Genetic associations with educational fields in >460,000 individuals

Rosa Cheesman[1], Ville Anapaz[2], Sjoerd van Alten[3], Abdel Abdellaoui[4], Ralph Porneso[1], Joakim C. Ebeltoft[1], Ziada Ayorech[1], Perline A. Demange[1], Espen Moen Eilertsen[1], Agnes Fauske[5], Alexandra Havdahl[6], Hannu Lahtinen[7], Torkild Hovde Lyngstad[5], Qi Qin[1], FinnGen, Andrea Ganna[2], Eivind Ystrom[1,6]

[1] PROMENTA Center, Department of Psychology, University of Oslo, Oslo, Norway

[2] Institute for Molecular Medicine Finland (FIMM), Helsinki, Finland

[3] School of Business and Economics, Economics, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

[4] Department of Psychiatry, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands

[5] Department of Sociology & Human Geography, University of Oslo, Oslo, Norway

[6] PsychGen Centre for Genetic Epidemiology and Mental Health, Child Health and Development, Norwegian Institute of Public Health, Oslo, Norway

[7] Helsinki Institute for Demography and Population Health, University of Helsinki, Finland

Abstract

The choice of a field of study is a significant decision influenced by a complex interplay of individual traits, interests, and contextual factors. Little is known about the genetic architecture of educational fields. Genetic methods make it possible to explore common influences on different field specialisations. First, we conducted genome-wide association study (GWAS) meta-analyses of 10 broad fields of education using population-wide administrative data from Finland and Norway (FinnGen and MoBa; total n=463,134). Measured genetic differences were associated with fields of study (17 independent genome-wide significant lead SNPs across 10 fields GWAS). SNP-based heritability estimates were 7% on average. Polygenic indices (PGIs) based on our GWAS results were significantly associated with their respective fields in Lifelines, an independent Dutch cohort, for 7 out of 10 fields (p<0.005; n=36,373). Second, we detected overlapping genetic influences on different field specialisations, summarised by a Technical versus Social trait (10 independent genome-wide significant lead SNP associations), and a second Creative versus Conformist trait (3 independent genome-wide significant lead SNPs). Technical versus Social tendencies are genetically correlated with personality traits Extraversion and Agreeableness, and Creative versus Conformist tendencies are genetically correlated with Openness to Experience and Occupational Creativity. Third, results were robust to controls for stratification, both in GWA analyses adjusting for birthplace and parents' fields and in within-family polygenic index analyses in the Lifelines (n=14,767). Weaving together genetics, complex traits, and contexts, we create a fuller picture of the underpinnings of educational qualifications, shifting the focus of social science genomics from the conventional hierarchy of attainments towards multidimensional tendencies and interests. We discuss socially mediated mechanisms by which genetic associations with fields of study arise.

Introduction

Fields of education and training, from fine arts to finance, are extremely diverse. They involve various amounts of cultural, economic, technical, and communicative skills [1]. Choosing a field of education sets a course for one's life, impacting on a person's health, skills, attitudes, and socioeconomic position [2,3]. Even within a level of education, there are large differences in socioeconomic outcomes across fields [4].

In studying why people gravitate towards and gain qualifications in certain fields, structural social forces have received much attention. For example, gender segregation in the labour market indicates that gender norms and beliefs influence young women and men to pursue caring versus technical fields, respectively [5–7]. There may be urban-rural disparities in choice norms and access to certain fields [8]. So-called 'vertical positions' of mothers and fathers (i.e., educational attainment) as well as 'horizontal positions' (i.e. parents' educational and occupational fields) are correlated with offspring field choices [9,10]. The intergenerational transmission of horizontal position (field of study) is particularly strong in Agriculture and Education [11], and seems to be independent from the transmission of vertical position [12]. However, this sociological literature has not yet considered that familial similarities in heritable traits could partly contribute to the intergenerational correlations.

Psychology research shows that educational choices are correlated with individuals' vocational interests and systematic tendencies for behaviour [13,14]. For example, more Extraverted people sort into fields which provide opportunities for social contact like healthcare, and higher levels of Openness to Experience are observed among students of the arts, humanities and psychology [15]. Efforts have been made to establish the overarching mechanisms of choice by measuring individuals' preferences (e.g., for fields with intrinsic versus extrinsic rewards, or for entrepreneurial versus bureaucratic characteristics [16]). However, preference measures are usually only available in small sample sizes, each explaining a small fraction of the choice process, and studies rarely include data on individuals' actual choices. An accurate, holistic, and hypothesis-free approach to the dimensionality of educational fields would be to factor analyse *actual* field choices. However, this is difficult because an individual usually only studies one field. To facilitate research on the causes and consequences of educational choices, it is critical to understand the structure of educational field choices.

Given that psychological differences linked to fields of education such as personality are known to be influenced by genetic factors [17], fields of study are also likely to be heritable. Genetic influences on field choices would represent *active gene-environment correlations (rGE)*, whereby people choose their experiences in line with their heritable traits [18]. Genetic influences on field choices would also be consistent with *evocative rGE* if individuals are encouraged into certain fields due to their heritable traits. Genetic influence has been demonstrated in twin studies of vocational interests and choices (e.g., creative professions) [19,20], and of school subject choices (heritability estimates were 50% for humanities, 60% for STEM) [21]. However, genetic associations with the wide range of possible educational field choices on the population level have not been studied, and new genomic tools have not been employed.

In the study of educational fields, genomic methods bring several novel advantages. First, they allow common dimensions underlying field choices to be investigated. The genetic covariance structure of traits can be estimated using genome wide association study (GWAS) summary statistics, even when GWAS samples are not overlapping (such as when individuals are only observed in one field) [22]. It is then possible to establish the number and nature of latent dimensions explaining genetic covariance patterns. Second, individuals' genetic data are valuable for causal inference when studied within the social context. Naive associations between genetic variants and educational outcomes include not only *direct genetic effects* (effects of an individuals' own DNA on their field choice, operating via active and evocative rGEs) but also confounding due to correlations with environmental influences (*passive rGE* [18]). Possible confounders include *indirect genetic effects* of relatives' genomes on the focal individual's field choice, geographical and social stratification due for example to regional educational policies, and population stratification (due to allele frequency differences across sub-populations) [23–25]. By placing individuals' genetic data within their family and geographical context, these mechanisms can be disentangled [25–27]. However, the relative contributions of direct versus non-direct genetic associations with field choices have yet to be established.

In turn, research on educational fields can enrich genomic studies of education and social stratification. Genomic research on attainment (defined simply as the number of years a person spends in education; EA), income [28] and occupational status [29,30] is valuable. However, the focus on these conventional measures of a person's position in a socioeconomic hierarchy, ignores the diversity of preferences, traits, and skills that educational pathways entail. Genetic associations with different fields of study are unlikely to be completely accounted for by the known genetic correlates of socioeconomic status, and thus may lead to novel insights on how individual and contextual factors combine to influence life chances.

Here, we study genetic influences on 10 broad fields of education as defined by the International Standard Classification of Education (ISCED) using large-scale population-wide data from Finland, Norway, and the Netherlands. Field specialisations are complex outcomes influenced by not only individual traits, interests, and skills, but numerous social barriers and supports. We perform our discovery genetic association analyses in the two Nordic cohorts. In these settings, education is free and social security is high, although there is pronounced gender segregation in the labour market [31]. Therefore, our findings are more easily interpreted as reflecting choices based on individual interests, preferences, and values rather than family resources or monetary incentives such as lifetime earnings. Nonetheless, the strong role of normative factors such as cultural beliefs about gender means that the study does not purely capture individual interests. After estimating genetic associations with fields and estimating SNP-based heritabilities, we pursue two lines of enquiry. First, we parse genetic associations from confounding factors using rich within-family and geographical data. Second, we reveal the key latent heritable traits underlying different field choices and characterise these through phenome-wide genetic correlation analyses.

Results

Genetic associations with 10 fields of education and training

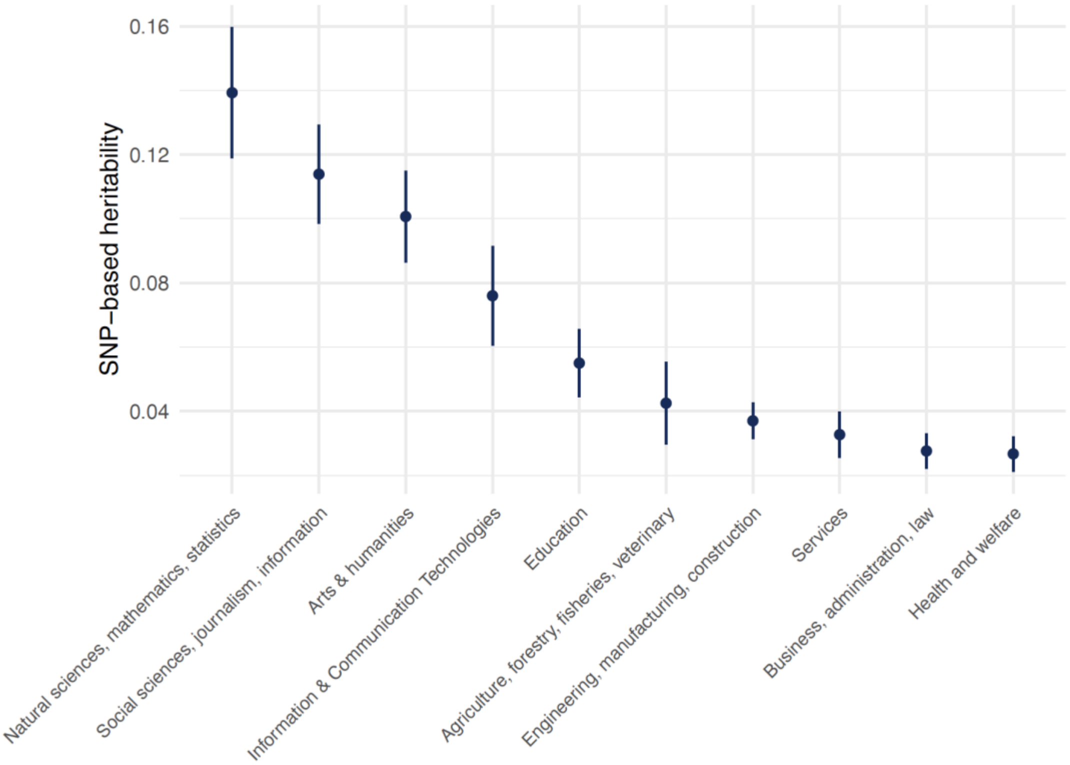
We harmonised data from Norwegian and Finnish education registers (using 2018 data on qualifications of adults older than 25, at various educational levels) to capture 10 broad fields of education as defined by the International Standard Classification of Education (ISCED) (https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-fields-of-education-and-training-2013-detailed-field-descriptions-2015-en.pdf). After linking the register data to genotype data in MoBa [32,33] and FinnGen [34] and performing genome-wide association analyses (GWAS), we performed sample-size-weighted meta-analyses with METAL [35]. The total sample size was 463,134 for all GWAS, with the number of 'cases' ranging from 10,252 for Natural sciences, mathematics and statistics to 102,874 for Engineering, manufacturing and construction. The sum of effective sample sizes ranged from 40,072 for Natural sciences, mathematics and statistics to 317,209 for Engineering, manufacturing and construction.

