Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income

W. David Hill[1,2*], Neil M. Davies[3,4], Stuart J. Ritchie[5], Nathan G. Skene[6,7,8], Julien Bryois[9], Steven Bell[10,11,12], Emanuele Di Angelantonio[10,11,12,13], David J. Roberts[14,15,16], Shen Xueyi[17], Gail Davies[1,2], David C.M. Liewald[1,2], David J. Porteous[1,18], Caroline Hayward[19], Adam S. Butterworth[10,11,12], Andrew M. McIntosh[1,17], Catharine R. Gale[1,2,20†], Ian J. Deary[1,2†]

1 Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK
2 Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh, EH8 9JZ, UK
3 Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, BS8 2BN, UK.
4 Bristol Medical School, University of Bristol, Bristol, BS8 2BN, UK
5 Social, Genetic and Developmental Psychiatry Centre, King's College London, London, UK.
6 Laboratory of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden
7 UCL Institute of Neurology, Queen Square, London, UK.
8 Division of Brain Sciences, Department of Medicine, Imperial College, London, UK
9 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.
10 The National Institute for Health Research Blood and Transplant Unit in Donor Health and Genomics at the University of Cambridge, University of Cambridge, Strangeways Research Laboratory, Wort's Causeway, Cambridge CB1 8RN, UK.
11 UK Medical Research Council/British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Wort's Causeway, Cambridge CB1 8RN, UK.
12 British Heart Foundation Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK.
13 NHS Blood and Transplant, Cambridge, UK.
14 Cambridge Substantive Site, Health Data Research UK, Wellcome Genome Campus, Hinxton, UK
15 BRC Haematology Theme and Radcliffe Department of Medicine, University of Oxford, Oxford, UK
16 NHS Blood and Transplant–Oxford Centre, Oxford, UK
17 Division of Psychiatry, University of Edinburgh, Edinburgh, EH10 5HF, UK

18 Centre for Genomic and Experimental Medicine, Institute of Genetics & Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom
19 MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom
20 MRC Lifecourse Epidemiology Unit, University of Southampton, SO16 6YD, UK

† These authors contributed equally.

*Corresponding author,

W. David Hill

Centre for Cognitive Ageing and Cognitive Epidemiology

Department of Psychology

University of Edinburgh

7 George Square

Edinburgh, EH8 9JZ

Scotland, UK

Telephone: +44 131 650 8405

David.Hill@ed.ac.uk

Abstract

Socio-economic position (SEP) is a multi-dimensional construct reflecting (and influencing) multiple socio-cultural, physical, and environmental factors. In a sample of 286,301 participants from UK Biobank, we identify 30 independent-loci associated with income. Using a recently-developed method to meta-analyze data from genetically-correlated traits, we identify an additional 120 income-associated loci. These loci show clear evidence of functionality, with transcriptional differences identified across multiple cortical tissues, and links to GABAergic and serotonergic neurotransmission. By combining our GWAS on income with data from eQTL studies and chromatin interactions, 24 genes are prioritized for follow up, 18 of which were previously associated with intelligence. We identify intelligence as one of the likely causal, partly-heritable phenotypes that might bridge the gap between molecular genetic inheritance and phenotypic consequence in terms of income differences. These results indicate that, in Great Britain in the modern era, genetic effects contribute towards some of the observed socioeconomic inequalities.

People living in advantaged socio-economic backgrounds, on average, live longer, and have better mental and physical health than those from more deprived environments.[1, 2, 3] An understanding of the causes underlying the association between socioeconomic position (SEP) and health is likely to be helpful to minimize social disparities in health and wellbeing.[4]

The link between SEP and health is typically thought to be due to environmental factors including, but not limited to: access to resources, exposure to harmful or stressful environments, adverse health behaviors such as smoking, poor diet, and excessive alcohol consumption, and a lack of physical exercise.[5] However, genetic factors (most likely via mediated pleiotropy, **Figure 1**) have been discussed as a partial explanation for the SEP-health association; for example, genetic predispositions towards certain diseases, and/or genetic influences on what foods people like, could lead to poor diet which in turn could lead to both lower SEP and poorer health.[6] It has recently been demonstrated that genome-wide association studies (GWASs) can capture shared genetic associations with both measures of health, and with SEP.[7] Potential pleiotropic effects are highlighted in the observed genetic correlations between SEP variables such as completed years of education, household income, and social deprivation, and physical and mental health traits including longevity.[7, 8]

Loci associated with two SEP phenotypes, education and household income, have been identified via GWASs[7, 9, 10, 11], but—consistent with other complex traits, such as height—these loci collectively account for only a small fraction of the total heritability of the traits in question. For household income, an analysis of a sample of 96,900 individuals from Great Britain found that additive genetic effects tagged by common SNPs accounted for approximately 11% (SE = 0.7%) of differences in household income.[7] Two loci attained genome-wide significance in that study, but they collectively accounted for less than 0.005% of the total SNP heritability.

Here, we use the UK Biobank dataset[12] to examine genetic associations with household income (N=286,301) in a contemporary British sample. We identify 30 independent genome wide significant loci, 29 of which are unreported in previous work. Using a method that leverages power from genetically-correlated traits, MTAG, an additional 120 loci are found to be associated with income. We identify neurogenesis and the components of the synapse as being associated with income. Furthermore, we link transcription differences across multiple cortical tissue types, as well as

both GABAergic and serotonergic neurotransmission, to income differences. We also show that the genes linked to differences in income are predominantly those that have been previously linked with intelligence,[8] and that intelligence is one of the likely causal factors leading to differences in income. We compare the genetic correlations derived using income with those derived using another measure of SEP, educational attainment, to show that the genetic variants associated with income are related to better mental health than those related to education. Finally, we were able to predict 2.5% of income differences using genetic data alone in an independent sample.

Results

Graphical representation of statistical analysis

A flow chart summarizing all statistical analyses conducted is displayed in **Figure 2**.

SNP-based analysis of income

For household income, 3,712 SNPs attained genome-wide significance ($P < 5\times10^{-8}$), across 30 independent loci (**Figure 3A** & **Supplementary Data 1**) which contained 68 independent significant SNPs and 31 lead SNPs. A total of 29 of these 30 loci were not identified in the previous UK Biobank analysis of income[7] (**Supplementary Data 2**). The 30 loci predominantly contained SNPs found within intronic regions (47%) as well as non-coding RNA introns (29%). A total of 17% of the SNPs within the independent loci were found in intergenic regions, and only 1.2% were found in exons (**Figure 3B**). Many of the loci contained SNPs showing evidence of influencing gene regulation with 33% having a Regulome-DB score of <2 (**Figure 3C**) and 86% having evidence of being in an open chromatin state (indicated by a minimum chromatin state of <8, in **Figure 3D**). Additionally, these loci were linked to intelligence (11 loci), mental health (schizophrenia, 1 locus; bipolar disorder, 2 loci; neuroticism, 4 loci), and neurological variables (corticobasal degeneration, 1 locus; subcortical brain volumes, 1 locus; and Parkinson's disease, 1 locus) (**Supplementary Data 3**).

Linkage disequilibrium score (LDSC) regression showed that the mean χ^2 statistic was 1.45 and the intercept of the LDSC regression was 1.04. These statistics indicate that around 92% of the inflation in the GWAS test statistics was due to a polygenic signal rather than residual stratification or

confounding. The LDSC regression estimate of the heritability of household income was 7.39% (SE=0.33%).

Gene prioritization

Three methods of mapping allelic variation to genes were used to better understand the functional consequences of the 30 independent loci linked to income (positional mapping, eQTL analysis, and chromatin mapping). Using positional mapping, SNPs from the GWAS were aligned to 117 genes. eQTL mapping was used to match cis-eQTL SNPs to 186 genes, and chromatin interaction mapping linked the SNPs to 277 genes (**Figure 3E** & **Supplementary Data 4 & Supplementary Figure 1**). These mapping strategies identified a total of 400 unique genes, of which 133 (**Figure 3E** cells 14+23+26+3+24+11+2+30) were implicated by at least two mapping strategies and 47 (**Figure 3E** cells 23+24) were implicated by all three. Of the 133 implicated by two mapping strategies, two showed evidence of a chromatin interactions with two independent genomic risk loci (**Supplementary Data 5**). Both *HOXB2* and *HOXB7* showed interactions with loci 24 and loci 25. *HOXB2* showed interactions in mesendoderm (an embryonic tissue layer) tissue and IMR90 (fetal lung fibroblasts) tissue, whereas *HOXB7* showed associations in the tissues of hESC (human embryonic stem cell), Mesenchymal (multipotent stromal cells which differentiate into a variety of different cell types) Stem Cell, IMR90, Left Ventricle, GM12878, and Trophoblast-like Cells.

Gene-based association analysis

Using MAGMA,[13] 118 genes were associated with income ($P < 2.662\times10^{-6}$) (**Supplementary Data 6** & **Figure 4A**). These genes overlapped with 24 of those implicated using positional, eQTL, and chromatin interaction modelling (**Figure 3E)**. Of the genes implicated by each of the three methods and the gene-based-GWAS, *BSN* was of particular note due to its being expressed primarily in the neurons of the brain and its role in the scaffolding protein involved in the organization of the presynaptic cytoskeleton. Also found in this overlap was the gene *CHST10*. The protein encoded by *CHST10* is a sulfotransferase that acts on HNK-1 which is involved in neurodevelopment and synaptic plasticity.

These 24 genes were then examined to determine if gene-based statistics had implicated them in intelligence due to the previously-reported, strong genetic correlations between income and intelligence.[7] We found that 18 were associated ($P<2.75\times10^{-6}$) with intelligence from the GWAS conducted by Hill et al. (2018).[8] This indicates that the genes with the most biological relevance to income were also linked to intelligence, again suggestive of the role that intelligence plays in SEP differences.

Gene-set and gene-property analysis

Gene-set analysis did not find evidence that any of the gene-sets included here were enriched for differences in household income (**Supplementary Data 7**). However, a gene-property analysis showed that genes that were more associated with household income were also more highly expressed in the brain ($P=1.31\times10^{-5}$) and the testis ($P=1.31\times10^{-5}$) than genes that were less associated with income (**Supplementary Table 1**). This relationship was found across tissues of the cerebellum ($P=5.61\times10^{-6}$), the cerebellar hemisphere ($P=5.99\times10^{-6}$), the frontal cortex BA9 ($P=9.68\times10^{-5}$), the cortex ($P=1.05\times10^{-4}$), the nucleus accumbens basal ganglia ($P=2.93\times10^{-4}$), and the anterior cingulate cortex BA24 ($P=6.81\times10^{-4}$) (**Supplementary Data 8** & **Figure 4B**).

Cell-type analysis conducted on household income indicated that, of the 24 cell types examined, two were statistically significant after controlling for 24 tests. The significant cell types include medium spiny neurons $P=7.67\times10^{-5}$, and serotonergic neurons P=0.002 (**Supplementary Table 2** & **Figure 4C**). Finally, gene-property analysis found little evidence that genes linked to household income were transcribed in the brain at any one of 11 developmental stages,[14] or across 29 different specific ages[14] (**Supplementary Table 3** & **Supplementary Table 4**).

Partitioned heritability

The partitioned heritability analysis describes whether or not the SNPs that capture the greatest proportion of the heritability of income, also cluster in regions of the genome that are united by a shared biological theme. We find that, consistent with the notion that intelligence and income are genetically linked,[15] the regions of the genome that have undergone purifying selection are those that

harbor the greatest proportion of heritability for income ($P=1.62\times10^{-10}$). Also enriched was the Conserved (GERP RS>=4) annotation providing additional evidence that conserved regions of the genome are enriched for the heritability of income. None of the other functional categories were significantly enriched for the heritability of income (**Figure 5A & Supplementary Data 9**).

The partitioned heritability analysis using the six continuous categories analysed by quintile showed that common variants that were in the first three quintiles for age (i.e. the younger three groupings) were associated with a greater proportion of the heritability of income (1st quintile $P=2.57\times10^{-4}$, 2nd quintile $P=3.33\times10^{-7}$, 3rd quintile $P=6.91\times10^{-16}$) as were SNPs in the upper two quintiles for background selection greater level of background selection (4th quintile $P=9.81\times10^{-8}$, 5th quintile P=0.001). The first three quintiles describing nucleotide diversity and the same quartiles describing the level of LD (LDD-AFR) were also significantly enriched for heritability (Nucleotide diversity, 1st quintile $P=2.47\times10^{-23}$, 2nd quintile $P=3.79\times10^{-20}$, 3rd quintile P=0.003, LDD-AFR, 1st quintile $P=5.38\times10^{-12}$, 2nd quintile $P=7.36\times10^{-16}$, 3rd quintile P=0.002) (**Figure 5B & Supplementary Table 5**). The enrichment found by examining the continuous annotations by quintile is consistent with the idea that negative selective pressure has acted on the partially heritable traits linked to income.

When examining cell-type specific enrichment using partitioned heritability we show that the greatest level of enrichment for cell type specific groupings comes from the brain and central nervous system. This is indicated by the fact that the 24 cell types that were significantly enriched using the gene expression data set were all cell types that are found within the brain and the rest of the central nervous system (**Figure 5C,** & **Supplementary Data 10**). Additionally, using the chromatin based sets, 32 of the 34 cell groupings that were significantly enriched were drawn from the brain and the central nervous system (**Figure 5D,** & **Supplementary Data 11**).

