bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Associations between common genetic variants and income provide insights about the socioeconomic health gradient

Data and code are available at
https://nodes.desci.com/node/MOUMrtGN0CkZkK3xqKzh16POxiKDGc95o-EF-ElvbJ4/v3

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Associations between common genetic variants and income provide insights about the socioeconomic health gradient

Hyeokmoon Kweon[1], Casper A.P. Burik[1], Yuchen Ning[1], Rafael Ahlskog[2], Charley Xia[3], Erik Abner[4], Yanchun Bao[5], Laxmi Bhatta[6], Tariq O. Faquih[7], Maud de Feijter[8], Paul Fisher[9], Andrea Gelemanović[10], Alexandros Giannelis[11], Jouke-Jan Hottenga[12], Bita Khalili[13,14], Yunsung Lee[15], Ruifang Li-Gao[7], Jaan Masso[16], Ronny Myhre[17], Teemu Palviainen[18], Cornelius A. Rietveld[19,20], Alexander Teumer[21,22], Renske M. Verweij[23], Emily A. Willoughby[11], Esben Agerbo[24,25,26], Sven Bergmann[13,14], Dorret I. Boomsma[12,27,28], Anders D. Børglum[24,29,30], Ben M. Brumpton[31,32,33], Neil Martin Davies[34,31,35], Tõnu Esko[4], Scott D. Gordon[36], Georg Homuth[37], M. Arfan Ikram[8], Magnus Johannesson[38], Jaakko Kaprio[18], Michael P. Kidd[39,40], Zoltán Kutalik[41,13], Alex S.F. Kwong[42,43], James J. Lee[11], Annemarie I. Luik[8,44], Per Magnus[15], Pedro Marques-Vidal[45,46], Nicholas G. Martin[36], Dennis O. Mook-Kanamori[8,47], Preben Bo Mortensen[24,25,26], Sven Oskarsson[2], Emil M. Pedersen[24,25,26], Ozren Polašek[10,48], Frits R. Rosendaal[7], Melissa C. Smart[49], Harold Snieder[50], Peter J. van der Most[50], Peter Vollenweider[45,46], Henry Völzke[21], Gonneke Willemsen[12,51], Jonathan P. Beauchamp[52], Thomas A. DiPrete[53], Richard Karlsson Linnér[54,1], Qiongshi Lu[55], Tim T. Morris[56,41], Aysu Okbay[1], K. Paige Harden[57], Abdel Abdellaoui[58,*], W. David Hill[59,3], Ronald de Vlaming[1], Daniel J. Benjamin[60,61,62], Philipp D. Koellinger[1,*]

[1]Department of Economics, School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [2]Department of Government, Uppsala University, Uppsala, Sweden. [3]Department of Psychology, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK. [4]Institute of Genomics, University of Tartu, Tartu, Estonia. [5]School of Mathematics, Statistics and Actuarial Sciences, University of Essex, Essex, UK. [6]HUNT Center for Molecular and Clinical Epidemiology, Department of Public Health and Nursing, NTNU Norwegian University of Science and Technology, Trondheim, Norway. [7]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands. [8]Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands. [9]Institute for Social and Economic Research, Wivenhoe Park, University of Essex, Essex, UK. [10]Department of Public Health, University of Split School of Medicine, Split, Croatia. [11]Department of Psychology, University of Minnesota Twin Cities, Minneapolis, USA. [12]Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [13]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [14]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [15]Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway. [16]School of Economics and Business Administration, University of Tartu, Tartu, Estonia. [17]Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway. [18]Institute for Molecular Medicine Finland - FIMM, University of Helsinki, Helsinki, Finland. [19]Department of Applied Economics, Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands. [20]Rotterdam Institute for Behavior and Biology, Erasmus University Rotterdam, Rotterdam, The Netherlands. [21]Institute for Community Medicine, University Medicine Greifswald, Germany. [22]Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany. [23]Department of Public Administration and Sociology, Erasmus University Rotterdam, Rotterdam, The Netherlands. [24]iPSYCH - The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Aarhus University, Aarhus, Denmark. [25]National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark, [26]School of Business and Social Sciences, Aarhus University, Aarhus, Denmark. [27]Amsterdam Public Health (APH), Amsterdam UMC, Amsterdam, The Netherlands. [28]Amsterdam Reproduction & Development (AR&D), Amsterdam UMC, Amsterdam, The Netherlands. [29]Department of Biomedicine, Aarhus University, Aarhus, Denmark. [30]Center for Genome Analysis and Personalized Medicine, Aarhus, Denmark. [31]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology, Trondheim, Norway. [32]HUNT Center for Molecular and Clinical Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Levanger, Norway. [33]Clinic of Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. [34]Division of Psychiatry, and Department of Statistical Sciences, University College London, London, UK. [35]Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [36]Genetic Epidemiology Lab, Queensland Institute of Medical Research, Brisbane, Australia. [37]Interfaculty Institute for Genetics and Functional Genomics, University

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Medicine Greifswald, Germany [38]Department of Economics, Stockholm School of Economics, Stockholm, Sweden. [39]Economics, RMIT University, Melbourne, Australia. [40]International School of Technology and Management, Feng Chia University, Taichung, Taiwan. [41]University Center for Primary Care and Public Health, Unisante, Lausanne, Switzerland. [42]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [43]Division of Psychiatry, University of Edinburgh, Edinburgh, UK. [44]Trimbos Institute - Netherlands Institute for Mental Health and Addiction, Utrecht, The Netherlands, [45]Department of Medicine, Internal Medicine, Lausanne University Hospital (CHUV), Lausanne, Switzerland. [46]Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland, [47]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands. [48]Algebra University, Zagreb, Croatia. [49]Institute for Social and Economic Research, University of Essex, Essex, UK. [50]Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [51]Faculty of Health, Sports and Wellbeing, Inholland University of Applied Sciences, Haarlem, The Netherlands. [52]Interdisciplinary Center for Economic Science and Department of Economics, George Mason University, Fairfax, VA, USA, [53]Department of Sociology, Columbia University, New York, USA. [54]Department of Economics, Leiden Law School, Universiteit Leiden, Leiden, The Netherlands. [55]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA. [56]Centre for Longitudinal Studies, Social Research Institute, University College London, London, UK. [57]Department of Psychology and Population Reseach Center, University of Texas at Austin, Austin, TX, USA. [58]Department of Psychiatry, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands. [59]Lothian Birth Cohort Studies, University of Edinburgh, Edinburgh, UK. [60]Anderson School of Management, UCLA, Los Angeles, CA, USA. [61]Human Genetics Department, UCLA, David Geffen School of Medicine, Los Angeles, CA, USA. [62]National Bureau of Economic Research, Cambridge, MA, USA.

