

## GENETIC CONSEQUENCES OF SOCIAL STRATIFICATION IN GREAT BRITAIN

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

<sup>1</sup> Department of Psychiatry, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands

<sup>2</sup> Department of Economics, University of East Anglia, Norwich, England, United Kingdom

<sup>3</sup> Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia

<sup>4</sup> Department of Biological Psychology, VU University, Amsterdam, the Netherlands

<sup>5</sup> School of Psychology, University of Queensland, Brisbane, Australia

<sup>6</sup> Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK

<sup>7</sup> Queensland Brain Institute, University of Queensland, Brisbane, Australia

\* Correspondence to: a.abdellaoui@amc.uva.nl; peter.visscher@uq.edu.au

### ABSTRACT

Human DNA varies across geographic regions, with most variation observed so far reflecting distant ancestry differences. Here, we investigate the geographic clustering of genetic variants that influence complex traits and disease risk in a sample of ~450,000 individuals from Great Britain. Out of 30 traits analyzed, 16 show significant geographic clustering at the genetic level after controlling for ancestry, likely reflecting recent migration driven by socio-economic status (SES). Alleles associated with educational attainment (EA) show most clustering, with EA-decreasing alleles clustering in lower SES areas such as coal mining areas. Individuals that leave coal mining areas carry more EA-increasing alleles on average than the rest of Great Britain. In addition, we leveraged the geographic clustering of complex trait variation to further disentangle regional differences in socio-economic and cultural outcomes through genome-wide association studies on publicly available regional measures, namely coal mining, religiousness, 1970/2015 general election outcomes, and Brexit referendum results.

### INTRODUCTION

The first law of geography states that “everything is related to everything else, but near things are more related than distant things”.<sup>1</sup> Humans living near each other tend to share more ancestry with each other than with humans that live further away, which is reflected in genome-wide patterns of genetic variation on a global scale<sup>2</sup> and on finer scales.<sup>3-5</sup> Regional differences in allele frequencies are driven by genetic drift (i.e., the random fluctuations of allele frequency each generation), natural selection pressures, migrations, or admixture (i.e., 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.<sup>3,6,7</sup> Migration is behavior, and since most behavioral traits have heritable components,<sup>8</sup> migration is likely to be associated with genetic variants that influence behavior. Long-distance migratory events may in turn result in admixture. Internal migrations (i.e., 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 neighborhood, housing, and inhabitant characteristics, and/or socio-economic factors (e.g., education or job-related considerations),<sup>9</sup> such as the mass migrations from rural to industrial areas during the industrialization.<sup>10</sup> These geographic movements may coincide with regional clustering of heritable social outcomes such as socio-economic status and major group ideologies (e.g., religion<sup>11</sup> and political preference<sup>12</sup>).

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 complex traits that reflect education, wealth, growth, health, and disease, may help understand why those traits are unevenly distributed across Great Britain. Besides the known regional differences in income and SES, significant regional differences have been reported for mental<sup>13</sup> and physical<sup>14</sup> health problems. Regional differences in wealth and health are likely linked to each other,<sup>15-17</sup> and have been shown to be partly driven by migration.<sup>14,18</sup> If genome-wide complex trait variation is geographically clustered, this should also 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.<sup>19</sup> Geographic clustering of genetic complex trait variation could introduce gene-environment correlations that violate this assumption.<sup>20</sup> Such gene-environment correlations could also introduce bias in heritability estimates in twin and family studies,<sup>21</sup> and could affect signals from genome-wide association studies (GWASs). Furthermore, studying the genetics of migration and geographically clustered cultural phenomena that are related to how society is organized, such as SES, political preference, and religiosity, may help us to further understand regional differences beyond what can be learned from standard observational data. For example, as we will show in this study using a novel regional GWAS approach, we can compute genetic correlations between these clustered social phenomena and a wide range of other traits for which GWASs have been conducted through their GWAS summary statistics.<sup>22</sup> This can teach us about how these regional differences are related to traits that have not been measured in the same dataset.

In this study, we first investigate whether genome-wide complex trait variation is geographically clustered after accounting for ancestry differences; if so, this may reflect the genetic consequences of more recent (internal) migration events. In addition, we investigate whether genome-wide complex trait variation is sufficiently clustered to capture the heritability of regional cultural outcomes such as coal mining, religiousness, and political preference by conducting GWASs on publicly available regional measures. We will then utilize the genetic signals from these GWASs to estimate genetic correlations between the regional measures and a wide range of complex traits.

## DATA AND ANALYSIS

We investigated the geographic clustering of ancestry and complex trait variation using genome-wide single-nucleotide polymorphism (SNP) data from ~450,000 British individuals of European ancestry from the UK Biobank project.<sup>23</sup> Ancestry within Great Britain was captured by conducting a principal component analysis (PCA)<sup>24</sup> on genome-wide SNPs, a method that has been shown to successfully capture ancestry differences within relatively homogeneous populations.<sup>3</sup> 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 30 published GWASs on traits related to psychiatric disease, substance use, personality, body composition, cardiovascular disease, diabetes, reproduction, and educational attainment (see Supplementary Table 1). Importantly, the 30 GWASs that produced the effect estimates did not include UK Biobank participants.<sup>25</sup> Geographic clustering of genetic variation was then investigated using 320,940 unrelated individuals and their birthplace by testing whether the spatial autocorrelation (Moran's  $I$ ) is significantly greater than zero for ancestry-informative principal components (PCs), polygenic scores, and the residuals of polygenic scores after regressing out the first 100 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) to 0 (spatially random) to 1 (spatially clustered).<sup>26</sup> Supplementary Figure 1 shows geographic locations of UK Biobank participants. Furthermore, we test whether polygenic scores that showed significant geographic clustering were associated with an index of economic deprivation of the neighborhood (the Townsend index) and migration into or out of the most economically deprived regions (coal mining areas), while accounting for ancestry differences (100 PCs).

We subsequently investigate whether geographic clustering of genome-wide complex trait variation is associated with regional cultural outcomes by running genome-wide association analyses on coal mining, regional estimates of the proportion of religious vs non-religious inhabitants, election outcomes of the Brexit referendum and of the 1970 & 2015 general elections. We estimate the degree to which these regional differences share genetic influences with a range of traits related to cognitive ability, socio-economic status (SES), personality, behavior, substance use, mental and physical health, well-being, reproduction, and body composition.

For more detailed descriptions of the data and analyses, see Online Methods.

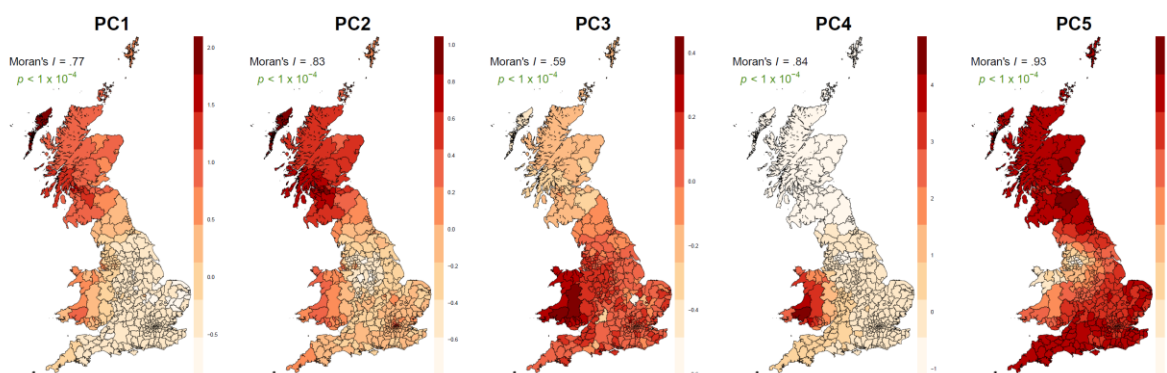


Figure 1: The geographic distributions (birthplace) of the first five PCs, Moran's  $I$  and empirical  $p$ -values for Moran's  $I$ .  $P$ -values denoted in green are significant after Bonferroni correction.

## GEOGRAPHIC CLUSTERING OF GENOME-WIDE ANCESTRY AND COMPLEX TRAIT VARIATION

In line with earlier studies,<sup>5</sup> British ancestry showed significant geographic clustering: the first 100 genetic PCs all show Moran's  $I$  statistics that are greater than 0, with 72 PCs showing an empirical  $p$ -value  $< .0005$ , the Bonferroni corrected threshold, and 95 PCs showing an empirical  $p$ -value  $< .05$ . Many PCs roughly capture the differentiation between Scotland, England, and Wales (see Figure 1 for the first 5 PCs; see

[https://holtzyan.shinyapps.io/UKB\\_geo/](https://holtzyan.shinyapps.io/UKB_geo/) for maps of all 100 PCs). The geographic distributions of the ancestry differences captured by the PCs are likely to reflect consequences of historical demographic events.<sup>5</sup> 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.

Without controlling for ancestry, 27 out of the 30 polygenic scores tested showed a Moran's  $I$  significantly greater than 0, indicating significant geographic clustering (Figures 2 & 3, see [https://holtzyan.shinyapps.io/UKB\\_geo/](https://holtzyan.shinyapps.io/UKB_geo/) for maps of all polygenic scores). Only age at menarche, agreeableness, and caffeine consumption were not significantly geographically clustered. Many clustered polygenic scores showed geographic distributions that were similar to the ancestry differences captured by the PCs. After regressing out the 100 ancestry-informative PCs, 16 polygenic scores remained significantly geographically clustered with FDR correction, with educational attainment (EA) showing the highest Moran's  $I$  (before PC correction: Moran's  $I$  = .57, empirical  $p < 10^{-4}$ ; after PC correction: Moran's  $I$  = .51, empirical  $p < 10^{-4}$ ; see Figures 3 & 4).

It has been argued that geographic clustering of complex trait genetic variation in UK Biobank is due to (subtle) ancestry differences or ascertainment bias.<sup>27</sup> We discuss in more detail in the Supplementary Material why these are unlikely to be the sole explanations of our observations (paragraph: *Population Stratification and Ascertainment Bias*). In the next paragraph, we explore the more likely explanation, namely recent internal SES-related migrations.

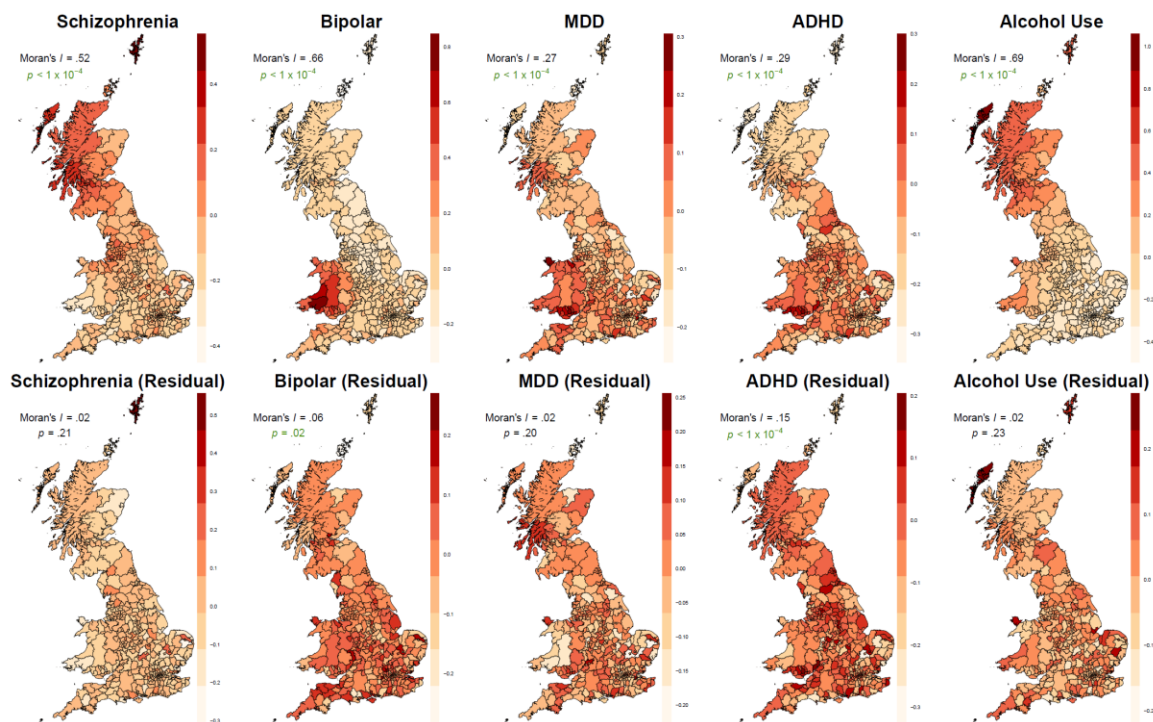


Figure 2: Geographic distribution (birthplace) and Moran's  $I$  values for polygenic scores of four major psychiatric disorders (based on GWASs from the Psychiatric Genomics Consortium (PGC): schizophrenia<sup>28</sup>, bipolar<sup>29</sup>, MDD<sup>30</sup>, and ADHD<sup>31</sup>) and alcohol use<sup>32</sup> before (top row) and after (bottom row) regressing out 100 ancestry-informative PCs. Green  $p$ -values are significant after FDR correction.

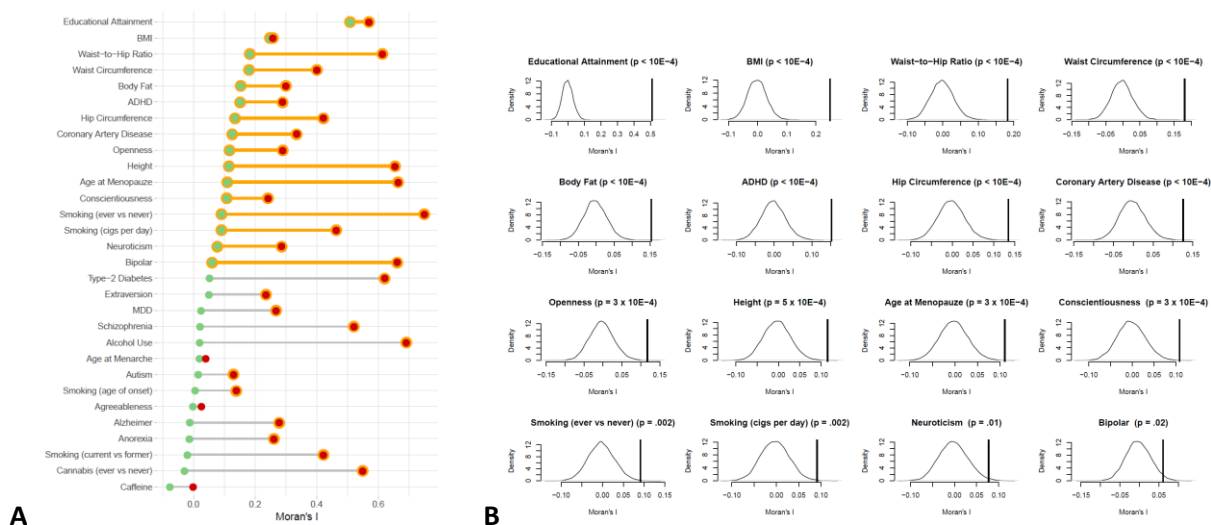


Figure 3: Moran's I of 30 SBLUP polygenic scores computed using the average polygenic score per region in 378 local authority regions. **A** shows the Moran's I of the polygenic scores unadjusted for PCs (red) and adjusted for 100 PCs (green), where orange means a significant FDR corrected p-value  $< .05$  (corrected for 30 tests). **B** shows the distribution of significant Moran's I statistics from 10,000 permutations that were conducted to obtain an empirical p-value for Moran's I. The vertical line to the right of the permutation distribution shows the observed Moran's I of the actual data.

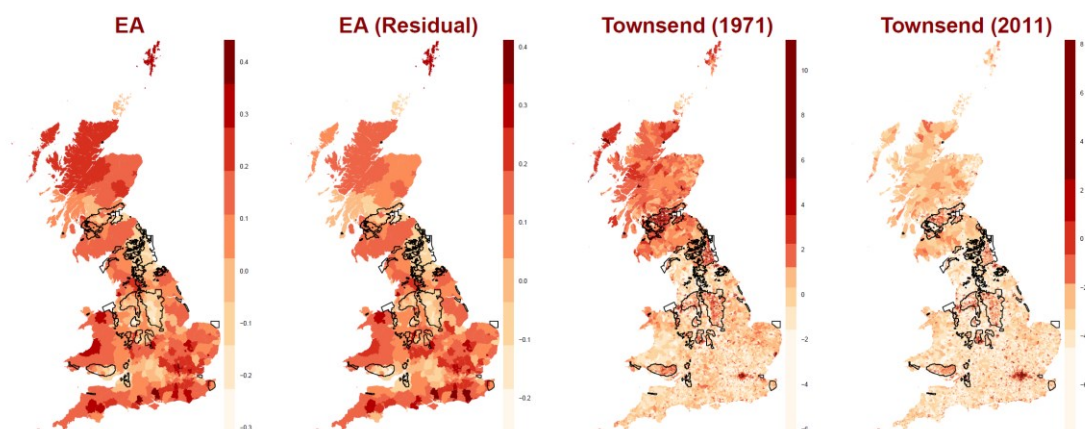


Figure 4: Geographic distribution (birthplace) of the educational attainment (EA) polygenic scores before and after regressing out 100 PCs, and the geographic distribution of Townsend indices from 1971 and 2011. The black lines indicate coal mining areas.

## CONSEQUENCES OF SES-RELATED MIGRATION

The geographic clustering of genome-wide trait-associated alleles after correcting for 100 PCs possibly reflects migration events that occurred more recently than the pre-modern demographic events that drove the regional ancestry differences captured by the PCs. Given the exceptionally strong geographic clustering of the EA score, we investigate here whether it reflects relatively recent internal migrations due to SES-related factors, which are known to especially motivate longer distance moves.<sup>33</sup> Two types of migration flows may have affected the geographic clustering of SES-related alleles: 1) laborers and farmers leaving the country-side during the Industrial Revolution to work in the geographically clustered industrial jobs,<sup>10</sup> 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 the mass-production that characterized the birth of the Industrial Revolution came from coal mines. The presence of coal and iron ore attracted large numbers of manual laborers. The Industrial Revolution and the later deindustrialization had a great impact on the economy of the coal mining areas.<sup>34</sup> 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 still remain visible in unemployment rates decades later.<sup>35</sup> Economic deprivation is widespread in coal mining areas: 43% of neighborhoods from coal mining areas fall into the 30% most economically deprived.<sup>34</sup> In our analysis, coal mining areas show more economic deprivation than the rest of Great Britain from 1971 to 2011 as measured with the Townsend index<sup>36</sup> (higher Townsend = more economic deprivation; all FDR corrected  $p$ -values  $< 10^{-32}$ ; see Figure 4 & Supplementary Figure 6). All regions have become less economically deprived over time, but the difference between coal mining areas and the rest remains highly significant.