Across the 10 GWAS, 17 independent genome-wide-significant lead single nucleotide polymorphisms (SNPs) were identified for Education (1 locus); Arts and humanities (3 loci); Social sciences, journalism and information (3 loci); Business, administration and law (2 loci); Natural sciences, mathematics and statistics (3 loci), Engineering, manufacturing and construction (1 locus), and Health and welfare (4 loci) Significant loci were field-specific (see Supplementary Table 1 for lead SNPs). Manhattan plots and QQ plots are shown in Supplementary Figures 1-14.

Liability-scale SNP heritability estimates, calculated using LD Score regression [22], were 7% on average (median 5%), and ranged from 3% (Health and welfare) to 14% (Natural sciences, mathematics and statistics) (Figure 1 and Supplementary Table 2). SNP heritability estimates were consistent across the Finnish and Norwegian cohorts (Supplementary Table 3). Genetic correlations within educational fields across cohorts were strong (rG >0.75) and statistically indistinguishable from 1 apart from for Business, administration and law (cross-cohort rG=0.45, 95% CIs: 0.25-0.65), and Services (cross-cohort rG=-0.06, with particularly wide 95% CIs: -0.76-0.88). See Supplementary Table 4 for genetic correlation results within and across the Finnish and Norwegian cohorts.

Sample sizes of educational fields within the Norwegian and Finnish populations, within each genotyped cohort, and effective sample sizes for genome-wide association analyses are shown in Supplementary Table 5. The narrow fields of study included within the broad ISCED field categories are shown in Supplementary Table 6.

Figure 1: SNP-based heritability estimates for educational fields.



Genetic associations with educational fields reflect more than predisposition to educational attainment

To explore the extent that genetic influences on educational fields reflect individuals' choice of a field *per se* rather than general liability for a longer education, we repeated the GWAS controlling for years of education and then calculated SNP heritability. Overall, results suggest that field-specific genetic influences exist after accounting for EA: that is, the effect of genetic factors on field choice does not simply operate by changing educational attainment. After controlling for EA, the mean SNP heritability went from 7% to 5%, and the median SNP heritability remained at 4% (Supplementary Table 7 and Supplementary Figure 15). Although several heritability estimates were attenuated by more than 70%, they remained significantly greater than zero: Social sciences, journalism and information (11% to 3%); and Natural sciences, mathematics and statistics (14% to 4%). Additionally, at the SNP level, 64% of the independent loci identified through our main GWAS of educational fields did not reach genome-wide significance in the latest GWAS of educational attainment [30]; see Supplementary Figure 16. Unless otherwise stated, we focus on GWAS results unadjusted for attainment in this article. We summarise the many possible causal interrelationships, and pros and cons of controlling for educational attainment, in Supplementary Figure 17. See Supplementary Figures 18 and 19 for the distributions of educational attainment per field in MoBa.

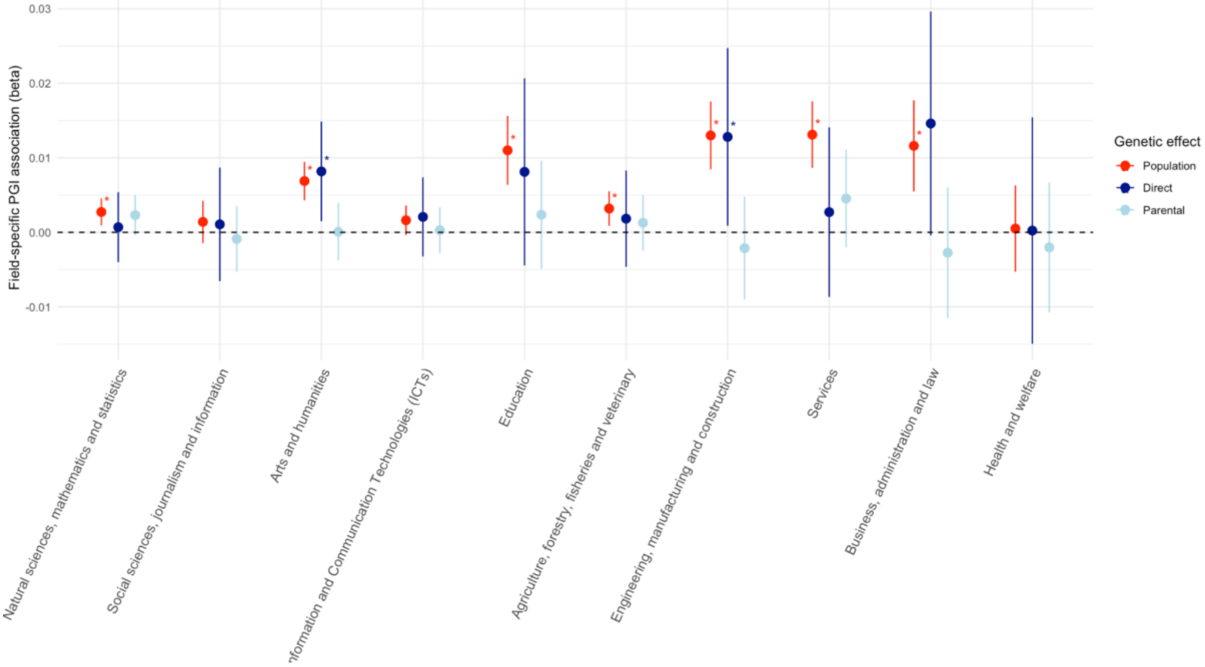
Genetic associations with educational fields capture 'direct' genetic effects

Population-level SNP associations with field choices might reflect not only direct genetic effects but several other mechanisms. These include indirect genetic effects, geographical influences, and population stratification. Although these non-direct genetic effects include interesting causal environmental effects in themselves, they are confounders with respect to estimating direct genetic effects.

We used two different approaches to understand the relative contributions of factors other than direct genetic effects to our main findings.

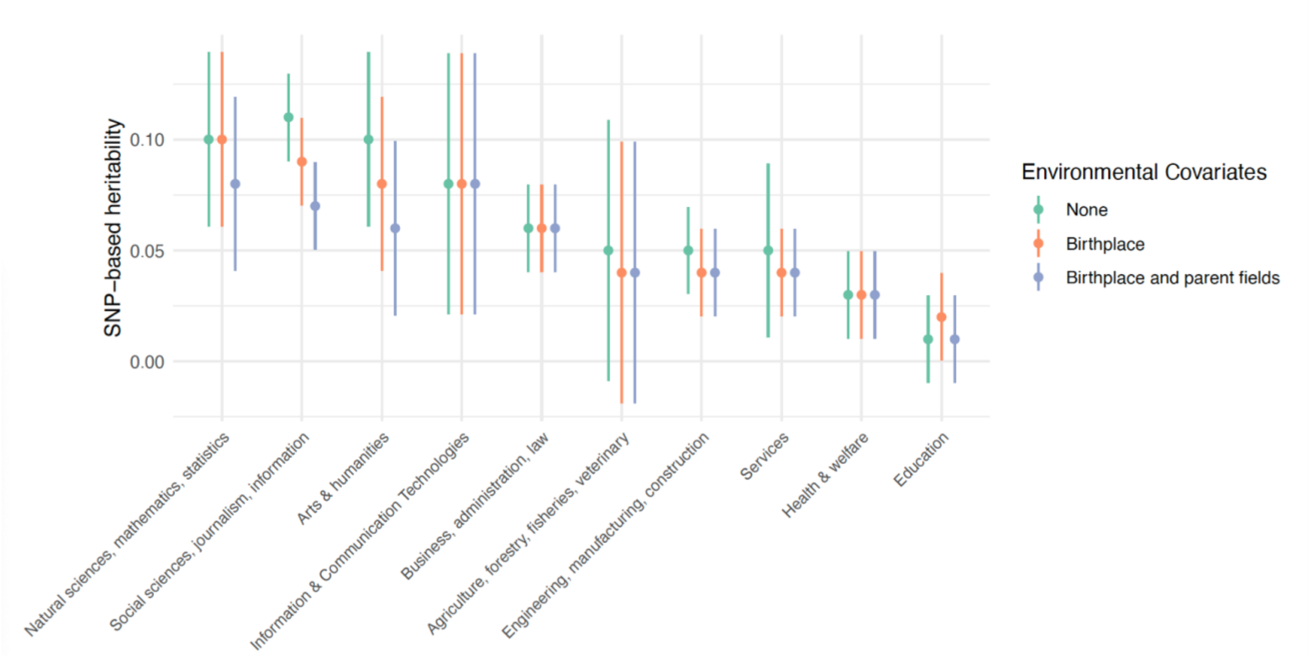
First, with the main meta-analytic summary statistics, we created polygenic indices (PGI) for educational fields in an independent sample of Lifelines Biobank respondents linked to Dutch administrative data and observed population-level and within-family PGI associations with educational field choices. We first validated the GWAS results, finding that whilst effect sizes were small, 7 out of 10 PGIs significantly predict their respective educational field on the population level (N=36,373; red data points in Figure 2; Supplementary Table 8). Next, in a subsample of 14,160 individuals, we included the imputed sum of the PGIs of their parents as a control variable. This exploits random within-family genetic variation such that direct genetic effects can be estimated without confounding. Direct genetic effect estimates (dark blue in Figure 2; average beta = 0.0052) were not much smaller than population estimates (red in Figure 2; average beta = 0.0065), suggesting minimal contributions from parental indirect genetic effects and population stratification to population associations (see Supplementary Table 9). However, power for within-family analyses in the subsample was low, such that only PGI associations for Arts, humanities and languages and Engineering, manufacturing and construction remained significant. Estimates of parental PGI associations (pale blue in Figure 2) were negligible and non-significant.

