$) and the individual EA3 GWAS without British cohorts ($n = 245{,}621$) is also 0.99. The red lines in each panel represent $y = x$.

phenotype) and social deprivation (postal code phenotype)[35]. An RGWAS would allow us to quantify genetic relationships between regional outcomes and a wider range of complex traits using linkage disequilibrium score regression—a method that computes genetic correlations between GWASs without bias from ancestry differences or sample overlap[36]. Unlike effect estimates from the polygenic scores association analyses, the estimates of genetic correlations are not dependent on the power of the GWAS; however, their precision is. This allows a direct comparison between the strengths of the genetic relationships without bias due to different GWAS sample sizes. RGWASs were run on the ~400,000 UK Biobank participants, corrected for relatedness, age, sex and ancestry (100 PCs). The highest SNP-based heritability estimates were observed for regional measures of EA obtained from census data ($h^2_{SNP} = 3.3$–6.6%, compared with 10% for the individual-level EA3 GWAS; Supplementary Tables 2 and 3), which showed genetic correlations almost identical to the individual-level EA3 GWAS that excluded all British cohorts, and replicated its genome-wide significant SNPs (Fig. 6). Nearly all RGWASs showed high genetic correlations with socioeconomic outcomes such as EA, IQ, income and age at first birth (Extended Data Figs. 8–10). The exception was BMI in 4- to 5-year-olds, with a SNP-based heritability of ~0 (in line with previous epidemiological observations showing BMI in that age range to be less associated with regional SES indicators than at older ages[37] and less influenced by parental EA at the age of 4 years than at older ages[38]). We describe the RGWAS results in more detail in the Supplementary Information section 'RGWAS and genetic correlation results'. The fact that RGWASs on regional measures for traits such as obesity, diabetes, BMI and height showed much higher genetic correlations with individual-level GWASs of EA than with individual-level GWASs of diabetes, BMI or height itself suggests that: (1) RGWASs should not be interpreted as an alternative for traditional individual-level GWASs; and (2) many of the regional differences in health-related outcomes are not likely to be caused by differences in genetic variants that are directly causal for the health outcomes themselves, and are more likely to be due to environmental factors correlated with regional SES and therefore also with the SES-related genes of inhabitants.

Discussion

Understanding the consequences of DNA variation in human populations is of major importance for medical, biological, forensic, behavioural and anthropological research. Since the advent of DNA measurement at a sequence level, studies have shown that the geographic distributions of alleles are not random, and have mapped striking geographic patterns of ancestry[2,3,5,39]. Here, we investigated geographic patterns of genome-wide complex trait variation and show that there are additional levels of geographic clustering beyond

the geographic patterns that reflect older ancestry differences. We show that the geographic clustering of genome-wide trait-associated alleles is consistent with the hypothesis that regional allele frequencies are changed by recent geographic movements of people. We show that the resulting regional genetic patterns are associated with regional socioeconomic, health and cultural outcomes.

To build polygenic scores, we used the effects of genome-wide alleles on complex traits that were estimated in independent datasets that excluded UK Biobank, and in the case of EA, excluded all data from Great Britain. Without controlling for ancestry, almost all traits that we examined showed significant geographic clustering, often resembling the geographic patterns of ancestry differences within Great Britain. This indicates that: (1) the allele frequencies were differentiated between the different ancestries due to genetic drift or natural selection; and/or (2) the GWASs that produced the SNP effect estimates did not sufficiently control for ancestry differences, resulting in SNP effect estimates that were biased towards certain ancestral backgrounds. Recently, it was shown that inflation due to population stratification in polygenic scores may lead to incorrect inference about natural selection when comparing multiple ancestries[40,41]. When we control for ancestry differences, 21 polygenic scores remain significantly clustered by geography. The strongest clustering was observed for EA. Among the rest of the geographically clustered traits were body dimensions, personality dimensions, substance use, and physical and mental health traits. There may be independent influences of those traits on non-random migration, but their clustering is likely to be at least partly driven by their genetic overlap with EA. The increased geographic clustering of EA-associated alleles is consistent with the clustering expected from relatively recent SES-driven migration, as is the simultaneous decrease in geographic clustering of the older geographic patterns of ancestry (Extended Data Figs. 1–3). A similar process was observed in the Dutch population, where the ancestry of offspring of higher-educated parents showed lower correlations with geography than the offspring of less mobile parents with lower education levels[42]. A recent study in a much smaller US sample also found EA to geographically cluster more than the other phenotypes they analysed, but for the polygenic scores they studied, height and smoking showed stronger geographic clustering in their sample than the EA polygenic score[43]. It is not clear whether this discrepancy was because of greater sampling error due to a much smaller sample size (n = ~8,600) and lower geographic resolution (US state level), or whether the situation is different in the United States compared with Great Britain.

The degree of geographic clustering of the polygenic scores of different traits was largely in line with the strength of the relationship of the traits with regional socioeconomic, health and cultural outcomes (Supplementary Fig. 12) and with migration out of economically deprived regions (Supplementary Fig. 10). People are reportedly more likely to migrate to improve their skills or employment prospects than for other area characteristics[9]. Many industrialized countries showed these types of migration flows during the late nineteenth and early twentieth century, when labourers and small farmers left the countryside to work in industrial jobs that were often highly clustered in geographic space (for example, coal mining areas)[12]. After deindustrialization, the dense, durable and affordable working-class houses and public transportation networks from the industrial revolution remained in these neighbourhoods and continued to attract poorer immigrants[12]. Our results show that people with a genetic predisposition to higher cognitive abilities are leaving these regions, probably attracted by better educational or occupational opportunities in other regions. In fact, the people who were born in coal mining areas and migrated to better neighbourhoods have higher average EA polygenic scores than people born outside of these regions. These demographic processes may influence GWAS signals as well, where (EA-related) alleles that increase the chances of living in the unhealthy circumstances of lower SES neighbourhoods (for example, in the presence of more fast food outlets)[44] may become part of the signal of an individual-level GWAS for health-related traits such as BMI or body fat.

Selective migration has led to geographic clustering of social and economic needs, which can coincide with collective attitudes towards how communities should be organized and governed. We successfully captured signals in our RGWASs on regional religiousness and regional political attitudes, both of which have been shown to be partly heritable on an individual level[45–50] and to cluster geographically[11,12]. From a regional genetic perspective, the election outcomes can be roughly divided into electorates of lower SES (Labour Party, UKIP, 'leave' votes for Brexit and non-voters) and higher SES (Green Party, Liberal Democrats and Conservatives), in line with previous reports on the geographic clustering of political preference in Great Britain[12,51]. Our findings suggest that previously reported heritability estimates of these traits on an individual level could possibly contain genetic effects on traits, such as EA, that influence which socioeconomic strata and geographic regions people end up living in. Regional religiousness shows higher genetic correlations with personality (openness and conscientiousness) and less with SES and health traits than political preferences do, which implies additional dimensions of geographic clustering beyond high versus low SES.