After correcting for ancestry differences (100 PCs), the Townsend index is significantly associated with the 16 geographically clustered polygenic scores, with the strongest associations for EA (higher EA polygenic score = lower Townsend index; see Supplementary Figures 7 & 8). These 16 polygenic scores also all show significant differences between coal mining areas and the rest of the regions, both based on birth place and current address (Supplementary Figure 9), with EA showing the strongest differences (FDR corrected  $p$ -value  $< 10^{-200}$ ). We further compared ancestry-corrected polygenic scores between four groups of unrelated individuals: 1) people born in coal mining areas who moved out of coal mining areas (N=35,024), 2) people born outside of coal mining areas and still live 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). ANOVAs for all 16 geographically clustered polygenic scores show significant differences between the four groups (Figure 5), with EA showing the largest and most significant differences ( $F_{3,323661} = 687.3$ , FDR corrected  $p$ -value  $< 10^{-200}$ ). The largest differences were between people born in coal mining areas who moved away versus those who remained in the coal mining areas. The people that moved away have significantly higher EA polygenic scores than all other groups combined ( $t_{43923} = 19.8$ ,  $p = 9 \times 10^{-87}$ ), while those that remained have significantly lower EA polygenic scores than all other groups combined ( $t_{230220} = 44.6$ ,  $p < 10^{-200}$ ). The degree of geographic clustering of polygenic scores, as measured by Moran's  $I$ , is significantly correlated with the strength of their associations with Townsend, coal mining areas, and migration groups; the strongest correlations were between Moran's  $I$  and the  $F$  statistic of the migration group differences:  $r = .95$ ,  $p = 2 \times 10^{-8}$  including EA and  $r = .73$ ,  $p = .002$  excluding EA (Supplementary Figure 10).

To get a better sense of the scale of regional differences in polygenic scores, and of how these change due to migration, we computed how much of the individual differences are explained by regional differences for both birthplace and current address (Supplementary Figure 15). The regional differences are greatest for the EA polygenic score, with about ~0.6-2.6% of individual differences being explained by regional differences, depending on how fine the regional scale is (the finer the scale, the more individual differences explained) and by whether the calculations are based on the birthplace or the current address. Regional differences are ~38-54% greater for the current address than for birthplace. The increase in variation explained by regional differences for the EA score (i.e., difference between birthplace and current address in % variance explained by region) is greater than the total variance explained by region for any other polygenic score. As would be expected from recent migration events, ancestry shows the opposite effect: comparing birthplace to current address, the variance explained by region has on average *decreased* by 37-73% for the first 30 PCs (Supplementary Figure 16).

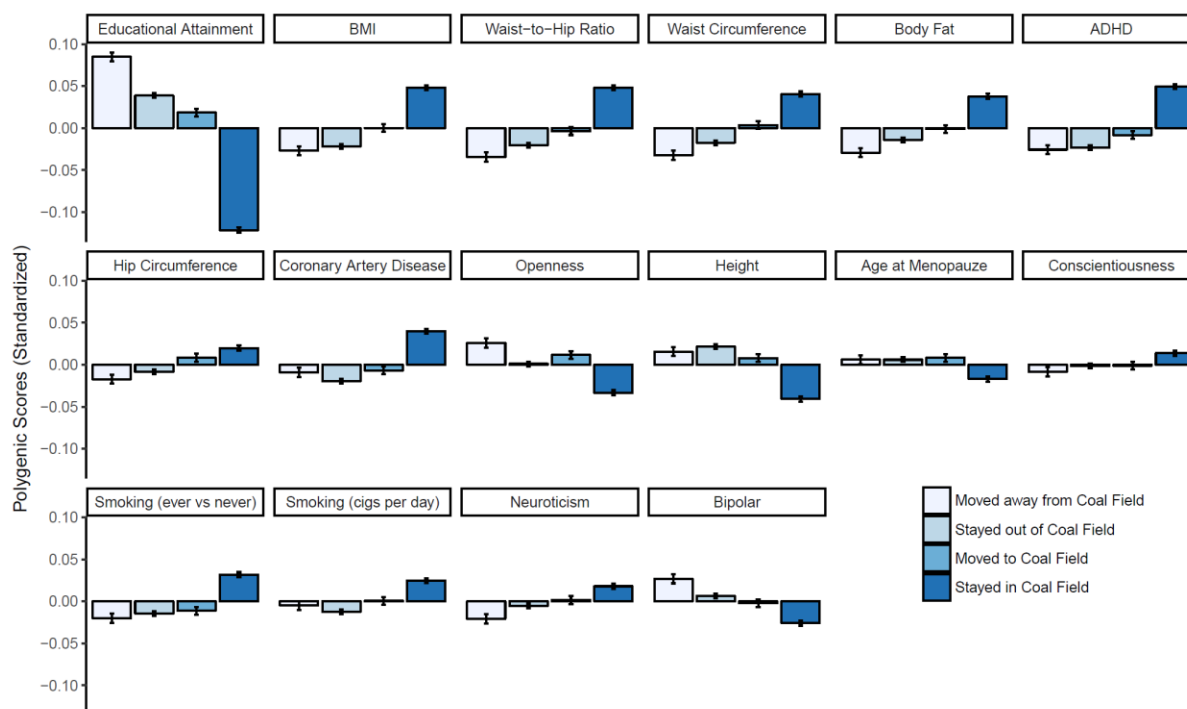


Figure 5: The average and standard errors of the 16 geographically clustered polygenic scores (ordered by Moran's I) for four migration groups: born in coal field area and moved out, born in coal mining area and stayed, born outside of coal mining area and moved to coal mining area, born outside of coal mining area and stayed out. All polygenic scores shown are standardized residuals after regressing out 100 ancestry-informative PCs. ANOVAs were conducted for every polygenic score to test the presence of group differences, which were all significant with the least significant FDR corrected p-value of  $1 \times 10^{-4}$  for conscientiousness.

## GENOME-WIDE ASSOCIATION STUDIES ON REGIONAL OUTCOMES

The geographic clustering of socio-economic resources and associated genetic variants may coincide with a range of regional collective views and attitudes. We examined this by leveraging the geographic clustering of genome-wide complex trait variation with GWASs on regional socio-economic and cultural outcomes, whereby all participants from the same region were assigned the same regional value as a phenotype (from here-on referred to as regional GWASs). The regional GWASs were run on the >400,000 UK Biobank participants, corrected for relatedness, age, sex, and ancestry (100 PCs). We first verified whether the approach works by running a regional GWAS on a regional measure of educational attainment (EA), obtained from census data, which showed genetic signals almost identical to an individual-level EA GWAS that excluded UK Biobank<sup>37</sup> (see Supplementary Materials). We then ran regional GWASs on the presence of coal in an area and regional measures of major ideological factors known to cluster geographically, namely religiousness<sup>11</sup> and political preference<sup>12</sup>. The regional socio-economic and cultural outcomes were defined as follows: whether the individual was born/lives in a coal mining area, the proportion of religious vs non-religious inhabitants in their region (based on current address), the proportion of "Leave" votes and non-voters in the 2016 Brexit referendum (current address), the proportion of non-voters and votes in the individuals' constituency for three major UK parties in the UK 1970 general elections (based on birthplace) and the five major UK parties in the UK 2015 general elections (current address). We used the genome-wide summary statistics of the regional GWASs to estimate genetic correlations with a wide range of complex traits using LD score regression, a

method that computes genetic correlations based on GWAS summary statistics without bias from sample overlap or ancestry differences.<sup>22</sup> We summarize the main results on their (genetic) relationship with other complex traits below, and additional results in the Supplementary Materials.

The regional outcomes showed striking and often highly significant genetic correlations with a wide range of other traits (Figure 6). EA, IQ, and age at first birth showed significant genetic correlations with every regional outcome. Overall, the strongest genetic correlations were observed for cognition & SES-related traits (IQ, EA, income, and Townsend). These suggest that the election outcomes can be divided roughly into higher SES and lower SES regions, with Green Party, Liberal Democrats, and Conservative regions containing more alleles associated with higher SES trait values, and the Labour Party, UKIP, “Leave” votes for Brexit, and non-voters reflecting regions with more alleles associated with lower SES trait values. The election outcomes that are genetically associated with higher cognition and SES outcomes generally also show negative genetic correlations with disease risk outcomes (ADHD, MDD, smoking, alcohol dependence, heart disease, type-2 diabetes, BMI, longevity, and self-rated health), except for alcohol consumption, cannabis use, autism, and psychiatric disorders that are characterized by delusions (schizophrenia, bipolar, anorexia), for which the genetic association is the other way around (higher SES = higher risk). The genetic correlations were largely similar between election outcomes and the coal mining regions, likely due to the same systematic regional SES differences. A different pattern was observed for the proportion of religious inhabitants, which showed weaker genetic correlations with the SES related traits (cognition and health) and stronger associations with the two personality dimensions that showed significant geographic clustering in our previous analyses, openness and conscientiousness (more religious people = lower openness and higher conscientiousness). Risk taking, schizophrenia, and autism also show the highest genetic correlations with being religious (more religious people = lower genetic risk).

The signals were largely consistent within parties over time (1970 & 2015) with respect to SES, but UKIP and Green Party did not yet exist in 1970. The genetic correlations between regional religiousness and the 1970 & 2015 election outcomes suggest that UKIP regions include former Labour Party regions with a more religious genetic profile (lower openness, higher conscientiousness), while the Green Party regions include former Conservative regions with a more non-religious genetic profile (higher openness, lower conscientiousness).

The genetic correlations between the regional outcomes were much stronger than the phenotypic correlations, and in some instances in opposite directions: Labour Party 2015 & Green Party (negative genetic, positive phenotypic correlation), Conservative Party 2015 & Green Party (positive genetic, negative phenotypic correlation), Conservative Party 2015 & Brexit (negative genetic, positive phenotypic correlation). A possible explanation is that the part of the regional variation that is explained by genetic differences is mostly related to regional socio-economic status (lower SES associated alleles in Labour and Brexit areas, higher SES in Conservative and Green areas), while environmental factors, which are responsible for most of the regional variation, are more characterized by ideology (Labour and Green areas being more left-wing, Conservative and Brexit more right-wing).



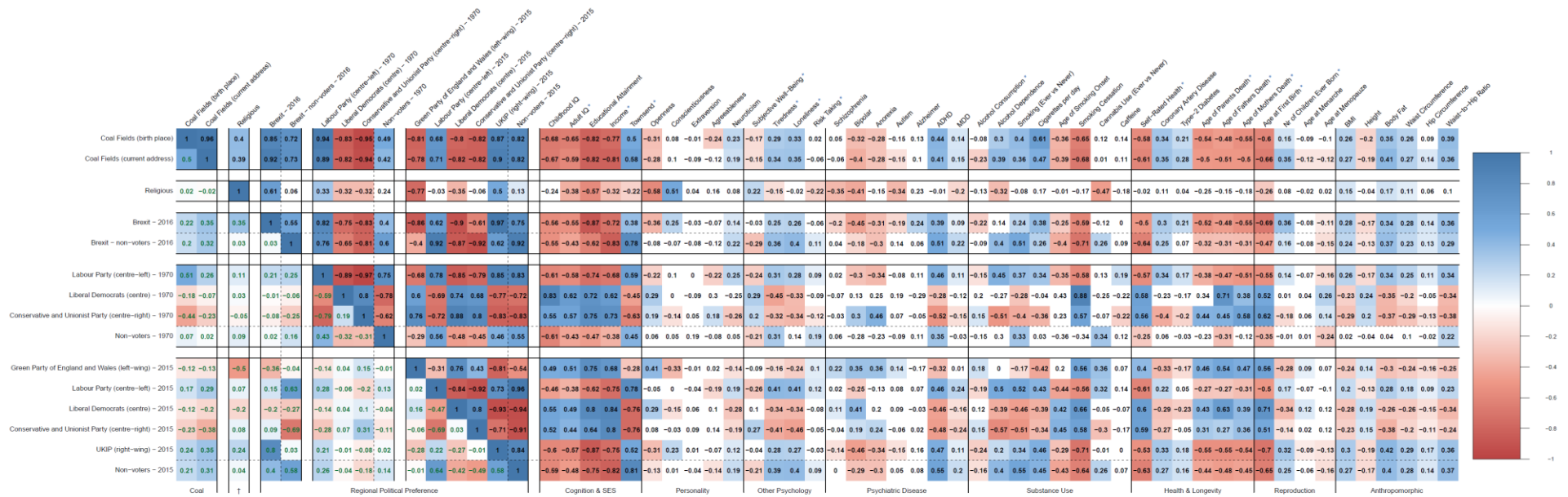


Figure 6: Genetic correlations based on LD score regression. Colored is significant after FDR correction. The green numbers in the left part of the Figure below the diagonal of 1's are the phenotypic correlations between the regional outcomes of coal mining, religiousness, and regional political preference. The blue stars next to the trait names indicate that UK Biobank was part of the GWAS of the trait. See Supplementary Figure 23 for the standard errors and Supplementary Table 4 for the list of GWASs that the summary statistics of the complex traits were derived from.

## DISCUSSION

Understanding the consequences of DNA variation in human populations is of major importance for medical, biological, forensic, behavioral, and anthropological research. Since we have been able to measure DNA at a sequence level, studies have shown that the geographic distributions of alleles are not random and have mapped striking geographic patterns of ancestry.<sup>2,3,5,38</sup> Here, we investigated geographic patterns of genome-wide complex trait variation and show that there are additional levels of genetic geographic clustering beyond the geographic patterns that reflect older ancestry differences. We show that the geographic clustering of genome-wide trait-associated alleles is related to recent geographic movement of people and that the resulting regional genetic patterns are associated with regional socio-economic and cultural outcomes.

Without controlling for ancestry, almost all traits we examined showed significant geographic clustering, often resembling the geographic patterns of ancestry differences within Great Britain. This indicates that either 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 are biased towards certain ancestral backgrounds. When we control for ancestry, 16 polygenic scores remain significantly clustered by geography. The strongest clustering was observed for EA. Among the rest of the geographically clustered traits are body dimensions, personality dimensions, and physical and mental health traits, which may reflect independent influences of them on non-random migration, and/or clustering that is (partly) driven by a genetic overlap with EA. The geographic clustering of complex trait variation seems to have increased due to relatively recent migration which is disrupting the older geographic patterns of ancestry (Supplementary Figures 15 & 16).

The degree of geographic clustering of the polygenic scores is largely in line with the strength of their relationship with regional economic deprivation and migration out of economically deprived regions (Supplementary Figure 10). People are more likely to migrate to improve their skills or employment prospects than for other area characteristics.<sup>9</sup> Many industrialized countries showed these types of migration flows during the late 19<sup>th</sup> and early 20<sup>th</sup> century, where poorer laborers and small farmers left the country-side to work in industrial jobs that were often highly clustered in geographic space (e.g., coal mining areas).<sup>12</sup> After the deindustrialization, the dense, durable, and affordable working-class houses and the public transportation networks from the industrial revolution remained in these neighborhoods and continued to attract poorer immigrants.<sup>12</sup> Our results show that people with a genetic predisposition for higher cognitive abilities are leaving these regions, likely 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 neighborhoods have higher average EA polygenic scores than people born outside of these regions. The regional clustering of cognitive abilities that follows may further affect the economic development of neighborhoods. These demographic processes may influence GWAS signals as well, where alleles that increase the chances of living in the unhealthy circumstances of lower SES neighborhoods may become part of the signal of a GWAS for a trait like BMI or body fat. There are for example significantly more McDonald's restaurants in lower SES neighborhoods in Great Britain.<sup>39</sup> This may be part of the explanation for why four out of the top five geographically clustered polygenic scores are related to body weight.

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 heritability signals for regional religiousness and regional political attitudes, both of which have been shown to be partly heritable on an individual level<sup>40-45</sup> and to cluster geographically<sup>11,12</sup>. From a regional genetic perspective, the election outcomes can be roughly divided into lower SES and higher SES electorates. Our findings suggest that the previously reported heritability estimates of these traits on an individual level may contain genetic effects on traits, such as EA, that influence which socio-economic strata and geographic regions people end up living in. Regional religiousness shows higher genetic correlations with personality (openness and conscientiousness) and less with the SES and health traits than the political parties do, which implies additional dimensions of geographic clustering beyond high versus low SES.

Our findings may largely reflect genetic consequences of social stratification, a key characteristic of human civilizations whereby society groups their people into strata based on SES. SES is generally based on occupation, income, and educational attainment, which are influenced by many environmental and genetic factors, and are associated with a wide range of physical and mental health outcomes. Socioeconomic status is not distributed randomly across geographic space, which leads to geographic clustering of alleles that are associated with SES-related traits such as educational attainment. Educational attainment is known for its high levels of assortative mating,<sup>46,47</sup> which may be further induced by geographic clustering. This may affect social inequalities across generations through expanding biological inequalities in cognitive abilities and susceptibility to disease. It is possible that the combination of recent increases in social mobility and an improved educational system accelerates this separation of higher and lower genetic predisposition for traits related to cognition, SES, and health. Even though the genetic effects we find explain only part of the observed regional differences, researchers and social policy makers should keep these effects in mind, as they seem to be growing due to migration and can lead to detectable regional differences in health and social and economic success. For example, the significant genetic correlations between educational attainment and traits related to disease risk or body composition may decrease in the presence of stronger social safety nets that are geared towards making inhabitants of lower SES regions live more economically prosperous and healthier lives. Social policies that increase the quality of life in lower SES regions may also help to decrease migration out of the currently more economically deprived regions by people with genetic predispositions for higher SES outcomes, and thereby possibly result in a less geographically stratified society.

## **ACKNOWLEDGEMENTS**

This research was supported by the Australian National Health and Medical Research Council (1107258, 1078901, 1078037, 1056929, 1048853, and 1113400), and the Sylvia & Charles Viertel Charitable Foundation (Senior Medical Research Fellowship). B.P.Z. received funding from The Australian Research Council (FT160100298). M.G.N. is supported by ZonMw grants 849200011 and 531003014 from The Netherlands Organisation for Health Research and Development. K.J.H.V is supported by the Foundation Volksbond Rotterdam. This study makes use of data from the UK Biobank Resource (Application Number: 12514) and dbGaP (Accession Number: phs000674).

## REFERENCES

- 1 Tobler, W. R. A computer movie simulating urban growth in the Detroit region. *Economic geography* **46**, 234-240 (1970).
- 2 Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98 (2008).
- 3 Abdellaoui, A. *et al.* Population structure, migration, and diversifying selection in the Netherlands. *European journal of human genetics* **21**, 1277 (2013).
- 4 Kerminen, S. *et al.* Fine-Scale Genetic Structure in Finland. *G3: Genes, Genomes, Genetics* **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. *Applied & translational genomics* **2**, 78-94 (2013).
- 7 Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS genetics* **10**, e1004412 (2014).
- 8 Turkheimer, E. Three laws of behavior genetics and what they mean. *Current directions in psychological science* **9**, 160-164 (2000).
- 9 Coulter, R. & Scott, J. What motivates residential mobility? Re-examining self-reported reasons for desiring and making residential moves. *Population, Space and Place* **21**, 354-371 (2015).
- 10 Long, J. Rural-urban migration and socioeconomic mobility in Victorian Britain. *The Journal of Economic History* **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. *Annual Review of Political Science* **13**, 321-340 (2010).
- 13 Lewis, G. & Booth, M. Regional differences in mental health in Great Britain. *Journal of Epidemiology & Community Health* **46**, 608-611 (1992).
- 14 Boyle, P. Population geography: migration and inequalities in mortality and morbidity. *Progress in Human Geography* **28**, 767-776 (2004).
- 15 Tyrrell, J. *et al.* Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank. *bmj* **352**, i582 (2016).
- 16 Marmot, M. The health gap: the challenge of an unequal world. *The 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. *Social Science & Medicine* **50**, 861-878 (2000).
- 19 Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology* **32**, 1-22 (2003).
- 20 Richards, J. B. & Evans, D. M. (British Medical Journal Publishing Group, 2017).
- 21 Verweij, K. J., Mosing, M. A., Zietsch, B. P. & Medland, S. E. in *Statistical Human Genetics* 151-170 (Springer, 2012).
- 22 Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature genetics* **47**, 1236-1241 (2015).
- 23 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 medicine* **12**, e1001779 (2015).