This enrichment of heritability in the central nervous system led us to examine brain regions and cell types. We found that gene expression in the cortex harbored an enriched proportion of the heritability of income (P=0.006), but no other regions were found to be enriched (**Figure 5E** & **Supplementary Table 6)**. Finally, gene expression associated with the category of neuron was found

206 to be enriched ($P=1.30\times10^{-9}$) but the two glia annotations of astrocyte and oligodendrocyte were not
207 linked to income (**Figure 5F** & **Supplementary Table 7)**.
208
209 Inference of causal links with intelligence
210 Mendelian randomization was performed using the genetic instrument derived using 19 SNPs
211 associated with intelligence from a meta-analysis of a GWAS of intelligence from the INTERVAL
212 BioResource[16, 17] as well as publicly-available sources (**Supplementary Methods**). Here we inferred a
213 strong, causal link between intelligence and income (Beta=0.213, SE=0.063, $P=7.63\times10^{-4}$)
214 (**Supplementary Table 8**). Should the assumptions of MR be met, this indicates that greater
215 intelligence causes a higher level of income. Sensitivity analyses revealed little evidence of
216 directional pleiotropy which can bias MR estimates (MR-Egger intercept=0.010, SE=0.007, P=0.189)
217 (**Supplementary Table 8**). The heterogeneity statistics indicate that the estimated size of the causal
218 effect of intelligence on income varies across the SNPs (**Supplementary Table 8**). However, since
219 there was little evidence of directional pleiotropy, the overall causal estimate based on all of the
220 genetic variants is unlikely to be biased if the MR-Egger assumptions hold (i.e. the InSIDE
221 assumption).
222
223 Genetic correlations
224 Genetic correlations were calculated between household income and a set of 27 data sets
225 covering psychological traits, mental health, health and wellbeing, anthropometric traits, metabolic
226 traits, and reproduction.
227 First, we build on the findings of Hill et al. (2016)[7] by using a larger, better-powered dataset
228 on income to show that the genetic variants associated with household income are linked with those
229 that influence intelligence, $r_g=0.69$, SE=0.02, $P<10\times10^{-200}$. We also show there are genetic
230 correlations between income with health (self-rated health, $r_g=0.60$, SE=0.03, $P=5.72\times10^{-73}$), mental
231 health (subjective wellbeing, $r_g=0.32$, SE=0.04, $P=4.99\times10^{-17}$), and longevity ($r_g=0.47$, SE=0.07,
232 $P=1.29\times10^{-10}$). Furthermore, we replicated the finding of Hill et al. (2019) by showing in our current
233 study that, whilst a general factor of neuroticism shows a negative genetic correlation with household

income (r_g=−0.36, SE=0.02, P=2.07×10^{-53}) the two special factors of neuroticism of anxiety/tension and worry/vulnerability each show positive genetic correlations with income (r_g=0.12, SE=0.03, P=7.19×10^{-5} and r_g=0.15, SE=0.03, P=5.61×10^{-7} respectively).

These findings show that many of the same genetic variants linked to higher SEP are also linked to better health. It should, however, be noted that income shows a positive genetic correlation with the mental health variables of anorexia nervosa (r_g=0.09, SE=0.03, P=9.53×10^{-3}) and bipolar disorder (r_g=0.11, SE=0.04, P=1.20×10^{-2}) (**Figure 6A & Supplementary Data 12**).

Second, as SEP is a multi-dimensional construct and each marker of SEP is imperfectly correlated with the others, the magnitude of the genetic correlations derived using income were compared with those derived using another measure of SEP, educational attainment. The goal of these analyses was to indicate if the genetic associations between income with health differed from those of education with health. As can be seen in **Figure 6A**, whereas the magnitude and direction of the genetic correlations derived using income and EA with the 27 health and wellbeing, anthropometric, mental health, and metabolic traits were highly similar, there were instances of divergence indicating unique genetic associations with the two SEP variables. Of note are the variables of autism and schizophrenia. As found in previous studies[8, 18, 19, 20, 21, 22, 23] schizophrenia showed a small positive genetic correlation with EA (r_g=0.06, SE=0.02, P=1.15×10^{-3}) whereas, in the present study, income showed a negative genetic correlation with schizophrenia (r_g=−0.14, SE=0.02, P=6.49×10^{-9}, P_{diff}=6.57×10^{-11}). Autism was positively genetically correlated with EA (r_g=0.27, SE = 0.03, P=1.10×10^{-15}) as previously,[8, 21, 24] whereas there was no detectable genetic correlation between income and autism (r_g=0.04, SE=0.05, P=0.37, P_{diff}=1.17×10^{-11}). There was evidence of differences between the income and education genetic correlations and nine other traits (subjective wellbeing, P_{diff}=1.42×10^{-5}, tiredness, P_{diff}=1.60×10^{-4}, age at first birth, P=1.24×10^{-3}, bipolar disorder, P_{diff}=1.41×10^{-2}, social deprivation, P_{diff}=1.72×10^{-2}, and chronotype, P=3.83×10^{-2}, the worry/vulnerability special factor of neuroticism P=1.17×10^{-2}, and a general factor of neuroticism P_{diff}=7.26×10^{-3}) (**Figure 6A** & **Supplementary Table 12)**.

Third, the role of intelligence in mediating the effect of genetic variation on income was explored by estimating the genetic correlation of income with each of the traits after conditioning the

income GWAS on a GWAS on intelligence. As can be seen in **Figure 6B**, after controlling for intelligence the genetic correlations between income and the 27 health and wellbeing, anthropometric, mental health, and metabolic traits remained largely similar. Two exceptions to this were age at first birth, where the genetic correlation with income decreased from r_g=0.58 (SE=0.03, P=8.81×10^{-99}) to r_g=0.45 (SE=0.04, P=1.20×10^{-35}, P_{diff}=0.003), and ADHD which decreased from r_g=−0.48 (SE=0.03, P=2.20×10^{-45}) to r_g=−0.36 (SE=0.04, P=1.86×10^{-17}, P_{diff}=0.03). This means that genetic variation that is associated with income, but not intelligence, shows much of the same overlap with the 27 traits used here, as the genetic variation that is common to both income and intelligence.

In **Figure 6C** however, 12 genetic correlations with intelligence changed after controlling for income. There was little evidence that subjective wellbeing was genetically correlated with intelligence (r_g=0.03, SE=0.03, P=0.31), as previously found[8]; however, subjective wellbeing was negatively genetically correlated after adjusting for income (r_g=−0.18, SE=0.04, P=3.11×10^{-5}), (P_{diff}=9.92×10^{-5}). The genetic correlation between intelligence and social deprivation (as measured by Townsend Scores) of r_g=−0.42 (SE=0.04, P=1.38×10^{-23}), attenuated to r_g= 0.04 (SE=0.05, P=0.38); (P_{diff} =1.29×10^{-13}). The genetic correlation between intelligence and neuroticism (r_g=−0.23, SE=0.02, P=1.83×10^{-23}) also attenuated to close to zero after conditioning on income (r_g=−0.02, SE=0.03, P=0.57), (P_{diff}=7.26×10^{-3}). This means that the genetic variation that is associated with intelligence, but not income, shows less overlap with the 27 traits used here, than the genetic variation that is common to both intelligence and income. The genetic correlations with intelligence once conditioning on income were different for the variables of self-rated health (P_{diff}=6.76×10^{-12}), age at first birth (P_{diff}=1.33×10^{-8}), fatigue or tiredness (P_{diff}=6.82×10^{-8}), ADHD (P_{diff}=5.55×10^{-4}), height (P_{diff}=2.59×10^{-4}), BMI (P_{diff}=0.013), obesity (P_{diff}=1.60×10^{-2}), longevity (P_{diff}=0.014), smoking (P_{diff}=0.032) (**Figure 6C, Supplementary Table 12)**.

Genetic prediction

Polygenic risk scores were derived using the summary statistics from our GWAS of household income and the GS:SFHS data on household income. When examining the polygenic risk scores within each of the five income groups in GS:SFHS we found those in category 5 (those earning

more than £70,000) had the highest PGR scores **(Figure 7A)**. The predicted income for the PGR scores was lower in each subsequent level of household income in GS:SFHS.

Those in the lowest quintile of the polygenic score for income were found on average to have the lowest predicted income (**Figure 7B**) with the mean level of household income rising across each quintile. Those in the three lowest quintiles for their genetic propensity for income were found to have an average level of household income between £10,000 and £30,000, whereas those in the top two quintiles were found to have a household income of between £30,000 and £50,000. Polygenic prediction conducted using the summary data from UK Biobank applied to the GS:SFHS data showed that between 1.2% and 2.0% of the variance in household income can be predicted using the polygenic score for income (**Supplementary Table 9** & **Figure 7C)** with the PGRS that was most predictive using a P-value cut off of 0.1.

Multi-trait analysis of genome-wide association studies

MTAG has previously been used to conduct the first well-powered GWAS on intelligence.[8] We used MTAG here to increase the power of our GWAS on income by meta-analysing it with another measure of SEP, educational attainment[10] as measured by the number of years of education a participant has completed. MTAG was conducted using the default settings and applied to increase the power in the GWAS of household income. Following the application of MTAG, the mean χ^2 statistic increased from 1.45 to 1.73 and increased the effective sample size to 505,541 for income.

The maxFDR derived was 0.003, over an order of magnitude lower than the commonly accepted standard of false discovery and comparable with those reported previously,[8, 25] indicating that the data set was capturing variance associated with income. We also find that the genetic correlation between our MTAG-income phenotype and a previous GWAS on income[7] was r_g=0.97 (SE=0.024), with a genetic correlation of r_g=0.94 (SE=0.004) with educational attainment. This indicates that the polygenic signal in the MTAG-income analysis is virtually identical to that found in previous GWAS of income, but also that it captures more of the variance that is shared between income and education.

Using this MTAG-income phenotype we identify 144 independent genomic risk loci (**Supplementary Figure 2A** & **Supplementary Data 13**). A total of 24 overlapped with the 30 found without using MTAG, meaning that by using MTAG an additional 120 independent loci were identified that were associated with income (**Supplementary Data 14)**. Functional annotation of these loci, as well as gene-based analyses and partitioned heritability analysis showed results that were consistent with a better-powered GWAS dataset on household income (**Supplementary Figure 2B - 2E)**. These results can be found in **Supplementary Note 1**.

Polygenic risk scores analysis using the MTAG phenotype, showed that between 1.7% and 2.5% the variance of income was predicted in an independent sample (**Supplementary Table 9** & **Figure 7C)** with the PGRS that was most predictive using a P-value cut off of 0.05.

Discussion

Using the UK Biobank data set, we identified 30 independent genetic loci associated with income levels in Great Britain today. This represents a considerable advance on the two loci previously identified by Hill et al. (2016).[7] The present study contributes to our understanding of the genetic contributions to SEP in seven major ways.

First, the loci associated with income showed clear evidence of functionality, particularly regarding their links to gene expression, regulatory regions of the genome, and open chromatin states. Second, by combining our GWAS data with eQTL data from BRAINEAC,[26] GTEx,[27] and others, along with chromatin interaction data[28, 29] we were able to prioritize which genes were likely to be causal based on the overlap of multiple lines of biological enquiry. Although income, as a biologically distal phenotype, will not be directly linked to genetic variation (**Figure 1**),[7] genes that may exert a causal influence are likely to do so through their effect on more proximal phenotypes.[30]

Using our GWAS data set on income, we identified 47 genes that were mapped to the 30 independent genomic loci using positional, eQTL, and chromatin mapping. In addition, we used the 118 genome-wide significant genes from our gene-based analysis of income to further refine this set to a total of 24 implicated genes. These 24 genes therefore should be prioritized in follow-up studies as they are located close to the associated loci, have expression correlated with genetic variation of the

SNPs in the independent genomic loci, have chromatin interactions taking place between these genes and the SNPs found in the independent loci (**Supplementary Data 4**) and, consistent with highly polygenic traits, these genes harbor many SNPs that show consistent associations with income (**Supplementary Data 6**). In addition, 18 of these genes have been associated with intelligence,[8] so efforts to ascertain how such genetic variation is associated with income differences should examine their associations with intelligence more closely.

Third, by broadening our analysis to include the polygenic signal that fell outside of the independent loci, we identified additional, functional elements of the genome linked to differences in income. By combining the gene-based statistics from MAGMA with gene expression data from the GTEx[27] database, we identified a positive association between expression in the brain, as well as several specific regions, and the level of association displayed by the gene-based statistics on income. This indicates that the higher the level of association between a gene and income, the higher that gene's level of expression specific to the brain will be.

Cell type specific analysis revealed that the expression that was specific to the serotonergic neurons and to medium spiny neurons was associated with income. Medium spiny neurons have previously been linked to schizophrenia[31] which has a strong cognitive component and has previously been linked to glutamatergic systems including the NMDA receptor signaling complex.[32] Medium spiny neurons are a sub-type of GABAergic inhibitory neurons. Future work should examine if, like other cognitive traits, income is linked to both GABAergic and glutamatergic systems.

Partitioned heritability analysis identified enrichment across cell types from the central nervous system and across the cortex as being significantly enriched for the heritability of income. These findings indicate that income in Great Britain today is associated with phenotypes that are associated with differences in the brain such as intelligence[8] or neuroticism.[33, 34]

These two approaches, gene-based statistics and LDSC regression, illustrate how combining the genetic data from GWAS with gene expression data can be informative as to the possible biological processes that are associated with income. This is of particular value for traits, like income, that have no clear biological analogue and are likely linked to genetic variation via mediated pleiotropy. This combination of data provides evidence that some of the individual differences in

income are related to gene expression differences in the brain (**Figure 4B & Figure 5C, D, and E**), as well as highlighting the role of specific classes of neuron (**Figure 4C & Figure 5F**). As importantly, we show the role for some tissue types outside of the central nervous system (**Figure 5D**) indicating that genetic factors associated with income differences may also lie outside of the phenotype of intelligence, and outside of cortical tissue types.

Fourth, using Mendelian Randomization, we provided evidence implicating intelligence as one of the potentially causal, partly-heritable, phenotypes that might be one bridge in the gap between molecular genetic inheritance and phenotypic consequence. This result, if the assumptions of MR are met, helps explain why individual differences in income are found to be partly heritable.

Fifth, our data show that income and education each have similar genetic correlations with many variables. However, some genetic correlations differ depending on whether income or education is used as a measure of SEP, and those that differed tend to be those related to mental health. In those, the income genetic correlations that are negative are of a greater magnitude than those derived using education, and where the income genetic correlations are positive, they are of smaller effect than the education derived genetic correlations (**Figure 6A & Supplementary Data 12**).

Together this implies that the genetic variants that are associated with higher income tend to be more strongly associated with better psychological health than the genetic variants associated with education. This could be a stage-of-life-course-specific-phenomenon, i.e. education tends to be completed earlier in the life course, before some illnesses appear that could affect earning capability. It should also be considered that these significantly different genetic correlations between education and income indicate that educational attainment serves to provide access to opportunities in the labor market, and those that have these opportunities are then better placed to engage in health-relevant behaviors. This would indicate that, whereas income may be a more distal phenotype from DNA than education, it is potentially closer to outcomes such as later-life health, as evidenced by differences between the genetic correlations. Future work should examine models where DNA -> neuronal properties -> intelligence -> education -> income -> health, using multivariable Mendelian randomization[35, 36, 37] to gauge the direct and indirect effects of income and education on health outcomes.