* Contacts: Philipp D. Koellinger (p.d.koellinger@vu.nl), Abdel Abdellaoui (a.abdellaoui@amsterdamumc.nl)

Abstract

We conducted a genome-wide association study (GWAS) on income among individuals of European descent and leveraged the results to investigate the socio-economic health gradient (*N*=668,288). We found 162 genomic loci associated with a common genetic factor underlying various income measures, all with small effect sizes. Our GWAS-derived polygenic index captures 1 - 4% of income variance, with only one-fourth attributed to direct genetic effects. A phenome-wide association study using this polygenic index showed reduced risks for a broad spectrum of diseases, including hypertension, obesity, type 2 diabetes, coronary atherosclerosis, depression, asthma, and back pain. The income factor showed a substantial genetic correlation (0.92, *s.e.* = .006) with educational attainment (EA). Accounting for EA's genetic overlap with income revealed that the remaining genetic signal for higher income related to better mental health but reduced physical health benefits and increased participation in risky behaviours such as drinking and smoking.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Income is a crucial determinant of individuals' access to resources and overall quality of life. Extensive evidence shows that higher income is positively correlated with increased subjective well-being, better health, and longer life expectancy.[1–5] For instance, the gap in life expectancy between the richest and poorest 1% of individuals in the US has been estimated to be 14.6 years for men (95% CI, 14.4 to 14.8 years) and 10.1 years for women (95% CI, 9.9 to 10.3 years).[6] Notably, higher income is associated with increased longevity and well-being across the entire income distribution, highlighting its broad relevance in current society.[3,6,7]

Income is a complex phenotype influenced by many factors, including environmental conditions and education.[8,9] Parents' socio-economic status shapes a child's developmental trajectory, including their skills, behaviours, educational attainment, career prospects, and eventual adult income.[10,11] Moreover, certain heritable individual characteristics, such as cognitive ability and personality traits,[12–14] are well-known predictors of income within contemporary Western societies. Twin studies have estimated income heritability in these societies to be around 40-50%.[15–17] However, the heritability of income and its associated genes are not fixed; rather, they reflect social realities shaped by technological, institutional, and cultural factors.[18] These factors are malleable and exhibit variations across different regions and historical epochs, which can lead to fluctuations in heritability estimates for socio-economic status (SES) over time[19,20] and imperfect genetic correlations across samples.[21]

The results from statistically well-powered GWAS of SES present numerous opportunities to shed light on these social realities. For example, they allow investigating questions about sex differences in labour market processes, cross-country comparisons in the genetic architecture of income, and investigating the processes contributing to intergenerational social mobility.[22] They also facilitate studies investigating the interaction effects between genetic and environmental factors. Furthermore, they enable the exploration of genetic correlations between income and health outcomes, potentially unveiling new insights into the socioeconomic health gradient.

Two previous GWAS have been conducted on household income.[23,24] The first was in a sample of 96,900 participants from the initial release of the UK Biobank (UKB)[25] and found two loci. The second was carried out in the full release of the UKB with 286,301 individuals and found 30 approximately uncorrelated loci. A meta-analysis of these results with the genetically correlated trait educational attainment increased the effective sample size to 505,541 individuals and identified 144 loci. A recent GWAS on occupational status in the UKB

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data identified cognitive skills, scholastic motivation, occupational aspiration, personality traits, and behavioural disinhibition (proxied by ADHD) as potential mediating factors linking genetics to occupational status.[26]

Building on these earlier contributions, we conducted a GWAS leveraging multiple income measures. We ran sex-stratified analyses and meta-analyzed results from 32 cohorts across 12 economically advanced countries and three continents, yielding the largest GWAS on income to date with an effective sample size of N = 668,288 (Table 1). Due to data availability and statistical power considerations, our analyses and conclusions are restricted to individuals carrying genotypes most similar to the EUR panel of the 1000 Genomes dataset, as compared to individuals sampled elsewhere in the world (1KG-EUR-like individuals).

The greater statistical power of our GWAS enabled us to conduct a series of follow-up analyses that investigate the socio-economic health gradient from a genetic perspective. In particular, we leveraged the data to compare the GWAS results for income and EA to disentangle their unique genetic correlates with health. Furthermore, our multi-sample approach and sex-specific GWAS results allowed us to test for possible differences in the genetic architecture of income across samples and sexes.

For a less technical description of the paper and how it should -- and should not -- be interpreted, see the **Frequently Asked Questions** document (FAQ) and **Box 1**.

Results

Multivariate GWAS of income

We used four measures of income (individual, occupational, household, and parental income) and conducted a GWAS meta-analysis of their shared genetic basis. **Supplementary Information section 2.1** discusses the differences between these measures and their relative advantages and disadvantages as proxies for individual income. Dropping parental income from the meta-analysis leads to a slight statistical power decrease but does not qualitatively change our results.

A sex-stratified GWAS was carried out on each available income measure in each cohort. We restricted our analyses to 1KG-EUR-like individuals who were not currently enrolled in an educational program or who were aged above 30 if their current enrollment status was unknown. The natural log transformation was applied to the income measures. We applied standardised quality control procedures to each cohort-level result (see **Supplementary**

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Information section 2.4 for details). For each sex and each income measure, we performed a sample-size-weighted meta-analysis with METAL.[27] We then meta-analyzed the male and female results of each income measure using MTAG,[28] which accounts for any potential genetic relatedness between the male and female samples.