There are several limitations to this study that have to be kept in mind when interpreting our results. First, there is an ascertainment bias in the UK Biobank sample, such that participants are healthier and more educated than the general population[52]. Our comparison with EA from census data suggests that some of the regional differences we report, such as between coal regions and the rest of the country (Supplementary Information), might be larger in the general population, but it is not clear to what extent this ascertainment bias has affected the composition of the four migration groups that we compared. Second, while it is unlikely that population stratification was solely responsible for our observations (Supplementary Information), we acknowledge that the possibility of residual population stratification in our signal is difficult to rule out completely. Replication in within-family analyses could help address this possibility. Third, SNP effect sizes on EA have been shown to have higher estimates in population-based samples compared with within-family designs[27,53]. This could be due to indirect genetic effects, gene–environment correlations and/or assortative mating, which have been shown to exist for EA in family studies[27,53,54]. It is likely that this plays a role in the regional associations that we find, but it is not clear to what extent. Finally, we used the RGWAS approach to help shed light on the relationship between regional differences in genome-wide complex trait variation and publicly available regional measures. It is not yet clear how its results are affected by ascertainment bias on both sides (in UK Biobank and the different sources of the publicly available regional measures), or how, or on what scale, the heritability and effect size estimates of the individual SNPs should be interpreted. However, the scale is not expected to affect the genetic correlation estimates if the scale is consistent across SNPs (which is confirmed in Fig. 6b,c). Our results also show that RGWASs should not be considered an alternative for individual-level GWASs, even though this seems to work exceptionally well for EA. This was probably due to strong geographic clustering of EA-associated alleles, which seems to be captured well by census data on EA.

Our findings possibly reflect genetic consequences of social stratification—a key characteristic of human civilizations[55] whereby society groups its people into strata based on SES. SES is generally based on occupation, income and EA, which are influenced by many environmental and genetic factors, and are associated with a wide range of physical and mental health outcomes. SES is not distributed randomly across geographic space, which can lead to

geographic clustering of alleles that are associated with SES-related traits such as EA. EA is known for its high levels of assortative mating[56,57], which may be further induced by geographic clustering. This may exacerbate social inequalities across generations. It is possible that the combination of recent increases in social mobility and an improved educational system accelerate this separation of higher and lower genetic predisposition for traits related to cognition, SES and health. Even though the genetic effects we find do not explain all of the observed regional differences, researchers and social policymakers may want to keep these genetic effects in mind. For example, the significant genetic correlations between EA and health-related traits may decrease in the presence of stronger social safety nets geared towards making inhabitants of lower SES regions live more economically prosperous and healthier lives. Increasing the quality of life in lower SES regions may also help decrease migration out of these regions by people with genetic predispositions for higher SES outcomes, and thereby possibly result in a less geographically stratified society.

Methods

Participants. The participants of this study were sourced from UK Biobank[22,58], which has received ethical approval from the National Health Service North West Centre for Research Ethics Committee (reference: 11/NW/0382). A total of 502,536 participants (273,402 females and 229,134 males) aged between 37 and 73 years were recruited in Great Britain between 2006 and 2010. The participants were recruited across 22 assessment centres throughout Great Britain to cover a variety of different settings providing socioeconomic and ethnic heterogeneity and urban–rural mix. They underwent a wide range of cognitive, health and lifestyle assessments, provided blood, urine and saliva samples, and will have their health followed longitudinally.

Genotypes and quality control. A total of 488,377 UK Biobank participants had their genome-wide single SNPs genotyped on either the UK BiLEVE array ($n = 49{,}950$) or the UK Biobank Axiom Array ($n = 438{,}423$). The genotypes were imputed using the Haplotype Reference Consortium panel as a reference set (pre-imputation quality control and imputation are described in more detail by Bycroft et al.[58]). To create polygenic scores, we extracted a set of 1,312,100 autosomal HapMap 3 SNPs with a minor allele count of >5, an info score of >0.3, Hardy–Weinberg equilibrium (HWE) P value (P_{HWE}) $< 10^{-6}$ and missingness < 0.05. For the GWAS, we used 5.8 million SNPs that survived quality control and have a minor allele frequency of >0.01.

Ancestry and PCA. To capture British ancestry, we first excluded individuals with non-European ancestry. Ancestry was determined using PCA in GCTA[59]. The UK Biobank dataset was projected onto the first two PCs from the 2,504 participants of the 1000 Genomes Project[60], using HapMap 3 SNPs with a minor allele frequency of >0.01 in both datasets. Next, participants from UK Biobank were assigned to one of five super-populations from the 1000 Genomes project: European, African, East Asian, South Asian or admixed. Assignments for European, African, East Asian and South Asian ancestries were based on a >0.9 posterior probability of belonging to the 1000 Genomes reference cluster, with the remaining participants classified as admixed. Posterior probabilities were calculated under a bivariate Gaussian distribution where this approach generalizes the k-means method to take account of the shape of the reference cluster. We used a uniform prior and calculated the vectors of means and 2×2 variance–covariance matrices for each super-population. A total of 456,426 subjects were identified to be of European ancestry.

A PCA was then conducted on individuals of European ancestry to capture ancestry differences within the British population. To capture ancestry differences in homogenous populations, genotypes should be pruned for linkage disequilibrium, and long-range linkage disequilibrium regions should be removed[3]. The linkage disequilibrium-pruned ($r^2 < 0.1$) UK Biobank dataset without long-range linkage disequilibrium regions consisted of 137,102 genotyped SNPs. The PCA to construct British ancestry-informative PCs was conducted on this SNP set for unrelated individuals using flashPCA version 2 (ref. [61]). PC SNP loadings were used to project the complete set of European individuals onto the PCs.

Polygenic scores. Polygenic scores—the genome-wide sum of alleles weighted by their estimated effect sizes—were computed for 33 traits. The effect size estimates were obtained from GWASs that were chosen to not have included the UK Biobank dataset, to avoid overestimation of the genetic predisposition of a trait[24]. The polygenic scores were computed using the summary-data based best linear unbiased prediction (SBLUP) approach[56], which maximizes the predictive power by creating scores with best linear unbiased predictor properties that account for linkage disequilibrium between SNPs. As a reference sample for the linkage disequilibrium, we used a random sample of 10,000 unrelated individuals from UK Biobank, imputed using the Haplotype Reference Consortium panel[62]. The traits included psychiatric disorders, substance use, anthropomorphic traits, personality dimensions, EA, reproduction, cardiovascular disease and type 2 diabetes. Supplementary Table 1 lists the 33 traits and the GWASs from which we obtained the genome-wide effect sizes.

To further investigate the robustness of our results, we also created polygenic scores using only independent SNPs that were associated with the trait with a P value < 0.05. The SNPs were clumped using PLINK[63], using an r^2 threshold of 0.1 and a window of 1 megabase as the physical distance threshold for clumping.

To examine the geographic clustering of polygenic scores beyond the clustering of ancestry, we created additional sets of polygenic scores that had the first 100 British ancestry-informative PCs regressed out.

Geographic clustering of ancestry, phenotypes and genome-wide complex trait variation. The geographic clustering of ancestry, phenotypes and genome-wide complex trait variation was investigated by testing whether the spatial autocorrelation (Moran's I) is significantly greater than zero for ancestry-informative PCs, polygenic scores and the residuals of polygenic scores after regressing out 100 ancestry-informative PCs. The spatial autocorrelation (Moran's I) is the correlation in a measure among nearby locations in space, and its values range between −1 (dispersed) and 1 (spatially clustered), where to 0 = spatially random[25]. Moran's Is were computed using the average PCs or polygenic scores per region based on the birthplace of the subjects (378 regions; Fig. 1), whereby the regions were defined according to the local authorities division, as provided by the UK Data Service InFuse database[64]. The empirical P values of Moran's I statistics were derived with 10,000 permutations in which the average PCs or polygenic scores were permuted across regions (Supplementary Fig. 4). To determine significance, we used an alpha level of 0.05 for FDR-corrected empirical P values, and all tests were two tailed. Moran's I was computed using the mc.moran function from the spdep package in R. Spatial weights were determined using binary weights (style = "B" in the nb2listw function from spdep).

To investigate differences between coal mining regions and the rest of the country, and the role of related migration flows herein, we used t-tests and ANOVAs (see the section 'SES-related migration'), whereby statistical significance was determined using an alpha level of 0.05 for FDR-corrected P values, and all tests were two tailed. The data distribution was assumed to be normal, but this was not formally tested due to the large sample size[65].