However, previous work using lotteries in Sweden as natural experiments to examine the causal effect of wealth on health differences[38] found that, in the 10 years after receiving a prize (either as a single payment or multiple instalments), winning participants did not have a longer life or fewer hospital admissions compared with those who did not win the lottery.[38] This indicates that, whereas high earners may be in better health and have a greater level of education than low earners, a high income might not be causal in such differences in affluent countries that have strong social support systems. Furthermore, children born to lottery winners were not found to be advantaged in terms of their level of scholastic performance compared to the children of those who did not win the lottery,[38] a finding that argues against a dynastic effect mediated via wealth. Although any causal effect of wealth on health are likely to differ across countries and times, should the results of this Swedish study generalize to the UK today, they would complement our results and together would support a model whereby genetic differences that are linked with health might be linked to partly heritable intermediary phenotypes, such as intelligence.

In this scenario, the similarities and differences between the genetic correlations derived using education and income might be accounted for in part by the differences in the intermediary phenotypes that give rise to each measure of SEP. Under this model, the observed differences between genetic correlations with mental health (**Figure 6A**) would be due to intermediary variables that make a greater contribution to both income and mental health than they do to education. The similarities between income and education genetic correlations and health potentially indicates a similar contribution from intermediary phenotypes to income, education, and health.

Using mtCOJO we found that, when the genetic associations that are shared between income and intelligence were removed, the genetic correlations with other traits were largely unchanged. The exceptions were with ADHD and with age of first birth, where the genetic correlations with income are both attenuated once conditioned on intelligence. However, 12 of the income-health genetic correlations were attenuated after adjusting the SNP-income associations for intelligence. These results indicate that the genetic variation associated with intelligence and income is also associated with many health and mental health traits, because, when this shared variance is removed, leaving only the variance that is unique to intelligence, the magnitude of the genetic link between intelligence

and health is reduced. In the case of the genetic link between intelligence, social deprivation, neuroticism, and height, this genetic association disappears entirely following adjustment for income. The exception is that subjective wellbeing shows a genetic correlation with intelligence only after the variance that is common to both income and intelligence is removed.

One interpretation of this finding is that the residual variance left in income after conditioning on intelligence still contains the genetic contributions to other partly-heritable traits (such as conscientiousness, or resistance to disease). These traits also contribute towards individual differences in income and so the association between income and health is, largely, intact following conditioning on intelligence. This would imply that intelligence is only one of a number of factors that contributes to variation in income, but income is a very important factor that mediates the associations between intelligence and health. Future work examining the genetic relationship between income and health, as well as intelligence and health, should focus on this genetic overlap between intelligence and income using tools such as genomic structural equation modelling (SEM) to partition the total variance of traits like income into the variance that is shared with intelligence and the variance that is separate from it.[39]

Sixth, we were able to predict up to 2% of income differences using polygenic risk scores. This shows that even for phenotypes that are not impacted directly by genetic effects, but rather are more biologically distal as is the case with income, that the link between genotype and phenotype is sufficient to make predictions, based on DNA alone.

Seventh, using MTAG, we increased our effective sample size from 286,301 participants to 505,541. With this increase in power we were able to raise the number of loci found to be associated with income from 30 to 144. Of these 144 associations, 120 of were not found to be genome-wide significant before the application of MTAG. These loci demonstrated the same patterns of functional enrichment as shown in the 30 loci identified using income alone. We also identified the same relationship between expression in the brain, and across multiple cortical structures, using the better powered MTAG-derived income phenotype (**Supplementary Note 1**). Furthermore, following meta-analysis with MTAG, we were able to increase our prediction accuracy of income by 25%.

The limitations of this study include that income was measured at the level of the household and was not an individual-level measure of income. However, previous GWASs examining household income variables have shown that income, measured on a household level, has a genetic correlation of 0.90 (S.E. = 0.04) with educational attainment, as measured on an individual level, indicating that the household-level effects are likely to be generalizable to individual persons.[7] Furthermore, GWASs conducted on regional measures of educational attainment show genetic correlations of >0.9 with education measured using an individual's own level of educational attainment.[40]

A limitation of the Mendelian Randomization analysis specifically are potential dynastic effects, which may violate the assumptions of MR. Dynastic effects are where genetic variants that the parent has but the child does not, are associated with parental behaviors, and these parenting behaviours are a causal factor in the SEP of the child.[41] An example of this would could be that parents with a greater predisposition towards intelligence are also those that are more likely to provide opportunities for their children to enter higher-income occupations. In this instance the second assumption of Mendelian Randomization, that the instrument must only affect income via their effect on intelligence, would be violated. The association of the offspring's SNPs and income would be partially due to the effects of the parents' genotype on their parents' intelligence which subsequently affects offspring's income. Whereas the current data cannot differentiate between causality and dynastic effects, it should be noted that, for another measure of SEP, educational attainment, there is evidence of indirect genetic effects which account for ~30% of the variance of the direct genetic associations.[42] Future work in multi-generational samples should examine the role that such indirect genetic effects play in individual differences in income, as well as if their presence (if established) could result in an inflation of the estimate for a causal effect using Mendelian Randomization.[43]

Furthermore, genetic variants associated with intelligence are likely to have pleiotropic effects.[44] However, to break the assumptions Mendelian randomization it is not sufficient for the genetic variants to have pleiotropic effects.[45] The genetic variants we use as instruments must have horizontally pleiotropic effects mediated via mechanisms other than intelligence. If the genetic variants have vertically pleiotropic effects, e.g. SNP->neuron->intelligence->income->health, then our Mendelian randomization estimates will not be biased. Equally, if the SNPs affect other

phenotypes, but these phenotypes do not affect outcome, then these effects will not result in bias in the Mendelian randomization estimates. It is possible that the genetic variants identified in intelligence GWAS have horizontally pleiotropic effects, however, it is unclear what mechanisms would mediate these effects. The genetic correlations between intelligence and personality traits are relatively low.[46] The genetic variants identified in the intelligence GWAS are likely to also affect a range of cognitive ability related traits. However again, these pleiotropic effect via related phenotypes are unlikely to cause bias if the results are interpreted as a test of general cognitive function. It is possible to investigate potentially horizontal pleiotropic effects further using multivariable Mendelian randomization.[47] If SNPs have been identified that explain sufficient independent variation in two or more two potential pathways, e.g. intelligence and education, it is possible to identify the direct effects of each exposure. Future research should use multivariable Mendelian randomization to investigate this further.

Another limitation is that the present study was restricted to examining common genetic effects. Should rare or less common genetic variation be associated with income, then these effects will be absent from this study. Future work should utilize methods that can capture these genetic effects,[48] as well as examine SEP variables using whole exome or whole genome sequencing. In addition, the participants of UK Biobank are drawn from the more educated and healthier individuals in the UK, which might introduce collider bias.[49] Whereas a comparison of the level of SEP between the individuals in UK Biobank and the census conducted in the UK indicates that SEP, as measured using the Townsend Deprivation Index,[50] was very similar,[7] future work aiming to quantify or control for collider bias would be of value in addressing this potential issue.

A further limitation is that molecular genetic analyses of phenotypes such as intelligence, income, or SEP, appear prone to being misinterpreted.[51] Such misunderstandings include describing associated variants as, genes for income, or the misinterpretation that any associated variant, and indeed any non-zero heritability estimate, is evidence for genetic determinism or the immutable nature of these phenotypes via environmental intervention. We include a figure (**Figure 1**) that illustrates that genetic variants do not act directly on income; instead, genetic variants are associated with partly heritable traits (such as intelligence, conscientiousness, health etc.) which have their own complex

gene-to-phenotype paths (including neural variables) and are ultimately associated with income. Therefore, the genetic variant-income associations discovered here are no more for income than they are for these other traits. For more discussion of the implications of these results, aimed at the general reader, we have provided a Frequently Asked Questions (FAQ) document in the **Supplementary Note 2**.

Finally, it should be noted that GWASs, like heritability estimates, describe differences that exist within populations. This means that, although we report here that those with a greater number of intelligence-associated genetic variants tend to be those who report higher incomes, it does not hold that this is true across other societies or times. Indeed, the links between markers of SEP and health are not consistent across all societies.[52] Research into genetic links to education has found indications that the genetic variants linked to higher educational attainment are less predictive of success in societies that have less meritocratic selection for education and occupation.[53] Future work examining the relative contribution of genetic and environmental associations with income, as well as the biological systems causally implicated in any GWAS conducted on a marker of SEP across many cultures, would be valuable in identifying more and less meritocratic societies.

In conclusion, this work adds to the growing body of evidence indicating that markers of socioeconomic position, and their links to health, are likely to be both genetic and environmental in origin.[6, 7] We found that SEP variation in the Great Britain is partially accounted for by genetic differences in the population.[54] We found little evidence that these genetic differences were attributable to population stratification, but rather that they indicated the unequal distribution of heritable traits, including intelligence, across different SEP groups. Using multiple forms of biological data, we showed that these genetic differences are predominantly found in regions of the genome that have undergone negative selection and are related to differences in gene expression in the brain, particularly in medium spiny neurons. We also prioritise 24 genes for further follow up as evidence from eQTL analysis, chromatin interactions, with previous associations with intelligence converging to implicate 18 of these genes. Furthermore, we identify intelligence as one of the likely causal psychological traits partly driving differences in income and SEP in Great Britain today.

Method

Participants

The primary sample used involved participants from UK Biobank, an open-access resource established to examine the determinants of disease in middle-aged and older adults living in the United Kingdom.[55] Recruitment to UK Biobank occurred between 2006 and 2010, targeting community-dwelling individuals from both urban and rural environments across a broad range of socio-economic circumstances. A total of 502,655 participants were assessed at baseline on a range of cognitive and other psychological measures, physical and mental health, and their socioeconomic position. They donated a number of biological samples, including DNA for genotyping. In order to reduce the effects of population stratification, only participants from a single ancestry group, those of White British ancestry, were included in the analysis. High quality genotyping was performed by UK Biobank on 332,050 participants. Ethical approval for UK Biobank was received from the Research Ethics Committee (REC reference 11/NW/0382). This work was conducted under UK Biobank application 10279.

Phenotype description

A total of 332,050 participants had genotype data and data on their level of household income. Self-reported household income was collected using a 5 point scale corresponding to the total household income before tax, 1 being less than £18,000, 2 being £18,000 - £29,999, 3 being £30,000 - £51,999, 4 being £52,000 – £100,000, and 5 being greater than £100,000. This 5 point scale was analysed by treating the categories of income as a continuous variable. Participants were removed from the analysis if they answered “do not know” (n = 12,721), or “prefer not to answer” (n = 31,947). This left a total number of 286,301 participants (138,425 male) aged 39-73 years (mean = 56.5, SD = 8.0 years) with genotype data who had reported, between 1 and 5, their level of household income.

UK Biobank genotyping

Full details of the UK Biobank genotyping procedure have been made available.[56] In brief, two custom genotyping arrays were used to genotype 49,950 participants (UK BiLEVE Axiom Array) and 438,427 participants (UK Biobank Axiom Array).[56, 57] Genotype data on 805,426 markers were available for 488,377 of the individuals in UK Biobank. Imputation to the Haplotype Reference Consortium (HRC) reference panel lead to 39,131,578 autosomal SNPs being available for 487,442 participants.[56] Allele frequency checks[58] against the HRC[59] and 1000G[60] site lists were performed, and variants with minor allele frequencies (MAF) differing more than +/- 0.2 from the reference sets were removed.

Additional quality control steps were conducted and described previously.[8, 34] These included the removal of those with non-British ancestry based on self-report and a principal components analysis, as well as those with extreme scores based on heterozygosity and missingness. Individuals with neither XX or XY chromosomes, along with those individuals whose reported sex was inconsistent with genetically inferred sex, were also removed. Finally, individuals with more than 10 putative third degree relatives (identified by Bycroft et al.[56] by estimating the kinship coefficients for all pairs of samples using the software KING[61]) were also removed. Following these exclusions, a sample of 408,095 individuals remained. Using GCTA-GREML on 131,790 reportedly-related participants,[62] related individuals were removed based on a genetic relationship threshold of 0.025. Following this quality control, household income data, and genetic data, were available on 286,301 participants. Following association analysis, SNPs with a minor allele frequency < 0.0005, and an imputation quality score < 0.1 were removed. Finally, only bi-allelic SNPs were retained, resulting in 18,485,882 autosomal SNPs.

Genome-wide association analysis (GWAS) in the UK Biobank sample

The level of household income as measured on the 5 point scale was subjected to a regression using income as the outcome as has been conducted previously,[7] and 40 genetic principal components (to control for population stratification), genotyping array, batch, age, and sex as predictors. The residuals from this model were then used in a GWAS assuming an additive genetic model as implemented in BGENIE.[56]

596

597 Functional annotation and loci discovery

598 Genomic risk loci were derived using the summary data from the data set of household
599 income derived in UK Biobank, using FUnctional Mapping and Annotation of genetic associations
600 (FUMA).[63] First, independent significant SNPs were defined using a P-value cut off of genome-wide
601 significant ($P < 5 \times 10^{-8}$), as well as being independent from each other ($r^2 < 0.6$) within a 1mb
602 window. Second, SNPs that were in LD with any independent SNP ($r^2 \geq 0.6$) and within a 1mb
603 window in addition to being in the HRC genomes reference panel with a MAF greater than 0.001,
604 were included for further annotation. Third, lead SNPs were identified using the independent
605 significant SNPs as defined above. Lead SNPs were a subset of the independent significant SNPs that
606 were in LD with each other at $r^2 < 0.1$, with a 1mb window. Fourth, genomic risk loci were created by
607 merging lead SNPs if they were closer than 250 kb apart. This means that a genomic risk locus could
608 contain multiple independent significant SNPs and multiple lead SNPs. Finally, all SNPs in LD of r^2
609 $\geq$0.6 with one of the independent significant SNPs formed the border, or edge, of the genomic risk
610 loci.