Across the four GWAS on different income measures, we identified 86 non-overlapping loci in the genome (see Supplementary Information section 2.6 for the definition of loci and lead SNPs). **Table 1** summarises the results. Occupational and household income produced the most genetic associations (59 and 41 loci, respectively), as expected based on sample sizes and SNP-based heritability estimates based on linkage disequilibrium score regression (LDSC) (h^2 = 0.08 (*s.e.* = 0.003) and 0.06 (*s.e.* = 0.003), respectively). The four income measures' pairwise genetic correlation (r_g) estimates demonstrated substantial shared genetic variance, with all pairwise r_g's at least 0.8 (**Fig. 1a**).

Table 1. GWAS summary

Measure	*N*	Female %	# SNP	Mean χ^2	# Loci	h^2 (s.e.)
Household	497,413	0.55	11,500,222	1.54	41	0.06 (0.003)
Individual	72,601	0.54	5,986,804	1.06	0	0.04 (0.007)
Occupational	443,064	0.57	11,500,419	1.64	59	0.08 (0.003)
Parental	128,724	0.50	6,144,179	1.11	1	0.05 (0.006)
INC factor	668,288*	-	9,131,507	1.94	162	0.07 (0.002)

Note: INC factor is from the meta-analysis across the four income measures: individual, occupational, household, and parental. * is the estimated effective sample size reported for the INC factor. Some individuals contributed multiple times to different income measures. The mean χ^2 was computed only with the HapMap 3 SNPs. The number of approximately independent loci (sixth column) was obtained using FUMA. The SNP heritability (h^2) was estimated with LDSC.

Next, we meta-analyzed the association results across the four income measures using MTAG (see Supplementary Information section 2.5 for details). We observed that the MTAG procedure yields nearly identical results to genomic SEM's common factor function.[29] Thus, we hereafter refer to the meta-analyzed income as 'the income factor' (INC factor). Since

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MTAG already applies a bias correction with the intercept from LDSC,[30] we did not apply further adjustments for cryptic relatedness and population stratification.

The INC factor GWAS was estimated to have an effective sample size of 668,288, based on occupational income's heritability scale (N_{eff} = 1,198,347 based on individual income). The meta-analysis across the income measures led to a substantial increase in power, which allowed us to identify 162 loci tagged by 207 lead SNPs (**Fig. 1b**). 88 of these loci were newly identified compared to the previously published GWAS household income result conducted in the UKB.[24] The genetic correlation of the previous household income GWAS result was 0.92 (*s.e.* = 0.008) with the INC factor and 0.94 (*s.e.* = 0.006) when we restrict our analysis to only our household income measure.

The effect sizes of the lead SNPs were small. Adjusting for the statistical winner's curse, one additional count in the effect allele of the median lead SNP was associated with an increase in income of 0.30% (these effect-size calculations require an assumption about the standard deviation of the dependent variable because MTAG yields standardised effect-size estimates; we use the standard deviation estimate of log hourly occupational wage from the UKB, which is 0.35). The estimated effects at the 5th and 95th percentiles were 0.18 and 0.60%, respectively (see **Supplementary Information section 2.7**). To put these estimates into perspective, the median annual earnings of full-time workers in the US was $56,473 in 2021.[31] A 0.3% increase would equal an additional annual income of $169. In terms of the variance explained (R^2), all of the lead SNPs each had R^2 lower than 0.011% after adjustment for the statistical winner's curse (**Supplementary Fig. 2**).

Cross-sex and cross-country heterogeneity

The heritability of income and its genetic associations may vary across different social environments or different groups within an environment. To investigate the potential heterogeneity of genetic associations with income, we examined cross-cohort genetic correlations. We found that the inverse-variance weighted mean genetic correlations across pairs of cohorts were 0.45 (*s.e.* = 0.22) for individual income, 0.52 (*s.e.* = 0.13) for household income, and 0.90 (*s.e.* = 0.24) for occupational income (**Supplementary Tables 28a-c**).

Next, we meta-analyzed cohorts from the same country with the same income measure available and estimated the genetic correlations across these countries (Estonia, Netherlands, Norway, United Kingdom, USA - **Extended Data Figure 1a)**. For most country-pairs, the genetic correlation of the same income measure is >0.8. While meta-analysis increases statistical power and yields more precise estimates of the average effect size, it also tends to

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mask non-random heterogeneity in effect size estimates across samples. Despite this latter point, we find that occupational income in Norway displayed lower genetic correlations with occupational or household income in other countries, ranging from 0.43 (*s.e.* = 0.23) to 0.82 (*s.e.* = 0.10). Similarly, occupational income's genetic correlation with educational attainment (EA) was also lower in Norway (r_g = 0.69, *s.e.* = 0.08) compared to the other countries. These findings align with phenotypic evidence that ranks Norway lowest among OECD countries in terms of financial returns for obtaining a college degree.[32] Next, we investigated whether the large number of samples from the United Kingdom in our meta-analysis could have skewed our results. To address this, we conducted a separate meta-analysis procedure for the British and non-British cohorts, comprising participants from 10 countries. We obtained two distinct GWAS results for the INC factor and found a perfect genetic correlation of 1.001 (*s.e.* = 0.03) between them. Thus, the average effect sizes of SNPs associated with the INC factor are almost identical in British and non-British cohorts.

We observed slight between-sex heterogeneity in the genetic associations of income, as supported by the evidence presented in **Extended Data Figure 1b**. The estimated between-sex genetic correlations based on meta-analysed GWAS results for individual, occupational, and household income were 1.06 (s.e. = 0.32), 0.91 (s.e. = 0.03), and 0.95 (s.e. = 0.03), respectively. Notably, the latter two estimates were statistically distinguishable from unity but remained above 0.9. Most cohort-specific cross-sex genetic correlations for income are too noisy to be interpreted (**Supplementary Tables 17b-d**). One exception is the UK Biobank sample, which shows a non-perfect genetic correlation between men and women for occupational income (r_g = 0.91, *s.e.* = 0.03). Another exception is the Danish iPsych cohort, where we estimated a genetic correlation of 0.76 (s.e. = 0.10) between maternal and paternal income. These findings are consistent with the hypothesis that men and women face non-identical labour market conditions. The lower genetic correlation between maternal and paternal income suggests that differences in labour market conditions were more pronounced in previous generations.