Figure 2: PGI associations with educational fields in an independent Dutch cohort. *Note: N=36,373 for estimating population effects and 14,160 for estimating direct genetic effects and parental indirect genetic effects; bars represent 99.5% confidence intervals; asterisks represent significance at the p<0.005 level to correct for testing 10 hypotheses.*



Second, we used population and education register data within the Norwegian dataset (MoBa) to perform GWAS of educational fields controlling for birthplace municipalities. We then added controls for the educational fields of participants' parents. This removes confounding factors that are correlated with birthplace and parental fields of education, thus approximating within-region and within-region-and-family GWAS. We estimated SNP heritability using the resulting summary statistics. The rationale was that a large drop in SNP heritability indicates geographical and parental environmental confounding. Figure 3 demonstrates a lack of evidence for this: for all fields, heritability estimates are similar. Modelling heritability ratios in Genomic SEM following [25] also showed that none of the adjusted estimates were significantly different from the original estimates, apart from for Social sciences, journalism and information. However, the drop was small (11% to 7% when adjusting for both birthplace and parents' fields of study) and the FDR-corrected ratio p-value was relatively large (p=0.03). See Supplementary Table 10 for heritability results and Supplementary Table 11 for SNP-heritability ratios and corresponding FDR-corrected p-values.

***Figure 3: SNP heritability estimates for educational fields controlling for birthplace and parental educational fields.** Note: bars = 95% confidence intervals.*



Even after environmental confounding is considered, direct genetic effects cannot be interpreted as purely genetic: they are mediated through the environment, via active and evocative gene-environment correlation processes. See Supplementary Figure 20 for an explanation of how gene-environment correlation mechanisms apply here.

Two key heritable latent traits linked to educational field choices

We explored whether genetic associations with the 10 field specialisations could be reduced to a smaller number of general latent dimensions. We first estimated pairwise genetic correlations among all fields using the GWAS summary statistics with LD score regression. We used EA-adjusted summary statistics to enable investigation of the key components of field specialisation preferences *beyond* genetic influences on general attainment.

Figure 4a shows positive genetic correlations between technical/STEM subjects (e.g., rG=0.59 (SE= 0.12) between Information and communications technology (ICT) and Natural sciences, mathematics and statistics), between Arts, humanities and languages and Social sciences, journalism and information (rG = 0.59, SE=0.10), and between Health and welfare and Education (rG = 0.39, SE=0.08). Notably, Health and welfare, Education, and Services were negatively genetically correlated with most other fields apart from each other. See Supplementary Table 12 for genetic correlations among fields.

To establish the number of components explaining patterns of genetic correlations between fields, we performed a principal components analysis (PCA). Two PCs were satisfactory, given that they collectively explained 69% of the variance, and that the individual fields GWAS showed strong correlations with the top two PCs but not the third (see Supplementary Table 13 for PCA results).

Figure 4b plots the contributions of the genetic influences on each field to the two PCs. The correlation between an educational field and a PC is used as the coordinates of the variable on the PC. Component 1 (horizontal axis in Figure 3b,), which we term 'Technical versus Social', reflects genetic variation correlated with pursuing Engineering, manufacturing and construction, Natural sciences, mathematics and statistics, and ICT, versus Education and Services. Component 2 (vertical axis in Figure 3b), which we term 'Creative versus Conformist', reflects genetic variation correlated with pursuing Arts and humanities, and Social sciences, journalism and information, versus Health and welfare, Services and Engineering, manufacturing and construction. Genetic variation associated with qualifications in Services, Health, and Education are strongly positively intercorrelated with each other, but negatively genetically correlated with the Technical and Creative fields (e.g., genetic influences on pursuing Health and welfare are negatively correlated with Arts, humanities and languages (rG=-0.31; SE=0.07) and ICT (rG=-0.51; SE=0.07). Without controlling for EA, the structure of genetic correlations between fields were relatively similar (see Supplementary Figure 21).

To further characterise the two heritable latent traits correlated with educational field choices, we compared genomic factor analytic models and then performed multivariate GWAS of the best-fitting model in Genomic SEM [36]. The best model had uncorrelated latent traits but extensive cross-loadings (see Supplementary Tables 14 and 15 and Supplementary Figure 22). The resulting GWAS summary statistics for the two independent 'Technical versus Social' and 'Creative versus Conformist' educational field factors had effective sample sizes 10,413 and 7353, respectively (calculated following Mallard et al. [37]). There were 10 independent genome-wide significant lead SNP associations with factor 1 and 3 with factor 2 (See Supplementary table 16 for the lead SNPs and Supplementary Figures 23-24 for manhattan plots). Only one of these 13 SNPs was also a lead significant SNP in the field-specific GWAS (rs6937215 for factor 2 and Business, administration and law; previously linked to alcohol use [38]). The SNP association with the lowest p-value was for rs7106434 and factor 1; this SNP has been mapped to the NCAM1 gene, previously linked to learning and memory [39].

Figure 4a) genetic correlations between educational fields adjusted for educational attainment.

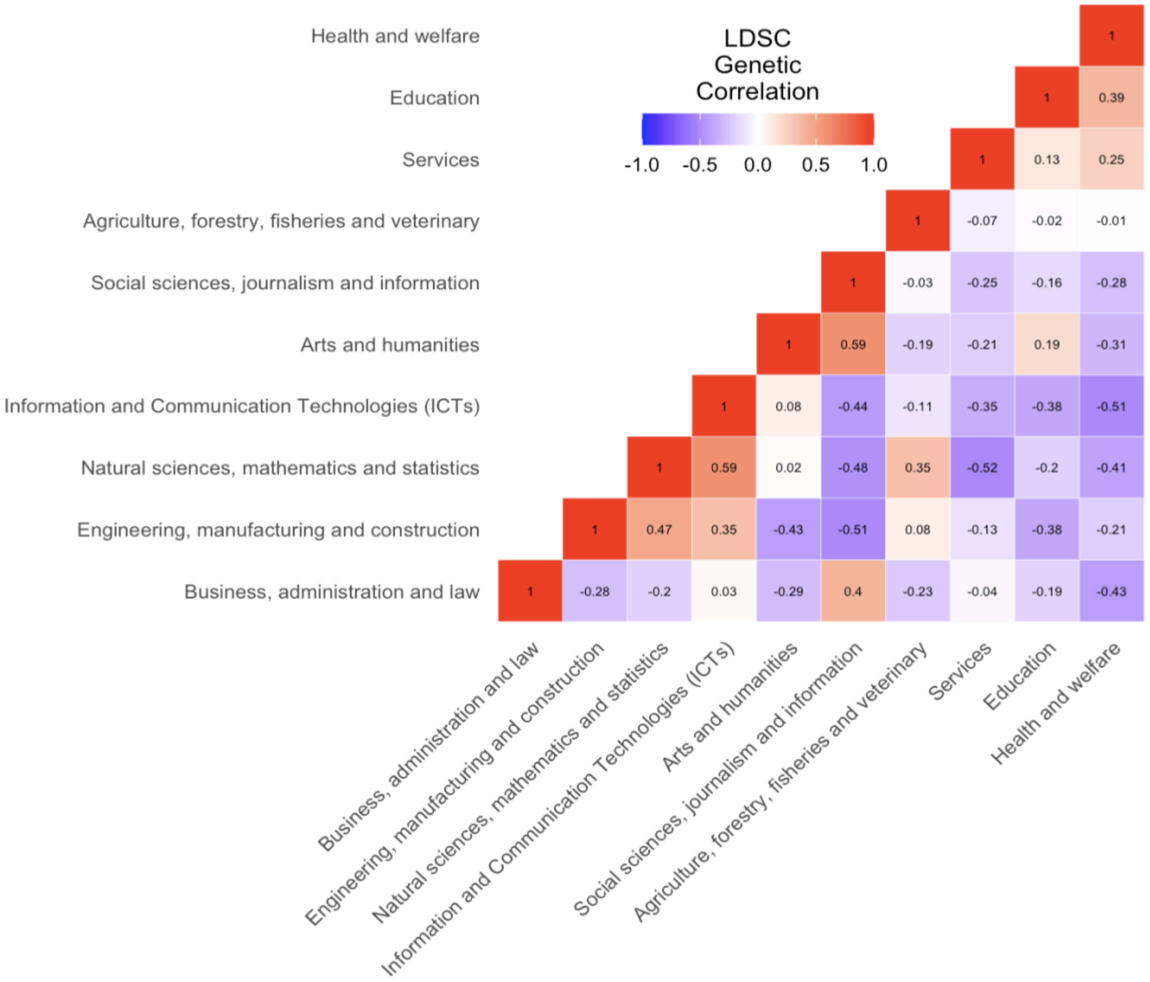
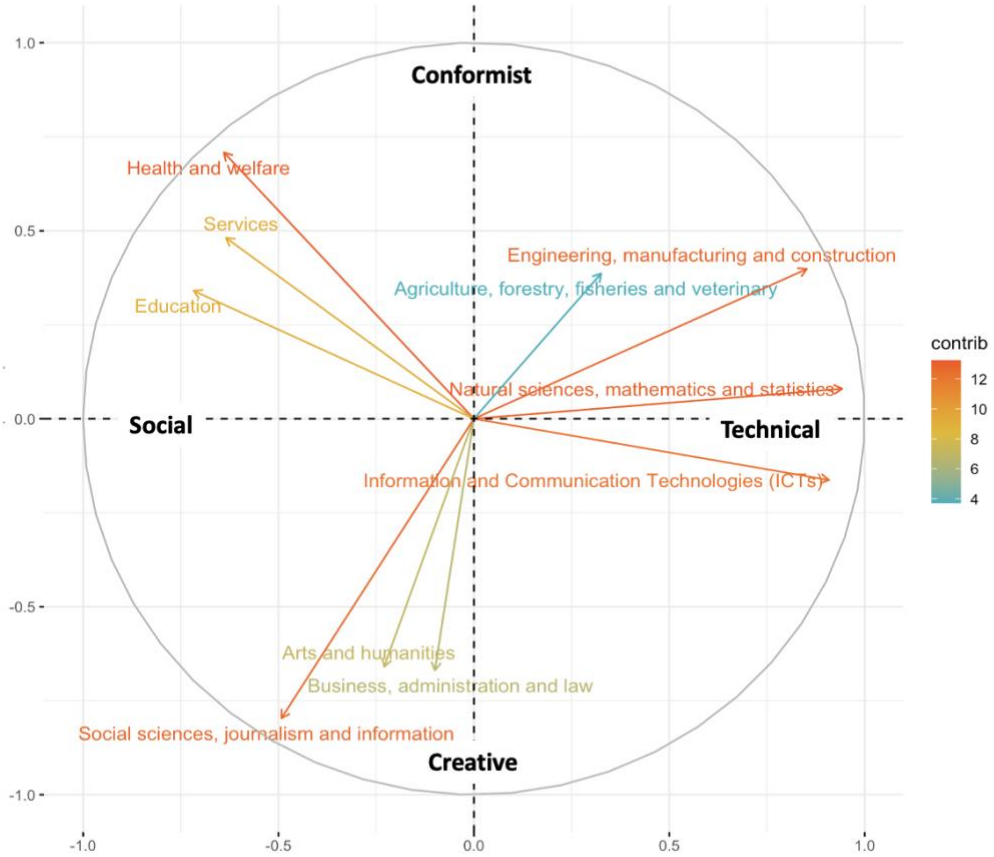