- 24 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904 (2006).
- 25 Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics* **14**, 507-515 (2013).
- 26 Moran, P. A. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17-23 (1950).
- 27 Haworth, S. *et al.* Common genetic variants and health outcomes appear geographically structured in the UK Biobank sample: Old concerns returning and their implications. *bioRxiv*, 294876 (2018).
- 28 Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421 (2014).
- 29 Ruderfer, D. M. *et al.* Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Molecular psychiatry* **19**, 1017 (2014).
- 30 Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *bioRxiv*, 167577 (2017).
- 31 Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for ADHD. *bioRxiv*, 145581 (2017).
- 32 Schumann, G. *et al.* KLB is associated with alcohol drinking, and its gene product  $\beta$ -Klotho is necessary for FGF21 regulation of alcohol preference. *Proceedings of the National Academy of Sciences* **113**, 14372-14377 (2016).
- 33 Niodomysl, T. How migration motives change over migration distance: evidence on variation across socio-economic and demographic groups. *Regional Studies* **45**, 843-855 (2011).
- 34 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 University* (2014).
- 35 Beatty, C., Fothergill, S. & Powell, R. Twenty years on: has the economy of the UK coalfields recovered? *Environment and Planning A* **39**, 1654-1675 (2007).
- 36 Townsend, P., Phillimore, P. & Beattie, A. *Health and deprivation: inequality and the North*. (Routledge, 1988).
- 37 Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539 (2016).
- 38 Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786-792 (1978).
- 39 Cummins, S. C., McKay, L. & MacIntyre, S. McDonald's restaurants and neighborhood deprivation in Scotland and England. *American journal of preventive medicine* **29**, 308-310 (2005).
- 40 Alford, J. R., Funk, C. L. & Hibbing, J. R. Are political orientations genetically transmitted? *American political science review* **99**, 153-167 (2005).
- 41 Benjamin, D. J. *et al.* The genetic architecture of economic and political preferences. *Proceedings of the National Academy of Sciences* **109**, 8026-8031 (2012).
- 42 Hatemi, P. K. & McDermott, R. The genetics of politics: discovery, challenges, and progress. *Trends in Genetics* **28**, 525-533 (2012).
- 43 Hatemi, P. K., Medland, S. E., Morley, K. I., Heath, A. C. & Martin, N. G. The genetics of voting: An Australian twin study. *Behavior genetics* **37**, 435 (2007).
- 44 Smith, K. *et al.* Biology, ideology, and epistemology: how do we know political attitudes are inherited and why should we care? *American journal of political science* **56**, 17-33 (2012).
- 45 Koenig, L. B., McGue, M., Krueger, R. F. & Bouchard, T. J. Genetic and environmental influences on religiousness: Findings for retrospective and current religiousness ratings. *Journal of personality* **73**, 471-488 (2005).

- 46 Robinson, M. R. *et al.* Genetic evidence of assortative mating in humans. *Nature Human Behaviour* **1**, 0016 (2017).
- 47 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).
- 48 Bycroft, C. *et al.* Genome-wide genetic data on ~ 500,000 UK Biobank participants. *Nature* **562**, 203-209 (2018).
- 49 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**, 76-82 (2011).
- 50 Consortium, G. P. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
- 51 Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776-2778 (2017).
- 52 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* **48**, 1279 (2016).
- 53 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559-575 (2007).
- 54 Sweet, D. (2011).
- 55 Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics* **47**, 284-290 (2015).
- 56 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. *Nature genetics* **46**, 100-106 (2014).
- 57 Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* **47**, 291-295 (2015).
- 58 Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics* **47**, 1228-1235 (2015).
- 59 Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459 (2010).
- 60 Kvale, M. N. *et al.* Genotyping informatics and quality control for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. *Genetics* **200**, 1051-1060 (2015).
- 61 Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS genetics* **9**, e1003348 (2013).
- 62 Fry, A. *et al.* Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American journal of epidemiology* **186**, 1026-1034 (2017).
- 63 De Moor, M. H. *et al.* Meta-analysis of genome-wide association studies for personality. *Molecular psychiatry* **17**, 337 (2012).
- 64 Lo, M.-T. *et al.* Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. *Nature genetics* **49**, 152 (2017).
- 65 Duncan, L. *et al.* Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa. *American Journal of Psychiatry* **174**, 850-858 (2017).
- 66 Consortium, A. S. D. W. G. o. T. P. G. *et al.* Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Molecular autism* **8**, 1-17 (2017).
- 67 Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature genetics* **45**, 1452 (2013).
- 68 Furberg, H. *et al.* Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature genetics* **42**, 441 (2010).

- 69 Stringer, S. *et al.* Genome-wide association study of lifetime cannabis use based on a large meta-analytic sample of 32 330 subjects from the International Cannabis Consortium. *Translational psychiatry* **6**, e769 (2017).
- 70 Cornelis, M. C. *et al.* Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. *Molecular psychiatry* **20**, 647 (2015).
- 71 Nikpay, M. *et al.* A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature genetics* **47**, 1121 (2015).
- 72 Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics* **44**, 981 (2012).
- 73 Elks, C. E. *et al.* Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nature genetics* **42**, 1077 (2010).
- 74 Day, F. R. *et al.* Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nature genetics* **47**, 1294 (2015).
- 75 Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics* **46**, 1173 (2014).
- 76 Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197 (2015).
- 77 Lu, Y. *et al.* New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nature communications* **7**, 10495 (2016).
- 78 Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187 (2015).
- 79 Benyamin, B. *et al.* Childhood intelligence is heritable, highly polygenic and associated with FBNP1L. *Molecular psychiatry* **19**, 253 (2014).
- 80 Sniekers, S. *et al.* Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nature Genetics* (2017).
- 81 Hill, W. D. *et al.* Molecular genetic contributions to social deprivation and household income in UK Biobank. *Current Biology* **26**, 3083-3089 (2016).
- 82 Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature genetics* **48**, 624 (2016).
- 83 Deary, V. *et al.* Genetic contributions to self-reported tiredness. *Molecular psychiatry* **23**, 609 (2018).
- 84 Churchhouse, C. & Neale, B. Rapid GWAS of Thousands of Phenotypes for 337,000 Samples in the UK Biobank. *Neale Lab* (2017).
- 85 Clarke, T.-K. *et al.* Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N= 112 117). *Molecular psychiatry* **22**, 1376 (2017).
- 86 Walters, R. K. *et al.* Trans-ancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *bioRxiv*, 257311 (2018).
- 87 Harris, S. E. *et al.* Molecular genetic contributions to self-rated health. *International journal of epidemiology* **46**, 994-1009 (2016).
- 88 Pilling, L. C. *et al.* Human longevity is influenced by many genetic variants: evidence from 75,000 UK Biobank participants. *Aging (Albany NY)* **8**, 547 (2016).
- 89 Barban, N. *et al.* Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nature genetics* **48**, 1462 (2016).



## ONLINE METHODS

### *Participants*

The participants of this study come from UK Biobank (UKB),<sup>23</sup> which has received ethical approval from the National Health Service North West Centre for Research Ethics Committee (reference: 11/NW/0382). A total of 502,655 participants aged between 37 and 73 years old were recruited in the UK between 2006 and 2010. 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 QC*

A total of 488,377 UKB participants had their genome-wide single nucleotide polymorphisms (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 (HRC) panel as a reference set (pre-imputation QC and imputation are described in more detail in Bycroft et al, 2018).<sup>48</sup> To create polygenic scores, we extracted a set of 1,312,100 autosomal HapMap 3 (HM3) SNPs with minor allele count (MAC) > 5, info score > 0.3,  $p_{HWE} < 10^{-6}$ , and missingness < .05. For the genome-wide association study, we used 5.8 million SNPs that survived QC and have a MAF > .01.

### *Ancestry & Principal Component Analysis*

To capture British ancestry, we first excluded individuals with non-European ancestry. Ancestry was determined using Principal Component Analysis (PCA) in GCTA<sup>49</sup>. The UKB dataset was projected onto the first two principal components (PCs) from the 2,504 participants of the 1000 Genomes Project,<sup>50</sup> using HM3 SNP with minor allele frequency (MAF) > 0.01 in both datasets. Next, participants from UKB 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 > 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 2x2 variance-covariance matrices for each super-population. A total of 456,426 subjects were identified to have a European ancestry.

A PCA was then conducted on individuals of European ancestry in order to capture ancestry differences within the British population. In order to capture ancestry differences in homogenous populations, genotypes should be pruned for LD and long-range LD regions removed.<sup>3</sup> The LD pruned ( $r^2 < .1$ ) UKB dataset without long-range LD 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 v2.<sup>51</sup> 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 30 traits. The effect size estimates were obtained from genome-wide association studies (GWASs) that were chosen to not have included the UKB dataset to avoid over-estimation of the genetic predisposition of a trait.<sup>25</sup> The polygenic scores were computed using the SBLUP approach,<sup>46</sup> which maximizes the predictive power by creating scores with best linear unbiased predictor (BLUP) properties that account for linkage disequilibrium (LD) between SNPs. As a reference sample for the LD, we used the a random sample of 10,000 unrelated individuals from UK Biobank that were imputed using the Haplotype Reference Consortium (HRC) reference panel.<sup>52</sup> The traits included psychiatric disorders, substance use, anthropomorphic traits, personality dimensions, educational attainment, reproduction, cardiovascular disease, and type-2 diabetes. Supplementary Table 1 lists the 30 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  $< .05$ . The SNPs were clumped using PLINK<sup>53</sup>, using an  $r^2$  threshold of 0.1 and a window of 1 Mb as the physical distance threshold for clumping.

In order 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.

### *Spatial autocorrelations (Moran's I)*

The geographic clustering of ancestry and of genome-wide complex trait variation was investigated by testing whether the spatial autocorrelation (Moran's  $I$ ) is significantly greater than zero for ancestry-informative principal components (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) to 0 (spatially random) to 1 (spatially clustered).<sup>26</sup> Moran's  $I$ 's were computed using the average PCs or polygenic scores per region based on the birthplace of the subjects (378 regions, see Figure 1), whereby the regions were defined according to the local authorities division as provided by the UK Data Service InFuse database.<sup>54</sup> 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 (Figure 3B).

### *Regional genome-wide association studies (GWASs)*

Genome-wide association studies (GWASs) were run on publicly available regional outcomes, whereby all subjects from the same regions had the same regional phenotypic value assigned. Supplementary Figure 16 & 21 show the distributions of all phenotypes analyzed, except for the coal mining phenotypes, which were binary traits (47% of the participants were born in a coal mining area, and 50% of the participants currently live in a coal mining area). 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://data.gov.uk/dataset/coal-mining-reporting-area>
- The regional educational attainment for 342 local districts and for 7,195 Middle Super Output Areas (MSOA) was measured as the 2011 estimates of the highest qualification of residents of England >16 years old (5 levels: level 1 qualifications, level 2 qualifications, apprenticeship, level 3 qualifications, level 4 qualifications) was obtained from the *Nomis* database of the Office of National Statistics: <https://www.nomisweb.co.uk/>
- The proportion of religious vs non-religious inhabitants were obtained for 7,195 Middle Super Output Areas (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>
- The 2015 general election outcomes were obtained for 633 constituencies from *data.parliament.uk*: <http://www.data.parliament.uk/dataset/general-election-2015>

All political parties were included that had a median proportion of votes > 0.

We ran linear mixed model (LMM) GWASs with BOLT-LMM<sup>55</sup> on participants with European ancestry, which controls for cryptic relatedness and population stratification by including a genetic relatedness matrix (GRM) in the model.<sup>56</sup> 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 LD score intercepts<sup>57</sup> shown in Supplementary Table 2). This is likely due to the fact that participants that share regional environmental influences, because they come from the same region, are all assigned the same phenotypic value. We controlled for this inflation with an LD score intercept-based genomic control,<sup>57</sup> i.e., we adjusted the standard errors (SE) of the estimated effect sizes as follows:  $SE_{GC} = \sqrt{LDSC\ intercept * SE^2}$  (see Supplementary Table 2).

### *LD Score Regression*

We partitioned the polygenic contributions to the heritability across genomic regions associated with histone modifications specific to ten cell-type/tissue groups using stratified LD score regression<sup>58</sup> (Supplementary Figure 23). Genetic correlations were also computed using LD-score regression (Figure 6).<sup>58</sup> 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 LD score and represents the genetic covariation between two traits based on all polygenic effects captured by the included SNPs. The genome-wide LD information used by these methods were based on European populations from the HapMap 3 reference panel.<sup>57,58</sup> All LD score regression analyses included the 1,290,028 million genome-wide HapMap SNPs used in the original LD score regression studies.<sup>57,58</sup>

Computing genetic correlations with LD score regression is robust to sample overlap, so we included summary statistics from GWAS studies that also included UK Biobank (denoted with a blue star in Figure 6). Where possible however, we decided to display results obtained from summary statistics without UK Biobank, even if the GWASs from the original studies included UK Biobank participants. This was the case for MDD<sup>30</sup> and educational attainment<sup>37</sup>, 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. The genetic correlations for MDD and educational attainment obtained with the summary statistics that did include UK Biobank however were almost identical.

## GENOME-WIDE ASSOCIATION STUDIES ON REGIONAL OUTCOMES (ADDITIONAL RESULTS)

In order to empirically validate the approach, we first ran regional GWASs on regional measures of average EA outcomes as obtained from census data, namely the weighted average of 2011 estimates of the highest qualification of residents >16 years old, obtained from the Office for National Statistics (Supplementary Figure 16). This resulted in genetic signals that were very close to those of an individual-level EA GWAS that excluded UK Biobank.<sup>37</sup> Most significant SNPs in the regional GWAS were at least nominally significant in the individual level GWAS and their effect sizes correlate .93 (Supplementary Figure 18). The genetic correlation between the regional EA GWAS and the individual level EA GWAS was .90, and the genetic correlations with 64 other complex traits were almost identical between the regional and individual level EA GWASs ( $r = .99$ , Supplementary Figures 19 & 20).

For the regional GWASs conducted on the presence of coal in the area, religiousness, and political preference, there were a total of 12 independent SNPs with  $p < 5 \times 10^{-8}$  and 5 independent SNPs with  $p < 1 \times 10^{-8}$  (Supplementary Figure 22 & Supplementary Table 3). The variance that could be accounted for by all SNPs (i.e., SNP heritability) ranged from 0.3% to 2.4% (see Supplementary Table 2), with the highest ( $\geq 2\%$ ) observed for Brexit, Green Party, UKIP, and non-voters in 2015. The heritability signals were significantly enriched for genetic variants that are active in hormonal pathways for the Green Party, and in the central nervous system for the Green Party, UKIP, 2015 non-voters, and Brexit (Supplementary Figure 23).

Regions with more non-voters genetically show a lower SES profile (i.e., strong negative genetic correlations with cognition and SES-related traits) and the largest positive genetic correlations with regions with more Labour party votes, up to .96 between the 2015 non-voters and the 2015 Labour voters. The 2015 non-voters regional GWAS shows the highest SNP heritability of the non-voters GWASs (2.2%). The genetic correlations also imply that regions with more non-voters and Labour voters show more risk-increasing alleles for mood-related traits (i.e., more MDD, higher neuroticism, more loneliness, and lower wellbeing), and no significant genetic correlation with conscientiousness, as opposed the other lower SES regions with more votes for UKIP and “Leave” votes for Brexit, which show a significant positive genetic correlation with conscientiousness.

In order to further examine what differentiates the parties within the higher SES and lower SES clusters from each other, we repeated the regional GWASs for the proportion of votes among only the Green Party, Liberal Democrats, and Conservatives votes, and the proportion of votes among only Labour Party and UKIP votes. The correlations with religiousness were consistently higher and more often significant for the differences within the higher and lower SES voters than the differences between them (Supplementary Figure 24). The genetic signals that differentiate Green Party regions from the other higher SES votes show the highest genetic correlation with regional religiousness (more Green Party votes = less religious:  $r_g = -.82$ ,  $SE = .06$ ). What differentiates Liberal Democrats from the other higher SES parties still seems to be largely SES-related, given the high positive genetic correlation with EA and income (both .77). The lower SNP heritability estimates of the within SES differences (0.3% - 1.2%, compared to 1.1% - 2.4%, see Supplementary Table 2) suggest that the differences within the higher and lower SES voters are less influenced by regional genetics than the differences between them.

## POPULATION STRATIFICATION AND ASCERTAINMENT BIAS

### *Population Stratification*

The largest patterns of genome-wide variation between and within human populations are due to differences in ancestry rather than trait variation. In genetic association studies, false positives due to population stratification occur when these systematic ancestry differences get mistaken for associations due to genetic variants that influence the trait that is being studied.<sup>59</sup> False positives due to population stratification can occur when trait differences are in line with ancestry differences, which could also occur due to non-genetic factors, such as regional differences in environmental exposures. Geographic location is known to strongly correlate with ancestry differences: the closer people live to each other, the more likely it is that they share more ancestors. The main focus of this study is the relationship between geographic location and genome-wide complex trait variation, which is why we had to be particularly rigorous in accounting for population stratification. We summarize below why it is unlikely that our observations are merely a result of ancestry differences or biased polygenic scores.

- The most widely used approach to account for ancestry differences is to quantify ancestry differences with a principal component analysis (PCA) on genome-wide SNP data and then account for the resulting principal components (PCs).<sup>24,59</sup> Instead of using the standard 40 PCs provided by UK Biobank, which capture both non-European and European ancestry differences,<sup>48</sup> we re-computed PCs to more effectively capture population stratification within the more homogeneous group of British participants with European ancestry (see Online Methods). While genome-wide association studies (GWASs) usually control for 10 to 40 PCs, we controlled for the first 100 PCs in all our analyses.
- We validated the effectiveness of the 100 PCs in accounting for geographic clustering due to population stratification using polygenic scores that reflect European ancestry differences as captured in an independent European-American dataset: the GERA cohort.<sup>60</sup> First, we conducted GWASs in GERA (N = 51,258) on the first 20 GERA PCs in order to get SNP effects that reflect genome-wide patterns of their European-American ancestry differences. We then used these SNP effects to build polygenic scores in UK Biobank. These all show significant geographic clustering as quantified with Moran's *I*. After controlling for 100 PCs from UKB all Moran's *I*'s dropped to being not significantly greater than 0 (see Online Methods and Supplementary Figures 3 and 4).
- The polygenic scores we analyzed were constructed from 1,312,100 autosomal SNPs, regardless of how significantly associated they are with the trait. The ensemble of non-significant SNPs contain a substantial amount of signals due to true causal relationships, and thus meaningful effect sizes, which increase the predictive power of polygenic scores.<sup>61</sup> Increasing the number of non-associated SNPs however may also increase the chances of including more stratified SNPs in the polygenic score. We therefore created a set of additional polygenic scores using only independent SNPs (i.e., clumped) that were at least nominally significantly associated with the trait at  $p < .05$ . This results in scores that are based on fewer SNPs that are more reliably associated with the trait, but also results in less predictive scores. With this approach, 8 out of 16 previously significant traits are significantly geographically clustered after FDR correction (see Supplementary Figure 2). These geographically clustered clumped scores also showed similar and significant associations with Townsend (Supplementary Figures 11 & 12), coal mining regions (Supplementary Figure 13) and migration out of coal fields (Supplementary Figures 14), with educational attainment showing the strongest effects.