We also conducted the INC factor GWAS for the male and female results separately and found that their genetic correlation was statistically indistinguishable from one (r_g = 0.98, *s.e.* = 0.02).

Comparison with educational attainment

To compare the GWAS results for the INC factor with those for EA, we first conducted an auxiliary GWAS on EA to obtain the most-powered GWAS result of EA with the summary statistics currently available to us: We first carried out a GWAS of EA in the UKB, based on

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the protocol of the latest EA GWAS (EA4).[33] We then meta-analyzed these GWAS results with the EA3 summary statistics[21] that did not include the UKB, using the meta-analysis version of MTAG. While previous GWASs on income found somewhat inconsistent results on the genetic correlation between educational attainment (EA)[21,33] and income (r_g = 0.90 (*s.e.* = 0.04)[23] and 0.77 (*s.e.* = 0.02)[24]), with much greater precision, we found a high genetic correlation that is very close to the first reported estimate (r_g = 0.917, *s.e.* = 0.006). Among the input income measures, the genetic correlation with EA was higher for occupational and parental income (r_g = 0.95 and 0.92; *s.e.* = 0.01 and 0.05 respectively) and lower for individual and household income (r_g = 0.81 and 0.82; *s.e.* = 0.07 and 0.01 respectively). Furthermore, 138 out of 161 loci for the INC factor overlapped with those for EA.

The r_g estimate of 0.917 between the INC factor and EA implies that only $1 - 0.917^2$ = ~16% of the genetic variance of the INC factor would remain once the genetic covariance with EA was statistically removed. We employed the GWAS-by-subtraction approach using Genomic SEM[29] to identify this residual genetic signal (referred to as 'NonEA-INC'). We identified one locus of genome-wide significance for NonEA-INC, marked by the lead SNP rs34177108 on chromosome 16 (**Extended Data Fig 2c**). This locus was previously found to be associated with vitamin D levels and hair and skin-related traits such as colour, sun exposure, and cancer, possibly picking up on uncontrolled population stratification or physical traits linked to discrimination in the labour market.

Polygenic prediction

We conducted polygenic index (PGI) analyses with individuals of European ancestry in the Swedish Twin Registry (STR), which was not included in our meta-analysis. We chose STR as the main prediction cohort because it has twins and administrative data on individual, occupational, and household income. In addition, we also used the UKB siblings (UKB-sib) and the Health and Retirement Study (HRS) from the US as prediction cohorts. For the UKB-sib, occupational and household income measures were available. For the HRS, a self-reported individual income measure was available. In the STR and the UKB-sib cohorts, except when examining within-family prediction, we randomly selected only one individual from each family.

After generating hold-out versions of GWAS on the INC factor and EA to remove the sample overlap with each prediction sample, we constructed PGIs for the INC factor and EA using LDpred2[34]. Before conducting prediction analyses, we residualised the log of income on demographic covariates, including a third-degree polynomial in age, year of observation, and

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

interactions with sex. We measured the prediction accuracy as the incremental R^2 from adding the PGI to a regression of the phenotype on a set of baseline covariates, which were the top 20 genetic principal components and genotype batch indicators.

A cohort-specific upper bound for the theoretically possible predictive accuracy of PGIs on income can be obtained by the GREML[35] estimate of the SNP-based heritability of income, which is close to 10% for the available income measures in the STR and UKB sibling sample (**SI Table 13**).

The actual prediction accuracy of PGIs for income is lower than the theoretical maximum, primarily due to finite GWAS sample size but also due to imperfect genetic correlations across meta-analysed cohorts and differences in measurement accuracy of income across samples.[36] In the STR (**Fig. 2**), the INC factor PGI predicted ΔR^2 = 1.3% (95% CI: 1.0-1.6) for individual income, 3.7% (95% CI: 3.1-4.2) for occupational income, and 1.0% (95% CI: 0.6-1.4) for household income. The EA PGI had predictive accuracy results in a similar range for individual and household income, except for occupational income, for which the accuracy was larger: ΔR^2 = 4.7% (95% CI: 4.0-5.4). In the UKB-sib, the predictive accuracy of the INC factor PGI was ΔR^2 = 4.7% (95% CI: 4.3-5.2) for occupational income and 3.9% (95% CI: 3.5-4.3) for household income. The EA PGI achieved a better predictive accuracy for occupational income (ΔR^2 = 6.9%, 95% CI: 6.3-7.4), while only slightly higher for household income (ΔR^2 = 4.4%, 95% CI: 3.9-4.8). In the HRS, INC factor PGI had ΔR^2 = 2.7% (95% CI: 2.1-3.3) for predicting individual income, which was close to the EA PGI's result (ΔR^2 = 3.1%, 95% CI: 2.4-3.8).

In terms of the coefficient estimates in the UKB-sib, one standard deviation increase in the INC factor PGI was associated with a 7.2% increase in the occupational income (95% CI: 6.7-7.7) and a 12.3% increase in the household income (95% CI: 11.4-13.2). These estimates were comparable to the effect of one additional year of schooling on income, whose estimates tend to range from 5 to 15%.[8,9,37]

The predictive power of the INC factor PGI approached zero once EA or the EA PGI was controlled for. In the UKB-sib, ΔR^2 decreased below 1% for occupational and household income, while the estimates were still statistically different from zero (**Extended Data Fig. 3** and **Supplementary Table 21**).