Figure 4b) the two main axes of genetic variation correlated with choice of educational fields. *Note that positively correlated variables are grouped together. Negatively correlated variables are positioned on opposite sides of the plot origin (opposed quadrants). The distance between variables and the origin measures the quality of the variables on the factor map. Variables that are away from the origin are well represented on the factor map.*



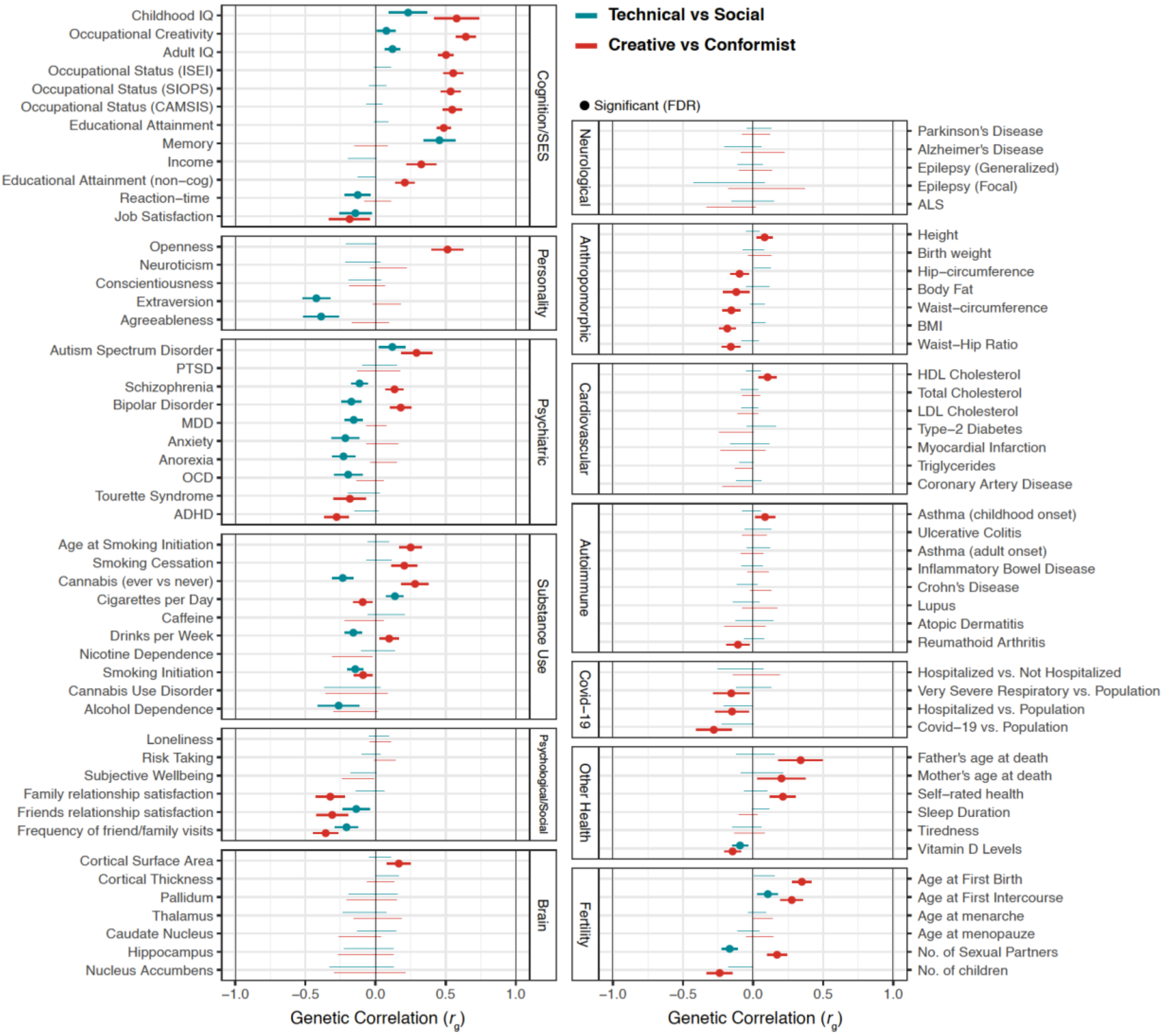
Genetic correlates of latent field choice traits

To characterise the two latent field choice-influencing traits (Technical versus Social and Creative versus Conformist), we estimated genetic correlations with 93 phenome-wide phenotypes studied in well-powered GWAS using LD score regression. Figure 5 presents genetic correlations with phenotypes spanning various domains including personality, mental health, substance use, health, and fertility.

The tendency towards Technical rather than Social fields was negatively genetically correlated with personality traits Extraversion and Agreeableness (rGs= -0.42 and -0.39, respectively). Significant negative genetic correlations were also observed with cannabis, alcohol use (rGs~ -0.25), as well as six psychiatric diagnoses (average rG= -0.18), and number of sexual partners (rG= -0.17). Notably, genetic factors associated with Technical qualifications were positively linked to memory and childhood IQ (rGs= 0.46 and 0.23, respectively).

The tendency towards Creative versus Conformist fields was positively genetically correlated with Openness to Experience (rG=0.48). Genetic correlations were also positive for schizophrenia, bipolar disorder, and autism spectrum disorder (rGs= 0.14, 0.18 and 0.29, respectively) but negative for ADHD (rG= -0.28). Regarding substance use, genetic factors associated with Creative qualifications were positively related to trying cannabis (rG= 0.28) but negatively related to cigarette smoking (rG= -0.09). Higher genetic predisposition to Creative fields was also linked to positive health outcomes such as paternal lifespan, reduced Covid-19 infection, and lower BMI (rGs= 0.34, -0.27, and -0.18, respectively). The Creative versus Conformist trait was positively genetically correlated with the number of sexual partners (rG= 0.17), but also with age at first birth (rG= 0.35). Importantly, Creative tendencies were most strongly associated with a continuous measure of Occupational Creativity (rG= 0.64), defined using detailed task descriptions of occupations [40]. Although educational attainment was adjusted for in our GWA analyses, the Creative versus Conformist trait remains positively genetically correlated with socioeconomic outcomes including educational attainment and childhood IQ (rGs= 0.49 and 0.58, respectively). See Supplementary Table 17 for all genetic correlation results.

Figure 5 genetic correlations between two latent educational field choice traits and 93 human phenotypes. *Note: Educational field GWAS were adjusted for EA; FDR= False Discovery Rate.*



Limited evidence for sex differences in the structure of genetic influences on educational fields

The Technical-Social latent trait is linked to sex and gender norms. In MoBa's genotyped sample, 84% of those with Engineering, manufacturing and construction qualifications are males, and 88% with Health and welfare qualifications are females (see Supplementary Figure 25 for sample sizes by sex in MoBa). To investigate the role of sex in the structure of genetic influences on educational fields, we performed sensitivity analyses. First, we repeated the PCA excluding fields where >=70% of cases were of one sex (Engineering, manufacturing and construction, Health and welfare, and Education). Second, we conducted sex-stratified GWAS in MoBa (sex defined from birth register data) and repeated the PCA with sex-specific indicators. We found that the 2-component genetic structure of field choices was consistent for the subset of fields with sex ratios closer to 50-50 (See Supplementary Figure 26), and when estimated within sex (Supplementary Figures 27-28). SNP heritability estimates were similar in males and females, but genetic correlations for educational fields between the sexes varied widely from 0.17 for Engineering, manufacturing and construction to 0.72 for Natural sciences, mathematics and statistics (Supplementary Table 18).