- We show that the geographically clustered polygenic scores are significantly associated with regional outcomes of economic deprivation and with migration out of the more economically deprived regions in the UK (coal mining regions). The strength of the geographic clustering is in line with the strength of the association with regional outcomes and, more importantly, migration (Supplementary Figure 10). In other words, 1) the traits that show significant geographic clustering are the traits that cluster in specific regions that are characterized by lower SES measures, and 2) we show that the processes that would result in these regional differences are measurable in the current dataset, namely migration out of the lower SES regions by individuals with a higher predisposition for SES-related traits such as higher educational attainment and lower body weight. Although these observations do not directly prove that ancestry differences cannot account for this geographic clustering, it does show that *if* subtle population stratification would be the cause of these regional differences (which is unlikely given our stringent control for ancestry differences), it would have to involve ancestry differences that are in line with genome-wide complex trait variation.
- SNPs that are in LD with many SNPs are more likely to tag a causal SNP (i.e., be correlated with a causal SNP), and are thus more likely to have a higher test-statistic in a GWAS. The amount of SNPs that is tagged by a SNP is quantified by its LD Score. LD Score regression is an approach that leverages the relationship between the LD score of a SNP and the GWAS test statistic to distinguish inflation of genome-wide test statistics due to variants that influence the complex trait under study from inflation due to confounding bias such as population stratification.<sup>57</sup> LD Score regression analyses show that the results from our regional GWASs all show an inflation of test statistics that is partly due to confounding (likely shared environmental influences) but also contains a considerable inflation due to variants that are associated with complex trait variation that is being captured with the regional measures. LD Score regression was then used to compare the parts of the genetic signals that were due to causal variants between our regional GWASs and GWASs from a wide range of other complex traits. Importantly, LD Score regression showed that the signals from our regional GWAS on EA contained almost the same signals as an individual level GWAS on EA that was conducted on non-UK Biobank datasets, which is in line with the geographic clustering of genome-wide alleles that have a causal influence on EA.

### *Ascertainment Bias*

The UK Biobank ascertainment strategy was designed to capture sufficient variation in socioeconomic, urban–rural, and ethnic background.<sup>23</sup> The participation rate however was 5.45% and was biased towards older, more healthy, and female residents.<sup>62</sup> The UK Biobank sample does reflect nationally representative data sources to a significant degree, making it likely that our observations would generalize to the population at large. We tested at the MSOA level how EA measurements in UK Biobank compare to nationally representative census data (the same EA census measurements that we used for the regional EA GWAS). The average EA per MSOA region as measured in UK Biobank is strongly predictive of MSOA-EA as measured from the nationally representative census data ( $p < 10^{-16}$ ,  $R^2 = 40\%$ ; Supplementary Figure 5). The average polygenic scores per MSOA region, with 100 PCs regressed out, are also highly predictive of MSOA-EA according to nationally representative census data ( $p < 10^{-16}$ ,  $R^2 = 19\%$ ; Supplementary Figure 5). Since UK Biobank has sampled healthier individuals as well as fewer individuals from more economically deprived areas as compared to the British population as a whole,<sup>62</sup> the regional differences that we report may turn out to be stronger in the real population than in the UK Biobank sample.

#### **SUPPLEMENTARY NOTE: ACKNOWLEDGEMENTS**

GERA: The Genetic Epidemiology Research on Adult Health and Aging study was supported by grant RC2 AG036607 from the National Institutes of Health, grants from the Robert Wood Johnson Foundation, the Ellison Medical Foundation, the Wayne and Gladys Valley Foundation and Kaiser Permanente. The authors thank the Kaiser Permanente Medical Care Plan, Northern California Region (KPNC) members who have generously agreed to participate in the Kaiser Permanente Research Program on Genes, Environment and Health (RPGEH).

UK Biobank: This study has been conducted using UK Biobank resource under Application Number 12514. UK Biobank was established by the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government and the Northwest Regional Development Agency. It has also had funding from the Welsh Assembly Government, British Heart Foundation and Diabetes UK.



Supplementary Table 1: Overview of GWASs that provided the effect size estimates that were used to compute polygenic scores

Trait Name	N <sub>GWAS</sub>	N <sub>cases</sub>	N <sub>controls</sub>	N <sub>effective</sub>	h <sup>2</sup> <sub>SNP</sub>	LDSC Intercept
Educational Attainment <sup>37</sup>	217,569	-	-	217,569	.10	1.02
<b>Personality:</b>						
Openness <sup>63,64</sup>	76,551	-	-	76,551	.10	.95
Conscientiousness <sup>63,64</sup>	76,551	-	-	76,551	.08	.96
Extraversion <sup>63,64</sup>	76,600	-	-	76,600	.14	.92
Agreeableness <sup>63,64</sup>	76,548	-	-	76,548	.07	.96
Neuroticism <sup>63,64</sup>	76,581	-	-	76,581	.09	.94
<b>Psychiatric Traits:</b>						
Schizophrenia <sup>28</sup>	150,064	36,989	113,075	111,486.6	.18	1.05
Bipolar <sup>29</sup>	63,766	11,974	51,792	38,902.1	.14	1.02
Anorexia <sup>65</sup>	14,477	3,495	10,982	10,605	.22	1.01
Autism <sup>66</sup>	46,350	18,381	27,969	44,366.62	.11	1.01
Alzheimer <sup>67</sup>	54,162	17,008	37,154	46,668.5	.08	1.04
ADHD <sup>31</sup>	55,374	20,183	35,191	51,306.4	.21	1.03
MDD <sup>30</sup>	431,394	116,404	314,990	161,089.68	.08	1.01
<b>Substance Use:</b>						
Alcohol Use <sup>32</sup>	70,493	-	-	70,493	.05	1.02
Smoking (ever vs never) <sup>68</sup>	74,035	41,969	32,066	72,710.4	.12	1.00
Smoking (cigs per day) <sup>68</sup>	38,181	-	-	38,181	.06	1.01
Smoking (current/former) <sup>68</sup>	41,278	23,969	17,309	40,203.4	.09	1.01
Smoking (age at onset) <sup>68</sup>	24,114	-	-	24,114	.06	1.00
Cannabis (ever vs never) <sup>69</sup>	32,330	14,387	17,943	31,938.9	.13	1.00
Caffeine <sup>70</sup>	91,462	-	-	91,462	.04	1.01
<b>CAD &amp; Diabetes:</b>						
Coronary Artery Disease <sup>71</sup>	86,995	22,233	64,762	66,204	.28	.87
Type-2 Diabetes <sup>72</sup>	69,033	12,171	56,862	40,100.7	.18	1.01
<b>Reproduction:</b>						
Age at Menarche <sup>73</sup>	87,802	-	-	87,802	.24	.98
Age at Menopause <sup>74</sup>	69,360	-	-	69,360	.13	.99
<b>Antropomorphic traits:</b>						
Height <sup>75</sup>	253,280	-	-	253,280	.31	1.33
BMI <sup>76</sup>	322,154	-	-	322,154	.13	.67
Body Fat <sup>77</sup>	100,716	-	-	100,716	.10	.91
Waist Circumference <sup>78</sup>	232,101	-	-	232,101	.13	.76
Hip Circumference <sup>78</sup>	213,038	-	-	213,038	.14	.78
Waist-to-hip-ratio <sup>78</sup>	212,248	-	-	212,248	.10	.84

Supplementary Table 2: Sample sizes and LD score regression results for the regional GWASs before and after LDSC-intercept based genomic control.

Regional Phenotype	N	Before Genomic Control				After Genomic Control (LDSC intercept=1; ratio=0)	
		SNP-h2 (SE) %	lambda	LDSC intercept (SE)	ratio	SNP-h2 (SE) %	lambda
2011 - Educational attainment (local authorities)	402,552	4.1 (.25)	1.49	1.22 (.01)	.39 (.02)	3.3 (.20)	1.23
2011 - Educational attainment (MSOA)	416,061	7.5 (.29)	1.63	1.14 (.01)	.18 (.01)	6.6 (.25)	1.44
Coal Fields (Birth Place)	422,757	1.5 (.20)	1.56	1.43 (.01)	.77 (.02)	1.0 (.12)	1.07
Coal Fields (Current Address)	449,972	2.1 (.18)	1.43	1.25 (.01)	.55 (.02)	1.7 (.13)	1.14
2011 - Proportion of non-religious individuals	416,061	1.2 (.17)	1.31	1.19 (.01)	.67 (.03)	1.0 (.13)	1.08
2016 - Brexit	446,910	3.0 (.22)	1.56	1.31 (.01)	.52 (.02)	2.3 (.16)	1.19
2016 - Non-voters	446,910	1.3 (.17)	1.20	1.12 (.01)	.50 (.04)	1.2 (.14)	1.09
1970 - Labour Party (centre-left)	422,189	2.3 (.22)	1.49	1.30 (.01)	.60 (.02)	1.8 (.15)	1.14
1970 - Liberal Democrats (centre)	207,045	0.9 (.30)	1.15	1.12 (.01)	.75 (.05)	0.8 (.25)	1.04
1970 - Conservative and Unionist Party (centre-right)	421,940	1.6 (.18)	1.37	1.27 (.01)	.66 (.02)	1.3 (.13)	1.10
1970 - Non-voters	422,415	0.6 (.17)	1.31	1.28 (.01)	.83 (.02)	0.5 (.12)	1.04
2015 - Green Party of England and Wales (left-wing)	449,553	2.2 (.16)	1.25	1.08 (.01)	.27 (.03)	2.0 (.14)	1.18
2015 - Labour Party (centre-left)	449,553	1.8 (.19)	1.37	1.22 (.01)	.57 (.02)	1.4 (.14)	1.13
2015 - Liberal Democrats (centre)	449,553	1.2 (.15)	1.20	1.08 (.01)	.42 (.05)	1.1 (.14)	1.09
2015 - Conservative and Unionist Party (centre-right)	449,553	1.5 (.17)	1.31	1.17 (.01)	.56 (.03)	1.2 (.14)	1.11
2015 - UKIP (right-wing)	449,553	3.1 (.21)	1.56	1.29 (.01)	.50 (.02)	2.4 (.15)	1.20
2015 - Non-voters	449,553	2.6 (.19)	1.43	1.19 (.01)	.43 (.02)	2.2 (.15)	1.19
2015 - Green/(Green+Liberals+Conservative)	449,553	1.2 (.14)	1.15	1.06 (.01)	.31 (.04)	1.2 (.13)	1.11
2015 - Liberals/(Green+Liberals+Conservative)	449,553	0.3 (.12)	1.10	1.08 (.01)	.72 (.07)	0.3 (.11)	1.03
2015 - Conservative /(Green+Liberals+Conservative)	449,553	0.9 (.13)	1.15	1.08 (.01)	.51 (.05)	0.8 (.11)	1.07
2015 - Labour/(UKIP+Labour)	449,553	0.7 (.14)	1.20	1.15 (.01)	.70 (.04)	0.6 (.11)	1.05
2015 - UKIP/(UKIP+Labour)	449,553	0.7 (.14)	1.20	1.15 (.01)	.70 (.04)	0.6 (.11)	1.05

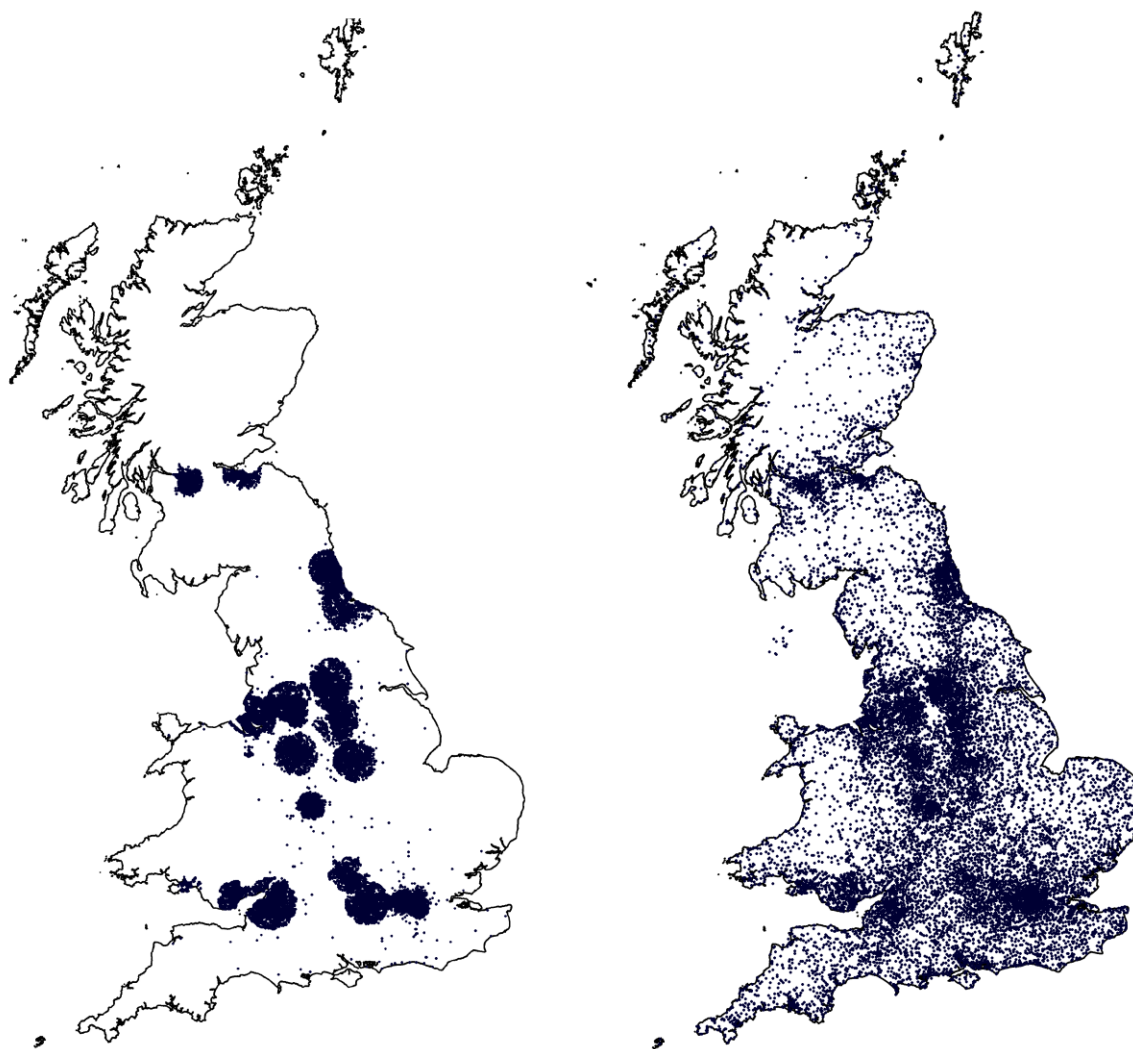
Supplementary Table 3: Independent SNPs with  $p < 5 \times 10^{-8}$ , with independence based on an  $r^2$  threshold of 0.1.

Election Phenotype	SNP	Chr	BP (hg19)	p-value	MAF	Gene
Coal Mining (Birth Place)	rs143252507	17	19970702	$2.5 \times 10^{-8}$	.01	SPECC1
2016 - Brexit	<b>rs7613360</b>	<b>3</b>	<b>49916710</b>	<b><math>9.3 \times 10^{-10}</math></b>	<b>.40</b>	<b>ACTBP13</b>
	rs1050450	3	49394834	$2.5 \times 10^{-8}$	.31	GPX1
	rs6130360	20	42010996	$3.1 \times 10^{-8}$	.15	
1970 - Labour Party (centre-left)	rs7159181	14	39229004	$1.4 \times 10^{-8}$	.44	LINC00639
2015 - Green Party of England and Wales (left-wing)	<b>rs2624838</b>	<b>3</b>	<b>50205642</b>	<b><math>8.3 \times 10^{-10}</math></b>	<b>.34</b>	<b>SEMA3F</b>
	<b>rs13135092</b>	<b>4</b>	<b>103198082</b>	<b><math>1.1 \times 10^{-9}</math></b>	<b>.08</b>	<b>SLC39A8</b>
2015 - Conservative and Unionist Party (centre-right)	rs61278749	10	106577845	$3.0 \times 10^{-8}$	.13	SORCS3
2015 - UKIP (right-wing)	<b>rs9827708</b>	<b>3</b>	<b>49649989</b>	<b><math>2.4 \times 10^{-11}</math></b>	<b>.30</b>	<b>BSN</b>
	rs2624848	3	50165101	$2.0 \times 10^{-8}$	.44	
2015 - Non-voters	<b>rs7548936</b>	<b>1</b>	<b>91207757</b>	<b><math>8.4 \times 10^{-11}</math></b>	<b>.37</b>	
	rs79003713	5	138135234	$3.6 \times 10^{-8}$	.06	CAP102

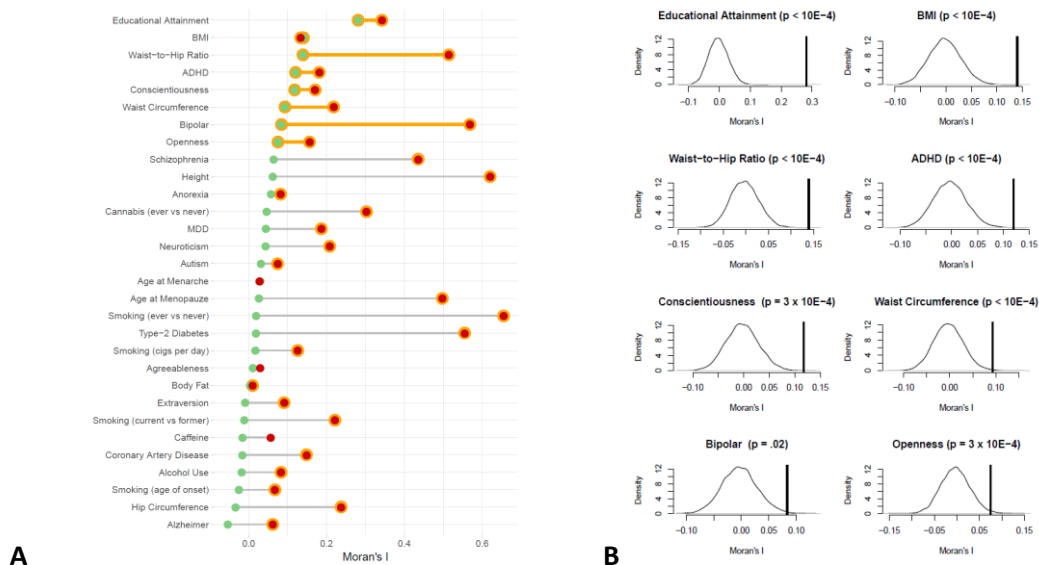
*Bold:  $p < 1 \times 10^{-8}$*

Supplementary Table 4: Overview of GWASs that provided the effect size estimates that were used to compute genetic correlations with regional GWASs using LD score regression.