Statistical analyses on regional outcomes. We investigated the genetic relationship between complex traits and publicly available regional outcomes, whereby all subjects from the same regions had the same regional phenotypic value assigned. We did this through association tests for the polygenic scores and genetic correlations for the RGWASs. For both, the significance was determined using an alpha level of 0.05 for FDR-corrected P values, and all tests were two tailed. The data distribution was assumed to be normal, but this was not formally tested due to the large sample size[65] (see Supplementary Fig. 13 for the distributions of the regional phenotypes). The associations with polygenic scores (Extended Data Figs. 4–7) were done on unrelated individuals and were computed with robust linear models using M-estimators. For the RGWASs, we ran linear mixed model GWASs with BOLT-LMM[66] on all participants with European ancestry, which controls for cryptic relatedness and population stratification by including a genetic relatedness matrix in the model[67]. Sex and age were included as covariates, as were the first 100 PCs as an additional control for population stratification. The results revealed a considerable inflation of test statistics that was not due to polygenic effects (this was captured by the linkage disequilibrium score intercepts (LDSC)[68] shown in Supplementary Table 2). This was probably due to the fact that participants who share regional environmental influences (because they come from the same region) were all assigned the same phenotypic value. We controlled for this inflation with an LDSC-based genomic control (GC)[68] (that is, we adjusted the s.e. of the estimated effect sizes as follows: $\text{s.e.}_{GC} = \sqrt{\text{LDSC intercept}} \times \text{s.e.}^2$ (Supplementary Table 2)).

The regional phenotypes were obtained from the following public resources:

- The borders of a total of 208 coal mining regions were obtained from the Coal Authority (https://www.gov.uk/guidance/using-coal-mining-information).
- The regional EA for 342 local districts and 7,195 middle layer super output areas (MSOA) was measured using the 2011 estimates of the highest qualification of residents of England who were >16 years old (five levels: level 1 qualifications; level 2 qualifications; apprenticeship; level 3 qualifications; and level 4 qualifications), obtained from the Nomis database of the Office of National Statistics (https://www.nomisweb.co.uk/).
- Regional measures on employment, income, health deprivation and disability, crime, barriers to housing and services, and living environment were subdomains of the English index of multiple deprivation in 2015 and were available at the lower layer super output area level, consisting of 32,844 areas in England. The raw data were downloaded from https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015.
- Children's height and BMI were measured as part of the National Child Measurement Programme in England at the local authority level. We analysed

the measurements from the school year 2015–2016 in the reception class (4–5 years old) and school year 6 (10–11 years old). The regional averages of height and BMI were calculated from the raw data, after adjusting the measures for age and sex. The raw data were downloaded from https://digital.nhs.uk/data-and-information/publications/statistical/national-child-measurement-programme/2015-16-school-year.

- The obesity, diabetes, physical activity, fruit consumption, vegetable consumption and suicide measures were obtained from the Public Health Outcomes Framework, Public Health England (https://fingertips.phe.org.uk/profile/public-health-outcomes-framework/data#page/0/gid/1000049/pat/6/par/E12000006/ati/101/are/E07000008). The obesity measure was defined as the percentage of adults (≥18 years) per local authority region who were classified as overweight or obese in 2015. The diabetes measure was defined as the estimated diagnosis rate for people with diabetes aged ≥17 years in 2015. The fruit/vegetable consumption measures were defined as the average number of portions of fruit/vegetables consumed daily by adults in 2015. The suicide measure was defined as the age-standardized mortality rate from suicide and injury of undetermined intent per 100,000 inhabitants in 2015.
- The fast food outlets density by local authority was obtained from Public Health England (https://www.gov.uk/government/publications/fast-food-outlets-density-by-local-authority-in-england) and was defined as the number of fast food outlets per 100,000 inhabitants in 2017.
- The electricity consumption in 2010 by lower layer super output area level was obtained from the Department for Business, Energy and Industrial Strategy (https://www.gov.uk/government/statistics/lower-and-middle-super-output-areas-electricity-consumption). Two measures of electricity consumption were defined as the average ordinary domestic consumption (kWh) and the average Economy 7 consumption (kWh).
- The proportions of religious versus non-religious inhabitants in 2011 were obtained for 7,195 MSOA regions from the Nomis database of the Office of National Statistics (https://www.nomisweb.co.uk/).
- The 2016 Brexit referendum results were obtained for 405 local authority districts from The Electoral Commission (https://www.electoralcommission.org.uk/find-information-by-subject/elections-and-referendums/past-elections-and-referendums/eu-referendum/electorate-and-count-information).
- The 1970 general election outcomes were obtained for 630 constituencies from Political Science Resources (http://www.politicsresources.net/area/uk/ge70/ge70index.htm). For analyses of 1970 election results, we used the birthplace of the participants instead of the current address to assign the phenotypic values.
- The 2015 general election outcomes were obtained for 633 constituencies from http://www.data.parliament.uk/dataset/general-election-2015.

All political parties were included from the 1970 and 2015 elections that had a median proportion of votes > 0.

Linkage disequilibrium score regression. Genetic correlations were computed using linkage disequilibrium score regression[36] (Extended Data Figs. 8–10)[36]. The genetic correlation between traits is based on the estimated slope from the regression of the product of *z* scores from two GWASs on the linkage disequilibrium score, and represents the genetic covariation between two traits based on all of the polygenic effects captured by the included SNPs. The genome-wide linkage disequilibrium information used by these methods was based on European populations from the HapMap 3 reference panel[36,68]. All linkage disequilibrium score regression analyses included the 1,290,028 million genome-wide HapMap SNPs used in the original linkage disequilibrium score regression studies[36,68].

Computing genetic correlations with linkage disequilibrium score regression is robust to sample overlap, so we included summary statistics from GWASs that also included UK Biobank (denoted with a blue star in Extended Data Figs. 8–10). For some traits, the results are displayed for summary statistics without UK Biobank, even if the GWASs from the original studies included UK Biobank participants. This was the case for major depressive disorder[69] and EA[26], for which we used the same summary statistics that we used for the polygenic scores; namely, from the GWASs that were re-run excluding UK Biobank. However, the genetic correlations for major depressive disorder and EA obtained with the summary statistics that did include UK Biobank were almost identical.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

This research was conducted using data from the UK Biobank resource (application number 12514) and dbGaP (accession number: phs000674). UK Biobank data can be accessed on request once a research project has been submitted and approved by the UK Biobank committee. dbGaP data can also be accessed on request once a research project has been submitted and approved by dbGaP. The regional measures that have been analysed are publicly available and can be downloaded using the links provided in the Methods.

Code availability

Custom R code used for statistical analyses (for example, the computation of Moran's *I*) is available from the corresponding authors on request.