Overall, several factors prevent us from drawing strong conclusions regarding the role of sex and gender in our results. First, the sample sizes were low for most sex-specific fields. Second, there are sex differences in qualifications *within* broad fields (e.g., within Engineering, manufacturing and construction, males are more likely than females to study construction), such that low genetic correlations capture different influences on different entities rather than on the same entity. As is the case for height, phenotypic differences by sex do not necessarily imply qualitative sex differences in genetic influences.

Discussion

Differentiation into fields of education is important for individuals and for society: it not only influences health, wellbeing and success but determines the knowledge and skills available in the labour supply. Using large-scale population-wide administrative and genetic data from Finland, Norway, and the Netherlands, we showed that genetic factors are correlated with educational field specialisations. Genetic influences relate to fields of study *per se,* independent of years of schooling, and seem to capture direct effects of individuals' own DNA rather than confounding factors. We discovered two key latent traits underlying the genetic associations with field choices, characterised by Technical versus Social tendencies (genetically correlated with Extraversion and Agreeableness), and Creative versus Conformist tendencies (genetically correlated with Openness to Experience and Occupational Creativity). In revealing patterns of genetic variation associated with field choices, we move the focus of social science genomic research beyond the conventional hierarchies of socioeconomic status, towards a multidimensional space of genetic influences capturing tendencies and interests.

We found that SNP heritability estimates for educational field choices were 7% on average. These are likely to be underestimated because our methodology only captures additive effects of common variants tagged by genotyping arrays, rather than the full heritability (the missing heritability problem [42]). Across two different approaches there was limited evidence for confounding of genetic effects, though lack of statistical power for these analyses precludes strong conclusions. Future studies should assess the magnitude of indirect genetic effects on fields of study by combining more large-scale datasets and applying family-based GWAS methods [30,43]. Given our own results and prior sociological evidence that the familial reproduction of field choices is an independent channel of transmission to that of attainment [12], educational field choices may be smaller than and show little overlap with parental genetic effects on educational attainment. Nonetheless, our low SNP heritability estimates leave a large role for environmental factors and random chance in educational field choices.

Genetic factors do not directly influence choice of field: they relate to individual tendencies that are correlated with choices. Our novel genetic approach enabled us to uncover what the key tendencies are without having to theorise about or measure individuals' preferences. Individual differences in educational field specialisations could be reduced to two heritable latent traits. The Technical-Social factor and Creative-Conformist factor correspond well with major theories in the social sciences. First, sociological work has emphasised the four main resources that individuals aim to acquire through education in specific fields: Communicative, Technical, Cultural, and Economic [2,44]. According to this theory, individuals invest in fields that provide resources they can capitalise on later in terms of status and lifestyle. Teachers and social workers are expected to choose programmes that will prepare them well for the social communication aspect of work, whereas others aim to obtain technical resources (mathematics skills, knowledge on physical laws). Second, Holland's RIASEC model of vocational interests includes similar dimensions of Social versus Realistic, and Conventional versus Artistic [45]. This model emphasises the behavioural tendencies that lead people to acquire certain competencies. Preferences for ambiguous and free (Artistic) versus systematised activities (Conventional) are held to be important when choosing art or psychology rather than engineering or nursing. Notably, the results also correspond with research on links between fields of study, RIASEC interests and the Big 5 [15,46].

Patterns of genetic correlations with traits and behaviours support the characterisation of the key tendencies influencing field choices as ‘Technical versus Social’ and ‘Creative versus Conformist’. First, the Technical versus Social trait clearly captures heritable preferences for activities involving things versus people. Indeed, significant negative genetic correlations were observed with the social phenotypes extraversion, agreeableness, relationship satisfaction, and number of sexual partners. The lower genetic risk for psychiatric disorders associated with Technical tendencies could reflect difficulties with completing technical qualifications whilst experiencing mental health problems [47,48]. Second, the Creative versus Conformist trait reflects preferences for creative and exploratory rather than conventional practical activities. Creativity is defined as the generation of new and useful ideas, and involves cognitive flexibility and persistence [49]. Genetic factors associated with Creative tendencies are most strongly linked to open personality and occupational creativity. The positive genetic relationships we observed with schizophrenia and bipolar disorder align with evidence that relatives of affected individuals are more likely to have creative jobs [50]. The negative genetic correlation with ADHD could reflect difficulties with the persistent attention and goal-directed behaviour involved in creative endeavours, or in education generally [49]. Interestingly, the ‘Creative versus Conformist’ construct shows genetic overlap with socioeconomic status indicators even though educational attainment was controlled for in the GWAS. These results could imply that the Creative trait captures the familial aggregation of cultural and economic capital. Those pursuing Creative fields, regardless of their educational level, may be more likely to possess various resources that are valued culturally (e.g., social connections, individual and parental genetic factors correlated with abstract thinking, verbal skills and curiosity), as well as an economic safety net. Future within-family genetic analyses will help to disentangle individual versus social background effects on creative tendencies.

What are the mechanisms behind genetic associations with outcomes as biologically distal as educational field qualifications? The Nordic context of our study is characterised by free education and universal stipends and student loans with low interest rates to cover living expenses. Our results are therefore likely to capture individuals’ choices based on their genetically influenced tendencies (likes, preferences, and skills linked to personality traits) more than tuition fees and family resources. Yet the mechanisms cannot be purely individual. Even in Finland and Norway, educational field choices are constrained and amplified by social factors. An individual would not study art if they had never heard of fine art or been exposed to encouragement suggesting that it is a suitable choice for them. Genetic contributions to field qualifications operate through active and evocative rGE mechanisms, whereby field-related interests and skills are selected and elicited by individuals based on their heritable traits. Such mechanisms are likely to begin early in life, involving parents, teachers, and other role models. Gender norms are a key social mediator, with stereotypes that influence choice of field of study beginning early. For instance, both girls and boys are steered away from female dominated educational tracks [7], and the gender gap in STEM degrees is partly because boys benefit from teacher biases [51]. Results could also capture social closure in admissions systems (e.g., if technical analytical skills are necessary to gain entry to engineering training and work as an engineer, then engineers will on average have higher genetic values for technical skills), and dropout due to poor person-environment fit or discrimination. The results reflect the interplay between individual tendencies, social norms and barriers driving selection into fields.

Determinist and essentialist interpretations of genetic associations with complex social outcomes like educational field qualifications are wrong [52]. First, genetic factors do not determine field choices but probabilistically influence individuals' tendencies, which, via interaction and correlation with the social and structural context, become correlated with educational decisions. Genetic factors correlated with field choices may look different if people were encouraged to explore a wider range of subjects, if the skills involved in certain fields were different, or if the gender norms or economic returns to fields changed. In countries with higher social inequality than in Nordic countries, where the socioeconomic consequences of some field choices are riskier, the heritability of field choices might be lower, and the links with individual interests and preferences might be less prominent. Second, the Technical and Creative factors found here are not meant to essentialise or categorise people. Individuals with more engineering-associated DNA variants 'should' not necessarily choose engineering and are not qualitatively different to people with more art-associated DNA variants. Indeed, whilst creativity and conformity have been thought to be diametrically opposed, recent work has highlighted that more conforming individuals can enhance the creative process [53].

Our study is limited in several ways. First, for reasons of statistical power, we conducted analyses on the level of broad fields rather than narrow or detailed fields. Some broad categories lump together rather different fields, and this heterogeneity could wash out genetic signals. For example, engineering and construction involve different tasks and are differently stratified by gender, so the heritable traits associated with choosing one or the other may be different. By increasing sample sizes in the future, it will become possible to study more homogeneous groups within narrow or even detailed field categories with GWAS methods. Second, although it is an advantage of the study that we used two strict approaches to control for passive gene-environment correlation and population stratification from genetic associations, these analyses were underpowered. Moreover, the SNP heritability estimates adjusted for birthplace and parental fields could still be confounded if parental fields do not have a perfect genetic correlation with offspring fields (e.g., if there are major cohort differences in the characteristics of educational fields and the influences on choosing them) or by social influences of other relatives such as aunts, uncles and cousins [54]. Third, our GWAS were limited to Norwegian and Finnish individuals with European-associated ancestries. It remains unclear how much the results generalise to people of diverse backgrounds. In the future, genetically sensitive studies of educational field choices should include underrepresented groups and countries, taking systematic social differences into account.

Several directions for future research are apparent. GWAS summary statistics for the two factors and for the individual educational fields could be used to further understand the mechanisms involved in, and consequences of, field choices. For example, PGIs could be used to trace how genetic predispositions manifest in young people's early interests, skills and choices, and how social contexts magnify or suppress certain predispositions (gene-environment interaction questions). Moreover, our GWAS summary statistics could also be used in mendelian randomisation analyses to provide a fresh look at the labour market returns to field qualifications [2]. Notably, interpretations of the PGI associations should include social norms as well as individual preferences. These research directions will be most informative in large-scale family based GWAS methods [43].

Overall, these findings offer a new direction for genetically informed education research incorporating individual differences in passions and preferences.

Methods

Contexts

Our main analyses are based on data from Finland and Norway. These are both social democratic welfare states [55] that fit the 'Scandinavian model' of *education for all*[56]. Compared to other wealthy nations, income inequality is low and access to education is less restricted by economic barriers. For example, Norway has free tuition, affordable loans and generous public subsidies for students. Paradoxically, Nordic labour markets are among the most gender segregated. Despite the reversal of the gender gap in educational attainment and the integration of women into previously male-dominated high-status fields of study like medicine and law, gender-typical segregation into fields of study persists [31,57].

We also analyse data from the Netherlands, which has been defined as a conservative welfare state [58]. Relative to the social demographic welfare states, social stratification in education is greater, partly due to early educational tracking and tuition fees.