We estimated the share of the direct genetic effect in the overall population effect captured by the INC factor PGI, following the recent approach that imputes parental genotypes from first-degree relatives.[33,38] Using the UKB-sib sample, we isolated the direct effect of the PGI from the population effect on occupational and household income by controlling for

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

parental PGIs. We found that the ratio of direct-to-population effect estimates is 0.51 (*s.e.* = 0.05) and 0.49 (*s.e.* = 0.05) for occupational and household income, respectively (**Supplementary Table 22**). These results imply that only 24.0% or 25.7% of the INC factor PGI's predictive power was due to direct genetic effects, which was very close to the result for the EA PGI estimated elsewhere (25.5%).[38]

We next conducted a phenome-wide association study of the INC factor PGI based on electronic health records in the UKB siblings holdout sample. We tested 115 diseases with sex-specific sample prevalence no lower than 1%. In total, 50 diseases from different categories were associated with the INC factor PGI after Bonferroni correction and 14 after controlling for parental PGI (**Fig. 3, Extended Data Fig. 4** and **Supplementary Tables 27a-b**). In all cases, a higher INC factor PGI value was associated with reduced disease risk, including reduced risk for hypertension, gastroesophageal reflux disease (GERD), type 2 diabetes, obesity, osteoarthritis, back pain, and depression. The strongest association of a higher INC factor PGI and a disease was found for essential hypertension.

Genetic correlations

We next explored the genetic correlations of the income (INC) factor, educational attainment (EA), and NonEA-INC with phenotypes related to behaviours, psychiatric disorders, and physical health (**Fig. 4**). LDSC estimates revealed that the genetic correlations of EA and the INC factor largely align. However, noticeable differences emerged for traits in the psychiatric and psychological domains. Specifically, NonEA-INC is associated with a reduced risk for certain psychiatric disorders previously reported to correlate positively with EA.[39–41] These discrepancies were observed for schizophrenia (r_g = -0.29, *s.e.* = 0.04), autism spectrum (r_g = -0.27, *s.e.* = 0.06), and obsessive-compulsive disorder (r_g = -0.22, *s.e.* = 0.08). One possible interpretation of these findings is that these psychiatric disorders have more severe negative effects on individual performance in the labour market than in the educational system.

Intriguingly, NonEA-INC exhibits a near-zero genetic correlation with cognitive performance (r_g = 0.03, *s.e.* = 0.03). At the same time, both EA and the general income (INC) factor display strong positive genetic correlations with it (r_g = 0.66, *s.e.* = 0.01 and r_g = 0.63, *s.e.* = 0.01, respectively). This may suggest that high cognitive performance primarily influences income through education. Furthermore, this result is consistent with high income being attainable through social connections, inherited wealth, entrepreneurial activities, or the pursuit of well-paying jobs that do not require high cognitive performance.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

While EA and the general INC factor have substantial negative genetic correlations with health-related behaviours such as age of smoking initiation, smoking persistence, cigarettes per day, and alcohol dependence, we found that NonEA-INC has near-zero genetic correlations with these traits (albeit the latter have substantially larger error margins of the point estimates).

NonEA-INC also displayed genetic correlations with other phenotypes that are similar to EA. Specifically, NonEA-INC had negative genetic correlations with major depressive disorder (r_g = -0.15, *s.e.* = 0.04), anxiety disorder (r_g = -0.19, *s.e.* = 0.05), and the related trait of neuroticism (r_g = -0.14, *s.e.* = 0.03), but positive genetic correlations with subjective well-being (r_g = 0.32, *s.e.* = 0.06), general risk tolerance (r_g = 0.13, *s.e.* = 0.04), and height (r_g = 0.11, *s.e.* = 0.03). The differences in correlations for neuroticism, subjective well-being, and risk tolerance were statistically significant when comparing EA and NonEA-INC, with NonEA-INC showing stronger positive correlations with well-being and risk tolerance and a less negative correlation with neuroticism.

Implicated genes and tissue-specific enrichments:

We used FUMA[42] to find genes implicated in INC factor GWAS. FUMA uses four mapping approaches: positional, chromatin interaction, expression quantitative trait locus (eQTL) mapping, and MAGMA gene-based association tests. In total, 2,385 protein-coding genes were implicated by at least one of the methods, out of which 225 genes were implicated by all four methods (**Extended Data Fig. 5a**). Only three of these commonly implicated genes were unique for the INC factor, compared to the genes implicated in EA GWAS by at least one of the four methods or previously prioritised for EA.[21]

We then performed tissue-specific enrichment analyses using LDSC-SEG[43] and MAGMA gene-property analyses[44] (see **Supplementary Information section 7**). Both methods indicated dominant enrichment for tissues of the central nervous system (**Extended Data Fig. 5b**), consistent with the previous results for household income and EA.[21,24]

Discussion

We conducted the largest GWAS on income to date, incorporating individual, household, occupational, and parental income measures. Our study design provided increased statistical power, identifying more genetic variants and improving the predictive power of the polygenic index (PGI) compared to previous income GWAS. Additionally, it allowed for comprehensive additional analyses.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We observed substantial heterogeneity in the genetic architecture of income across cohorts and high, but non-perfect genetic correlations of income across sexes. This underlines that the genetic associations we report here are averages across different groups and environments that should not be interpreted as fixed or universal.

Furthermore, we found a strong genetic correlation between income and educational attainment (EA).

Our analyses highlighted numerous associations between better health and higher income that are influenced by genetic differences among individuals. These better health outcomes include lower BMI, blood pressure, type-2 diabetes, depression, and reduced stress-related disorders. We note that the genetic overlap between income and health could be driven by different causal mechanisms, including pleiotropic effects of genes, limited income opportunities for individuals with health problems, or health advantages for individuals with higher income. Investigating these causal mechanisms is outside the scope of this study.

Interestingly, the genetic components of income not shared with EA (NonEA-INC factor) showed weaker associations with better physical health and health-related behaviour, such as drinking and smoking. One possible interpretation of this finding is that better health outcomes of higher socioeconomic status in wealthy countries are more due to their association with education rather than with income or wealth, consistent with findings from quasi-experimental studies.[45–47]

In contrast, we found negative genetic correlations of the NonEA-INC factor with schizophrenia, bipolar disorder, autism, and obsessive-compulsive disorder, while EA exhibited positive genetic correlations with these psychiatric outcomes. This may indicate that the educational system is more accommodating to individuals with these disorders than the labour market and/or that talents associated with these genetic risks (e.g., higher IQ with autism[48] or creativity with bipolar disorder and schizophrenia[49]) are more advantageous in school than in the labour market.