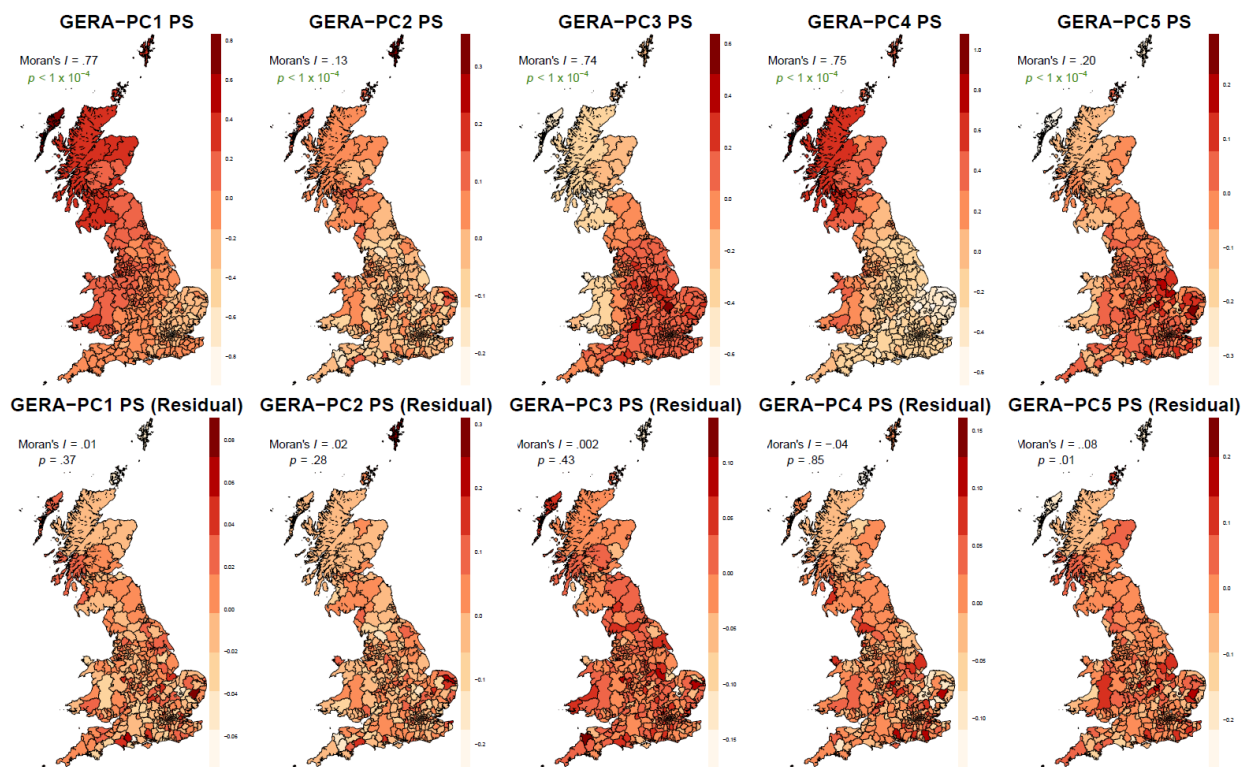
Trait Name	NGWAS	Ncases	Ncontrols	Neffective	h <sup>2</sup> <sub>SNP</sub>	LDSC Intercept
<b>Cognition and SES:</b>						
Childhood IQ <sup>79</sup>	12,441	-	-	12,441	.28	1.00
Adult IQ <sup>80</sup>	78,308	-	-	78,308	.19	1.02
Educational Attainment <sup>37</sup>	217,569	-	-	217,569	.10	1.02
Income <sup>81</sup>	112,151	-	-	112,151	.06	1.03
Townsend <sup>81</sup>	112,151	-	-	112,151	.04	1.01
<b>Personality:</b>						
Openness <sup>63,64</sup>	76,551	-	-	76,551	.10	.95
Conscientiousness <sup>63,64</sup>	76,551	-	-	76,551	.08	.96
Extraversion <sup>63,64</sup>	76,600	-	-	76,600	.14	.92
Agreeableness <sup>63,64</sup>	76,548	-	-	76,548	.07	.96
Neuroticism <sup>63,64</sup>	76,581	-	-	76,581	.09	.94
<b>Other Psychology:</b>						
Subjective Well-being <sup>82</sup>	298,420	-	-	298,420	.03	1.00
Tiredness <sup>83</sup>	108,976	-	-	108,976	.07	.99
Loneliness <sup>84</sup>	332,991	-	-	332,991	.08	1.01
Risk Taking <sup>84</sup>	332,991	-	-	332,991		
<b>Psychiatric Traits:</b>						
Schizophrenia <sup>28</sup>	150,064	36,989	113,075	111,486.6	.18	1.05
Bipolar <sup>29</sup>	63,766	11,974	51,792	38,902.1	.14	1.02
Anorexia <sup>65</sup>	14,477	3,495	10,982	10,605	.22	1.01
Autism <sup>66</sup>	46,350	18,381	27,969	44,366.62	.11	1.01
Alzheimer <sup>67</sup>	54,162	17,008	37,154	46,668.5	.08	1.04
ADHD <sup>31</sup>	55,374	20,183	35,191	51,306.4	.21	1.03
MDD <sup>30</sup>	431,394	116,404	314,990	161,089.68	.08	1.01
<b>Substance Use:</b>						
Alcohol Consumption <sup>85</sup>	112,117	-	-	112,117	.08	1.01
Alcohol Dependence <sup>86</sup>	46,568	11,569	34,999	34,779.54	.18	1.02
Smoking (ever vs never) <sup>68</sup>	74,035	41,969	32,066	72,710.4	.12	1.00
Smoking (cigs per day) <sup>68</sup>	38,181	-	-	38,181	.06	1.01
Smoking (current/former) <sup>68</sup>	41,278	23,969	17,309	40,203.4	.09	1.01
Smoking (age at onset) <sup>68</sup>	24,114	-	-	24,114	.06	1.00
Cannabis (ever vs never) <sup>69</sup>	32,330	14,387	17,943	31,938.9	.13	1.00
Caffeine <sup>70</sup>	91,462	-	-	91,462	.04	1.01
<b>Health &amp; Longevity:</b>						
Self-rated health <sup>87</sup>	111,749	-	-	111,749		
Coronary Artery Disease <sup>71</sup>	86,995	22,233	64,762	66,204	.28	.87
Type-2 Diabetes <sup>72</sup>	69,033	12,171	56,862	40,100.7	.18	1.01
Age of Parents Death <sup>88</sup>	45,627	-	-	45,627	.05	1.02
Age of Fathers Death <sup>88</sup>	63,775	-	-	63,775	.05	1.02
Age of Mothers Death <sup>88</sup>	52,776	-	-	52,776	.05	1.01
<b>Reproduction:</b>						
Age at First Birth <sup>89</sup>	251,151	-	-	251,151	.05	.96
Nr of Children Ever Born <sup>89</sup>	343,072	-	-	343,072	.02	.97
Age at Menarche <sup>73</sup>	87,802	-	-	87,802	.24	.98
Age at Menopause <sup>74</sup>	69,360	-	-	69,360	.13	.99
<b>Antropomorphic:</b>						
BMI <sup>76</sup>	322,154	-	-	322,154	.13	.67
Height <sup>75</sup>	253,280	-	-	253,280	.31	1.33
Body Fat <sup>77</sup>	100,716	-	-	100,716	.10	.91
Waist Circumference <sup>78</sup>	232,101	-	-	232,101	.13	.76
Hip Circumference <sup>78</sup>	213,038	-	-	213,038	.14	.78
Waist-to-hip-ratio <sup>78</sup>	212,248	-	-	212,248	.10	.84



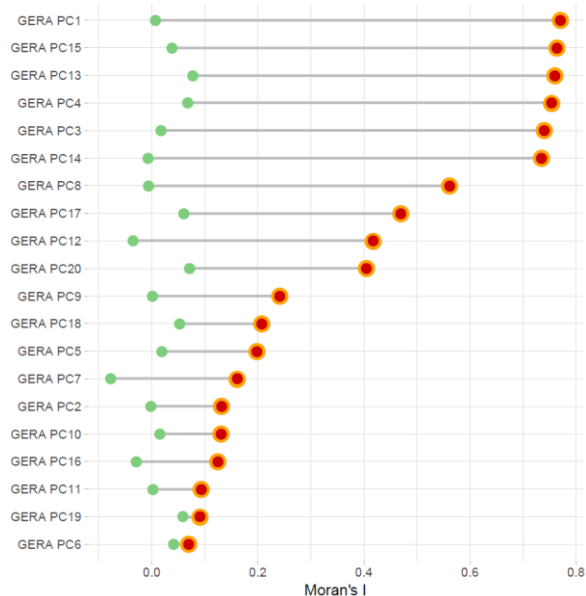
*Supplementary Figure 1: The geographic locations of the UK Biobank participants (each dot represents a participant). The left map shows the current living address (N=497,673). The right plot shows the birth places (N=444,992).*



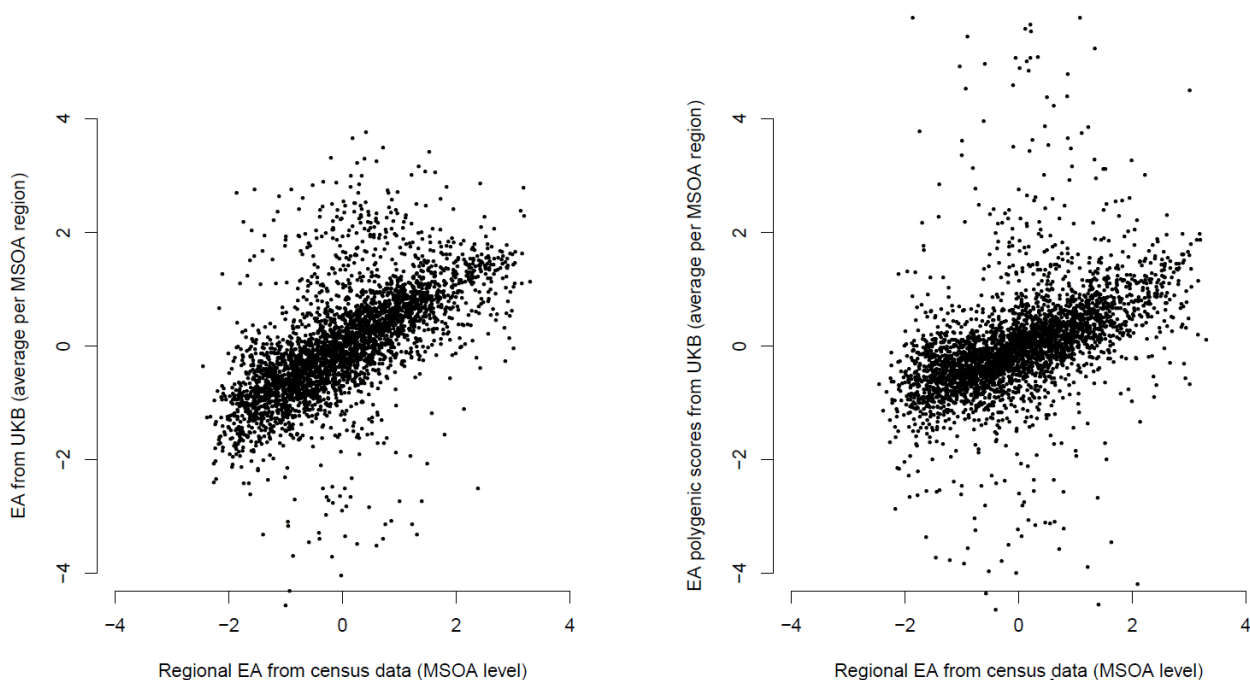
Supplementary Figure 2: Moran's I of 30 polygenic scores based on independent SNPs with  $p < .05$ . Moran's I were computed using the average polygenic score per region in 378 local authority regions. **A** shows the Moran's I of the polygenic scores uncorrected for PCs (red) and corrected for 100 PCs (green), where orange means a significant FDR corrected p-value  $< .05$  (corrected for 30 tests). **B** shows the permutation distributions for the SBLUP polygenic scores that have an FDR corrected p-value  $< .05$  compared to the observed Moran's I (vertical line to the right of the permutation distribution).



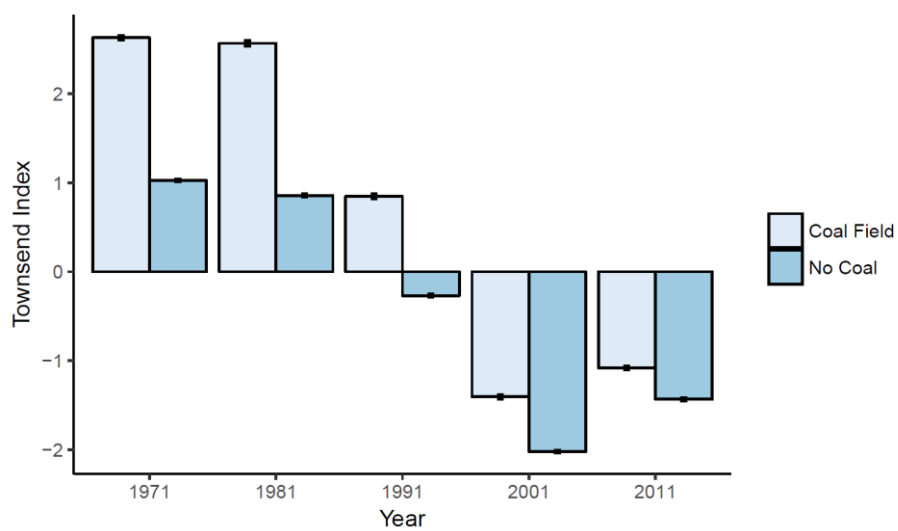
Supplementary Figure 3: Geographic distribution and Moran's I values for polygenic scores (PS) based on GWASs on the first 5 ancestry-informative PCs from the GERA dataset. The upper five maps display uncorrected polygenic score, while the five maps below display the residuals of the polygenic scores after regressing out 100 PCs. Green p-values are significant after FDR correction.



Supplementary Figure 4: Moran's  $I$  of polygenic scores based on GWASs on the first 5 ancestry-informative PCs from the GERA dataset computed using the average polygenic score per region in 378 local authority regions. The Figure shows the Moran's  $I$  of the polygenic scores uncorrected for PCs (red) and corrected for 100 PCs (green), where orange means a significant FDR corrected  $p$ -value  $< .05$  (corrected for 20 tests).

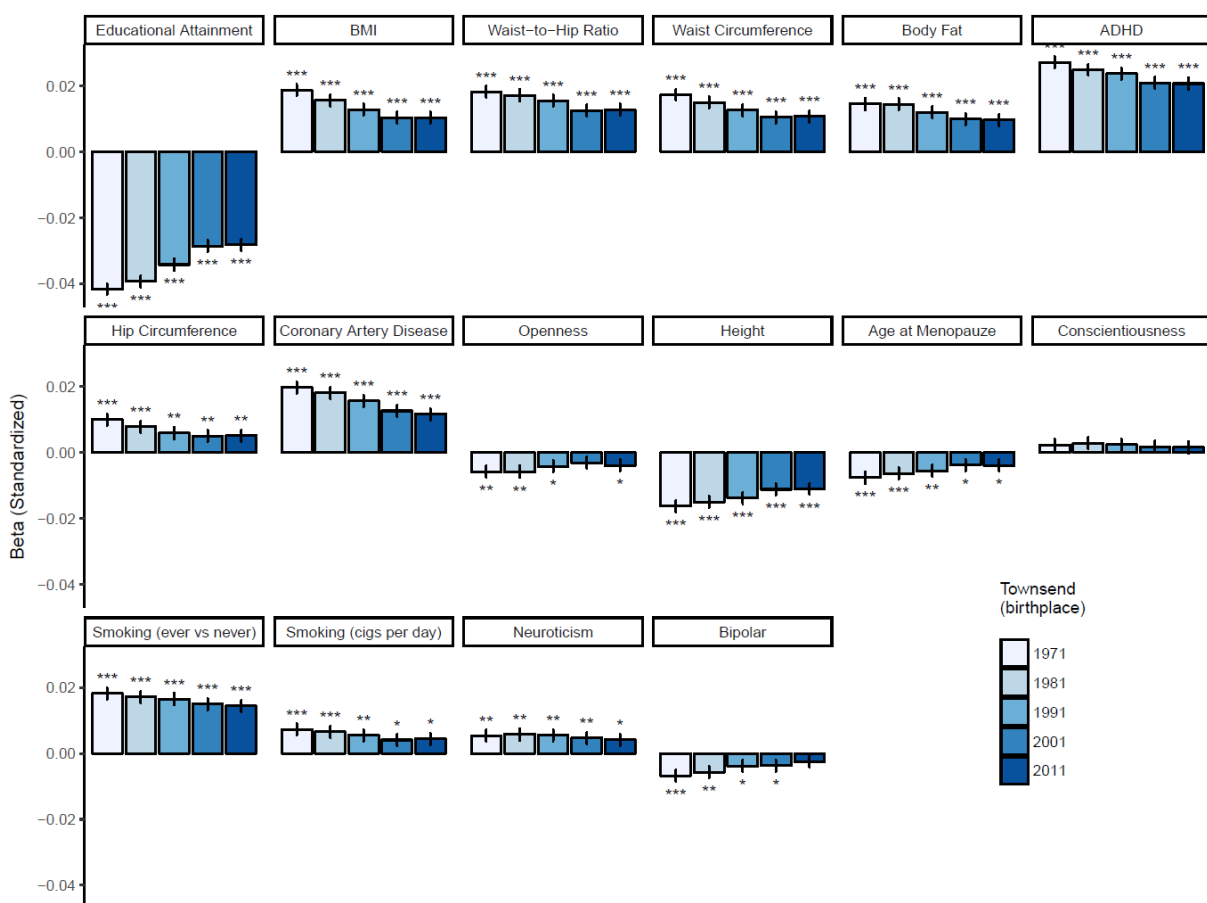


Supplementary Figure 5: Regional educational attainment (EA) on MSOA level obtained from census data (Office of National Statistics) plotted against the average EA in UK Biobank corrected for age, year of birth, and sex (left plot;  $R^2 = 40\%$ ) and EA polygenic scores corrected for 100 PCs (right plot;  $R^2 = 20\%$ ). All measures have been standardized to have mean 0 and SD 1.

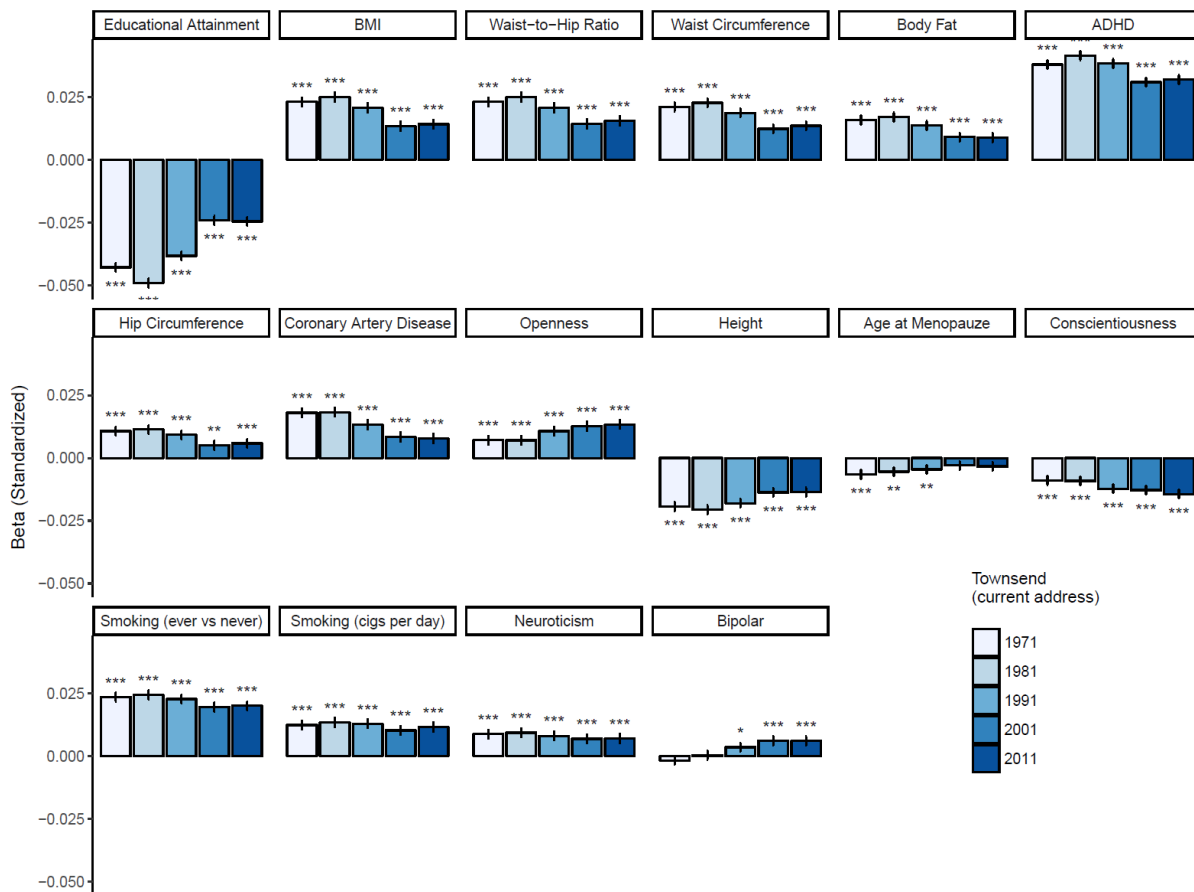


Supplementary Figure 6: The average Townsend indices from 1971 to 2011 for coal fields and regions without coal.