Samples

FinnGen

"FinnGen (https://www.finngen.fi/en) launched in 2017, is a public-private research project, combining genome and digital healthcare data on about 500,000 Finns. The nation-wide research project aims to provide novel medically and therapeutically relevant insight into human diseases. FinnGen is a pre-competitive partnership of Finnish biobanks and their background organisations (universities and university hospitals) and international pharmaceutical industry partners and Finnish biobank cooperative (FINBB). All FinnGen partners are listed here: https://www.finngen.fi/en/partners." The project utilises data from the nationwide longitudinal health register collected since 1969 from every resident in Finland. Analyses were conducted on individuals aged >25 with complete data for genome-wide genotyping, and complete educational records. The list of FinnGen flagship authors is provided in Supplementary Table 19.

The Norwegian Mother, Father, and Child Cohort Study (MoBa)

We studied adults who participated in the Norwegian Mother, Father, and Child Cohort Study. The Norwegian Mother, Father, and Child Cohort Study (MoBa) is a prospective population-based pregnancy cohort study conducted by the Norwegian Institute of Public Health [33]. Pregnant women were recruited from across Norway from 1999 to 2009. The women consented to initial participation in 41% of the pregnancies. Of fathers invited to participate, 82.9% consented. The total cohort includes approximately 114,500 children, 95,200 mothers and 75,200 fathers. Analyses were conducted on MoBa parents aged >25 with complete data for genome-wide genotyping, and complete administrative records linked to MoBa through the Norwegian national ID number system.

Dutch Lifelines

Lifelines is a multi-disciplinary prospective population-based cohort study examining in a unique three-generation design the health and health-related behaviours of 167,729 persons living in the North of the Netherlands [59]. It employs a broad range of investigative procedures in assessing the biomedical, socio-demographic, behavioural, physical and psychological factors which contribute to the health and disease of the general population, with a special focus on multi-morbidity and complex genetics. Participants were sampled from the northern population of the Netherlands, which is about 10% of the region's population. Between 2006 and 2013, randomly selected general practitioners invited all their listed patients aged 25-49 years to participate in the study. We restricted our sample to genotyped Lifelines respondents that were 25 years or older (N=63,927). PGIs and the first ten principal components of the genetic data were linked to an administrative data file containing educational fields ("HOOGSTEOPLTAB 2022, V1''), housed by Statistics Netherlands.

Genetic data quality control

FinnGen

FinnGen release 11 contains genotype data for 473,681 individuals after quality control (QC). A total of 387,601 individuals were genotyped with a FinnGen ThermoFisher Axiom custom array v2. Data on 86,080 additional individuals were derived from legacy collections [34] Further information is available at: https://finngen.gitbook.io/finngen-handbook/finngen-data-specifics/red-library-data-individual-level-data/genotype-data/affymetrix-chip-and-its-design.

MoBa

Blood samples were obtained from both parents during pregnancy and from mothers and children (umbilical cord) at birth. Quality-controlled genotyping array data for the full 207,569 unique MoBa participants was recently generated [32]. Phasing and imputation were performed with IMPUTE4.1.2_r300.3, using the publicly available Haplotype Reference Consortium release 1.1 panel as a reference. To identify a sub-population of European-associated ancestry, principal component analysis (PCA) was performed with 1000 Genomes phase 1 after LD-pruning. During post-imputation quality control, the following thresholds were used for SNP removal: imputation quality (INFO) score $\leq$ 0.8;MAF<1%; call rate<95%.

Dutch Lifelines

Blood samples were collected from Lifelines participants at the first assessment visit. Genotypes were released as part of two separate cohorts. The CytoSNP cohort was measured on the Illumina CytoSNP-12v2 array, measuring ~300,000 SNPs. The UGLI cohort was measured on the Infinium Global Screening Array®(GSA) MultiEthnic Disease VErsion, measuring ~700,000 SNPs. Quality-controlled data for both cohorts was released. The quality control reports for CytoSNP and UGLI are available from http://wiki.lifelines.nl/doku.php?id=gwas and http://wiki.lifelines.nl/lib/exe/fetch.php?media=qc_report_ugli_r1.pdf, respectively. Prior to PGI construction, and in each cohort, we dropped multiallelic SNPs, SNPs with MAF < 1%, SNPs with an info score < 0.8, or SNPs that were not in HWE ($P < 10^{-6}$). We also dropped individuals with homozygosity rates of +/- 3 standard deviations (removing 655 respondents). We further dropped 1,289 respondents from the CytoSNP cohort that were also available in the UGLI cohort. After all these QC steps were completed, we merged the CytoSNP and UGLI cohorts into a single data file, using only SNPs that both cohorts had in common after QC (~6.4 million SNPs in total).

Measures

Broad educational fields - the International Standard Classification of Education (ISCED)

In all 3 cohorts we extracted register data on broad educational field codes representing the field of education of each person's highest qualification completed by the year 2018.

To harmonise the data and facilitate future replication studies in other cohorts, we convert broad field codes from national-level coding systems to broad field codes as defined by the International Standard Classification of Education 2013 (https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-fields-of-education-and-training-2013-detailed-field-descriptions-2015-en.pdf).

In FinnGen, we used linked administrative data from Statistics Finland to define individuals' educational qualifications. The Finnish educational field records are described here: https://www2.stat.fi/fi/luokitukset/koulutusala/. In MoBa, we used linked administrative data from the Norwegian Standard Classification of Education (NUS2000). The administrative data are of high quality, and do not suffer from attrition. More information on the NUS coding and the conversion to ISCED is available at: http://www.ssb.no/en/utdanning/norwegian-standard-classification-of-education. In the present study, missing data only occurs for individuals whose fields do not map exactly on to the ISCED system, for instance because they are interdisciplinary (e.g., 16000 genotyped individuals in MoBa with qualifications termed 'Interdisciplinary programmes and qualifications involving Health and welfare'). Note that Dutch administrative records on education are incomplete, such that we only had educational field measures available for 56% of the original Lifelines sample (N=36,373).

We create a dummy binary variable for each of the 10 ISCED broad field specialisation codes, scoring individuals as 1 if they chose the field, and 0 otherwise. The 0 category included people studying Generic programmes (not a specialisation *per se).* This includes a wide range of qualifications e.g., unspecialised high school diploma, professional development skills training.

We also created harmonised educational attainment variables in all datasets. We took educational level information from the exact variable containing the field code, and converted it to the ISCED EduYears (years of completed education) categories, as per international GWA meta-analyses.

Geographical and parental data in MoBa

To explore the degree that genetic associations with field choices in the main analyses were due to factors other than direct genetic effects (e.g., familial and geographical stratification), we created covariates using data from the Norwegian population register and the education register.

Norway is a useful context for studying geography and educational choice, because the population is dispersed across diverse areas, tuition is free and there are generous student subsidies. Despite efforts to decentralise education, there are geographical differences in the kind of education offered e.g., with old and specialised universities still located in the major cities, offering more prestigious degrees qualifying for elite professions [8]

First, we created dummy variables representing birthplace municipalities. Norway's municipalities are the lowest administrative and elected level in Norway and divide the country into several hundred geographical areas (definitions of municipalities change over time, so we use municipality codes from 2018). The genetic-geographical data linkage and structure has been described previously [60]. There were birthplace dummy codes for 216 municipalities with at least MoBa 'cases' per code. Including these covariates therefore holds constant the effects of *all* environments shared by people born in the same municipality. This approach does not fully capture confounding from temporally unstable municipality factors, or more proximal lower-level geographical influences.

Second, we created dummy variables for the educational fields of the GWAS participants' mothers and fathers, in the same way as described in the section above. This will control for diverse effects of parental environments/behaviour on offspring field choices, to the extent that these effects are correlated with parental fields of education. Parents' educational fields may equip them with education-specific skills, cultural resources, social networks and economic resources which are passed onto offspring through environmental channels.

Polygenic indices (PGI) in Dutch Lifelines

To validate our GWA results, we tested associations between polygenic indices for educational field choices and actual educational field choices in an independent cohort. PGIs were constructed using SBayesR [61]. SBayesR uses Bayesian shrinkage to correct for linkage disequilibrium, using linkage disequilibrium scores of European-associated Ancestry individuals estimated in UK Biobank. To test for the direct effect of the PGI, we added the mid-parental PGI as a control variable. The parental PGI was constructed using a combination of observed and imputed genotypes (from parents and siblings), constructed using *snipar*. *Snipar* uses sibling data or the data of available parents to impute genotypes for unobserved parents [62]. The parental PGI could be imputed for individuals who had at least one sibling or at least one parent that was also genotyped in Lifelines (N=14,160).

Analyses

Genome-wide association meta-analyses (GWA)

We performed genome-wide association analyses (GWA) for 10 broad educational fields in MoBa and FinnGen. To allow relatives to be incorporated in analyses, we use mixed models controlling for genome-wide genetic relatedness (FastGWA software in MoBa[63], REGENIE In FinnGen [64].

To enable meta-analysis of MoBa and FinnGen GWAS summary statistics, we performed quality control and harmonisation. We removed variants with low minor allele frequencies <1%, poor imputation quality (INFO<0.8), multiallelic variants, variants with ambiguous alleles (e.g., alleles other than A, C, G or T), and we resolved strand and sign flips. The datasets were harmonised based on chr:pos from genome build 37 as SNP identifier. Sample-size-weighted meta-analyses of MoBa and FinnGen were then performed using METAL software [35].

To identify independent genome-wide significant associations in the meta-analytic results, we performed clumping using standard parameters in FUMA software: leadP = 5e-8, gwasP = 0.05, r2 = 0.6, r2_2 = 0.1, refpanel = 1KG/Phase3 [65].

In MoBa we also used FastGWA for the following additional analyses: GWAS of 10 fields with controls for educational attainment; GWAS of 10 fields with controlling for geographical and parental variables described above; GWAS of 10 fields in males and females separately; and GWAS of the Big 5 personality traits.

In all GWA (except sex-stratified) analyses, we controlled for sex, age, 20 principal components of genetic ancestry, and batch identifiers. In the case of the FinnGen study, we controlled for sex, age, the first ten principal components, educational attainment, and batch identifiers.