Received: 30 October 2018; Accepted: 18 September 2019;
Published online: 21 October 2019

References

1. Tobler, W. R. A computer movie simulating urban growth in the Detroit region. *Econ. Geog.* **46**, 234–240 (1970).
2. Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
3. Abdellaoui, A. et al. Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet.* **21**, 1277–1285 (2013).
4. Kerminen, S. et al. Fine-scale genetic structure in Finland. *G3 (Bethesda)* **7**, 3459–3468 (2017).
5. Leslie, S. et al. The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).
6. Zhang, G., Muglia, L. J., Chakraborty, R., Akey, J. M. & Williams, S. M. Signatures of natural selection on genetic variants affecting complex human traits. *Appl. Transl. Genom.* **2**, 78–94 (2013).
7. Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS Genet.* **10**, e1004412 (2014).
8. Turkheimer, E. Three laws of behavior genetics and what they mean. *Curr. Dir. Psychol. Sci.* **9**, 160–164 (2000).
9. Coulter, R. & Scott, J. What motivates residential mobility? Re-examining self-reported reasons for desiring and making residential moves. *Popul. Space Place* **21**, 354–371 (2015).
10. Long, J. Rural–urban migration and socioeconomic mobility in Victorian Britain. *J. Econ. Hist.* **65**, 1–35 (2005).
11. Park, C. *Sacred Worlds: An Introduction to Geography and Religion* (Routledge, 2002).
12. Rodden, J. The geographic distribution of political preferences. *Annu. Rev. Polit. Sci.* **13**, 321–340 (2010).
13. Boyle, P. Population geography: migration and inequalities in mortality and morbidity. *Prog. Hum. Geog.* **28**, 767–776 (2004).
14. Lewis, G. & Booth, M. Regional differences in mental health in Great Britain. *J. Epidemiol. Community Health* **46**, 608–611 (1992).
15. Tyrrell, J. et al. Height, body mass index, and socioeconomic status: Mendelian randomisation study in UK Biobank. *Br. Med. J.* **352**, i582 (2016).
16. Marmot, M. The health gap: the challenge of an unequal world. *Lancet* **386**, 2442–2444 (2015).
17. Beard, E. et al. Healthier central England or North–South divide? Analysis of national survey data on smoking and high-risk drinking. *BMJ Open* **7**, e014210 (2017).
18. Brimblecombe, N., Dorling, D. & Shaw, M. Migration and geographical inequalities in health in Britain. *Soc. Sci. Med.* **50**, 861–878 (2000).
19. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
20. Richards, J. B. & Evans, D. M. Back to school to protect against coronary heart disease? *Br. Med. J.* https://www.bmj.com/content/358/bmj.j3849 (2017).
21. Verweij, K. J., Mosing, M. A., Zietsch, B. P. & Medland, S. E. in *Statistical Human Genetics* 151–170 (Springer, 2012).
22. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
23. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
24. Wray, N. R. et al. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).
25. Moran, P. A. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).
26. Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
27. 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).
28. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18**, 117–127 (2017).
29. Haworth, S. et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, 333 (2019).
30. Niedomysl, T. How migration motives change over migration distance: evidence on variation across socio-economic and demographic groups. *Reg. Stud.* **45**, 843–855 (2011).

31. Foden, M., Fothergill, S. & Gore, T. *The State of the Coalfields: Economic and Social Conditions in the Former Mining Communities of England, Scotland and Wales* (Centre for Regional Economic and Social Research, Sheffield Hallam Univ., 2014).
32. Beatty, C., Fothergill, S. & Powell, R. Twenty years on: has the economy of the UK coalfields recovered? *Environ. Plan. A* **39**, 1654–1675 (2007).
33. Townsend, P., Phillimore, P. & Beattie, A. *Health and Deprivation: Inequality and the North* (Routledge, 1988).
34. Kong, A. et al. Selection against variants in the genome associated with educational attainment. *Proc. Natl Acad. Sci. USA* **114**, E727–E732 (2017).
35. Hill, W. D. et al. Molecular genetic contributions to social deprivation and household income in UK Biobank. *Curr. Biol.* **26**, 3083–3089 (2016).
36. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
37. Cetateanu, A. & Jones, A. Understanding the relationship between food environments, deprivation and childhood overweight and obesity: evidence from a cross sectional England-wide study. *Health Place* **27**, 68–76 (2014).
38. Silventoinen, K. et al. Parental education and genetics of BMI from infancy to old age: a pooled analysis of 29 twin cohorts. *Obesity* **27**, 855–865 (2019).
39. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786–792 (1978).
40. Berg, J. J. et al. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725 (2019).
41. Sohail, M. et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* **8**, e39702 (2019).
42. Abdellaoui, A. et al. Educational attainment influences levels of homozygosity through migration and assortative mating. *PLoS One* **10**, e0118935 (2015).
43. Domingue, B. W., Rehkopf, D. H., Conley, D. & Boardman, J. D. Geographic clustering of polygenic scores at different stages of the life course. *RSF* **4**, 137–149 (2018).
44. Cummins, S. C., McKay, L. & MacIntyre, S. McDonald's restaurants and neighborhood deprivation in Scotland and England. *Am. J. Prev. Med.* **29**, 308–310 (2005).
45. Alford, J. R., Funk, C. L. & Hibbing, J. R. Are political orientations genetically transmitted? *Am. Polit. Sci. Rev.* **99**, 153–167 (2005).
46. Benjamin, D. J. et al. The genetic architecture of economic and political preferences. *Proc. Natl Acad. Sci. USA* **109**, 8026–8031 (2012).
47. Hatemi, P. K. & McDermott, R. The genetics of politics: discovery, challenges, and progress. *Trends Genet.* **28**, 525–533 (2012).
48. Hatemi, P. K., Medland, S. E., Morley, K. I., Heath, A. C. & Martin, N. G. The genetics of voting: an Australian twin study. *Behav. Genet.* **37**, 435–448 (2007).
49. Smith, K. et al. Biology, ideology, and epistemology: how do we know political attitudes are inherited and why should we care? *Am. J. Polit. Sci.* **56**, 17–33 (2012).
50. Koenig, L. B., McGue, M., Krueger, R. F. & Bouchard, T. J. Genetic and environmental influences on religiousness: findings for retrospective and current religiousness ratings. *J. Pers.* **73**, 471–488 (2005).
51. Alabrese, E., Becker, S. O., Fetzer, T. & Novy, D. Who voted for Brexit? Individual and regional data combined. *Eur. J. Polit. Econ.* **56**, 132–150 (2019).
52. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
53. Selzam, S. et al. Comparing within-and between-family polygenic score prediction. *Am. J. Hum. Genet.* **105**, 351–363 (2019).
54. Kong, A. et al. The nature of nurture: effects of parental genotypes. *Science* **359**, 424–428 (2018).
55. Llobera, J. R. *An Invitation to Anthropology: the Structure, Evolution and Cultural Identity of Human Societies* (Berghahn Books, 2003).
56. Robinson, M. R. et al. Genetic evidence of assortative mating in humans. *Nat. Hum. Behav.* **1**, 0016 (2017).
57. Hugh-Jones, D., Verweij, K. J., Pourcain, B. S. & Abdellaoui, A. Assortative mating on educational attainment leads to genetic spousal resemblance for polygenic scores. *Intelligence* **59**, 103–108 (2016).
58. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
59. 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).
60. 1000 Genomes Project Consortium.A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
61. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
62. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
63. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
64. Office for National Statistics; National Records of Scotland; Northern Ireland Statistics and Research Agency. 2011 Census aggregate data. *UK Data Service* https://doi.org/10.5257/census/aggregate-2011-2 (Edition: February 2017).
65. Altman, D. G. & Bland, J. M. Statistics notes: the normal distribution. *Br. Med. J.* **310**, 298 (1995).
66. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
67. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
68. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
69. Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).