While our GWAS results contribute to constructing an income-specific PGI with improved predictive accuracy, the EA PGI remains a comparable or even better predictor of income and socio-economic status. This is due to even larger sample sizes in recent GWAS on EA ($N \sim 3$ million), lower measurement error in educational attainment compared to measures of income and the high genetic correlation between income and EA.

It is important to point out that the results of our study reflect the specific social realities of the analysed samples and are not universal or unchangeable. This is exemplified by the substantial heterogeneity in the genetic architecture of income that we found across cohorts.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We emphasise that our results are limited to individuals whose genotypes are genetically most similar to the EUR panel of the 1000 Genomes reference panel compared to people sampled in other parts of the world. Our results have limited generalizability and do not warrant meaningful comparisons across different groups or predictions of income for specific individuals (**FAQ**). To increase the representation of individuals from diverse backgrounds, cohort and longitudinal studies should seek to sample more diverse and representative samples of the global population.

Studies of genetic analyses of behavioural phenotypes have been prone to misinterpretation, such as characterising identified associated variants as 'genes for income.' Our study illustrates that such characterisation is incorrect for many reasons: The effect of each individual SNP on income is minimal, capturing less than 0.01% of the overall variance in income. Furthermore, the genetic loci we identified correlate with many other traits, including education and a wide range of health outcomes. Finally, the finding that only one-quarter of the genetic associations we identified are due to direct genetic effects suggests the potential importance of family-specific factors, including potential resemblance between parents, and environmental factors as important drivers of income inequality.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Box 1. Understanding Genetics and Income: A Cautionary Overview

Given the frequent misunderstanding of research on genetics and human behaviour, it is important to recognize the complexities underlying connections between genes and social outcomes and to communicate what our findings mean clearly and with appropriate nuance.

What did we do and why?

Several types of 'luck' help shape an individual's life trajectory, such as their society of birth, parents, and the genetic variants they inherit. Our study captures elements of this by examining the relationship between millions of genetic variants and income through a genome-wide association study (GWAS). GWASs of income can provide valuable insights into the genetic factors associated with income and how they interact with environmental factors, enhancing our understanding of intergenerational mobility and socioeconomic disparities.

GWASs of income can shed light on societal processes that favour certain genetic predispositions, providing insights into our socioeconomic system, but also into the relationships between income and health disparities. Recent GWASs have shown that socio-economic outcomes share genetic overlap with various health outcomes, with a considerable portion mediated through social environments.[50]

What did we find?

We identified numerous genetic variants associated with income, each with minor effects but collectively correlating with education, cognition, behaviour, and health. We found notable differences between income and educational attainment in their genetic associations with health outcomes. For several psychiatric disorders - namely autism, schizophrenia, and OCD - the genetic relationships acted in opposing directions. Shared genetic effects between income and health may stem from various causes. Genes might affect both income and health. Alternatively, higher income could lead to better health outcomes, not only directly but also indirectly through improved living conditions from family-members or neighbourhoods. Conversely, existing health problems may limit income opportunities, potentially due to reduced work capacity or increased healthcare costs.

When predicting differences between siblings, the overall predictive strength of these genetic effects diminishes significantly — by approximately 75%. Possible explanations for this include that the direct causal effects of the genetic variants are smaller compared to the causal effects of environmental factors that correlate with these genetic variants (e.g., the effects of parental nurture on their children) and that the way parents resemble each other (assortative mating) magnifies the predictive power of genetic effects.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We observed some variability in the genetic factors influencing income across the Western countries we analysed and between genders, underscoring that the genetic associations we report here should not be interpreted as fixed or universal.

Neither genetic nor environmental determinism is warranted

Historically, misinterpreting the role of genetics in shaping social outcomes has occasionally fueled controversial ideologies with far-reaching consequences. It is important to mitigate the risk of such misunderstandings, particularly the notions of genetic or environmental determinism. In this context, we emphasise the following:

One's genetic makeup or the family and societal environment into which they are born does not dictate their intrinsic value. The genetic variants that matter for income, and their effects, depend on the environment, i.e., on what skills are valued by the labour market and by society. As the labour market changes, or as government policies change, so can the variants and their effects.

It is important to recognize how genetics can impact income through diverse pathways, affecting one's own or one's parents' health, cognition, skills, and productivity-related behavioural tendencies, such as creativity, risk taking, or adaptability. Additionally, genetics can influence characteristics favoured or discriminated against in the labour market due to societal preferences.

As with previous genetic studies on social outcomes like educational attainment, the findings of this study have limited generalisability across different populations.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Ridley, M., Rao, G., Schilbach, F. & Patel, V. Poverty, depression, and anxiety: Causal evidence and mechanisms. *Science* **370**, eaay0214 (2020).
2. Braveman, P. A., Cubbin, C., Egerter, S., Williams, D. R. & Pamuk, E. Socioeconomic Disparities in Health in the United States: What the Patterns Tell Us. *Am. J. Public Health* **100**, S186–S196 (2010).
3. Stevenson, B. & Wolfers, J. Subjective well-being and income: Is there any evidence of satiation? *Am. Econ. Rev.* **103**, 598–604 (2013).
4. Wilkinson, R. G. & Marmot, M. *Social determinants of health: the solid facts*. (World Health Organization, 2003).
5. Stringhini, S. *et al.* Socioeconomic status and the 25× 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1· 7 million men and women. *The Lancet* **389**, 1229–1237 (2017).
6. Chetty, R. *et al.* The association between income and life expectancy in the United States, 2001-2014. *Jama* **315**, 1750–1766 (2016).
7. Adler, N. E. *et al.* Socioeconomic status and health: The challenge of the gradient. *Am. Psychol.* **49**, 15–24 (1994).
8. Card, D. Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica* **69**, 1127–1160 (2001).
9. Trostel, P., Walker, I. & Woolley, P. Estimates of the economic return to schooling for 28 countries. *Labour Econ.* **9**, 1–16 (2002).
10. Becker, G. S. & Tomes, N. An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility. *J. Polit. Econ.* **87**, 1153–1189 (1979).
11. Bowles, S. & Gintis, H. The Inheritance of Inequality. *J. Econ. Perspect.* **16**, 3–30 (2002).
12. Bowles, S., Gintis, H. & Osborne, M. The Determinants of Earnings: A Behavioral Approach. *J. Econ. Lit.* **39**, 1137–1176 (2001).
13. Acemoglu, D. Technical change, inequality, and the labor market. *J. Econ. Lit.* **40**, 7–72 (2002).
14. Corak, M. Income inequality, equality of opportunity, and intergenerational mobility. *J. Econ. Perspect.* **27**, 79–102 (2013).
15. Hyytinen, A., Ilmakunnas, P., Johansson, E. & Toivanen, O. Heritability of lifetime earnings. *J. Econ. Inequal.* **17**, 319–335 (2019).
16. Taubman, P. The determinants of earnings: Genetics, family, and other environments: A study of white male twins. *Am. Econ. Rev.* **66**, 858–870 (1976).
17. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
18. Harden, K. P. & Koellinger, P. D. Using genetics for social science. *Nat. Hum. Behav.* **4**, 567–576 (2020).
19. Silventoinen, K. *et al.* Genetic and environmental variation in educational attainment: an individual-based analysis of 28 twin cohorts. *Sci. Rep.* **10**, 12681 (2020).
20. Rimfeld, K. *et al.* Genetic influence on social outcomes during and after the Soviet era in Estonia. *Nat. Hum. Behav.* **2**, 269–275 (2018).
21. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112 (2018).
22. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424–