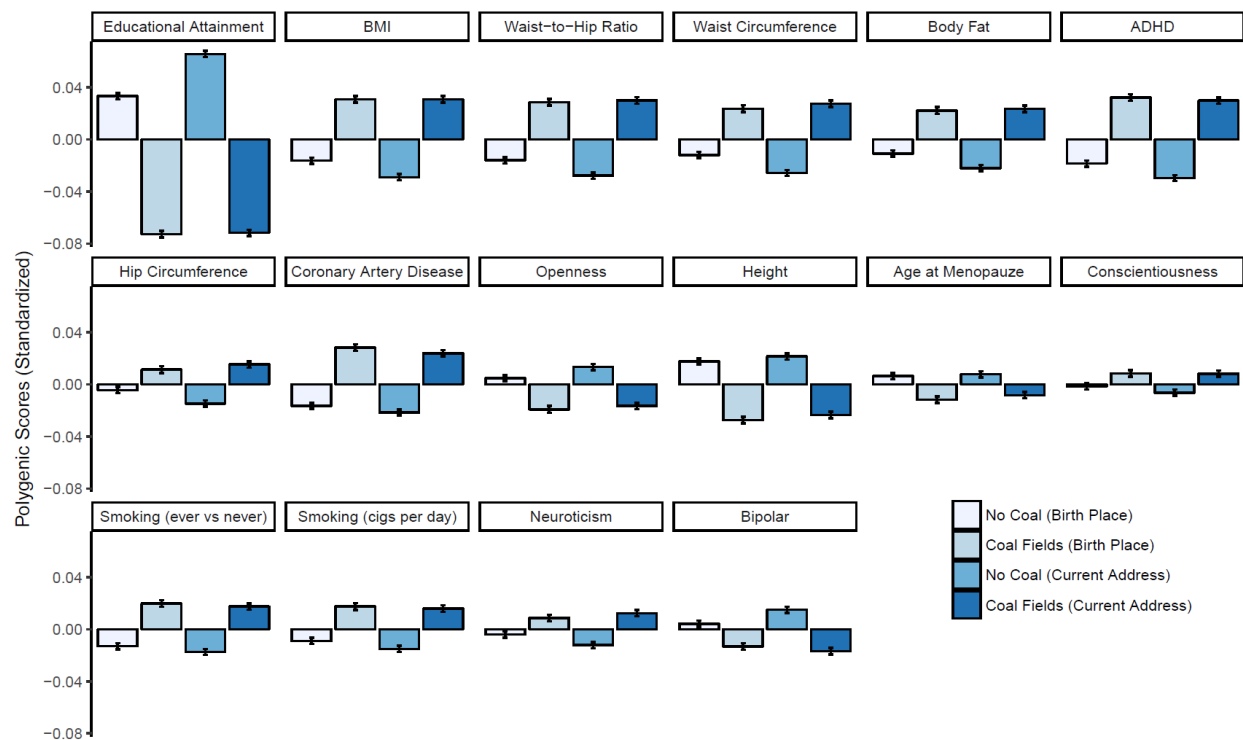




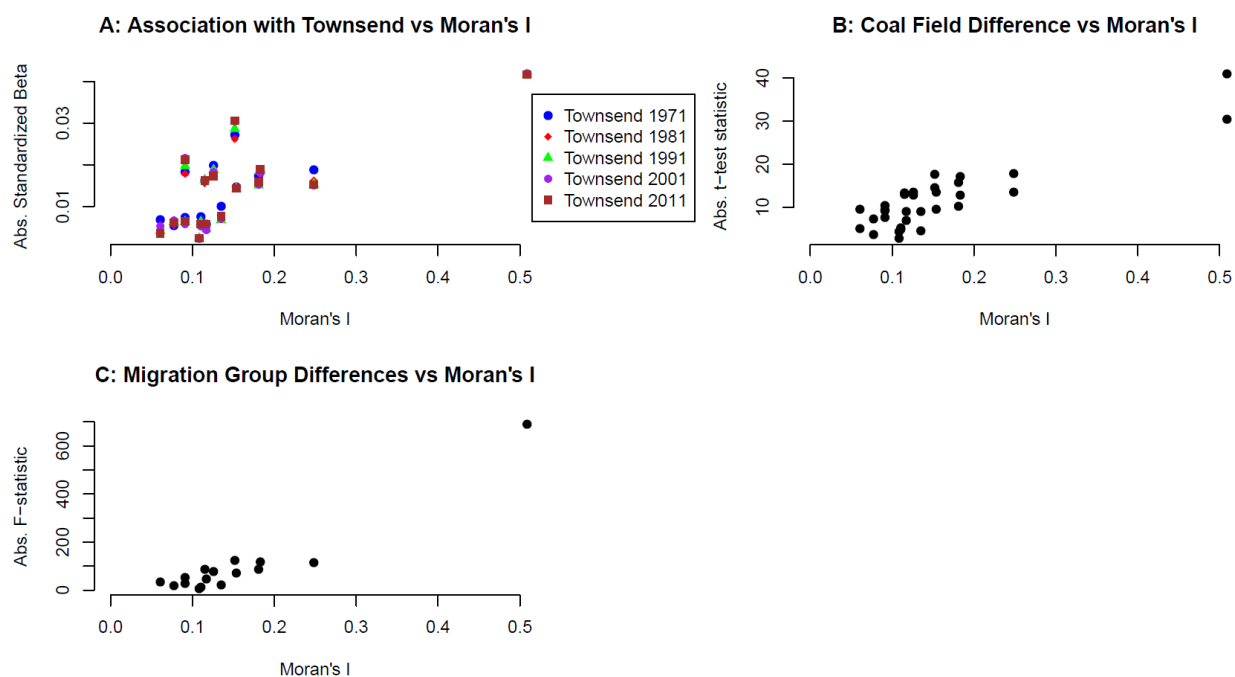
Supplementary Figure 7: The standardized regression coefficients and standard errors for the 16 geographically clustered polygenic scores (ordered by Moran's I) of the associations with the Townsend indices of 1971 to 2011 of the birth places of the subjects ( $N = 349,982$  unrelated subjects). All polygenic scores shown are standardized residuals after regressing out 100 PCs. \* =  $p < .05$ , \*\* =  $p < .01$ , \*\*\* =  $p < .001$ .



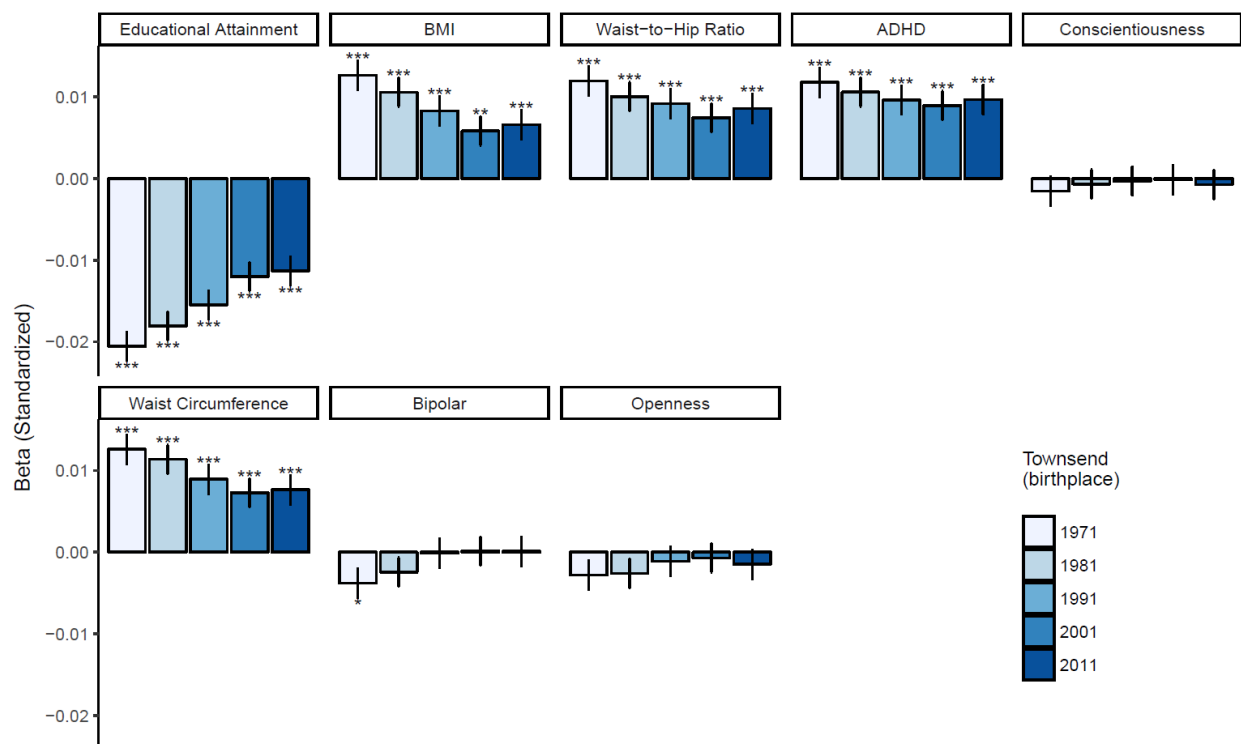
Supplementary Figure 8: The standardized regression coefficients and standard errors for the 16 geographically clustered polygenic scores (ordered by Moran's I) of the associations with the Townsend indices of 1971 to 2011 of the current address of the subjects ( $N = 349,982$  unrelated subjects). All polygenic scores shown are standardized residuals after regressing out 100 PCs. \* =  $p < .05$ , \*\* =  $p < .01$ , \*\*\* =  $p < .001$ .



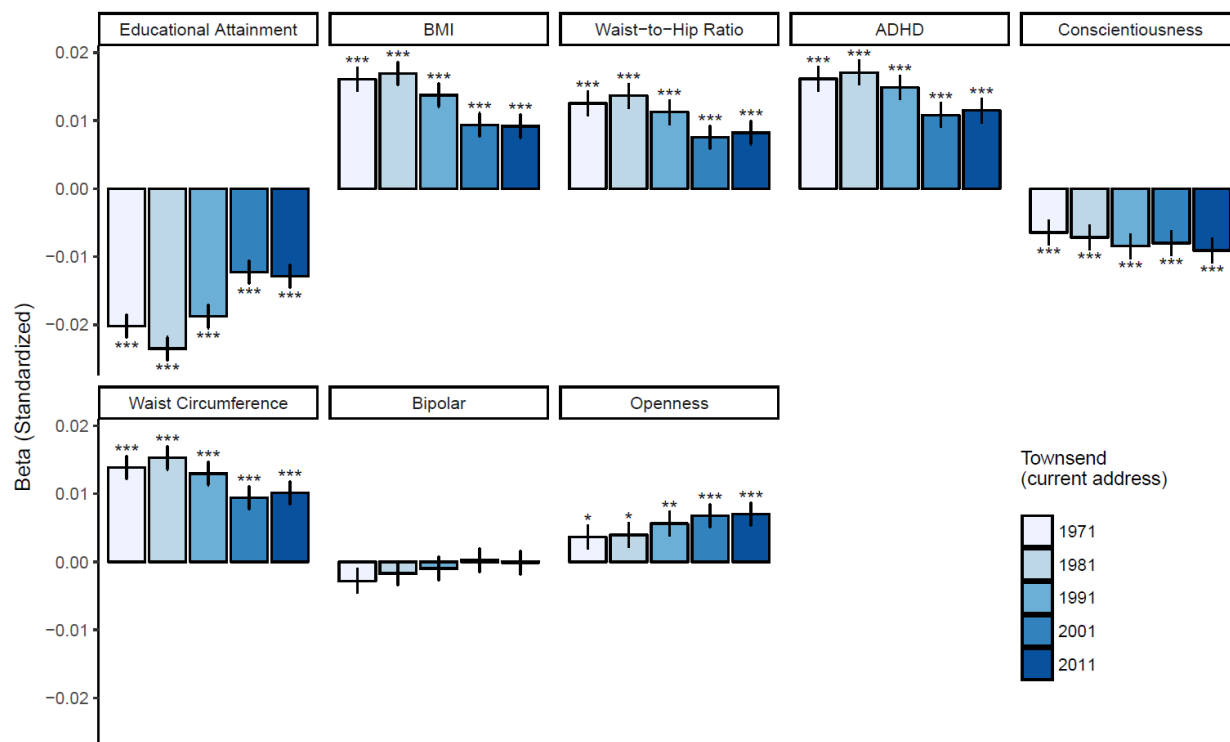
Supplementary Figure 9: The average and standard errors for the 16 geographically clustered polygenic scores (ordered by Moran's I) within coal mining regions vs the rest of Great Britain based on the birth place of the participants and on the current addresses separately. All polygenic scores shown are standardized residuals after regressing out 100 ancestry-informative PCs. The differences between Coal Fields and No Coal are all significant with an FDR corrected p-value < .05.



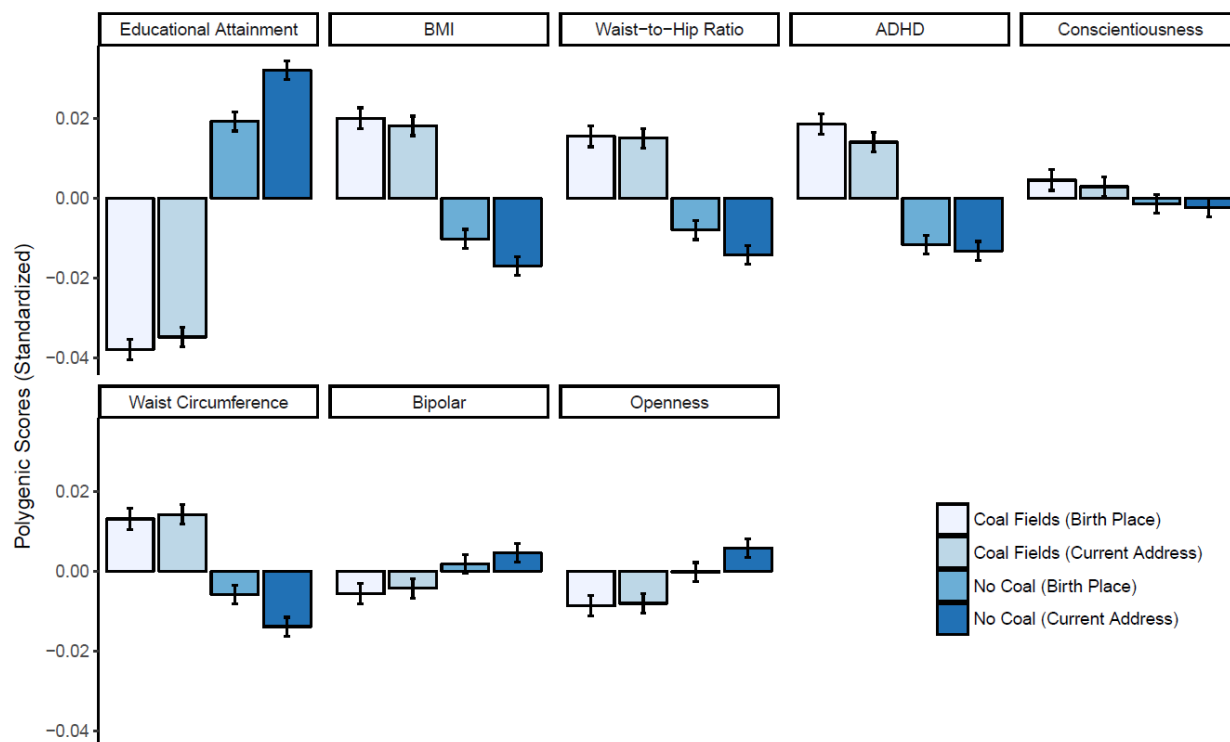
Supplementary Figure 10: Scatterplots showing the relationships between Moran's I of the significantly clustering polygenic scores and their association with indicators of economic deprivation (Townsend and the presence of coal fields) and migration. **A** shows on the y-axis the absolute regression coefficients of the regressions with birthplace Townsend shown in Supplementary Figure 3. The correlations between the absolute regression coefficients and Moran's I are .82 for 1971 ( $p = 9 \times 10^{-5}$ ), .82 for 1981 ( $p = 9 \times 10^{-5}$ ), .78 for 1991 ( $p = 4 \times 10^{-4}$ ), .75 for 2001 ( $p = 7 \times 10^{-4}$ ), .77 for 2011 ( $p = 5 \times 10^{-4}$ ). Excluding the outlier educational attainment, the correlations between the absolute regression coefficients and Moran's I are .57 for 1971 ( $p = .03$ ), .52 for 1981 ( $p = .04$ ), .47 for 1991 ( $p = .08$ ), .42 for 2001 ( $p = .12$ ), .46 for 2011 ( $p = .08$ ). **B** shows on the y-axis the absolute test statistic of the t-test for group differences between coal fields and the rest of Great Britain (based on birth place) from Supplementary Figure 5. The correlation between the absolute test statistic and Moran's I is .89 ( $p = 9 \times 10^{-12}$ ). Excluding the outlier educational attainment, the correlations between the absolute test statistic and Moran's I are .65 ( $p = 1 \times 10^{-4}$ ). **C** shows on the y-axis the absolute test statistic of the ANOVA for group differences between the four migration groups from Figure 5. The correlation between the absolute test statistic and Moran's I is .95 ( $p = 2 \times 10^{-8}$ ). Excluding the outlier educational attainment, the correlations between the absolute test statistic and Moran's I are .73 ( $p = .002$ ).