Acknowledgements

This research was supported by the Australian National Health and Medical Research Council (1107258, 1078901, 1078037, 1056929, 1048853 and 1113400) and the Sylvia and Charles Viertel Charitable Foundation (Senior Medical Research Fellowship). A.A. and K.J.H.V. are supported by the Foundation Volksbond Rotterdam. A.A. and M.G.N. are supported by ZonMw grants 849200011 and 531003014 from The Netherlands Organisation for Health Research and Development. B.P.Z. received funding from the Australian Research Council (FT160100298). The research was conducted using data from the UK Biobank Resource (application number: 12514) and dbGaP (accession number: phs000674). The Genetic Epidemiology Research on Adult Health and Aging study was supported by grant RC2 AG036607 from the National Institutes of Health, as well as grants from the Robert Wood Johnson Foundation, Ellison Medical Foundation, Wayne and Gladys Valley Foundation and Kaiser Permanente. The authors thank the Kaiser Permanente Medical Care Plan, Northern California Region members who participated in the Kaiser Permanente Research Program on Genes, Environment and Health. This study was conducted using UK Biobank resources under application number 12514. UK Biobank was established by the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government and Northwest Regional Development Agency. It also received funding from the Welsh Assembly Government, British Heart Foundation and Diabetes UK. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

A.A., D.H.-J. and P.M.V. conceived and designed the study. A.A., D.H.-J., L.Y. and K.E.K. analysed the data. A.A. wrote the manuscript and produced the figures. D.H.-J., L.Y., K.E.K., M.G.N., L.V., Y.H., B.P.Z., T.M.F., N.R.W., J.Y., K.J.H.V. and P.M.V. provided significant feedback on the analyses and the manuscript. P.M.V. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

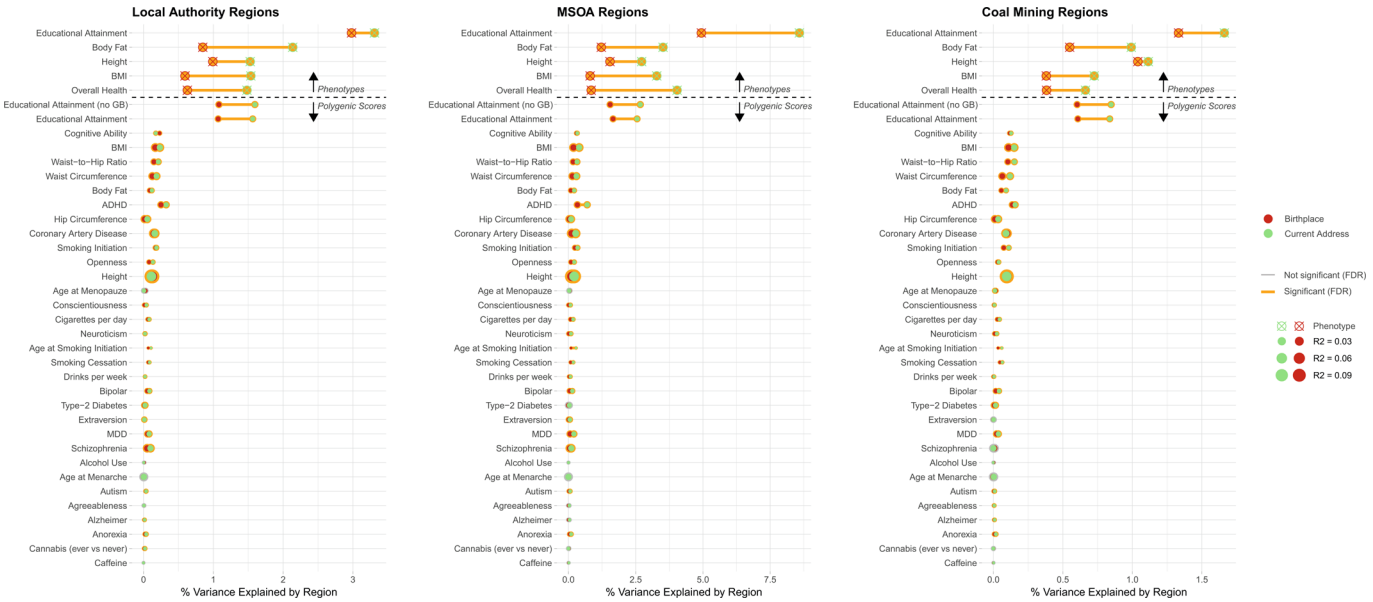
Extended data is available for this paper at https://doi.org/10.1038/s41562-019-0757-5.

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0757-5.

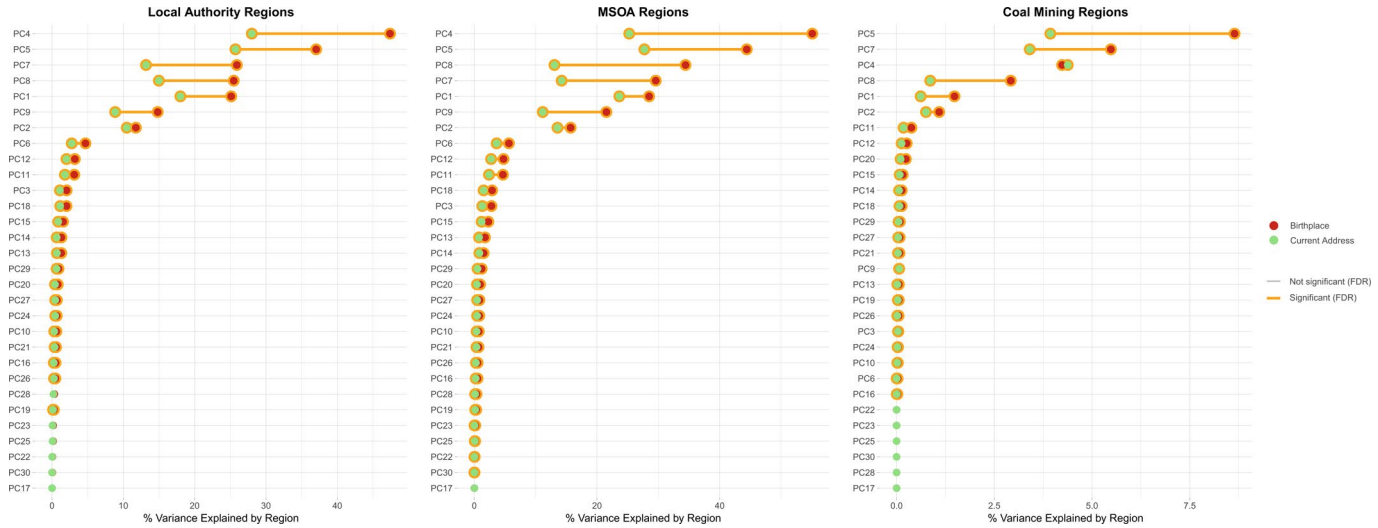
Correspondence and requests for materials should be addressed to A.A. or P.M.V.

Peer review information Primary Handling Editor: Stavroula Kousta.

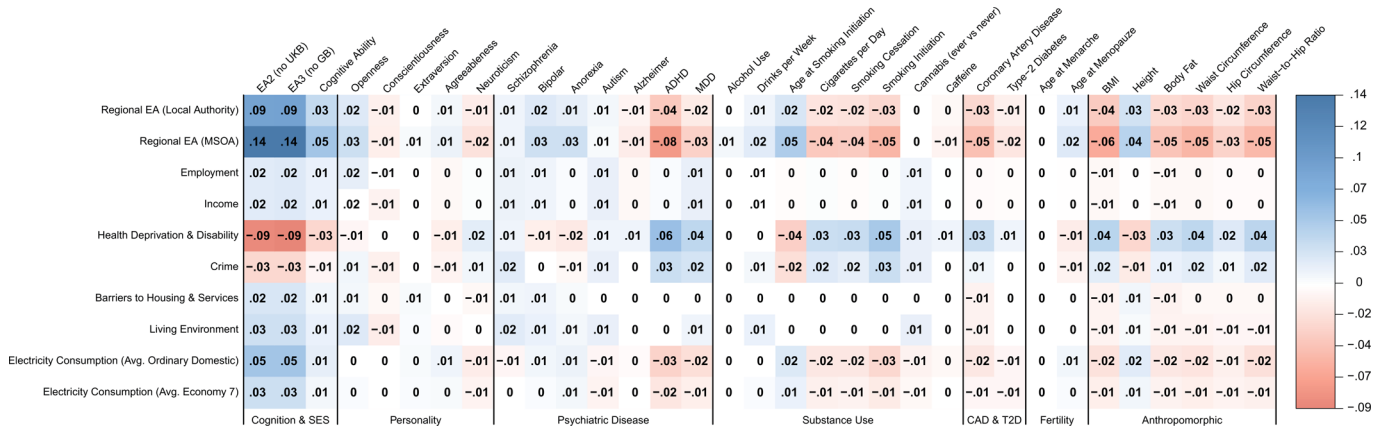
Extended Data Fig. 1 | Variation explained by regional differences of uncorrected polygenic scores. Linear mixed model results, with phenotype or polygenic score (without regressing out 100 PCs) as a dependent variable and region as random effect (N = 320,940 unrelated individuals). Left: Local Authorities (~380 regions); Middle: MSOA (~5,300 regions), Right: Coal mining Regions (fitted as a binary variable). Red: Birth Place; Green: Current Address; Yellow = significant after FDR correction.



