Originally published 30 May 2013; corrected 20 June 2013; corrected 1 November 2013

www.sciencemag.org/cgi/content/full/science.1235488/DC1

Supplementary Materials for

GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment

See the main paper for the full author list.

Corresponding author. E-mail: db468@cornell.edu (D.J.B.); dac12@nyu.edu (D.C.); koellinger@ese.eur.nl (P.D.K.); peter.visscher@uq.edu.au (P.M.V.)

Published 30 May 2013 on *Science* Express
DOI: 10.1126/science.1235488

This PDF file includes:

Correction: The revised file incorporates some minor text changes and a paragraph shift from the end of Section 7 to the end of Section 8.

Correction: Min A. Jhun, M.S., was originally omitted from the Additional Acknowledgements (Section 13). Her name has been added to this section, as she should have been included as a coauthor, based on her contributions in conducting genome-wide association analyses for the Genetic Epidemiology Network of Arteriopathy (GENOA) study.

Supplemental Online Materials

Contents

Materials and Methods

1. Conceptual design GWA study

The GWA study consisted of three parts: the discovery stage, the replication stage, and the combined stage. As predefined in the analysis plan, SNPs with p-values $< 10^{-6}$ in the discovery stage were eligible for bringing forward to the replication stage. The study cohorts (Table S1) in the discovery stage were recruited from February 2011 – June 2011, and the GWA summary results were uploaded before the end of July 2011. The replication cohorts were recruited from November 2011 – April 2012, and the results were uploaded before the end of May 2012. The combined-stage analysis used results from both the discovery and replication stages. All participants provided written informed consent, and the studies were performed in accordance with the respective Local Research Ethics Committees or Institutional Review Boards. The descriptive statistics and study designs are provided in Table S1.

a. Phenotype

Two measures of Educational Attainment (EA) were defined in accordance with the 1997 International Standard Classification of Education (ISCED) of the United Nations Educational, Scientific and Cultural Organization (UNESCO). This classification transforms each country-specific educational system into seven internationally comparable categories of EA *(14)*. In each study, EA of the subjects was first transformed into the appropriate ISCED level of the country. Thereafter the equivalent to US years of schooling was imputed, as described in Table S2. In some countries the measures did not differentiate between levels 5 and 6. In these cases everyone with a tertiary education was coded as ISCED 5, and 20 years of schooling was imputed instead of 19. The resulting continuous measure of EA as US-schooling-year equivalents is abbreviated as *EduYears* throughout the manuscript.

We also analyzed the binary outcome, *College*, which differentiates between individuals who hold a tertiary degree and those who do not. This binary variable was imputed taking the value 1 if the individual had completed a college degree (ISCED level 5 or above of the ISCED classification), and 0 if the individual had not completed a college degree (ISCED level 4 or below).

EduYears may provide more information about individual differences within a country, but *College* may be more comparable across countries. Nonetheless, the point biserial correlation between the two measures is relatively high, e.g., 0.82 (in the STR sample), 0.74 (RS-I), 0.88 (RS-II) and 0.91 (RS-III). Note, however, that the *EduYears* analysis focuses on the effects at the mean of the phenotype distribution, whereas the *College* analysis focuses on differences between the upper tail of the phenotype distribution and the remaining values.

The study-specific phenotype measurements and distributions are summarized in Table S3. All studies used a self-report of educational attainment, except STR. In STR, official register-based results for educational attainment were available. The descriptive statistics for the basic study-specific age and birth years are provided in Table S4.