611 The lead SNPs and those in LD with the lead SNPs were then mapped to genes based on their
612 functional consequences, as described using ANNOVAR[64] and the Ensemble genes build 85.
613 Intergenic SNPs were annotated as the two closest flanking genes which can result in them being
614 assigned to multiple genes.

615

616 Gene-mapping

617 Three strategies were used to link the income-associated independent genomic loci to genes.
618 First, positional mapping[65] was used to map SNPs to genes based on physical distance. SNPs were
619 mapped to genes if they were within a 10kb from a known protein gene found in the human reference
620 assembly (hg19).

621 Second, expression quantitative trait loci (eQTL) mapping was carried out by mapping SNPs
622 to genes if allelic variation at the SNP is associated with expression levels of a gene. For eQTL
623 mapping, information on 45 tissue types from three data bases (GTEx v7, Blood eQTL browser, BIOS

QTL browser) based on cis-QTLs was used and SNPs were mapped to genes up to 1Mb away. A false discovery rate (FDR) of 0.05 was used as a cut off to define significant eQTL associations.

Finally, chromatin interaction mapping was carried out to map SNPs to genes when there is a three-dimensional DNA-DNA interaction between the SNP and gene. No distance boundary was applied as chromatin interactions can be long-ranging and span multiple genes. Hi-C data of 14 tissue types was used for chromatin interaction mapping.[66] In order to reduce the total number of genes mapped using chromatin interactions and to increase the likelihood that those mapped are biologically relevant, an additional filter was added. We only retained interaction mapped genes if one region involved with the interaction overlapped with a predicted enhancer region in any of the 111 tissue/cell types found in the Roadmap Epigenomics Project,[67] and the other region was located in a gene promoter region (i.e., 250bp upstream and 500bp downstream of the transcription start site and also predicted to be a promoter region by the Roadmap Epigenomics Project [67]). An FDR of 1×10^{-5} was used to define a significant interaction.

Gene-based GWAS

Gene-based analyses have been shown to increase the power to detect association due to the multiple testing burden being reduced, in addition to the effects of multiple SNPs being combined.[68] Gene-based GWAS was conducted using MAGMA.[69] All SNPs located within protein coding genes were used to derive a P-value describing the association found with household income. The NCBI build 37 was used to determine the location and boundaries of 18,782 autosomal genes and linkage disequilibrium within and between genes was gauged using the HRC panel. In order to control for multiple testing, a Bonferroni correction was applied using each gene as an independent statistical unit ($0.05 / 18{,}782 = 2.66 \times 10^{-6}$). The gene-based statistics derived using MAGMA were then used to conduct the gene-set analysis, the gene-property analyses, and the cell type enrichment analysis.

Gene-set analysis

In order to understand the biological systems vulnerable to perturbation by common genetic variation, a competitive gene-set analysis was performed. Competitive testing, conducted in

MAGMA,[69] examines if genes within the gene-set are more strongly associated with the trait of interest than other genes, and differs from self-contained testing by controlling for type 1 error rate as well as being able examine the biological relevance of the gene-set under investigation.[70]

A total of 10,891 gene-sets (sourced from Gene Ontology,[71] Reactome,[72] and, MSigDB[73]) were examined for enrichment of household income. A Bonferroni correction was applied to control for the multiple tests performed on the 10,891 gene-sets available for analysis.

Gene-property analysis

In order to identify the relative importance of particular tissue types which may indicate the intermediary biological phenotypes that might act between genetic variation and SEP outcomes, a gene property analysis was conducted using MAGMA. The goal of this analysis was to determine if, in 30 broad tissue types, and 53 specific tissues, tissue specific differential expression levels were predictive of the association of a gene with household income. Tissue types were taken from the GTEx v7 RNA-seq database[74] with expression values being log2 transformed with a pseudocount of 1 after Winsorising at 50 with the average expression value being taken from each tissue. Multiple testing was controlled for using Bonferroni correction.

An additional gene property analysis was conducted to determine if transcription in the brain at 11 developmental stages,[14] or across 29 different age groupings,[14] was associated with a gene's link to household income. These RNA-Seq GEncode v10 summarized to genes data were accessed using the following link: http://www.brainspan.org/api/v2/well_known_file_download/267666525. The detailed descriptions of the normalization processes used can be found in the technical white paper at: http://help.brain-map.org/download/attachments/3506181/Transcriptome_Profiling.pdf?version=1&modificationDate=1382036562736&api=v2, where a total of 524 samples were available for analysis. The developmental stages were assigned to each two groups (11 developmental stages and 29 age groupings) based on the age of the sample. The groupings of 25 post-conception weeks (pcw) and 35 pcw were excluded from the age groups as they contained fewer than three samples. Next, the 52,376 annotated genes were filtered so that the average Reads Per Kilobase (RPKM) is >1. This was

680 performed in the developmental group and in the age group separately. This resulted in 19,601 genes
681 for the developmental stage groupings and 21,001 genes for the age groupings. RPKM was then
682 winsorized at 50 (RPKM>50 was replaced with 50). Then, the average of log transformed RPKM
683 with a pseudocount 1 ($\log_2$(RPKM+1)) per group (for either 11 developmental stages or 29 age
684 groups) was used as a covariate conditioning on the average across all the labels. To control for
685 multiple tests a Bonferroni correction was used to control for 11 and 29 tests separately.
686
687 Cell type enrichment
688 As previous studies had indicated the importance of cortical tissues to differences in SEP,[7, 10]
689 a gene property analysis was also conducted to examine a broad array of brain specific cell types.
690 Enrichment of heritability was tested against 173 types of brain cells (24 broad categories of cell
691 types), which were calculated following the method described in Skene et al., (2018)[31] Briefly, brain
692 cell-type expression data were drawn from single-cell RNA sequencing data from mouse brains. For
693 each gene, a specificity value for each cell-type was calculated by dividing the mean Unique
694 Molecular Identifier (UMI) counts for the given cell type by the summed mean UMI counts across all
695 cell types. MAGMA[69] was used to calculate cell type enrichments where specificity values were then
696 divided into 40 equal sized bins for each cell type for the MAGMA analysis. A linear model was
697 fitted over the 40 specificity bins (with the least specific bin indexed as 1 and the most specific as 40).
698 This was done by passing the bin values for each gene using the '*--gene-covar onesided*' argument.
699
700 Univariate linkage disequilibrium score
701 Univariate LDSC regression was performed on the summary statistics from the GWAS on
702 household income in order to quantify the degree to which population stratification may have
703 influenced these results.
704 For the GWAS on household income, LD score regression was carried out by regressing the
705 GWA test statistics (χ^2) from each GWAS onto the LD score (the sum of squared correlations
706 between the minor allele frequency count of a SNP with the minor allele frequency count of every

other SNP) of each SNP. This regression allows for the estimation of heritability from the slope, and a means to detect residual confounders using the intercept.

LD scores and weights were downloaded from (http://www.broadinstitute.org/~bulik/eur_ldscores/) for use with European populations. SNPs were included if they had a minor allele frequency of > 0.01 and an imputation quality score of > 0.9. Following this, SNPs were retained if they were found in HapMap 3 with MAF > 0.05 in the 1000 Genomes EUR reference sample. Following this indels and structural variants were removed along with strand ambiguous variants. SNPs whose alleles did not match those in the 1000 Genomes were also removed. As the presence of outliers can increase the standard error in LDSC regression and so SNPs where $\chi^2 > 80$ were also removed.

Partitioned heritability

Partitioned heritability was performed using stratified linkage disequilibrium score (LDSC) regression.[75, 76] Stratified LD Scores were calculated from the European-ancestry samples in the 1000 Genomes project (1000G) and only included the HapMap 3 SNPs with a minor allele frequency (MAF) of >0.05. The model was constructed using 60 overlapping, functional categories. In addition, 10 minor allele frequency bins, and 6 continuous annotations, were included to control for LD-related bias in the partitioned heritability analysis by modelling regional LD, as well as MAF. Correction for multiple testing was performed using a Bonferroni test on the 60 functional categories ($\alpha = 0.00083$). The continuous annotations were also analyzed by examining the enrichment of each quintile for the six continuous categories of predicted allele age, background selection, recombination rate, nucleotide diversity, low levels of linkage disequilibrium in African populations, and CpG content. Here, control for multiple testing was performed using a Bonferroni correction within each of the six annotations ($\alpha=0.05/5=0.01$).

Cell type analysis was conducted using the method of Finucane et al. (2018)[77]. Here, four data sets were used and examined for enrichment of household income. The first data set (gene expression) contained gene expression data from across 205 tissue or cell types taken from the GTEx[74] data base and from Franke lab data set[78, 79] from Finucane et al. (2018)[77]. The second data set (Chromatin)

contained data on 489 tissue and cell types taken from Roadmap Epigenomics consortium[67] and from EN-TEx, a subgroup of ENCODE.[80 77]. Data pertaining to expression in 13 regions the brain was taken from GTEx[74] and gene expression specific to the neuron, the astrocytes and the oligodendrocytes was taken from mouse data from the work of Cahoy et al.(2008).[81]

Multiple testing for the partitioning of the heritability by cell types was conducted using a Bonferroni correction across the 13 brain regions (α=0.05/13=0.004) and across the three types of neuron (α=0.05/3=0.017). For the gene expression and chromatin groupings a false discovery rate (FDR)[82] was applied to the 205 tests performed to look at enrichment using gene expression (α=0.006) and to the 489 tests examining chromatin based annotations (α=0.003).

Mendelian Randomization

The causal effects of intelligence (termed the exposure in an MR analysis) on income (termed the outcome in an MR analysis) were investigated using univariate Mendelian Randomization (MR) analysis. Here, the total causal effect of intelligence on income was examined by combining summary GWAS test statistics for intelligence and for income using an inverse-variance-weighted (IVW) regression model.[83] This is equivalent to a weighted regression of the SNP-outcome coefficients on the SNP-exposure coefficients, with the intercept constrained to zero (i.e. assuming no unbalanced horizontal pleiotropy).

The results of the IVW regression model were compared with the results obtained using MR-Egger regression.[84] MR analyses which use multiple SNPs are more likely to include invalid SNPs with horizontally pleiotropic effects.[85] By not constraining the intercept to zero (as done using inverse variance weighted regression) MR-Egger relaxes the assumption that the effects of genetic variants on the outcome act solely through the exposure (in this case intelligence). The intercept parameter of the MR-Egger regression indicates the average directional pleiotropic effects of the SNPs on the outcome. As such, the direct pleiotropic effect that the SNPs have on the outcome, independent of the exposure, can be quantified, where a non-zero intercept provides evidence for bias due to directional pleiotropy and a violation of the MR IVW estimator assumptions. Of note is that the MR-Egger regression estimates only remain consistent if the magnitude of the gene-exposure associations, across all

763 variants, are independent of their horizontally pleiotropic effects on the outcome (i.e. the InSIDE
764 assumption holds).[84] In addition, power is almost always lower for MR-Egger and it requires variation
765 in the size of effect of the SNPs on the exposure (i.e. if all SNPs have similar sized effects on the
766 exposure, then MR-Egger will have very low power).

767 For use with Mendelian Randomization, two-independent groups (n = 95,521 for intelligence
768 and n = 271,732 for income) were created whereby the GWAS on income was re-run using only those
769 participants whose data were not included in the interim release of the UK Biobank genotype data. A
770 GWAS data set on intelligence was created by meta-analysing publicly-available data on intelligence
771 with a GWAS (conducted for this study) on intelligence using data from the INTERVAL
772 BioResource[16, 17] (**Supplementary Methods**) where 19 SNPs were identified as being genome-wide
773 significant and independent. These 19 SNPs were used as instrumental variables for intelligence in the
774 MR analysis.

775

776 Genetic correlations

777 Genetic correlations were derived using bi-variate LDSC regression. A total of 27 GWAS
778 data sets on health, anthropometric, psychiatric, and metabolic traits were examined for a genetic
779 correlation with income (**Supplementary Table 16**). Genetic correlations were also derived between
780 household income with education and intelligence. There were three objectives to our analysis
781 examining genetic correlations using household income. First, we sought to replicate the results of
782 Hill et al. (2016)[7] who found genetic correlations between household income and other variables in a
783 smaller data subset from the UK Biobank sample used here. Second, SEP is multi-dimensional in
784 nature: it is composed of multiple measures, each of which are correlated imperfectly with the others.
785 Because of this, different measures of SEP may have genetic variance that is both unique to them, and
786 differentiates them from the others in the way it associates with health. To examine this, we compare
787 how genetic correlations with household income and 27 health, anthropometric, psychiatric, cognitive,
788 and metabolic traits differed compared to the genetic correlations derived using a different,
789 individual-level measure of SEP, i.e. educational attainment as measured by the number of years one
790 has spent in education.[11] Third, Hill et al. (2016) also found that the phenotypes with the strongest

genetic correlations with income are those that are cognitive (verbal numerical reasoning, childhood IQ, and years of education) in nature.[7] The magnitude of these genetic correlations might indicate the phenotypes that occur as potential mediators between molecular genetic inheritance and household income.

In addition, intelligence is known to be genetically correlated with many physical and mental health traits.[18, 21, 86] The role that intelligence might play in accounting for some of the genetic links between household income and 27 health and wellbeing, anthropometric, mental health, and metabolic traits was examined using genetic correlations. Here, the GWAS of income was conditioned on a GWAS on intelligence using Multi-trait-based conditional & joint analysis (mtCOJO). mtCOJO is used to perform conditional GWAS whereby the genetic effects from one GWAS are controlled for in another GWAS. Importantly, the mtCOJO method avoids well-known issues of collider bias that can occur by including heritable covariates.[87] In the current study, the GWAS on income was conditioned on a GWAS on intelligence (and the intelligence GWAS was conditioned on the income GWAS) before the genetic correlations between income (and intelligence) and 27 variables mentioned above were re-ran.