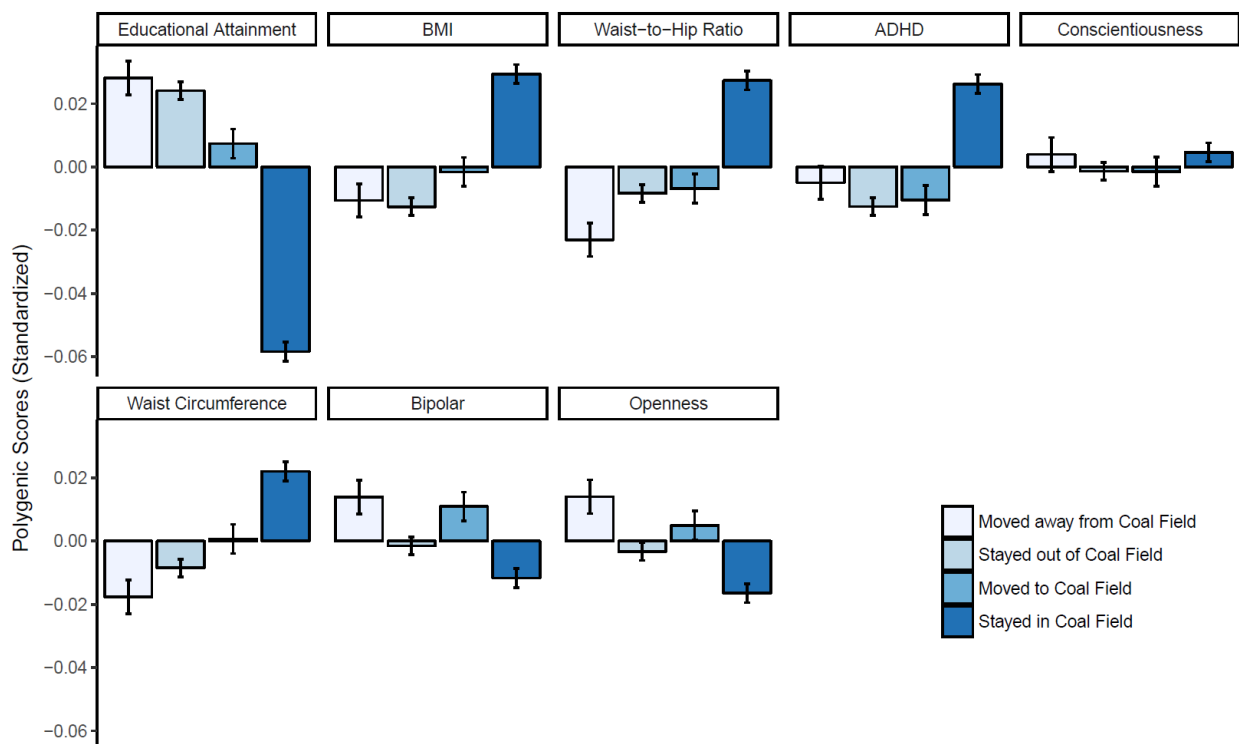
Supplementary Figure 11: The standardized regression coefficients and standard errors for the 8 significantly clustering polygenic scores (clumped, i.e., based on independent SNPs with  $p$ -values  $< .05$ ; ordered by Moran's  $I$ ) of the associations with the Townsend indices of 1971 to 2011 of the birth places of the subjects ( $N = 349,982$  unrelated subjects). All polygenic scores shown are standardized residuals after regressing out 100 PCs.



Supplementary Figure 12: The standardized regression coefficients and standard errors for the 8 significantly clustering polygenic scores (clumped, i.e., based on independent SNPs with  $p$ -values  $< .05$ ; ordered by Moran's  $I$ ) of the associations with the Townsend indices of 1971 to 2011 of the current address of the subjects ( $N = 349,982$  unrelated subjects). All polygenic scores shown are standardized residuals after regressing out 100 PCs.

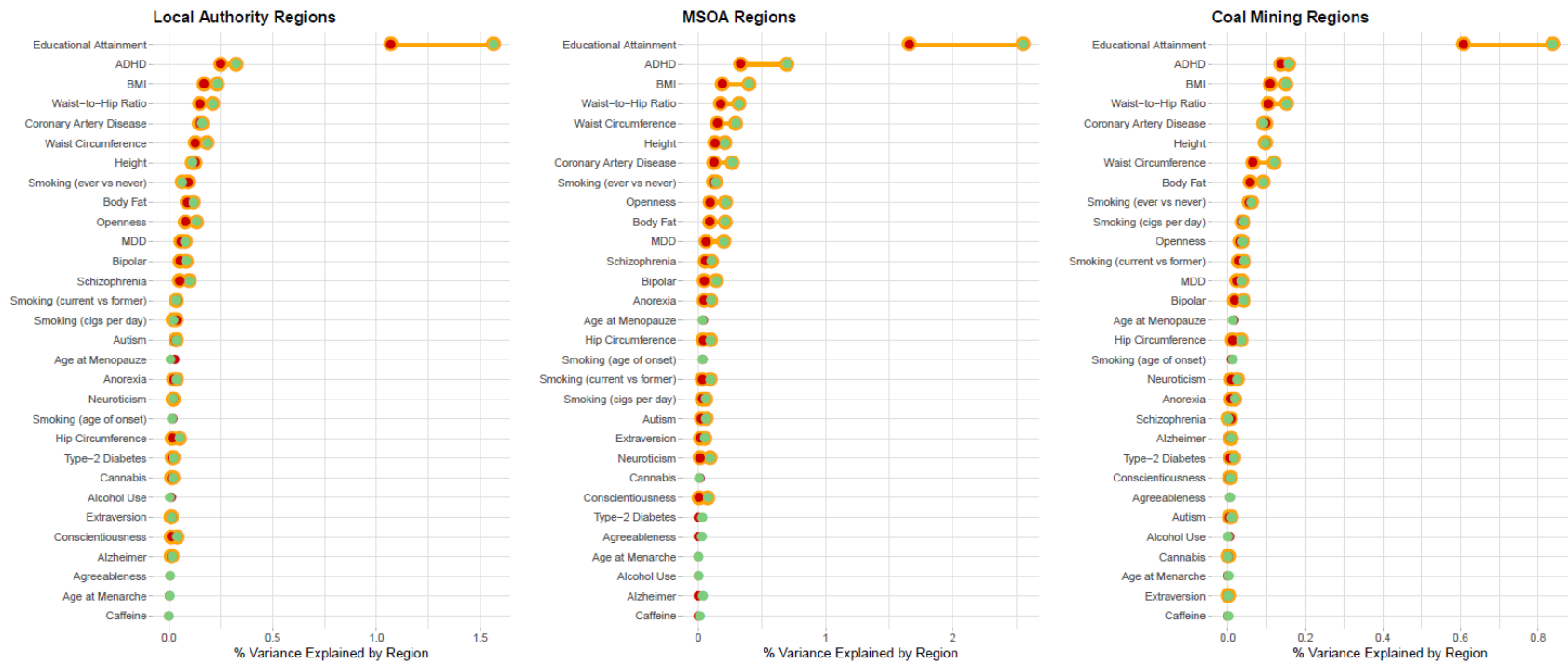


Supplementary Figure 13: The average and standard errors for 8 significantly clustering polygenic scores (clumped, i.e., based on independent SNPs with  $p$ -values  $<.05$ ; ordered by Moran's  $I$ ) within coal mining regions vs the rest of Great Britain for the birth place and the current address. All polygenic scores shown are standardized residuals after regressing out 100 ancestry-informative PCs. The differences between Coal Fields and No Coal are significant for all scores except conscientiousness for both current address and birth place with an FDR corrected  $p$ -value  $<.05$ .

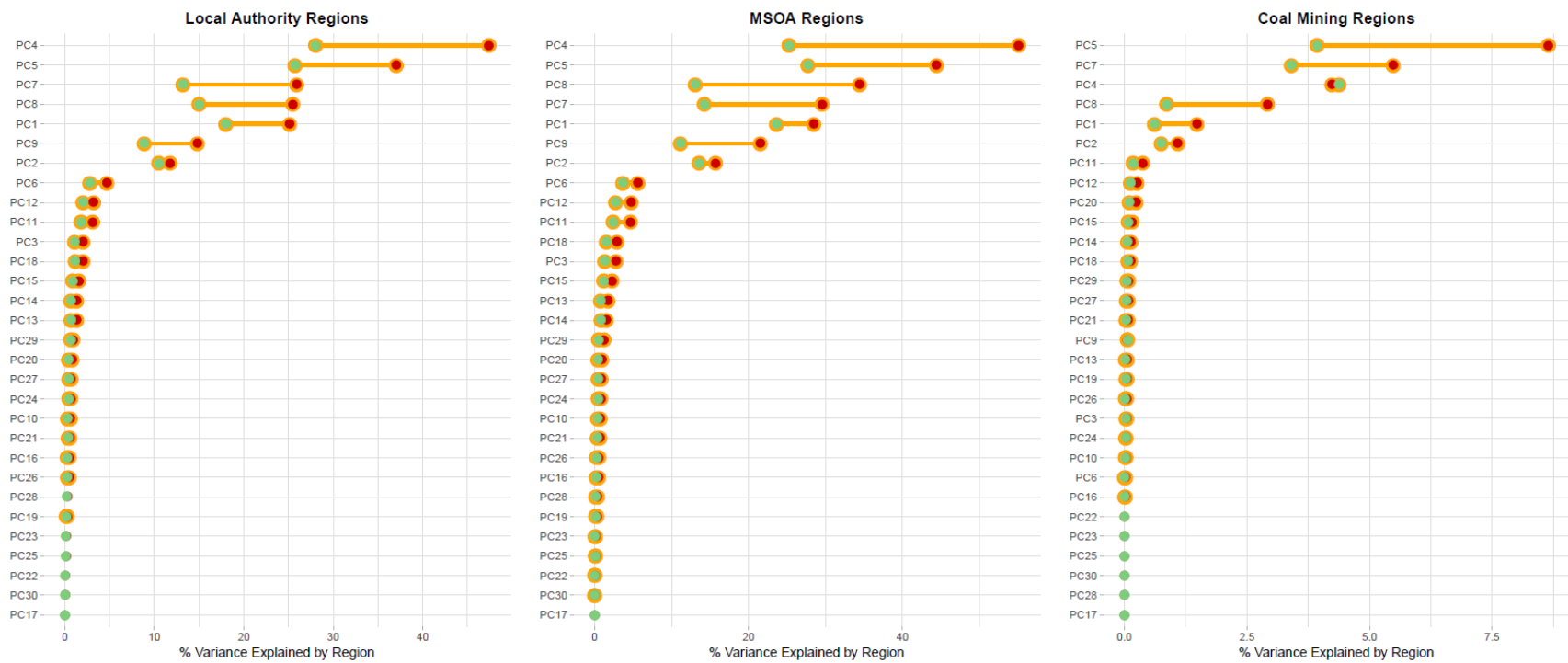


Supplementary Figure 14: The average and standard errors for 8 significantly clustering polygenic scores (clumped, i.e., based on independent SNPs with  $p$ -values  $< .05$ ; ordered by Moran's  $I$ ) for four migration groups: born in coal field area and moved out, born in coal field area and stayed, born outside of coal field area and moved to coal field area, born outside of coal field area and stayed out. All polygenic scores shown are standardized residuals after regressing out 100 ancestry-informative PCs. All polygenic scores show significant group differences after (FDR corrected  $p < .05$ ), except conscientiousness.

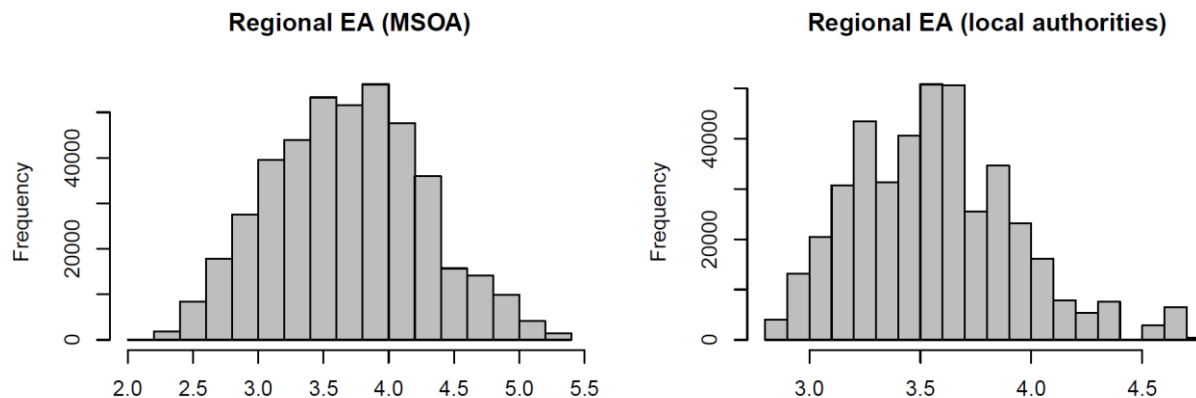




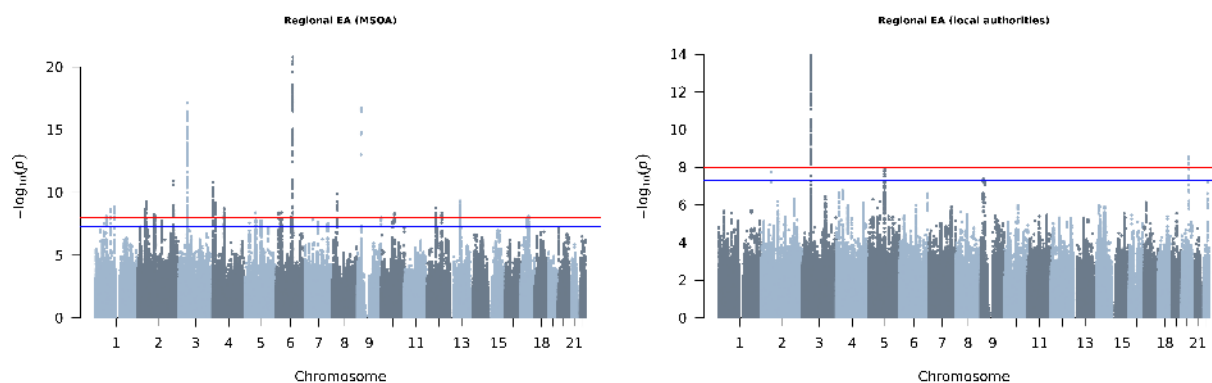
Supplementary Figure 15: Linear Mixed Model results, with polygenic score (after regressing out 100 PCs) as a dependent variable and region as random effect ( $N = 320,940$  unrelated individuals). Left: Local Authorities (~380 regions); Middle: MSOA (~5,300 regions), Right: Coal mining Regions (fitted as a binary variable). Red: Birth Place; Green: Current Address; Yellow = significant after FDR correction.



Supplementary Figure 16: 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.

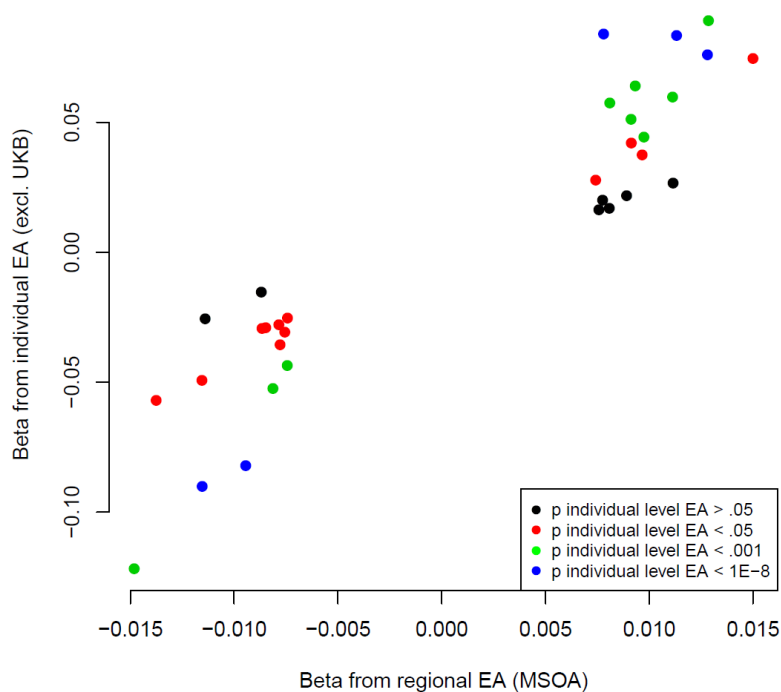


Supplementary Figure 17: Distributions of the regional-level educational attainment (EA) phenotypes. The distributions show all subjects included in the GWASs, where all subjects from the same region were assigned the same phenotypic value.

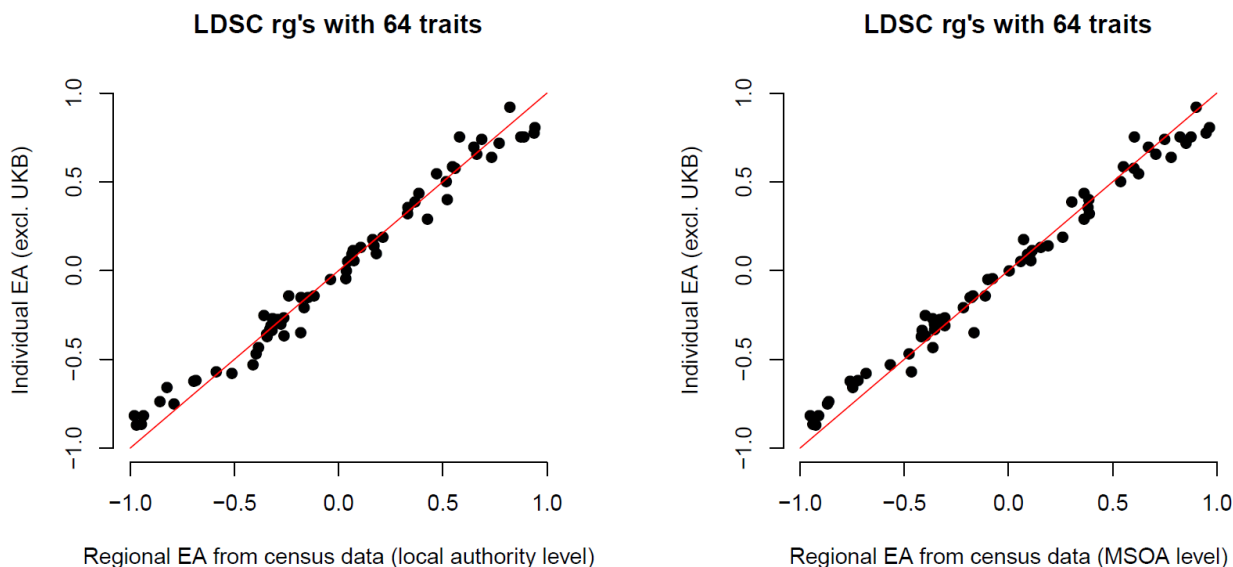


Supplementary Figure 18: Manhattan plots of the two GC-corrected GWASs on regional educational attainment (EA). The suggestive significance threshold (blue line) is set at  $5 \times 10^{-8}$ , and the genome-wide significance threshold (red line) is set at  $1 \times 10^{-8}$ .