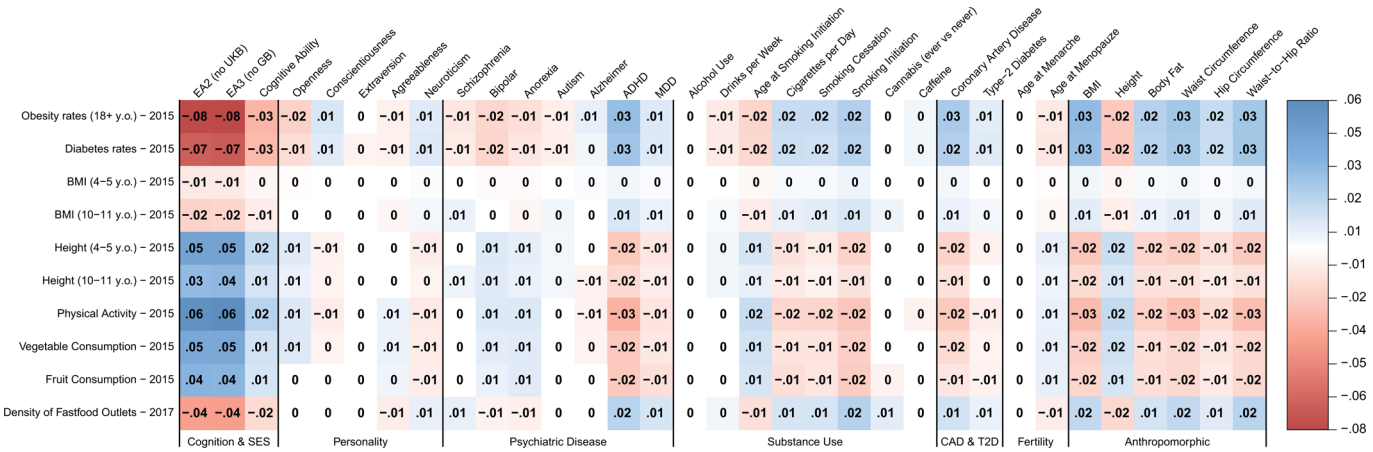
Extended Data Fig. 2 | Variation explained by regional differences of ancestry-corrected polygenic scores. Linear mixed model results, with phenotype or polygenic score (after regressing out 100 PCs) as a dependent variable and region as random effect (N = 320,940 unrelated individuals). Left: Local Authorities (~380 regions); Middle: MSOA (~5,300 regions), Right: Coal mining Regions (fitted as a binary variable). Red: Birth Place; Green: Current Address; Yellow = significant after FDR correction.



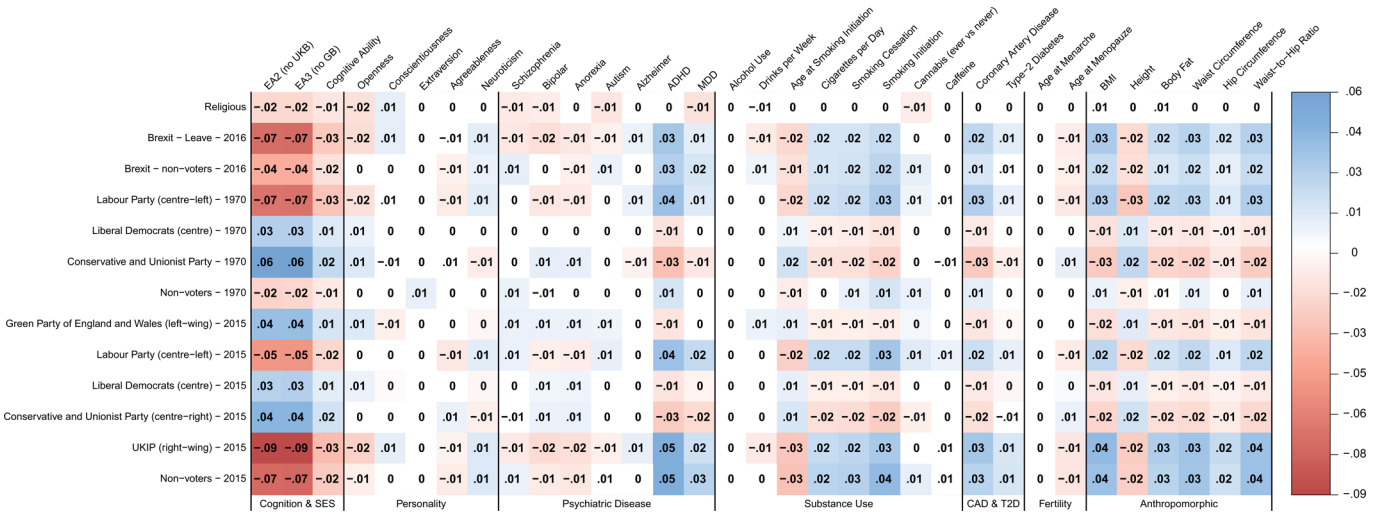
Extended Data Fig. 3 | Variation explained by regional differences of ancestry-informative PCs. Linear mixed model results, with PCs as a dependent variable and region as random effect (N = 320,940 unrelated individuals). Left: Local Authorities (~380 regions); Middle: MSOA (~5,300 regions), Right: Coal mining Regions (fitted as a binary variable). Red: Birth Place; Green: Current Address; Yellow = significant after FDR correction.



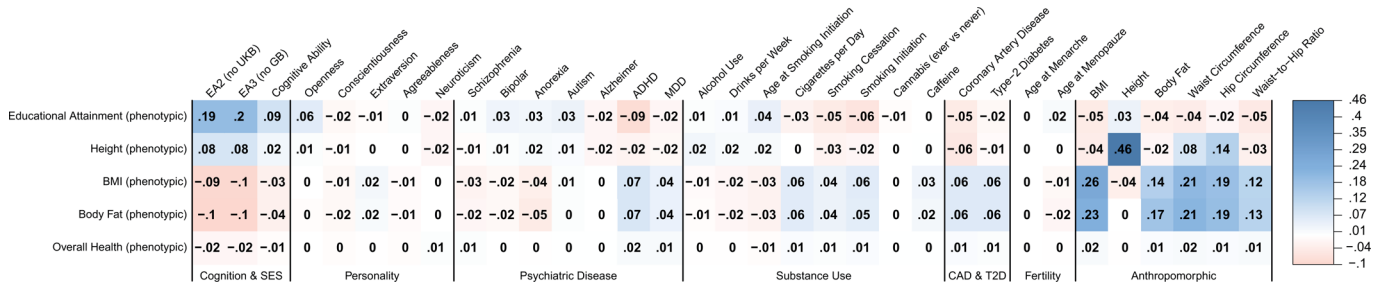
Extended Data Fig. 4 | Associations between polygenic scores and regional measures of socio-economic outcomes. The standardized effect size estimates of robust linear regressions of polygenic scores on regional measures of socio-economic outcomes in unrelated UK Biobank participants of European descent (N ~320k). The polygenic scores are all standardized residuals after regressing out 100 PCs. Every individual was given the value of their region. Significant effects are colored, whereby the significance threshold is based on FDR correction across all tests shown in all four panels. All SEs were ≤ .002.