Genetic prediction

Using the summary statistics from our GWAS of household income polygenic risk scores (PGRS) were derived using PRSice-2[88] and the Generation Scotland: Scottish Family Health Study (GS:SFHS) cohort. The recruitment protocol and sample characteristics of GS:SFHS are described in full elsewhere.[89, 90] In brief, 23,690 participants were recruited through their GP from across Scotland. Participants were all aged 18 and over and were not ascertained based on the presence of any specific disease. Following the removal of individuals who preferred not to answer, income was assessed in GS:SFHS by 5 point scale (1 less than £10,000, 2 between £10,000 and £30,000, 3 between £30,000 and £50,000, 4 between £50,000 and £70,000, 5 more than £70,000). Individuals who preferred not to answer were excluded from the analysis. Individuals who had taken part in UK Biobank were also removed from the GS:SFHS data set (n = 174). SNPs were included in the data if they had a MAF of ≥ 0.01 and Hardy-Weinberg P-value > 0.000001. Finally, one from every pair of related individuals

were removed from the data set by creating a genetic relationship matrix using GCTA[91] and removing individuals who are related at ≥ 0.025. This yielded a final sample size of 6,680 participants who had genotype data and income data.

The participant's level of income was then used as a predictor in a regression analysis with age, sex, and 20 principal components included to control for population stratification. The standardized residuals from this model were then used as each participant's income phenotype. PGRS were created using the income phenotype derived using UK Biobank.

In each instance PRSice-2 was used to create five PGRS corresponding to one of five P-value cut-offs ($P \leq 0.01$, $P \leq 0.05$, $P \leq 0.1$, $P \leq 0.5$, $P \leq 1$) applied to the association statistics from the summary data. The polygenic risk scores for each threshold were then standardized and used in a regression model to predict the income phenotype in GS:SFHS.

Multi-Trait Analysis of GWAS (MTAG)

MTAG[25] can be used to meta-analyze genetically correlated traits in order to increase power to detect loci in any one of the traits. Only summary data are required in order to carry out MTAG and bivariate LD score regression is carried out as part of an MTAG analysis to account for (possibly unknown) sample overlap between the GWAS data sets.[25] The goal of this analysis was to increase the power to detect loci associated with income, and so our income GWAS was meta-analysed with the GWAS on years of education by Okbay et al.[92] using MTAG. Both the Okbay data set and the income data set from UK Biobank had a similar level of power (Okbay mean $\chi^2 = 1.65$, UK Biobank income mean $\chi^2 = 1.45$) and they showed a genetic correlation of $r_g = 0.77$ (SE = 0.02), confirming that both income and education, as measured using these data sets, have a highly similar genetic etiology. Functional annotation and loci discovery, gene-mapping, gene-based GWAS, gene-set and gene-property analysis, were also performed using the MTAG derived data set on income. In addition, following the removal of UK Based cohorts from the educational attainment summary statistics, genetic prediction was performed using the MTAG derived income phenotype and the GS:SFHS as described above.

Data availability

The household income association results, and the multi-variate analysis conducted using MTAG can be downloaded from The Lothian Birth Cohorts of 1921 and 1936 data sharing resource; https://www.lothianbirthcohort.ed.ac.uk/content/gwas-summary-data and at http://www.phenoscanner.medschl.cam.ac.uk/.

References

1. Batty GD, Der G, Macintyre S, Deary IJ. Does IQ explain socioeconomic inequalities in health? Evidence from a population based cohort study in the west of Scotland. *Bmj* **332**, 580-584 (2006).

2. Calixto O-J, Anaya J-M. Socioeconomic status. The relationship with health and autoimmune diseases. *Autoimmunity Reviews* **13**, 641-654 (2014).

3. Marmot MG, *et al.* Health inequalities among British civil servants: the Whitehall II study. *The Lancet* **337**, 1387-1393 (1991).

4. Acheson D. Independent inquiry into inequalities in health: report. (ed^(eds Office S) (1998).

5. Wilkinson RG, Marmot MG. *Social determinants of health: the solid facts* (2003).

6. Marmot MG, Shipley MJ, Rose G. INEQUALITIES IN DEATH—SPECIFIC EXPLANATIONS OF A GENERAL PATTERN? *The Lancet* **323**, 1003-1006 (1984).

7. Hill WD, *et al.* Molecular genetic contributions to social deprivation and household income in UK Biobank. *Current Biology* **26**, 3083-3089 (2016).

8. Hill W, *et al.* A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Molecular psychiatry*, 1 (2018).

9. Rietveld CA, *et al.* GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science* **340**, 1467-1471 (2013).

10. Okbay A, *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539-542 (2016).

11. Lee JJ, *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics* **50**, 1112-1121 (2018).

12. 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**, 1-10 (2015).

13. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS computational biology* **11**, e1004219 (2015).

14. Kang HJ, *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483 (2011).

15. Hill WD, *et al.* Molecular genetic aetiology of general cognitive function is enriched in evolutionarily conserved regions. *Translational Psychiatry* **6**, e980 (2016).

16. Di Angelantonio E, *et al.* Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *The Lancet* **390**, 2360-2371 (2017).

17. Astle WJ, *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429.e1419 (2016).

18. Hill WD, Harris SE, Deary IJ. What genome-wide association studies reveal about the association between intelligence and mental health. *Current Opinion in Psychology* **27**, 25-30 (2019).

19. Hill WD. Comment on 'Large-Scale Cognitive GWAS Meta-Analysis Reveals Tissue-Specific Neural Expression and Potential Nootropic Drug Targets' by Lam et al. *Twin Research and Human Genetics* **21**, 84-88 (2018).

20. Hill WD. A Further Comment on 'Large-Scale Cognitive GWAS Meta-Analysis Reveals Tissue-Specific Neural Expression and Potential Nootropic Drug Targets' by Lam et al. *Twin Research and Human Genetics*, 1-8 (2018).

21. Hill WD, Davies G, Liewald DC, McIntosh AM, Deary IJ. Age-Dependent Pleiotropy Between General Cognitive Function and Major Psychiatric Disorders. *Biological Psychiatry* **80**, 266-273 (2016).

22. Lam M, *et al.* Pleiotropic Meta-Analysis of Cognition, Education, and Schizophrenia Differentiates Roles of Early Neurodevelopmental and Adult Synaptic Pathways. *The American Journal of Human Genetics* **105**, 334-350 (2019).

23. Bulik-Sullivan B, *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**, 1236 (2015).

24. Weiner DJ, *et al.* Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nature Genetics* **49**, 978 (2017).

25. Turley P, *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, (2018).

26. Ramasamy A, *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature Neuroscience* **17**, 1418 (2014).

27. GTEx Consortium, *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204 (2017).

28. Schmitt Anthony D, *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports* **17**, 2042-2059 (2016).

29. Roadmap Epigenomics C, *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317 (2015).

30. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. *Nature reviews Genetics* **14**, 483 (2013).

31. Skene NG, *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nature Genetics* **50**, 825-833 (2018).

32. Hill WD, *et al.* Human cognitive ability is influenced by genetic variation in components of postsynaptic signalling complexes assembled by NMDA receptors and MAGUK proteins. *Translational Psychiatry* **4**, e341 (2014).

33. Hill WD, *et al.* Genetic contributions to two special factors of neuroticism are associated with affluence, higher intelligence, better health, and longer life. *Molecular Psychiatry*, (2019).

34. Luciano M, *et al.* Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nature Genetics* **50**, 6-11 (2018).

35. Anderson EL, *et al.* Education, intelligence and Alzheimer's disease: Evidence from a multivariable two-sample Mendelian randomization study. *bioRxiv*, (2018).

36. Sanderson E, Macdonald-Wallis C, Davey Smith G. Negative control exposure studies in the presence of measurement error: implications for attempted effect estimate calibration. *International Journal of Epidemiology* **47**, 587-596 (2018).

37. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLOS Genetics* **13**, e1007081 (2017).

38. Cesarini D, Lindqvist E, Östling R, Wallace B. Wealth, Health, and Child Development: Evidence from Administrative Data on Swedish Lottery Players *. *The Quarterly Journal of Economics* **131**, 687-738 (2016).

39. Grotzinger AD, *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour* **3**, 513-525 (2019).

40. Abdellaoui A, *et al.* Genetic correlates of social stratification in Great Britain. *Nature Human Behaviour*, (2019).

41. Koellinger PD, Harden KP. Using nature to understand nurture. *Science* **359**, 386-387 (2018).

42. Kong A, *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424-428 (2018).

43. Brumpton B, *et al.* Within-family studies for Mendelian randomization: avoiding dynastic, assortative mating, and population stratification biases. *bioRxiv*, 602516 (2019).

44. Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics* **48**, 709 (2016).

45. Hemani G, Bowden J, Davey Smith G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human Molecular Genetics* **27**, R195-R208 (2018).

46. Trampush JW, *et al.* GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: a report from the COGENT consortium. *Molecular Psychiatry* **22**, 336 (2017).

47. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology*, (2018).

48. Hill WD, *et al.* Genomic analysis of family data reveals additional genetic effects on intelligence and personality. *Molecular Psychiatry*, (2018).

49. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology* **47**, 226-235 (2018).

50. Townsend P. Deprivation. *Journal of Social Policy* **16**, 125-146 (2009).

51. Leake J. Scientists find 24 'golden' genes that help you get rich. *The Times*, Retrieved from https://www.thetimes.co.uk (2019).

52. Smith GD. Life-course approaches to inequalities in adult chronic disease risk: Boyd Orr Lecture. *Proceedings of the Nutrition Society* **66**, 216-236 (2007).

53. Rimfeld K, *et al.* Genetic influence on social outcomes during and after the Soviet era in Estonia. *Nature Human Behaviour* **2**, 269-275 (2018).

54. Abdellaoui A, *et al.* Genetic Consequences of Social Stratification in Great Britain. *bioRxiv*, 457515 (2018).

55. 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).

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

57. Wain LV, *et al.* Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nature genetics* **49**, 416 (2017).

58. Winkler TW, *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nature protocols* **9**, 1192 (2014).

59. Haplotype Reference C. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* **48**, 1279-1283 (2016).

60. Genomes Project C. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56 (2012).

61. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).

62. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88**, 76-82 (2011).

63. Watanabe K, Taskesen E, Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nature communications* **8**, 1826 (2017).

64. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164-e164 (2010).

65. Butterworth AS, *et al.* ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Research* **47**, e3-e3 (2018).

66. Schmitt AD, *et al.* A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell reports* **17**, 2042-2059 (2016).

67. Kundaje A, *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).

68. Liu JZ, *et al.* A versatile gene-based test for genome-wide association studies. *Am J Hum Genet* **87**, 139-145 (2010).

69. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comp Biol* **11**, (2015).

70. de Leeuw CA, Neale BM, Heskes T, Posthuma D. The statistical properties of gene-set analysis. *Nature Reviews Genetics* **17**, 353 (2016).

71. Ashburner M, *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 4 (2000).

72. Fabregat A, *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* **44**, D481-D487 (2015).

73. Subramanian A, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

74. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648-660 (2015).

75. Gazal S, *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature Genetics* **49**, 1421 (2017).

76. Finucane HK, *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228 (2015).

77. Finucane HK, *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics* **50**, 621-629 (2018).

78. Pers TH, *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications* **6**, 5890 (2015).

79. Fehrmann RSN, *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature Genetics* **47**, 115 (2015).

80. The Encode Project Consortium, *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).

81. Cahoy JD, *et al.* A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. *The Journal of Neuroscience* **28**, 264-278 (2008).

82. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society Series B (Methodological)*, 289-300 (1995).

83. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* **40**, 304-314 (2016).

84. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512-525 (2015).

85. Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey Smith G. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *Am J Clin Nutr* **103**, 965-978 (2016).

86. Deary IJ, Harris SE, Hill WD. What genome-wide association studies reveal about the association between intelligence and physical health, illness, and mortality. *Current Opinion in Psychology* **27**, 6-12 (2019).

87. Zhu Z, *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature Communications* **9**, 224 (2018).

88. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466-1468 (2015).

89. Smith BH, *et al.* Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *International Journal of Epidemiology* **42**, 689-700 (2013).

90. Smith BH, *et al.* Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Medical Genetics* **7**, 74 (2006).

91. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**, 76-82 (2011).

92. Okbay A, *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet*, (2016).

93. Gale CR, Čukić I, Batty GD, McIntosh AM, Weiss A, Deary IJ. When is higher neuroticism protective against death? Findings from UK Biobank. *Psychological science* **28**, 1345-1357 (2017).

Acknowledgements

This work was undertaken in The University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology (CCACE), supported by the cross-council Lifelong Health and Wellbeing initiative (MR/K026992/1).

Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006] and from the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award "STratifying Resilience and Depression Longitudinally" (STRADL) Reference 104036/Z/14/Z). Funding from the Biotechnology and Biological Sciences Research Council (BBSRC), the Medical Research Council (MRC), and the University of Edinburgh and gratefully acknowledged. CCACE funding supports IJD. CH was supported by an MRC University Unit Programme Grant MC_UU_00007/10 (QTL in Health and Disease). The Medical Research Council (MRC) and the University of Bristol support the MRC Integrative Epidemiology Unit [MC_UU_12013/1, MC_UU_12013/9, MC_UU_00011/1]. The Economics and Social Research Council (ESRC) support NMD via a Future Research Leaders grant [ES/N000757/1]. This work is part of a project entitled 'social and economic consequences of health: causal inference methods and longitudinal, intergenerational data', which is part of the Health Foundation's Efficiency Research Programme.

Participants in INTERVAL were recruited with the active collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk). DNA extraction and genotyping was co-funded by the National Institute for Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk/) and the NIHR [Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust] [*]. The academic coordinating centre for INTERVAL was supported by core funding from: NIHR Blood and Transplant Research Unit in Donor Health and Genomics (NIHR BTRU-2014-10024), UK Medical Research Council (MR/L003120/1), British Heart Foundation (RG/13/13/30194) and the NIHR [Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust]. A complete list of the investigators and contributors to the INTERVAL trial is provided in reference[16]. The academic coordinating centre would like to thank blood donor centre staff and blood donors for participating in the INTERVAL trial.

*The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

WDH is supported by a grant from Age UK (Disconnected Mind Project). We thank George Davey Smith for his comments on an early version of this manuscript.