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

428 (2018).
23. Hill, W. D. *et al.* Molecular Genetic Contributions to Social Deprivation and Household Income in UK Biobank. *Curr. Biol.* **26**, 3083–3089 (2016).
24. Hill, W. D. *et al.* Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income. *Nat. Commun.* **10**, 1–16 (2019).
25. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
26. Akimova, E. T., Wolfram, T., Ding, X., Tropf, F. C. & Mills, M. C. Polygenic predictions of occupational status GWAS elucidate genetic and environmental interplay for intergenerational status transmission, careers, and health. *bioRxiv* 2023.03. 31.534944 (2023).
27. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
28. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
29. Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* **3**, 513–525 (2019).
30. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
31. Semega, J. & Kollar, M. *US Census Bureau, Current Population Reports, P60-276, Income in the United States: 2021*. (Washington, DC: US Government Publishing Office, 2022).
32. OECD. *Education at a Glance 2021: OECD Indicators*. (Organisation for Economic Co-operation and Development, 2021).
33. Okbay, A. *et al.* Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat. Genet.* 1–13 (2022) doi:10.1038/s41588-022-01016-z.
34. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2020).
35. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
36. Vlaming, R. de *et al.* Meta-GWAS Accuracy and Power (MetaGAP) Calculator Shows that Hiding Heritability Is Partially Due to Imperfect Genetic Correlations across Studies. *PLOS Genet.* **13**, e1006495 (2017).
37. Dickson, M. The Causal Effect of Education on Wages Revisited*. *Oxf. Bull. Econ. Stat.* **75**, 477–498 (2013).
38. Young, A. I. *et al.* Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nat. Genet.* **54**, 897–905 (2022).
39. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
40. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
41. Lam, M. *et al.* Pleiotropic Meta-Analysis of Cognition, Education, and Schizophrenia Differentiates Roles of Early Neurodevelopmental and Adult Synaptic Pathways. *Am. J. Hum. Genet.* **105**, 334–350 (2019).
42. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1–11 (2017).
43. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
44. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).
45. Cesarini, D., Lindqvist, E., Östling, R. & Wallace, B. Wealth, health, and child development: Evidence from administrative data on Swedish lottery players. *Q. J. Econ.* **131**, 687–738 (2016).
46. Cutler, D. M. & Lleras-Muney, A. *Education and health: evaluating theories and evidence*. (National bureau of economic research Cambridge, Mass., USA, 2006).
47. Lleras-Muney, A. The relationship between education and adult mortality in the United States. *Rev. Econ. Stud.* **72**, 189–221 (2005).
48. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
49. Power, R. A. *et al.* Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nat. Neurosci.* **18**, 953–955 (2015).
50. Abdellaoui, A., Dolan, C. V., Verweij, K. J. & Nivard, M. G. Gene–environment correlations across geographic regions affect genome-wide association studies. *Nat. Genet.* **54**, 1345–1354 (2022).
51. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
52. Demange, P. A. *et al.* Investigating the genetic architecture of noncognitive skills using GWAS-by-subtraction. *Nat. Genet.* **53**, 35–44 (2021).
53. Lichtenstein, P. *et al.* The Swedish Twin Registry: a unique resource for clinical, epidemiological and genetic studies. *J. Intern. Med.* **252**, 184–205 (2002).
54. Sonnega, A. *et al.* Cohort profile: the health and retirement study (HRS). *Int. J. Epidemiol.* **43**, 576–585 (2014).
55. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52 (2010).
56. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med. Inform.* **7**, e14325 (2019).

Contribution

H.K., C.A.P.B., T.A.D., R.K.L., A.O., R.D.V., and P.D.K. designed the GWAS meta-analysis. P.D.K oversaw the study. C.A.P.B. was the lead analyst for the meta-analysis, responsible for GWAS, quality control, and meta-analysis. H.K. was the lead analyst for the follow-up analyses, including heterogeneity, MiXeR, GWAS-by-subtraction, genetic correlation, PGI prediction, and biological annotation analyses. N.Y. assisted with the follow-up analyses. R.A conducted PGI prediction and heritability analyses in the STR sample. C.X and W.D.H contributed to several analyses, including cross-country heterogeneity and biological annotation. H.K., D.J.B., and P.K. drafted the manuscript. A.A. wrote Box 1. J.P.B., T.A.D., R.K.L., Q.L., T.T.M., A.O., K.P.H., A.A., W.D.H., and R.D.V. provided important input and feedback on various aspects of the study design and the manuscript. All authors contributed to and critically reviewed the manuscript. The individual contributions of all authors according to the CRediT taxonomy are listed in Supplementary Table 29.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

This section provides the overall summary of the analysis methods. Further details are available in the Supplementary Information.