SNP-based heritability analyses

The overall contribution of common SNPs to field choices was estimated from GWA summary statistics by Linkage disequilibrium (LD) score regression, via the GenomicSEM R package [22,36]. On average, SNPs with higher LD Scores (more correlations with other SNPs) are more likely to be correlated with a true causal variant. As such, when genome-wide association test statistics (χ^2) are regressed on LD scores, the slope provides an estimate of the heritability that can be explained by common SNPs.

Genetic analyses adjusting for birthplace and parents' educational fields

To study how much the genetic influences on broad field choices are mediated through unobserved social factors in the geographical area individuals were born in, we repeat the GWAS analyses in MoBa only, controlling for municipality codes and parental field codes as dummy variables and test for attenuated heritability with LD Score regression. We did this in an iterative fashion, first controlling for the most distal factor (birthplace municipality) then adding parental fields. This approach was informed by previous work in UK Biobank [25].

Polygenic index analyses in an independent cohort

In Dutch Lifelines, we tested whether each of the 10 PGIs was predictive for its respective educational field using genetic data at the population level. In these regressions, we controlled for the first 10 PCs, a cube in age, sex, and an interaction between sex and a cube in age. To estimate direct genetic effects of PGIs on field qualifications, we then introduced controls for imputed parental PGIs and observed the within-family effect sizes that are adjusted for parental genetic effects and population stratification. The drop in effect size compared to population level analyses indicates the impact of factors other than direct genetic effects [66]. Bonferroni correction (p< 0.005) was used to correct for testing 10 hypotheses.

Genetic structure of fields

We explored the structure of the fields GWAS (meta-analytic results) by calculating genetic correlations using LD Score regression within Genomic Structural Equation Modelling software (Genomic SEM) [36]. For pairs of traits, the product of the genomewide association z-scores at each SNP can be regressed on the LD score, providing an estimate of the genetic correlation between the two traits. We then explored the dimensionality of genetic components of the 10 educational fields. We estimated how many axes of variation were required to explain genetic influences on the 10 educational fields by applying principal component analysis (PCA) to the genetic correlation matrix and observing variance explained and loadings. We used the fviz_pca_var() function from the R package factoextra to plot the PCA results. We then applied confirmatory factor analysis (CFA) in Genomic SEM based on the number of orthogonal components identified in PCA. We examined multiple models by comparing how inclusion of cross-loadings influenced model fit (see Supplementary Table 13 and 14 and Supplementary Figure). We also tested the performance of a common factor model. We ran all models using diagonally weighted least squares estimation as this accounts for differences in GWAS sample size and is optimal for modelling binary outcomes. Finally, we performed GWAS of the common factors in Genomic SEM (https://github.com/GenomicSEM/GenomicSEM/wiki/5.-Multivariate-GWAS).

Genetic correlations

We estimated genetic correlations between latent educational fields factors and 93 human phenotypes using LD score regression. We used GWAS summary statistics that were well powered and covered a comprehensive range of domains of human variation. See Supplementary Table 20 for the GWAS study reference list with sample sizes.

GWAS are all publicly available, except Big 5 Personality, for which we conducted GWA in MoBa using FastGWA. The Big-5 personality traits were assessed among fathers at recruitment in the 15th week of pregnancy and among mothers when the child was 5 years old by using The International Personality Item Pool (IPIP) Big-Five Factor Markers [67]. The IPIP is self-reported and consists of 10 items for each of the Big-5 personality traits Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. We allowed up to two of the ten items to be missing among participants and calculated an individual’s score on a trait as the average of the valid items. Among participants with genome-wide data, 42436 individuals aged 30-85 (M = 47.32, SD = 5.07; 47% female) were available for the present analyses.

Data availability

Individual level data: The Finnish biobank data can be accessed through the Fingenious® services (https://site.fingenious.fi/en/) managed by FINBB. Finnish Health register data can be applied from Findata (https://findata.fi/en/data/).

Summary statistics: The code used in this study and GWAS summary statistics are available upon request from the first author. Summary statistics from each data release will be made publicly available after a one year embargo period and can be accessed freely from: www.finngen.fi/en/access_results.

Acknowledgements

RC ZA JCE EME RP QQ and EY are supported by the Research Council of Norway (288083). RC is also supported by the Jacobs Foundation grant no. 2023-1510-00. SA was supported by the National Institute on Aging of the National Institutes of Health (R01AG078522) and the Dutch National Science Foundation (016.VIDI.185.044). Research reported in this publication was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement (ESSGN 101073237). PAD, JCE, and QQ are also supported by the European Union Grant agreement No. 101045526. THL and AF were supported by the ERC consolidator grant OPENFLUX "Societal openness, normative flux and the social modification of heritability" (ERC CoG 818420). HL was supported by European Research Council under the European Union's Horizon 2020 research and innovation program (#101019329). EY is also supported by the ERC consolidator grant EQOP "Socioeconomic gaps in language development and school achievement: Mechanisms of inequality and opportunity" (ERC CoG 818425), the RCN (grant no. 336078), and by the European Union (Grant agreement No. 101045526). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. AH was supported by the Research Council of Norway (#274611; #336085; #300668) and the South-Eastern Norway Health Authority (#2020022; 2019097). The Lifelines initiative has been made possible by subsidy from the Dutch Ministry of Health, Welfare and Sport, the Dutch Ministry of Economic Affairs, the University Medical Center Groningen (UMCG), Groningen University and the Provinces in the North of the Netherlands (Drenthe, Friesland, Groningen). The authors wish to acknowledge the services of the Lifelines Cohort Study, the contributing research centres delivering data to Lifelines, and all the study participants. We are grateful to all the participating families in Norway who take part in the on-going MoBa cohort study.

Ethics

The establishment of MoBa and initial data collection was based on a licence from the Norwegian Data Protection Agency and approval from The Regional Committees for Medical and Health Research Ethics. The MoBa cohort is now based on regulations related to the Norwegian Health Registry Act. The current study was approved by The Regional Committees for Medical and Health Research Ethics (project # 2017/2205).

Study subjects in FinnGen provided informed consent for biobank research, based on the Finnish Biobank Act. Alternatively, separate research cohorts, collected prior the Finnish

Biobank Act came into effect (in September 2013) and start of FinnGen (August 2017), were collected based on study-specific consents and later transferred to the Finnish biobanks after approval by Fimea (Finnish Medicines Agency), the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) statement number for the FinnGen study is Nr HUS/990/2017.

The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers: THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019 and THL/1524/5.05.00/2020), Digital and population data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA 134/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA 16/522/2020), Findata permit numbers THL/2364/14.02/2020, THL/4055/14.06.00/2020, THL/3433/14.06.00/2020, THL/4432/14.06/2020, THL/5189/14.06/2020, THL/5894/14.06.00/2020, THL/6619/14.06.00/2020, THL/209/14.06.00/2021, THL/688/14.06.00/2021, THL/1284/14.06.00/2021, THL/1965/14.06.00/2021, THL/5546/14.02.00/2020, THL/2658/14.06.00/2021, THL/4235/14.06.00/2021, Statistics Finland (permit numbers: TK-53-1041-17 and TK/143/07.03.00/2020 (earlier TK-53-90-20) TK/1735/07.03.00/2021, TK/3112/07.03.00/2021) and Finnish Registry for Kidney Diseases permission/extract from the meeting minutes on 4th July 2019.

The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 11 include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71, BB2019_7, BB2019_8, BB2019_26, BB2020_1, BB2021_65, Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, HUS/248/2020, HUS/430/2021 §28, §29, HUS/150/2022 §12, §13, §14, §15, §16, §17, §18, §23, §58 and §59, Auria Biobank AB17-5154 and amendment #1 (August 17 2020) and amendments BB_2021-0140, BB_2021-0156 (August 26 2021, Feb 2 2022), BB_2021-0169, BB_2021-0179, BB_2021-0161, AB20-5926 and amendment #1 (April 23 2020) and it´s modification (Sep 22 2021), BB_2022-0262, BB_2022-0256, Biobank Borealis of Northern Finland_2017_1013, 2021_5010, 2021_5018, 2021_5015, 2021_5015 Amendment, 2021_5023, 2021_5023 Amendment, 2021_5017, 2022_6001, 2022_6006 Amendment, BB22-0067, 2022_0262, Biobank of Eastern Finland 1186/2018 and amendment 22§/2020, 53§/2021, 13§/2022, 14§/2022, 15§/2022, 27§/2022, 28§/2022, 29§/2022, 33§/2022, 35§/2022, 36§/2022, 37§/2022, 39§/2022, 7§/2023, Finnish Clinical Biobank Tampere MH0004 and amendments (21.02.2020 & 06.10.2020), 8§/2021, 9§/2021, §9/2022, §10/2022, §12/2022, 13§/2022, §20/2022, §21/2022, §22/2022, §23/2022, 28§/2022, 29§/2022, 30§/2022, 31§/2022, 32§/2022, 38§/2022, 40§/2022, 42§/2022, 1§/2023, Central Finland Biobank 1-2017, BB_2021-0161, BB_2021-0169, BB_2021-0179, BB_2021-0170, BB_2022-0256, and Terveystalo Biobank STB 2018001 and amendment 25th Aug 2020, Finnish Hematological Registry and Clinical Biobank decision 18th June 2021, Arctic biobank P0844: ARC_2021_1001.

The Lifelines protocol was approved by the UMCG Medical ethical committee under number 2007/152.