Author contributions

W.D.H conceptualization, data curation, data analysis, data visualization, interpretation of results, writing−original draft, writing−review and editing, project administration and co-ordination. N.M.D. data analysis, interpretation of results, writing−review and editing. S.J.R. interpretation of results, writing−review and editing. N.G.S. data analysis, interpretation of results, writing−review and editing. J.B. writing−review and editing. S.B. data analysis, interpretation of results, writing−review and editing. E. Di A. writing−review and editing. D.J.R. writing−review and editing. S.X. data analysis, data visualization, writing−review and editing. G.D. writing−review and editing. D.C.M.L writing−review and editing. D.J.P. data curation, writing−review and editing. C.H. data curation, writing−review and editing. A.S.B. data curation, writing−review and editing. A.M.M. writing−review and editing. C.R.G. writing−review and editing. I.J.D funding acquisition, data interpretation, writing−review and editing.

Competing interests

All other authors declare no competing interests.

Figure legends

Figure 1. Illustrating a possible pathway from genetic inheritance to income. In this pathway there are no direct effects of genetic variants on income. Rather, mediated pleiotropy (also termed vertical

pleiotropy shown in blue) is used to understand, in part, the link between genetic variation and more biologically distal phenotypes such as income and education. Mediated pleiotropy describes instances where genetic variation is linked to a phenotype (in this case income) through genetic effects that act on another partly-heritable trait. These partly-heritable traits would also be associated with income, and so the genetic effects that act on them would also be associated with income. For simplicity, this schematic illustrates only two possible pathways between genetic variation and income. In reality there may be, and are likely to be, many links between genetic variation, including bi-directional causality between the phenotypes in the pathway, and the more biologically distal phenotypes such as income.

Figure 2. Flow chart for the statistical analysis carried out using the GWAS data on household income in 286,301 White British participants in UK Biobank. Blue indicates a type of analysis conducted (i.e. LDSC to derive a heritability estimate) and gold indicates a subtype of this type of analysis (i.e. global heritability or the heritability of a stratified subset of the SNPs). Green indicates the result of an analysis (i.e. the global heritability was 7.39%).

Figure 3. SNP level associations for income and mapping of the SNPs in independent genomic loci. **Figure 3A**. Manhattan plot for income; negative log10 transformed P-values for each SNP are plotted against chromosomal location. The red line indicates genome-wide significance and the black line indicates suggestive associations (1×10^{-5}). **Figure 3B**. Functional annotation carried out on the independent genomic loci identified. The percentage of SNPs found in each of the nine functional categories is listed. **Figure 3C**. The percentage of SNPs from the independent genomic loci that fell into each of the Regulome DB scores categories. A lower score indicates greater evidence for that SNPs involvement in gene regulation. **Figure 3D**. The percentage of SNPs within the independent genomic loci plotted against the minimum chromatic state for 127 tissue/cell types. **Figure 3E**. Venn diagram illustrating the overlap of the genes implicated using positional mapping, eQTL mapping, chromatin interaction mapping, that was conducted on the independent significant loci identified in

1227 the SNP-based GWAS. Also shown is how these implicated genes overlap with those identified using
1228 the gene-based statistics derived using MAGMA.
1229
1230
1231 **Figure 4**. Gene level associations for income and links to transcription in the brain and cortical cell
1232 types. **Figure 4A**. A Manhattan plot of income using the gene-based statistics derived using
1233 MAGMA; negative log10 transformed P-values for each gene are plotted against chromosomal
1234 location. The red line indicates genome-wide significance. **Figure 4B.** The results of a gene-property
1235 analysis linking transcription differences in the brain with income differences. Significant links
1236 between expression differences in cerebellar hemisphere, at Brodmann area 9 (BA9) of the frontal
1237 cortex, the nucleus accumbens and at Brodmann area 24 of the anterior cingulate cortex (BA24) are
1238 illustrated. Dark blue indicates low $-\log_{10}$ P-values (a lower level of association) describing the link
1239 between gene expression and household income and light blue indicates high $-\log_{10}$ P-values (a higher
1240 level of association) describing the same relationship. The full results found in **Supplementary Data**
1241 **8)** with the gene based statistics produced using MAGMA. **Figure 4C.** Shows the results of a cell type
1242 specific gene-property analysis where the relationship between the gene-based statistics from
1243 MAGMA and the degree to which gene expression was specific to the annotations was examined. A
1244 Bonferroni correction was applied to control for the 24 tests conducted. The red line indicates
1245 statistical significance indicating that expression that is specific to the annotation is associated with
1246 the gene-based statistics for income.
1247 Embr. DA Neurons, Embryonic Dopaminergic Neurons; Pyramidal (SS), Pyramidal
1248 (Somatosensory); Hypoth. GABAergic Neurons, Hypothalamic GABAergic Neurons; DA Neuroblast,
1249 Dopaminergic Neuroblast.
1250
1251 **Figure 5.** Partitioned heritability of income. **Figure 5A.** Enrichment analysis for income using the 60
1252 functional categories as well as 10 minor allele frequency groupings and 6 continuous annotations (27
1253 categories describing enrichment within these categories is shown. This analysis differs from that
1254 presented in **Figure 3** and **Figure 4** as here all SNPs are used, not only those that reached genome

wide significance (**Figure 3**) or SNPs that were located within protein coding genes (**Figure 4**). The enrichment statistic is the proportion of heritability found in each functional group divided by the proportion of SNPS in each group ($Pr(h^2)/Pr(SNPs)$). The red line indicates no enrichment found when $Pr(h^2)/Pr(SNPs) = 1$. Error bars represent ± 1 standard error. A Bonferroni correction controlling for 52 tests was used to ascertain statistical significance which is indicated by an asterisk. **Figure 5B.** Enrichment analysis for the six continuous annotations by quintile. Shading represents quintile with light colours corresponding to low quintiles and dark colours to high quintiles. Groupings that contained a significantly greater proportion of heritability proportional to the number of SNPs they contain are marked with an asterisk. Multiple testing was performed within each of the annotations resulting in an alpha level of $\alpha=0.05/5=0.01$ with a red line indicating no enrichment. Error bars represent ± 1 standard error. **Figure 5C & 5D.** Shows the enrichment of 205 tissue of cell types assembled using gene expression data and 489 groupings assembled using chromatin data. In each instance these were arranged into 9 tissue type groupings with correction for multiple testing been performed using false discovery rate (FDR)[82] conducted separately for the gene expression and the chromatin groupings indicated by a red line. **Figure 5E.** Shows if the genes that are expressed 13 brain regions are enriched for the heritability of income. A Bonferroni correction was used to control for 13 tests and the alpha level was set at 0.004 with the brain regions that crossed the red line being those that were statistical significant. **Figure 5D.** Shows the level of enrichment for three brain cell types. A Bonferroni correction was used to control for three tests ($\alpha=0.05/3=0.017$) and groupings that crossed the red line were those that were statistically significant. The full results for each of these analyses can be found in **Supplementary Data 9, Supplementary Table 5, Supplementary Data 10-11, and Supplementary Tables 6-7**.

Figure 6. Pairs of genetic correlations are compared. **Figure 6A** compares genetic correlations derived using income and with those derived using income conditioned on intelligence, **Figure 6B** compares genetic correlations derived using intelligence with those derived using intelligence conditioned on income, and **Figure 6C** compares genetic correlations derived using income with

those derived using education. In each instance 27 pairs of genetic correlations are compared. Genetic correlations that were significantly different between within each of the three comparisons described above using a two-sided test (2*pnorm(-abs(abs(r_{gi} – r_{gj}) / sqrt(SE_i^2+SE_j^2)))) are indicated by an asterisk next to the phenotype label. Abbreviations, MDD, major depressive disorder; ADHD, attention deficit hyperactivity disorder; T2D, type 2 diabetes; CAD, coronary artery disease; SRH, self-rated health; SWB, subjective wellbeing; BMI, body mass index. Worry/vulnerability and anxiety/tension were derived as special factors of neuroticism.[33, 93] Full results for each of the genetic correlations derived can be found in **Supplementary Table 12**. Error bars represent ± 1 standard error.

Figure 7. Polygenic risk score analysis of income. **Figure 7A.** Violin plot showing the level of household income in GS:SFHS plotted against the standardized polygenic score of income in each group. Median and interquartile range are plotted. Summary data from the income GWAS performed in UK Biobank was used to derive PGRSs. Red line indicates a standardized polygenic score of 0. **Figure 7B.** The average level of household income for the five PGRSs is shown. Summary data from the income GWAS performed in UK Biobank was used to derive PGRSs. The y-axis corresponds to the 5 point classification of household income in Generation Scotland. Above the red line indicates a level of household income between £30,000 and £50,000 and below indicates a level of household income between £10,000 and £30,000 in Generation Scotland. Error bars represent ± 1 standard error. **Figure 7C**. The variance accounted for by each of the five P-value cut offs for the PGRS. Light orange indicates that the income phenotype derived in UK Biobank was used to generate polygenic risk scores along with Generation Scotland. The dark orange bars indicate instances of where the MTAG phenotype derived using income and educational attainment was used to derive polygenic risk scores in Generation Scotland. Summary data from the income GWAS performed in UK Biobank, and the MTAG analysis of income was used to derive PGRSs. All results can be found in **Supplementary Table 9**.

1310

Figure 1

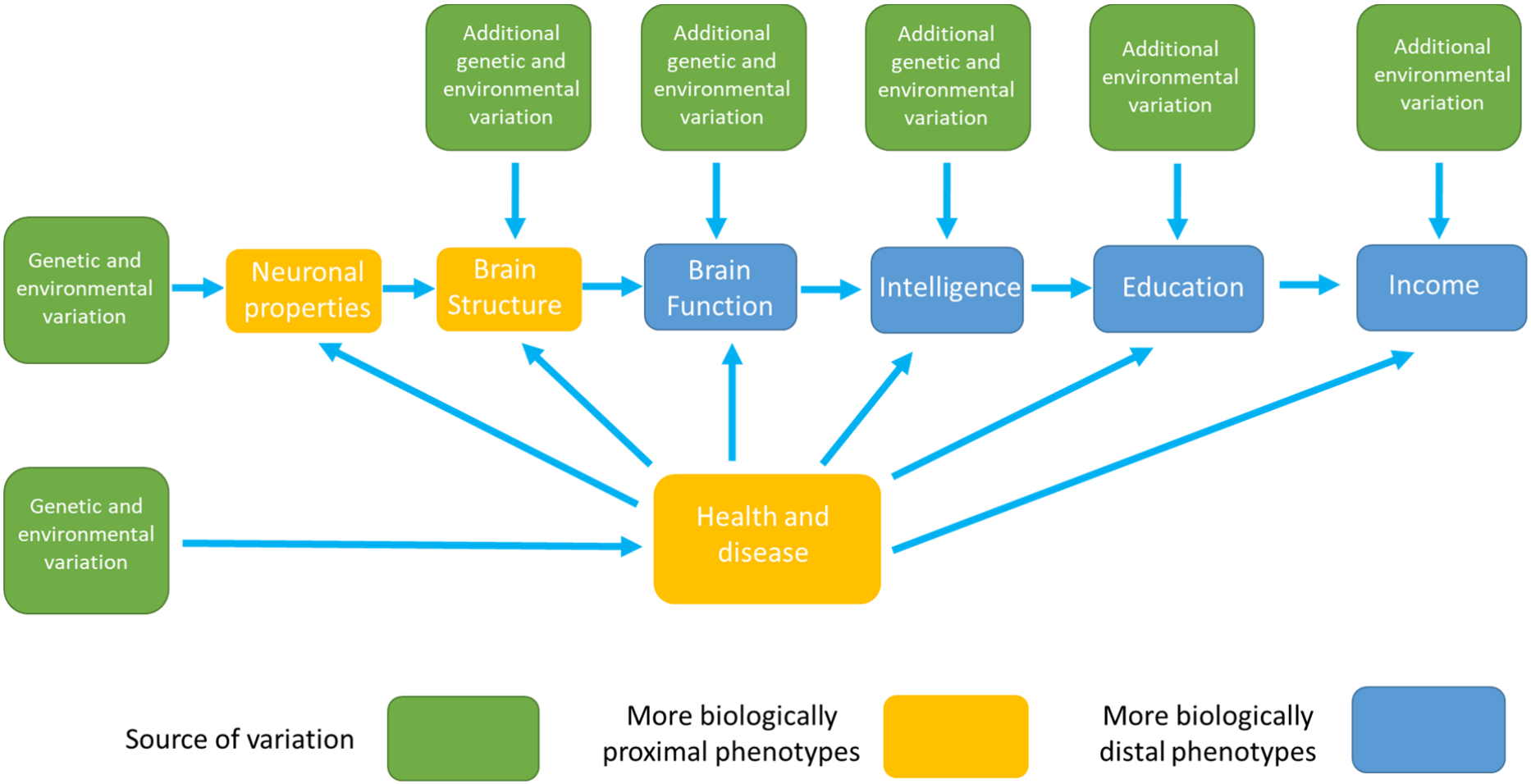

Additional genetic and environmental variation
Additional genetic and environmental variation
Additional genetic and environmental variation
Additional environmental variation
Additional environmental variation
Genetic and environmental variation
Neuronal properties
Brain Structure
Brain Function
Intelligence
Education
Income
Genetic and environmental variation
Health and disease
Source of variation
More biologically proximal phenotypes
More biologically distal phenotypes