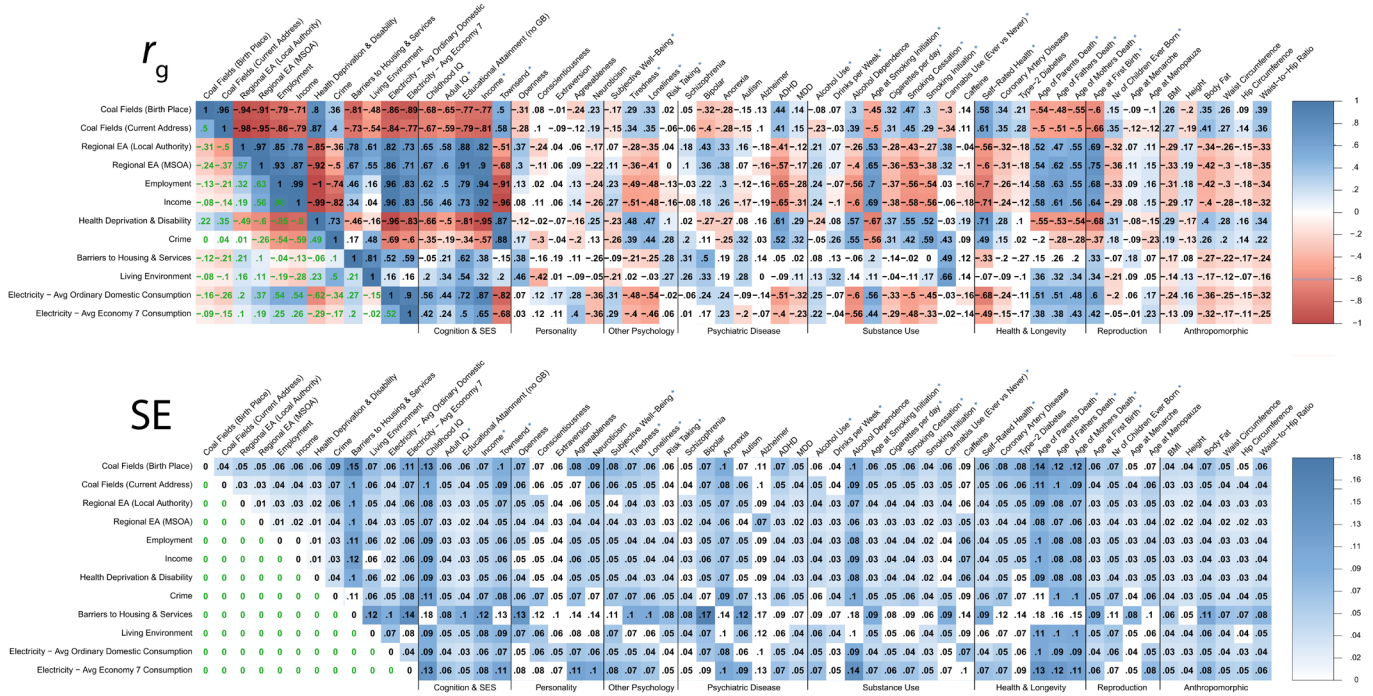
Extended Data Fig. 5 | Associations between polygenic scores and regional measures of nutrition and health. The standardized effect size estimates of robust linear regressions of polygenic scores on regional measures of nutrition and health outcomes in unrelated UK Biobank participants of European descent (N ~320k). The polygenic scores are all standardized residuals after regressing out 100 PCs. Every individual was given the value of their region. Significant effects are colored, whereby the significance threshold is based on FDR correction across all tests shown in all four panels. All SEs were ≤ .002.



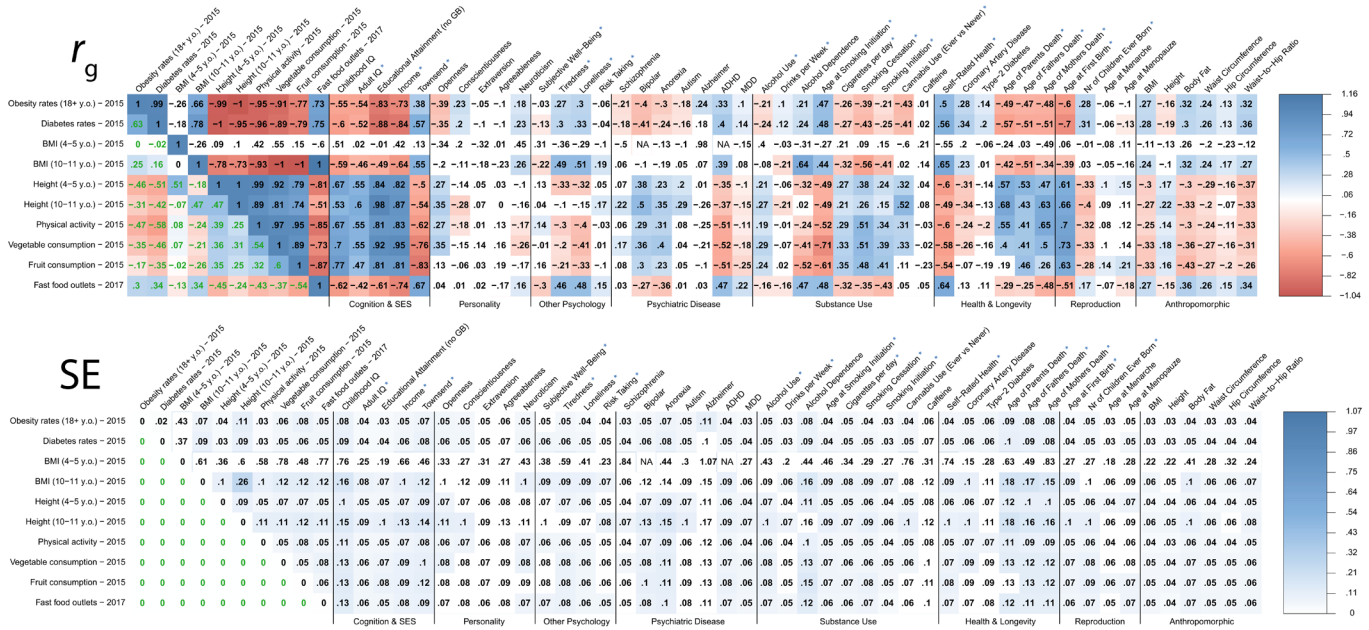
Extended Data Fig. 6 | Associations between polygenic scores and regional measures of religiosity and political preference. The standardized effect size estimates of robust linear regressions of polygenic scores on regional measures of religiosity and election outcomes in unrelated UK Biobank participants of European descent (N ~320k). The polygenic scores are all standardized residuals after regressing out 100 PCs. Every individual was given the value of their region. Significant effects are colored, whereby the significance threshold is based on FDR correction across all tests shown in all four panels. All SEs were ≤ .002.