GWAS meta-analysis

We pre-registered our analysis plan for the main income GWAS meta-analysis on August 30 2018 (https://osf.io/rg8sh/). We used four measures of income (individual, occupational, household, and parental income) and conducted a multivariate GWAS to combine these different measures. In total, we recruited 32 cohorts. Some of these cohorts contributed to multiple income measures. **Supplementary Tables 1 and 2** summarise the income measures used for each cohort. **Supplementary Section 2.1** provides details on the phenotype definition. The study was limited to 1KG-EUR-like individuals who were not enrolled in an educational program at the time of survey or who were above the age of 30 if their current enrollment status was unknown.

Each cohort conducted the additive association analysis as follows. The log-transformed income measure was regressed on the count of effect-coded alleles of the given SNP, controlling for any sources of variation in income that do not reflect individual earning potential according to the data availability of each cohort. This included hours worked (with square and cubic terms), year of survey, indicators for employment status (retired, unemployed), self-employment, and pension benefits (see Supplementary Table 4). In addition, the covariates included at least the top 15 genetic principal components and cohort-specific technical covariates related to genotyping (genotyping batches and platforms). This analysis was performed for male and female samples separately.

We applied a stringent QC protocol based on the EasyQC software package[51] to the GWAS results from each cohort (see **Supplementary Information section 2.4** for more detail). In order to combine multiple GWAS results on different income measures collected from multiple cohorts, we performed the meta-analysis in several steps. First, for each income measure and each sex, we meta-analyzed the cohort-level GWAS results with METAL[27] using sample-size weighting. Second, for each income measure, we meta-analyzed the male and female results by using the meta-analysis version of MTAG.[28] To extract the common genetic

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

factor from the four GWAS results with different income measures, we again leveraged MTAG, allowing for different heritabilities among the input traits.

Cross-sex and cross-country heterogeneity

We investigated the potential environmental heterogeneity in the GWAS of income by estimating the cross-cohort genetic correlations by sex or by country with LDSC.[39] Sex-specific meta-analysis results for each income measure were available as intermediary outputs from the meta-analysis procedure. In addition, we conducted INC factor GWAS on the sex-specific results, which yielded an effective sample size of 360,197 for men and 353,429 for women.

To derive country-specific GWAS meta-analyses, we only used occupational and household income, for which we were able to obtain a sufficiently large sample size for multiple countries. We obtained the household income GWAS for the USA (N_{eff} = 30.855), the UK (N_{eff} = 387,579), and the Netherlands (N_{eff} = 40,533); and the occupational income GWAS for Estonia (N_{eff} = 75,682), Norway (N_{eff} = 42,204), the UK (N_{eff} = 279,883), and the Netherlands (N_{eff} = 24,425).

Comparative analysis with EA

We compared our INC factor GWAS results with the GWAS of EA by examining genetic correlation with LDSC and using the GWAS-by-subtraction approach.[52] Here, we used a version of EA summary statistics slightly different from publicly available ones. The latest EA GWAS study revised the coding of the years of schooling in the UKB[33] to better reflect the educational qualification of the participants. Based on the new coding, we conducted a GWAS of EA in the UKB. Then, by using MTAG with the meta-analysis option, we meta-analyzed the UKB result with EA3 summary statistics[21] that did not include the UKB.

We then statistically decomposed the estimated genetic association of the INC factor into the indirect effect due to EA and the direct effect unexplained by EA (NonEA-INC), leveraging the GWAS-by-subtraction approach in genomic SEM.[29,52] We implemented this method in the form of a mediation model.

PGI analysis

We conducted three sets of analyses based on the polygenic index (PGI): 1) prediction analysis, 2) direct genetic effect estimation, and 3) phenome-wide association study of common diseases.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.09.574865; this version posted January 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For the PGI prediction analysis, we used the Swedish Twin Registry (STR),[53] UKB siblings (UKB-sib), and the Health and Retirement Study (HRS).[54] We constructed PGIs using the meta-analysis results of income excluding a prediction cohort at a time, as well as a PGI based on the EA GWAS summary statistics constructed in the same way for comparison. PGIs were created only with HapMap3 SNPs[55] as these SNPs have good imputation quality and provide good coverage for 1KG-EUR-like individuals. We derived PGIs based on a Bayesian approach implemented in the software LDpred2.[34]

We measured the prediction accuracy based on incremental R^2, which is the difference between the R^2 from a regression of the phenotype on the PGI and the baseline covariates and the R^2 from a regression on the baseline covariates only. Because income typically contains substantial demographic variation, we pre-residualized the log of income for demographic covariates. Then, as baseline covariates, we only included the top 20 genetic PCs and genotype batch indicators. Because income data was available for multiple years for the STR and the HRS, we residualised the log of income for age, age^2, age^3, sex, and interactions between sex and the age terms within each year and obtained the mean of residuals for each individual. For the UKB-sib, which only had cross-sectional data, we residualised the log of income for age, age^2, age^3, sex, dummies for survey year, and interactions between sex and the rest. For the EA measure (years of education), we applied the same procedure with birth year dummies. We constructed confidence intervals for the incremental R^2 by bootstrapping the sample 1,000 times.

To estimate the direct genetic effect of the INC factor PGI, we used snipar[38] to impute missing parental genotypes from sibling and parent-offspring pairs. Parental PGIs were then created with the imputed SNPs. We estimated the direct genetic effect of the PGI by controlling for the parental PGI. This analysis was conducted only with the UKB-sib sample. See Supplementary Information 5.2 for further details.

To explore the clinical relevance of the INC factor PGI for common diseases, we carried out a phenome-wide association study, using the in-patient electronic health records for 115 diseases with sex-specific sample prevalence no lower than 1% in the UKB-sib sample. We derived case-control status according to the phecode scheme by mapping the UKB's ICD-9/10 records to phecodes v1.2.[56] We fitted a linear regression of case-control status on the INC factor PGI while controlling for the parental PGIs to capture the direct genetic effects of income PGI. As covariates, we also included the year of birth, its square term, and its interactions with sex, genotype batch dummies, and 20 genetic PCs. Standard errors were clustered by family.