Figure 2

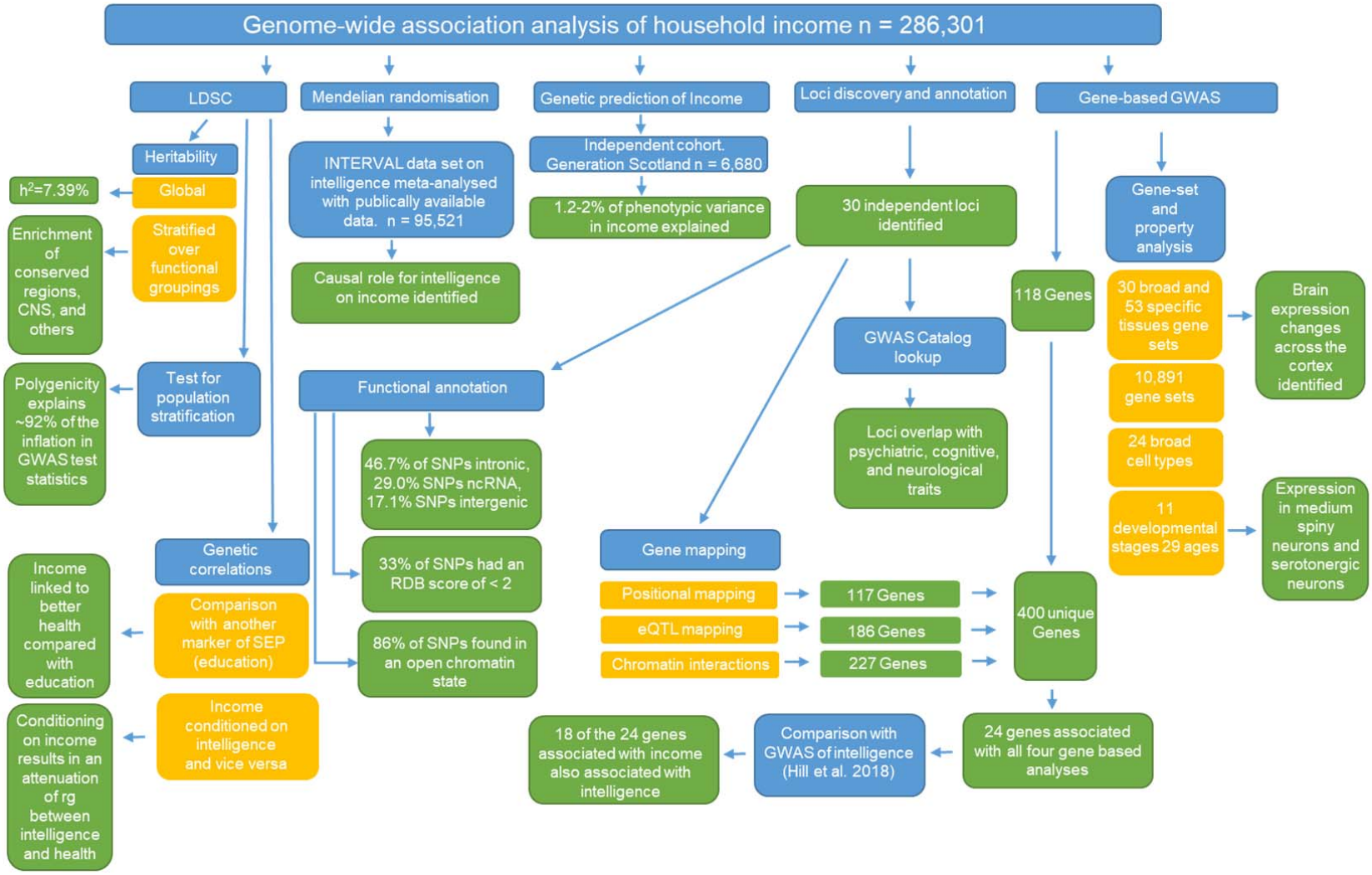

Genome-wide association analysis of household income n = 286,301
LDSC
Mendelian randomisation
Genetic prediction of Income
Loci discovery and annotation
Gene-based GWAS
Heritability
Global
h²=7.39%
Stratified over functional groupings
Enrichment of conserved regions, CNS, and others
Test for population stratification
Polygenicity explains ~92% of the inflation in GWAS test statistics
INTERVAL data set on intelligence meta-analysed with publically available data. n = 95,521
Causal role for intelligence on income identified
Independent cohort. Generation Scotland n = 6,680
1.2-2% of phenotypic variance in income explained
30 independent loci identified
GWAS Catalog lookup
Loci overlap with psychiatric, cognitive, and neurological traits
Functional annotation
46.7% of SNPs intronic, 29.0% SNPs ncRNA, 17.1% SNPs intergenic
33% of SNPs had an RDB score of < 2
86% of SNPs found in an open chromatin state
Genetic correlations
Comparison with another marker of SEP (education)
Income linked to better health compared with education
Income conditioned on intelligence and vice versa
Conditioning on income results in an attenuation of rg between intelligence and health
Gene mapping
Positional mapping
eQTL mapping
Chromatin interactions
117 Genes
186 Genes
227 Genes
400 unique Genes
24 genes associated with all four gene based analyses
Comparison with GWAS of intelligence (Hill et al. 2018)
18 of the 24 genes associated with income also associated with intelligence
118 Genes
Gene-set and property analysis
30 broad and 53 specific tissues gene sets
Brain expression changes across the cortex identified
10,891 gene sets
24 broad cell types
11 developmental stages 29 ages
Expression in medium spiny neurons and serotonergic neurons

Figure 3

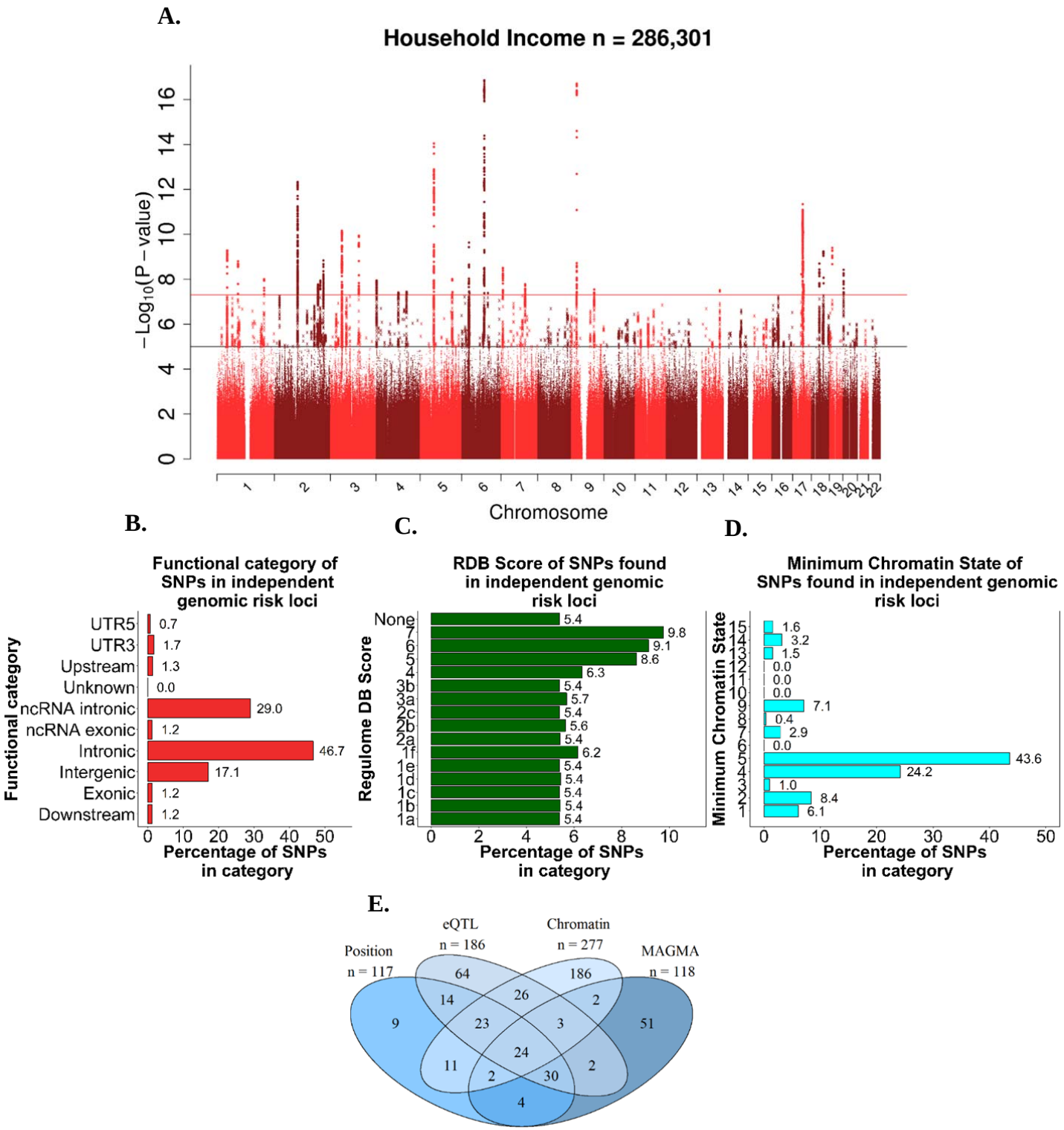

A.
Household Income n = 286,301
−Log10(P−value)
Chromosome
B.
Functional category of SNPs in independent genomic risk loci
Functional category
UTR5 0.7
UTR3 1.7
Upstream 1.3
Unknown 0.0
ncRNA intronic 29.0
ncRNA exonic 1.2
Intronic 46.7
Intergenic 17.1
Exonic 1.2
Downstream 1.2
Percentage of SNPs in category
C.
RDB Score of SNPs found in independent genomic risk loci
Regulome DB Score
None 5.4
7 9.8
6 9.1
5 8.6
4 6.3
3b 5.4
3a 5.7
2c 5.4
2b 5.6
2a 5.4
1f 6.2
1e 5.4
1d 5.4
1c 5.4
1b 5.4
1a 5.4
Percentage of SNPs in category
D.
Minimum Chromatin State of SNPs found in independent genomic risk loci
Minimum Chromatin State
15 1.6
14 3.2
13 1.5
12 0.0
11 0.0
10 0.0
9 7.1
8 0.4
7 2.9
6 0.0
5 43.6
4 24.2
3 1.0
2 8.4
1 6.1
Percentage of SNPs in category
E.
Position n = 117
eQTL n = 186
Chromatin n = 277
MAGMA n = 118
64
186
14
26
2
9
23
3
51
24
11
2
30
2
4

Figure 4

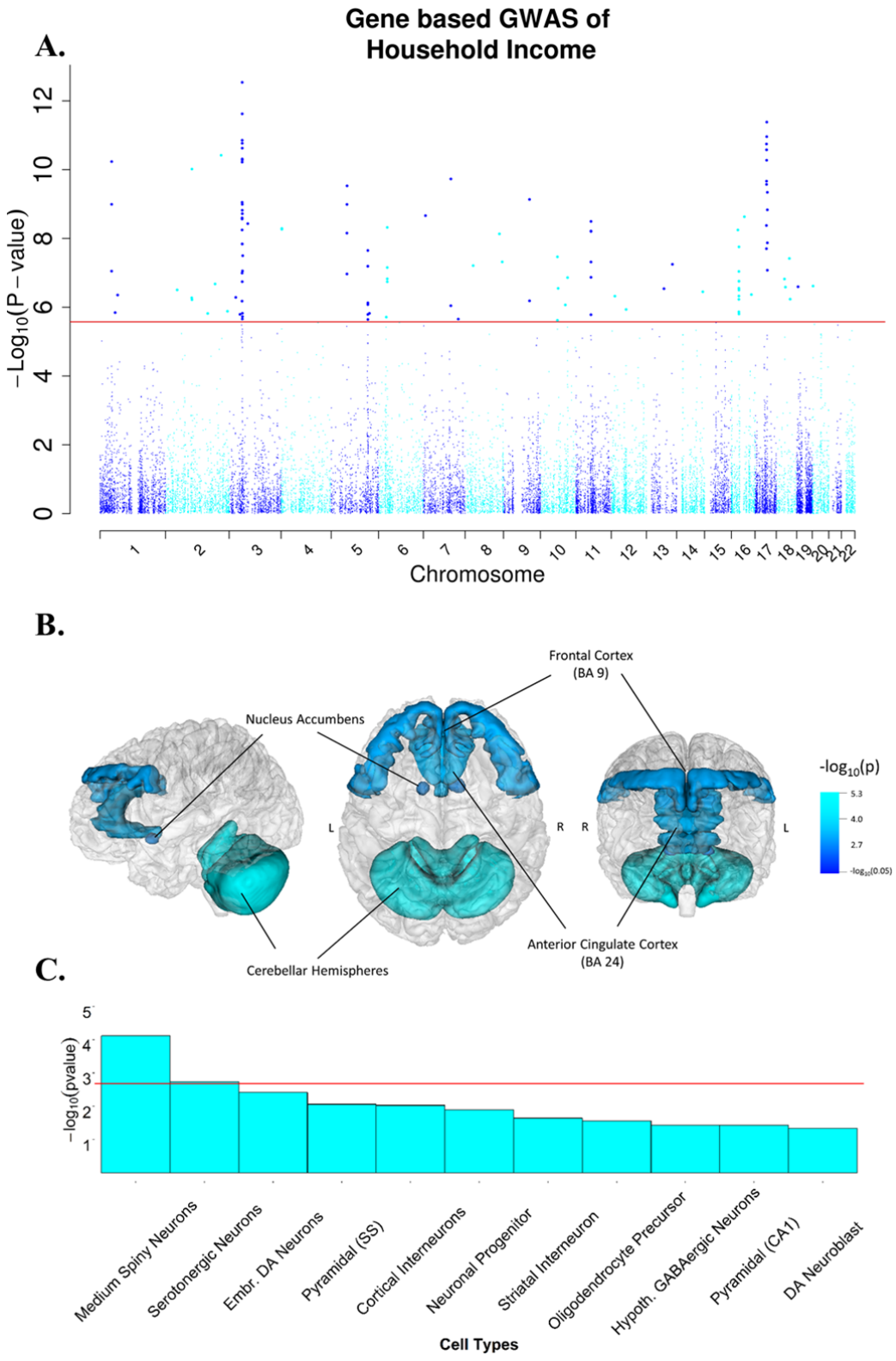

A.
Gene based GWAS of Household Income
−Log10(P−value)
0
2
4
6
8
10
12
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome
B.
Frontal Cortex (BA 9)
Nucleus Accumbens
L
R
R
L
-log10(p)
5.3
4.0
2.7
-log10(0.05)
Anterior Cingulate Cortex (BA 24)
Cerebellar Hemispheres
C.
−log10(pvalue)
5
4
3
2
1
Medium Spiny Neurons
Serotonergic Neurons
Embr. DA Neurons
Pyramidal (SS)
Cortical Interneurons
Neuronal Progenitor
Striatal Interneuron
Oligodendrocyte Precursor
Hypoth. GABAergic Neurons
Pyramidal (CA1)
DA Neuroblast
Cell Types