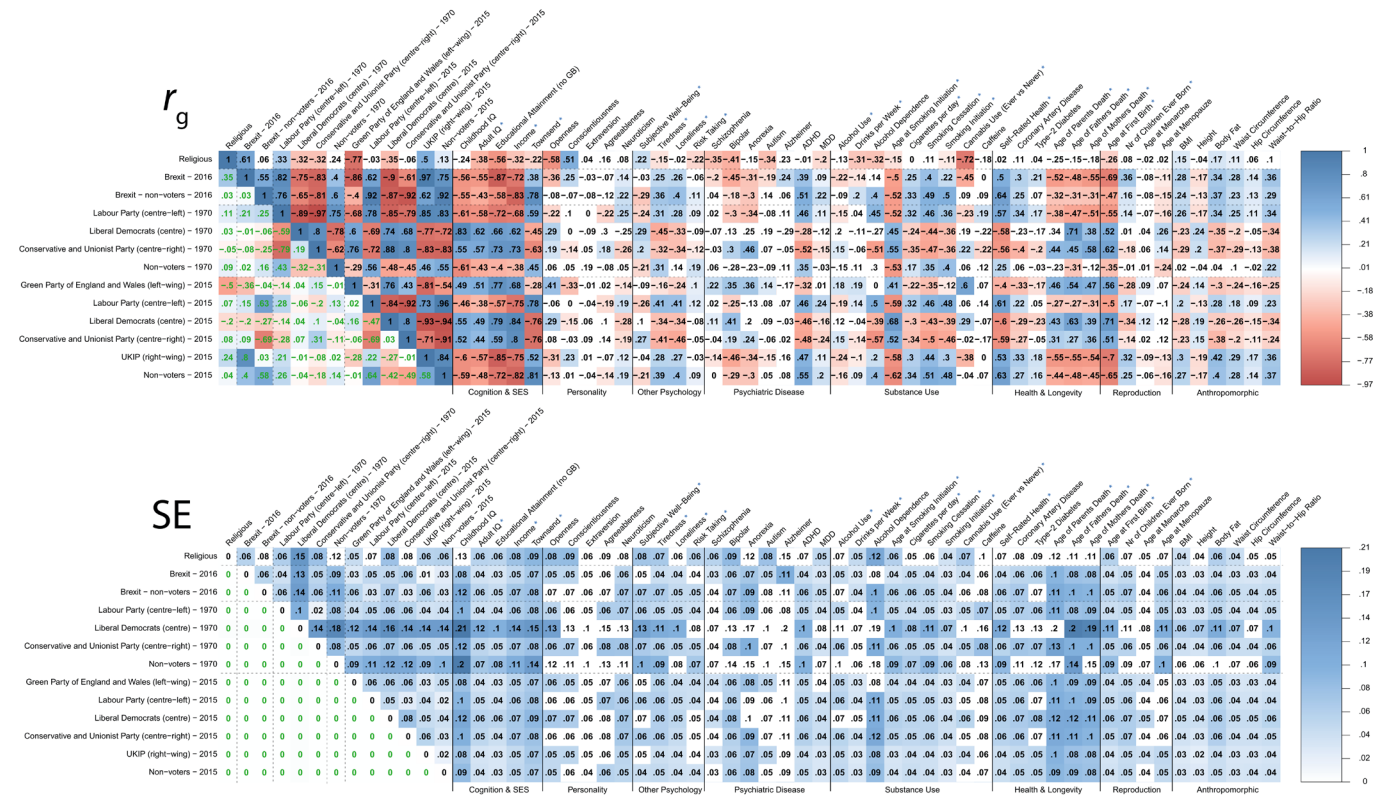
Extended Data Fig. 7 | Associations between polygenic scores and individual-level phenotypes. The standardized effect size estimates of robust linear regressions of polygenic scores on individual level phenotypes in unrelated UK Biobank participants of European descent (N ~320k). The polygenic scores are all standardized residuals after regressing out 100 PCs. Significant effects are colored, whereby the significance threshold is based on FDR correction across all tests shown in all four panels. All SEs were ≤ .002.



Extended Data Fig. 8 | Genetic correlations between regional measures of socio-economic outcomes and a range of complex traits and diseases. Genetic correlations (above) and their SEs (below) based on LD score regression for the RGWASs on SES-related traits. Colored is significant after FDR correction. The green numbers in the left part of the Figure below the diagonal of 1's are the phenotypic correlations between the regional outcomes. The blue stars next to the trait names indicate that UK Biobank was part of the GWAS of the trait. See Supplementary Table 3 for the list of GWASs that the summary statistics of the complex traits were derived from.



Extended Data Fig. 9 | Genetic correlations between regional measures of health- and nutrition and a range of complex traits and diseases. Genetic correlations (above) and their SEs (below) based on LD score regression for the RGWASs on health- and nutrition-related traits. Colored is significant after FDR correction. The green numbers in the left part of the Figure below the diagonal of 1's are the phenotypic correlations between the regional outcomes. The blue stars next to the trait names indicate that UK Biobank was part of the GWAS of the trait. See Supplementary Table 3 for the list of GWASs that the summary statistics of the complex traits were derived from.



Extended Data Fig. 10 | Genetic correlations between regional measures of religiosity and political preference and a range of complex traits and diseases. Genetic correlations (above) and their SEs (below) based on LD score regression for the RGWASs on ideology-related traits (religion and political preference). Colored is significant after FDR correction. The green numbers in the left part of the Figure below the diagonal of 1's are the phenotypic correlations between the regional outcomes. The blue stars next to the trait names indicate that UK Biobank was part of the GWAS of the trait. See Supplementary Table 3 for the list of GWASs that the summary statistics of the complex traits were derived from.

nature research

Corresponding author(s): Abdel Abdellaoui & Peter M. Visscher

Last updated by author(s): May 8, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The data used was collected by UK Biobank; see for more details Bycroft, C. et al (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203.
Data analysis	GCTA, PLINK, fastPCA v2, BOLT-LMM, and R. R-scripts are available upon request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

This study makes use of data from the UK Biobank Resource (Application Number: 12514) and dbGaP (Accession Number: phs000674).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was determined by UK Biobank; see for more details Bycroft, C. et al (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203.
Data exclusions	In order to avoid bias due to populations stratification, we excluded participants with non-European ancestry as determined by PCA on genome-wide SNPs (details are described in the Online Methods).
Replication	The current study was successful because of the exceptionally large sample size of UK Biobank, allowing for sufficient participants per geographic region in the UK to make inferences about polygenic scores. There are no genotype datasets of similar size and geographic spread as this one, making replication not (yet) feasible.
Randomization	The participants were chosen across 22 assessment centers throughout Great Britain in order to cover a variety of different settings providing socioeconomic and ethnic heterogeneity and urban–rural mix ; see for more details Bycroft, C. et al (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203.
Blinding	Blinding was not relevant as this was not an experimental design.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	A total of 502,536 participants (273,402 females and 229,134 males) aged between 37 and 73 years old were recruited in the UK between 2006 and 2010. The participants were recruited across 22 assessment centers throughout Great Britain in order to cover a variety of different settings providing socioeconomic and ethnic heterogeneity and urban–rural mix. See for more details Bycroft, C. et al (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203.
Recruitment	The UK Biobank ascertainment strategy was designed to capture sufficient variation in socioeconomic, urban–rural, and ethnic background. The participation rate however was 5.45% and was biased towards older, more healthy, and female residents. The UK Biobank sample does reflect nationally representative data sources to a significant degree.
Ethics oversight	The participants of this study come from UK Biobank (UKB), which has received ethical approval from the National Health Service North West Centre for Research Ethics Committee (reference: 11/NW/0382).

Note that full information on the approval of the study protocol must also be provided in the manuscript.