Bibliography

1. van de Werfhorst, H. G. Intergenerational resemblance in field of study in the netherlands. *Eur. Sociol. Rev.* **17**, 275–293 (2001).
2. van de Werfhorst, H. G. & Kraaykamp, G. Four Field-Related Educational Resources and Their Impact on Labor, Consumption, and Sociopolitical Orientation. *Sociol. Educ.* **74**, 296 (2001).
3. Kelly, E., O'Connell, P. J. & Smyth, E. The economic returns to field of study and competencies among higher education graduates in Ireland. *Econ. Educ. Rev.* **29**, 650–657 (2010).
4. Gerber, T. P. & Cheung, S. Y. Horizontal stratification in postsecondary education: forms, explanations, and implications. *Annu. Rev. Sociol.* **34**, 299–318 (2008).
5. Dryler, H. Parental role models, gender and educational choice. *Br. J. Sociol.* **49**, 375–398 (1998).
6. Ochsenfeld, F. Why do women's fields of study pay less? A test of devaluation, human capital, and gender role theory. *Eur. Sociol. Rev.* **30**, 536–548 (2014).
7. Reisel, L. & Seehuus, S. Unpacking the logics of gendered educational choices: 10th graders' evaluation of appropriate educational tracks. *Educational Review* 1–21 (2023) doi:10.1080/00131911.2023.2182762.
8. Helland, H. & Heggen, K. Regional differences in higher educational choice? *Scandinavian Journal of Educational Research* **62**, 1–16 (2017).
9. van de Werfhorst, H. G. & Luijkx, R. Educational field of study and social mobility: disaggregating social origin and education. *Sociology* **44**, 695–715 (2010).
10. van der Vleuten, M., Jaspers, E., Maas, I. & van der Lippe, T. Intergenerational transmission of gender segregation: How parents' occupational field affects gender differences in field of study choices. *Br. Educ. Res. J.* **44**, 294–318 (2018).
11. van de Werfhorst, H. G. Cultural capital: strengths, weaknesses and two advancements. *Br. J. Sociol. Educ.* **31**, 157–169 (2010).
12. Jackson, M., Luijkx, R., Pollak, R., Vallet, L.-A. & van de Werfhorst, H. G. Educational fields of study and the intergenerational mobility process in comparative perspective. *Int. J. Comp. Sociol.* **49**, 369–388 (2008).
13. Verbree, A.-R., Maas, L., Hornstra, L. & Wijngaards-de Meij, L. Personality predicts academic achievement in higher education: Differences by academic field of study? *Learn. Individ. Differ.* **92**, 102081 (2021).
14. Lykken, D. T., Bouchard, T. J., McGue, M. & Tellegen, A. Heritability of interests: a twin study. *J. Appl. Psychol.* **78**, 649–661 (1993).
15. Vedel, A. Big Five personality group differences across academic majors: A systematic review. *Pers. Individ. Dif.* **92**, 1–10 (2016).
16. Johnson, M. K., Mortimer, J. T., Lee, J. C. & Stern, M. J. Judgments About Work. *Work Occup.* **34**, 290–317 (2007).
17. Vukasović, T. & Bratko, D. Heritability of personality: A meta-analysis of behavior genetic studies. *Psychol. Bull.* **141**, 769–785 (2015).
18. Plomin, R., DeFries, J. C. & Loehlin, J. C. Genotype-environment interaction and correlation in the analysis of human behavior. *Psychol. Bull.* **84**, 309–322 (1977).
19. Moloney, D. P., Bouchard, T. J. & Segal, N. L. A Genetic and environmental analysis of the vocational interests of monozygotic and dizygotic twins reared apart. *J. Vocat. Behav.* **39**, 76–109 (1991).

20. Roeling, M. P., Willemsen, G. & Boomsma, D. I. Heritability of working in a creative profession. *Behav. Genet.* **47**, 298–304 (2017).
21. Rimfeld, K., Ayorech, Z., Dale, P. S., Kovas, Y. & Plomin, R. Genetics affects choice of academic subjects as well as achievement. *Sci. Rep.* **6**, 26373 (2016).
22. Bulik-Sullivan, B. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
23. Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**, 1396–1400 (2019).
24. Morris, T. T., Davies, N. M., Hemani, G. & Smith, G. D. Population phenomena inflate genetic associations of complex social traits. *Sci. Adv.* **6**, eaay0328 (2020).
25. Abdellaoui, A., Dolan, C. V., Verweij, K. J. H. & Nivard, M. G. Gene-environment correlations across geographic regions affect genome-wide association studies. *Nat. Genet.* **54**, 1345–1354 (2022).
26. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424–428 (2018).
27. Cheesman, R., Ayorech, Z., Eilertsen, E. M. & Ystrom, E. Why we need families in genomic research on developmental psychopathology. *JCPP Advances* **3**, e12138 (2023).
28. Kweon, H. *et al.* Associations between common genetic variants and income provide insights about the socioeconomic health gradient. *BioRxiv* (2024) doi:10.1101/2024.01.09.574865.
29. Akimova, E. T., Wolfram, T., Ding, X., Tropf, F. C. & Mills, M. C. Genome-wide association study of occupational status and prestige identifies 106 genetic variants and defines their role for intergenerational status transmission and the life course. *BioRxiv* (2023) doi:10.1101/2023.03.31.534944.
30. Okbay, A. *et al.* Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics* (2022).
31. Østbakken, K. M., Reisel, L., Schøne, P. & Barth, E. Kjønnssegregering og mobilitet i det norske arbeidsmarkedet. (2017).
32. Corfield, E. C. *et al.* The Norwegian Mother, Father, and Child cohort study (MoBa) genotyping data resource: MoBaPsychGen pipeline v.1. *BioRxiv* (2022) doi:10.1101/2022.06.23.496289.
33. Magnus, P. *et al.* Cohort profile update: the norwegian mother and child cohort study (moba). *Int. J. Epidemiol.* **45**, 382–388 (2016).
34. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
35. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
36. 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).
37. Mallard, T. T. *et al.* Multivariate GWAS of psychiatric disorders and their cardinal symptoms reveal two dimensions of cross-cutting genetic liabilities. *Cell Genomics* **2**, (2022).
38. Saunders, G. R. B. *et al.* Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature* **612**, 720–724 (2022).
39. Vukojevic, V. *et al.* Evolutionary conserved role of neural cell adhesion molecule-1 in

memory. *Transl. Psychiatry* **10**, 217 (2020).

40. Kim, H. *et al.* Genome-wide association analyses using machine learning-based phenotyping reveal genetic architecture of occupational creativity and overlap with psychiatric disorders. *Psychiatry Res.* **333**, 115753 (2024).
41. Begall, K. & Mills, M. C. The influence of educational field, occupation, and occupational sex segregation on fertility in the netherlands. *Eur. Sociol. Rev.* **29**, 720–742 (2013).
42. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
43. Howe, L. J. *et al.* Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nat. Genet.* **54**, 581–592 (2022).
44. Fields of Study, Acquired Skills and the Wage Benefit from a Matching Job on JSTOR. https://www.jstor.org/stable/4194947.
45. Holland, J. L. *Making vocational choices: A theory of vocational personalities and work environments*. (psycnet.apa.org, 1997).
46. Usslepp, N. *et al.* RIASEC interests and the Big Five personality traits matter for life success-But do they already matter for educational track choices? *J. Pers.* **88**, 1007–1024 (2020).
47. Hjorth, C. F. *et al.* Mental health and school dropout across educational levels and genders: a 4.8-year follow-up study. *BMC Public Health* **16**, 976 (2016).
48. Husky, M. M. *et al.* Self-reported mental health problems and performance in mathematics and reading in children across Europe. *European Journal of Developmental Psychology* **17**, 704–726 (2020).
49. Hoogman, M., Stolte, M., Baas, M. & Kroesbergen, E. Creativity and ADHD: A review of behavioral studies, the effect of psychostimulants and neural underpinnings. *Neurosci. Biobehav. Rev.* **119**, 66–85 (2020).
50. Kyaga, S. *et al.* Creativity and mental disorder: family study of 300,000 people with severe mental disorder. *Br. J. Psychiatry* **199**, 373–379 (2011).
51. Burgess, S., Hauberg, D. S., Rangvid, B. S. & Sievertsen, H. H. The importance of external assessments: High school math and gender gaps in STEM degrees. *Econ. Educ. Rev.* **88**, 102267 (2022).
52. Harden, K. P. Genetic determinism, essentialism and reductionism: semantic clarity for contested science. *Nat. Rev. Genet.* **24**, 197–204 (2023).
53. Goncalo, J. A. & Duguid, M. M. Follow the crowd in a new direction: When conformity pressure facilitates group creativity (and when it does not). *Organ. Behav. Hum. Decis. Process.* **118**, 14–23 (2012).
54. Avdeev, S., Ketel, N., Oosterbeek, H. & Klaauw, B. Spillovers in fields of study: siblings, cousins, and neighbors. *SSRN Journal* (2023) doi:10.2139/ssrn.4578416.
55. Esping-Andersen, G. *The Three Worlds of Welfare Capitalism*. (1990).
56. Barth, E., Moene, K. O. & Willumsen, F. The Scandinavian model—An interpretation. *J. Public Econ.* **117**, 60–72 (2014).
57. Seehuus, S. Social class background and gender-(a)typical choices of fields of study in higher education. *Br. J. Sociol.* **70**, 1349–1373 (2019).
58. Willemse, N. & de Beer, P. Three worlds of educational welfare states? A comparative study of higher education systems across welfare states. *J. Eur. Soc. Policy* **22**, 105–117 (2012).
59. Scholtens, S. *et al.* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* **44**, 1172–1180 (2015).

60. Cheesman, R. *et al.* A population-wide gene-environment interaction study on how genes, schools, and residential areas shape achievement. *NPJ Sci. Learn.* **7**, 29 (2022).
61. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).
62. Young, A. I. *et al.* Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nat. Genet.* **54**, 897–905 (2022).
63. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* **51**, 1749–1755 (2019).
64. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
65. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
66. Wang, B. *et al.* Robust genetic nurture effects on education: A systematic review and meta-analysis based on 38,654 families across 8 cohorts. *Am. J. Hum. Genet.* **108**, 1780–1791 (2021).
67. Goldberg, L. R. A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several Five-Factor models. in *Personality psychology in Europe* (ed. Tilburg, The Netherlands: Tilburg University Press.) 7–28 (1999).