Figure 5

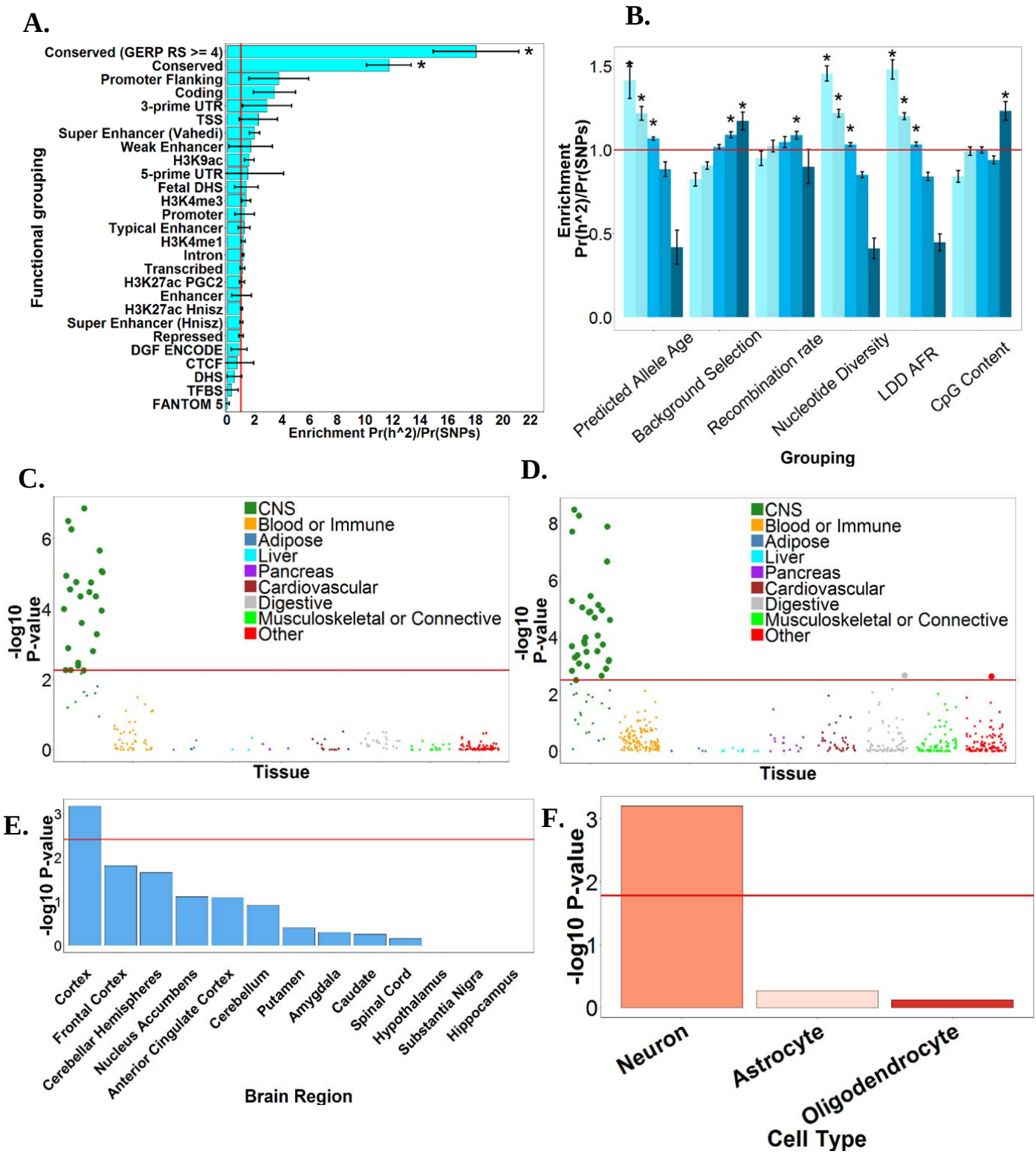

A.
Conserved (GERP RS >= 4)
Conserved
Promoter Flanking
Coding
3-prime UTR
TSS
Super Enhancer (Vahedi)
Weak Enhancer
H3K9ac
5-prime UTR
Fetal DHS
H3K4me3
Promoter
Typical Enhancer
H3K4me1
Intron
Transcribed
H3K27ac PGC2
Enhancer
H3K27ac Hnisz
Super Enhancer (Hnisz)
Repressed
DGF ENCODE
CTCF
DHS
TFBS
FANTOM 5
Functional grouping
0 2 4 6 8 10 12 14 16 18 20 22
Enrichment Pr(h^2)/Pr(SNPs)
B.
Enrichment Pr(h^2)/Pr(SNPs)
0.0 0.5 1.0 1.5
Predicted Allele Age
Background Selection
Recombination rate
Nucleotide Diversity
LDD AFR
CpG Content
Grouping
C.
-log10 P-value
0 2 4 6
CNS
Blood or Immune
Adipose
Liver
Pancreas
Cardiovascular
Digestive
Musculoskeletal or Connective
Other
Tissue
D.
-log10 P-value
0 2 4 6 8
CNS
Blood or Immune
Adipose
Liver
Pancreas
Cardiovascular
Digestive
Musculoskeletal or Connective
Other
Tissue
E.
-log10 P-value
0 1 2 3
Cortex
Frontal Cortex
Cerebellar Hemispheres
Nucleus Accumbens
Anterior Cingulate Cortex
Cerebellum
Putamen
Amygdala
Caudate
Spinal Cord
Hypothalamus
Substantia Nigra
Hippocampus
Brain Region
F.
-log10 P-value
0 1 2 3
Neuron
Astrocyte
Oligodendrocyte
Cell Type

Figure 6

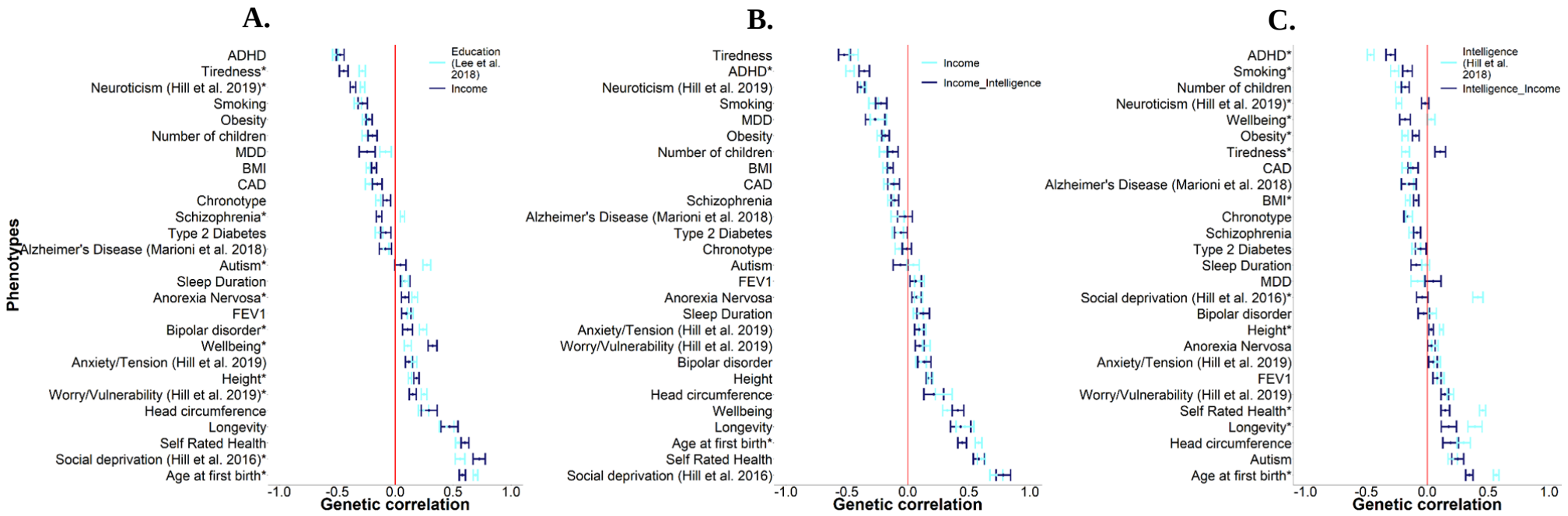
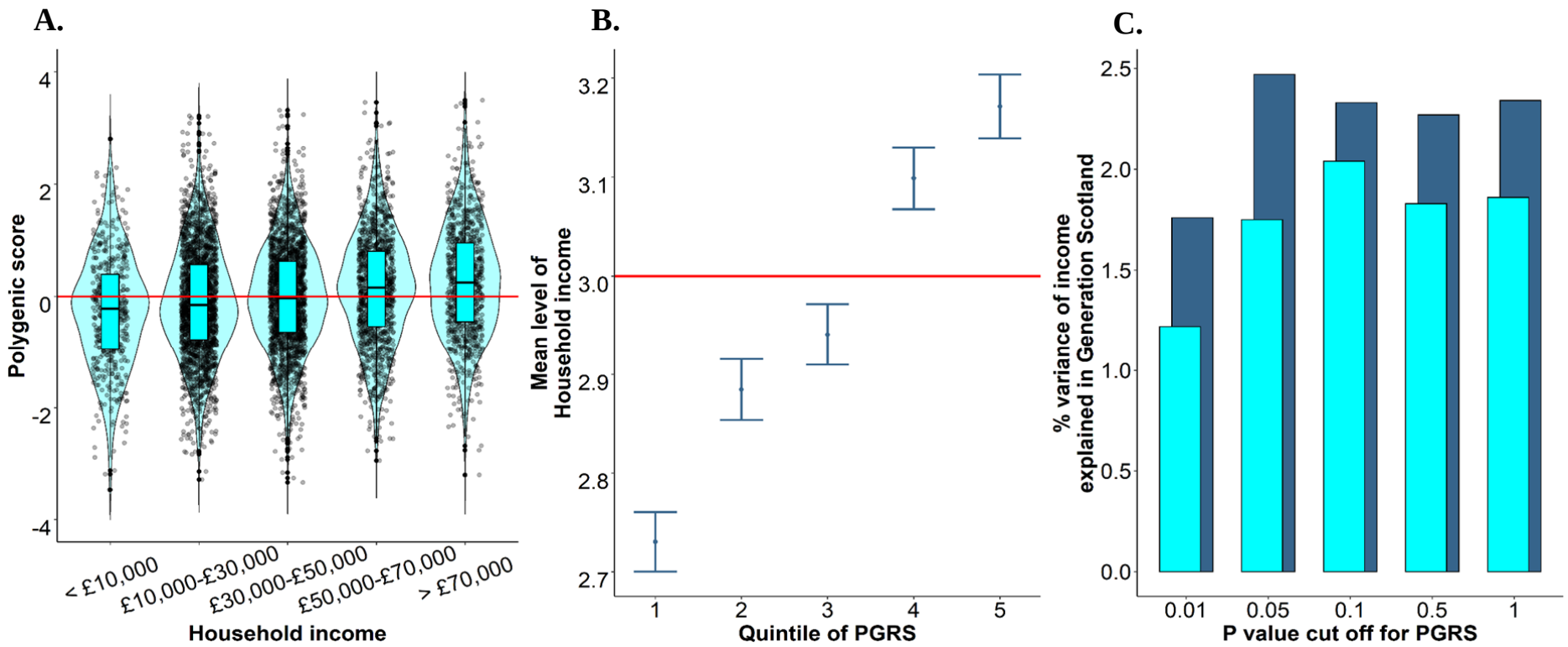

A.
B.
C.
Phenotypes
ADHD
Tiredness*
Neuroticism (Hill et al. 2019)*
Smoking
Obesity
Number of children
MDD
BMI
CAD
Chronotype
Schizophrenia*
Type 2 Diabetes
Alzheimer's Disease (Marioni et al. 2018)
Autism*
Sleep Duration
Anorexia Nervosa*
FEV1
Bipolar disorder*
Wellbeing*
Anxiety/Tension (Hill et al. 2019)
Height*
Worry/Vulnerability (Hill et al. 2019)*
Head circumference
Longevity
Self Rated Health
Social deprivation (Hill et al. 2016)*
Age at first birth*
Education (Lee et al. 2018)
Income
-1.0
-0.5
0.0
0.5
1.0
Genetic correlation
Tiredness
ADHD*
Neuroticism (Hill et al. 2019)
Smoking
MDD
Obesity
Number of children
BMI
CAD
Schizophrenia
Alzheimer's Disease (Marioni et al. 2018)
Type 2 Diabetes
Chronotype
Autism
FEV1
Anorexia Nervosa
Sleep Duration
Anxiety/Tension (Hill et al. 2019)
Worry/Vulnerability (Hill et al. 2019)
Bipolar disorder
Height
Head circumference
Wellbeing
Longevity
Age at first birth*
Self Rated Health
Social deprivation (Hill et al. 2016)
Income
Income_Intelligence
-1.0
-0.5
0.0
0.5
1.0
Genetic correlation
ADHD*
Smoking*
Number of children
Neuroticism (Hill et al. 2019)*
Wellbeing*
Obesity*
Tiredness*
CAD
Alzheimer's Disease (Marioni et al. 2018)
BMI*
Chronotype
Schizophrenia
Type 2 Diabetes
Sleep Duration
MDD
Social deprivation (Hill et al. 2016)*
Bipolar disorder
Height*
Anorexia Nervosa
Anxiety/Tension (Hill et al. 2019)
FEV1
Worry/Vulnerability (Hill et al. 2019)
Self Rated Health*
Longevity*
Head circumference
Autism
Age at first birth*
Intelligence (Hill et al. 2018)
Intelligence_Income
-1.0
-0.5
0.0
0.5
1.0
Genetic correlation

Figure 7


A.
Polygenic score
4
2
0
-2
-4
< £10,000
£10,000-£30,000
£30,000-£50,000
£50,000-£70,000
> £70,000
Household income
B.
Mean level of Household income
3.2
3.1
3.0
2.9
2.8
2.7
1
2
3
4
5
Quintile of PGRS
C.
% variance of income explained in Generation Scotland
2.5
2.0
1.5
1.0
0.5
0.0
0.01
0.05
0.1
0.5
1
P value cut off for PGRS