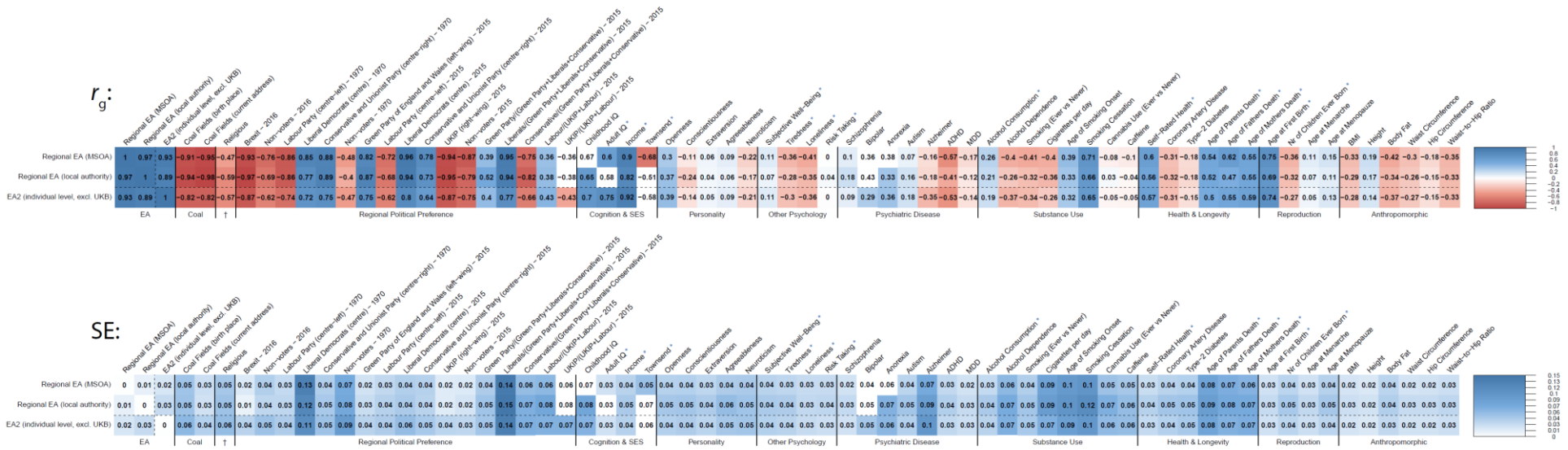
### Beta's for significant hits regional EA GWAS (MSOA)



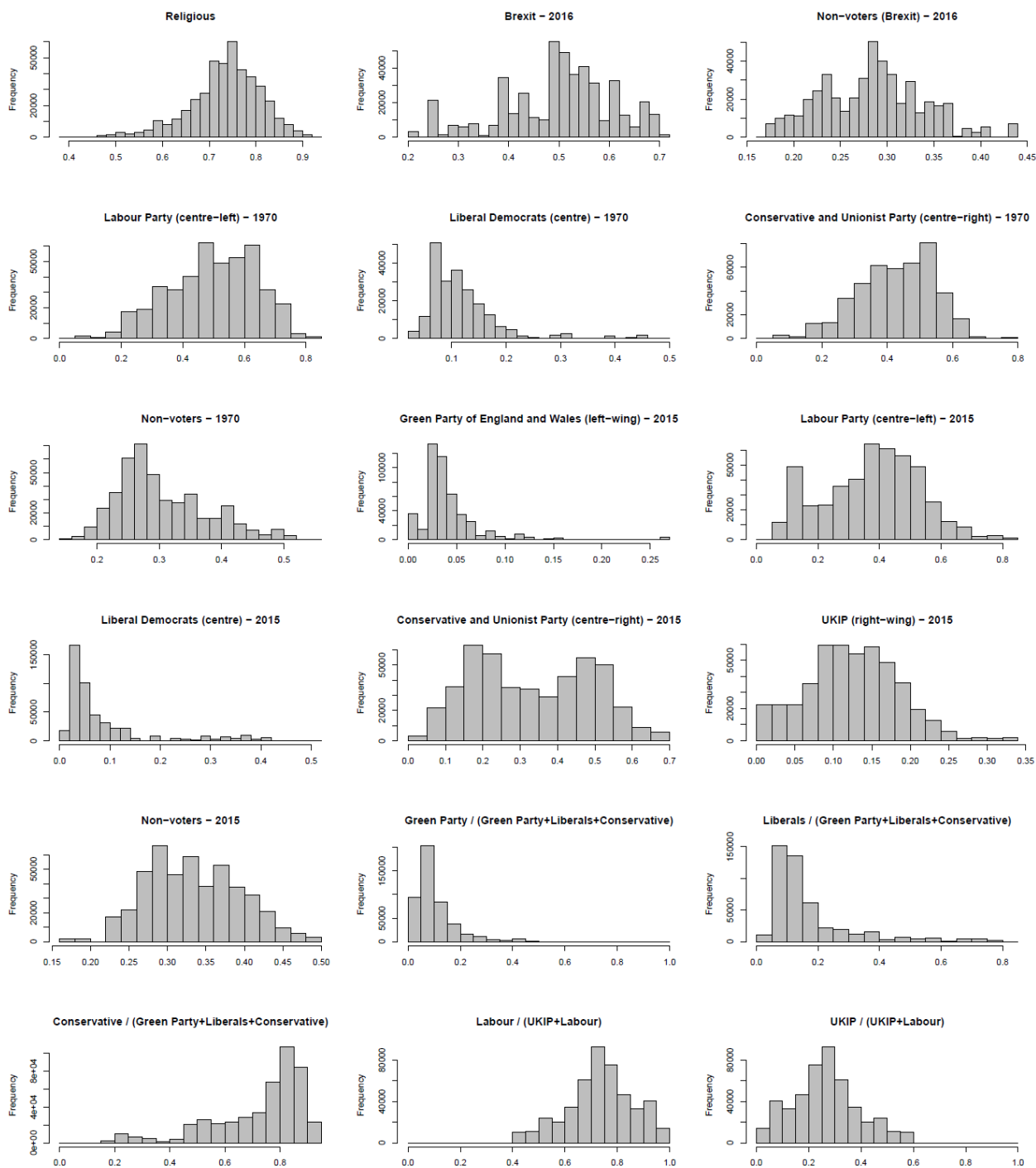
Supplementary Figure 19: Effect sizes of 33 independent SNPs ( $r^2 < .1$  and at least 1MB apart) that reached  $p < 5 \times 10^{-8}$  in the regional EA GWAS (MSOA) plotted against their effect sizes from the GWAS on individual EA that excluded UK Biobank participants. The correlation between the beta's is .93.



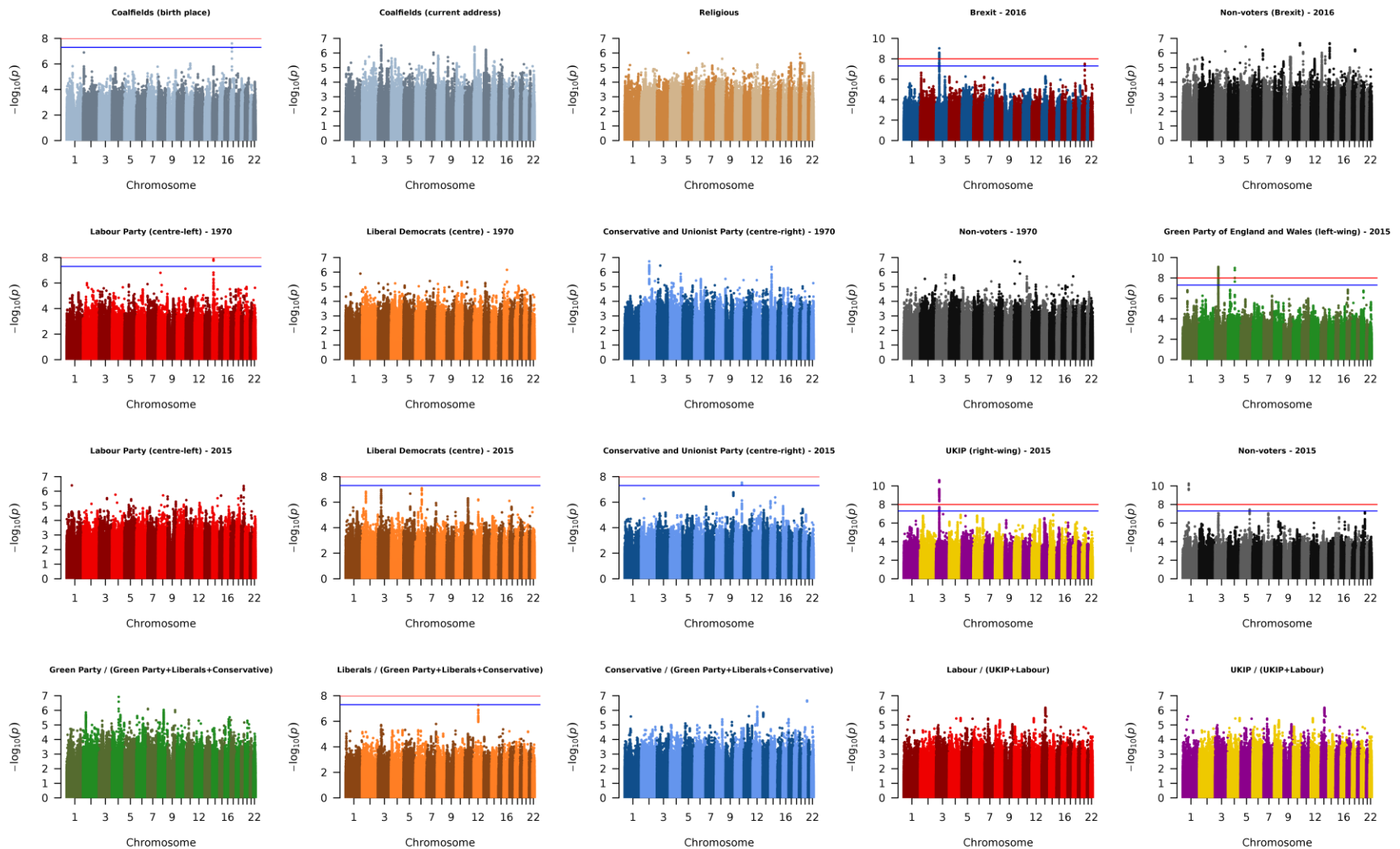
Supplementary Figure 20: Scatterplots of the genetic correlations with 64 complex traits from Supplementary Figure 16. Each dot represents a complex trait. The correlations between the  $rg$ 's from the regional EA GWAS at local authority level and the individual EA GWAS (plot to the left) is .99. The correlations between the  $rg$ 's from the regional EA GWAS at MSOA level and the individual EA GWAS (plot to the right) is also .99.



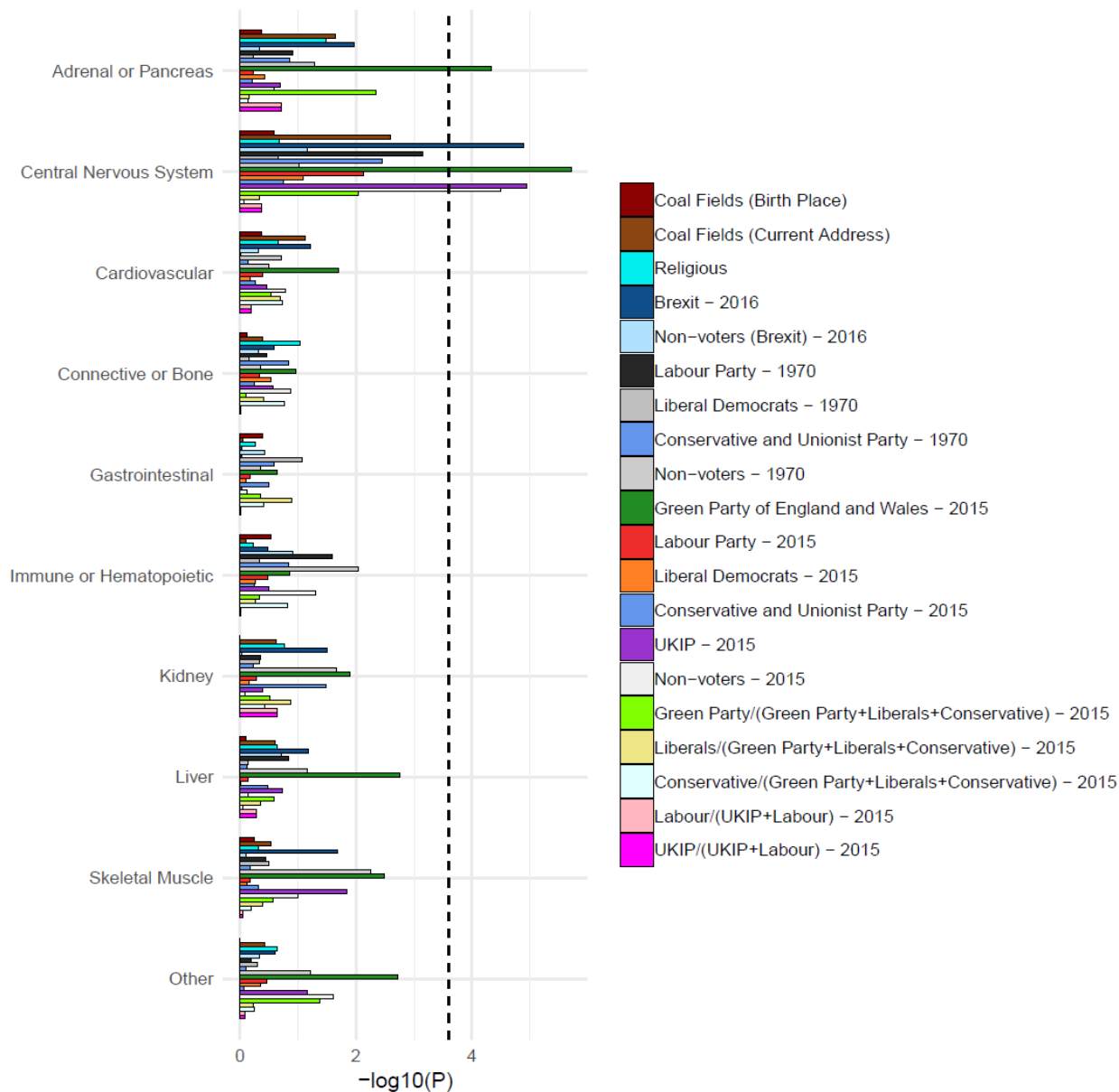
Supplementary Figure 21: Genetic correlations ( $r_g$ ) and their standard errors ( $SE$ ) based on LD score regression, for regional EA outcomes based on census data and the individual-level EA GWAS excluding UK Biobank. Colored is significant after FDR correction. The blue stars next to the trait names indicate that UK Biobank was part of the GWAS of the trait.



Supplementary Figure 22: Distributions of regional-level phenotypes. The distributions show all subjects included in the GWASs, where all subjects from the same region were assigned the same phenotypic value.

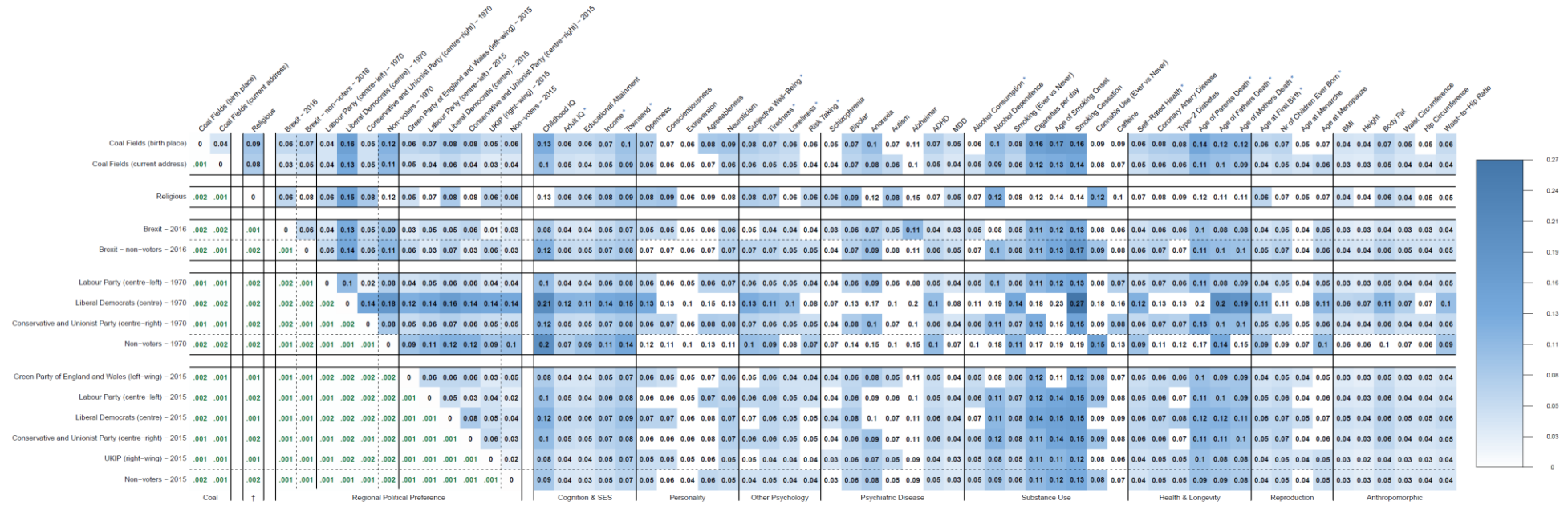


Supplementary Figure 23: Manhattan plots of the GC-corrected GWAS on the presence of coal fields in the birthplace and the current address, the proportion of religious vs non-religious inhabitants, the Brexit referendum of 2016, and the regional outcomes on general election results of 1970 and 2015. The suggestive significance threshold (blue line) is set at  $5 \times 10^{-8}$ , and the genome-wide significance threshold (red line) is set at  $1 \times 10^{-8}$ .

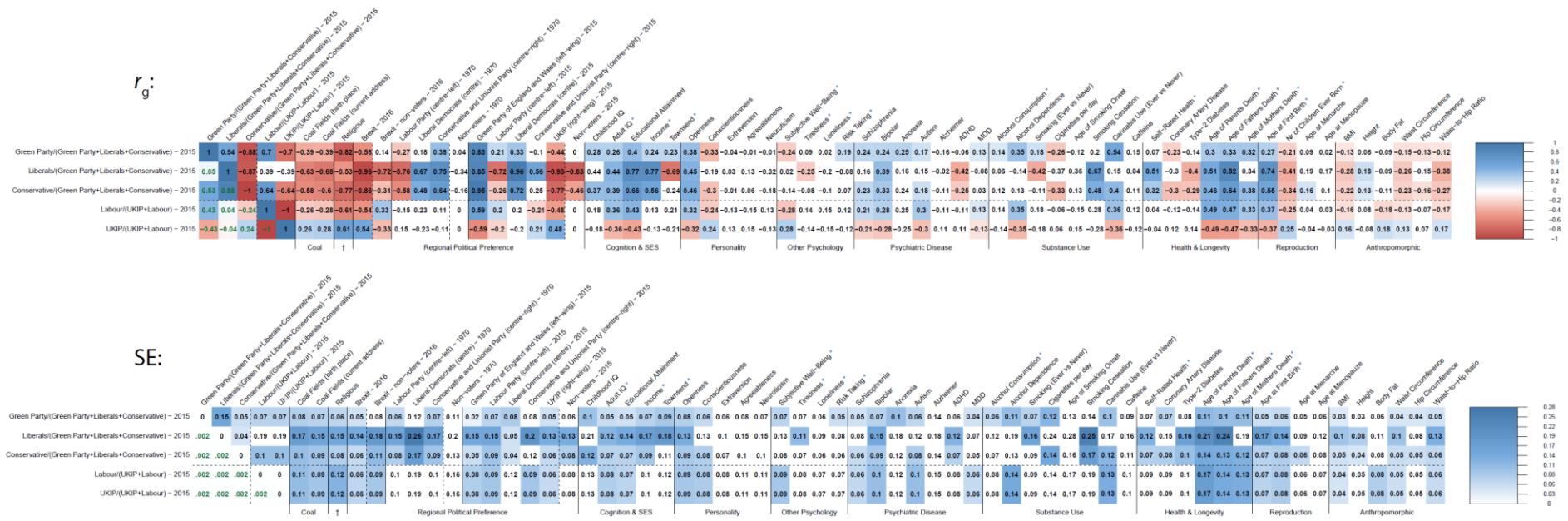


Supplementary Figure 24: The significance of enrichment for ten cell type groups based on heritability partitioning with LD score regression. The dashed line at  $-\log_{10}(P) = 3.53$  is the Bonferroni adjusted significance threshold (adjusted for 170 tests).





Supplementary Figure 25: Standard errors of the genetic correlations based on LD score regression from Figure 6. Colored is significant after FDR correction. The green numbers in the left part of the Figure below the diagonal of 1's are the standard errors of the phenotypic correlations between the regional outcomes of coal mining, religiousness, and regional political preference. The blue stars next to the trait names indicate that UK Biobank was part of the GWAS of the trait.



Supplementary Figure 26: Genetic correlations ( $r_g$ ) and their standard errors (SE) based on LD score regression between regional political outcomes, within higher and lower SES groups. Colored is significant after FDR correction. The blue stars next to the trait names indicate that UK Biobank was part of the GWAS of the trait.