The combined discovery sample comprises 101,069 individuals for *EduYears* and 95,427 individuals for *College*. Analyses were performed at the cohort level according to a pre-specified analysis plan, which restricted

the sample to Caucasians (to help reduce stratification concerns). Educational attainment was measured after subjects were very likely to have completed their education (over 95% of the sample was aged at least 30). While there are three exceptions to the age cut-off of 30 years, none of them are driving the results. The ALSPAC cohort includes 3,998 women aged < 30 years in the discovery meta-analysis. There were two reasons to deviate from the age-inclusion threshold for this cohort. First, because ALSPAC is a pregnancy cohort recruited in the early 1990s, few participants are likely to have obtained additional education following the peri-delivery questionnaire. Second, both for *EduYears* and *College*, the data collected after delivery are highly predictive of that collected at the latest time point available. A detailed description of the ALSPAC cohort can be found under Cohort Specific Acknowledgements below. In GENOA, 8 women between aged 25-30 years and 3 men aged of 26-30 were included in the analyses, since GENOA is a study of sibships. In ORCADES 27 people aged < 30 years (19 females and 8 males) were included in the analysis, with an average age of 28.2.

b. Genotyping

All cohorts were genotyped using commercially available Illumina (Illumina, Inc., San Diego, CA, USA), Affymetrix (Affymetrix, Inc., Santa Clara, CA, USA), or Perlegen (Perlegen Sciences, Inc. Mountain View, CA, USA) genotyping arrays. The quality controls were performed independently for each study. Each study imputed genotype data to HapMap 2 CEU (r22.b36) references using Beagle *(23)*, BIMBAM *(24)*, IMPUTE *(25)*, MaCH *(26)* or PLINK *(27)*. The study-specific details are provided in Table S5.

c. Analysis

For the *EduYears* (respectively, *College*) analysis, each study provided sex-stratified summary results of the ordinary least squares (logistic) regression of *EduYears* (*College*) on the imputed SNPs, the first four principal components of the Identity-by-State (IBS) matrix (to control for subtle population stratification), and [(birth year – 1900)/10], $[(\text{birth year} - 1900)/10]^2$ and $[(\text{birth year} - 1900)/10]^3$ (to control for age). In addition, appropriately-defined dummy variables were included when the educational attainment of some subjects was affected by significant country-specific events, such as World War II, the Vietnam War in the US, or changes in the educational system (Table S5). The family studies also provided GWAS results for males and females pooled, including as additional covariates in their analyses: sex and three interaction terms between sex and [(birth year – 1900)/10], $[(\text{birth year} - 1900)/10]^2$ and $[(\text{birth year} - 1900)/10]^3$.

d. Simulation study of Type I error rate

We performed two simulation experiments to determine whether the skewed, pseudo-continuous distribution of *EduYears* might inflate Type-I errors in the ordinary least squares (OLS) regression. See Figure S1 for a typical example of the distribution of *EduYears* using the data of one of the largest contributing cohorts, the Rotterdam Study I.

In the first simulation experiment we simulated a common (MAF = 0.331) and a rare (MAF = 0.026) SNP for 1,000 individuals using PLINK*(27)*. In each simulation run we generated a phenotype distribution from a multinomial distribution, with probabilities for each category equal to those in RS-I (Figure S1). We performed OLS using PLINK, and we stored the *p*-values of the two regression coefficients. We repeated these calculations 100,000 times and plotted the *p*-values in histograms. Assuming a significance level of 5%, we expected to obtain 5,000 (5%×100,000) *p*-values smaller than 0.05, as there was no association between the genotype and

phenotype due to the random data-generation process. Furthermore, we expected a uniform distribution of the *p*-values. For both the common and rare SNP, the number of regression coefficients with a *p*-value smaller than 0.05 was close to 5,000—5,132 and 4,922, respectively—and the *p*-values appear nearly uniformly distributed (Figure S2A and S2B).

Potential problems with the inflated Type-I error rate might be amplified in the tails of the distribution, however, due to linkage disequilibrium. These effects were not captured in experiment 1. Therefore, we performed a second simulation experiment using the observed (imputed) genotype data from the Rotterdam Study I. For each of five simulation runs, we permuted the phenotype values of the individuals in the sample and performed a full GWAS on the data. Because the null hypothesis is true in the simulation runs, the expected *p*-values follow a uniform distribution.

We obtained approximately 5×2.5 million = 12.5 million *p*-values and plotted the results in a Quantile-Quantile-plot (Figure S2C). No genome-wide significant associations ($p < 5\times10^{-8}$) were observed, and there was no excess of *p*-values in the tail of the distribution.

From the two simulation studies, we concluded that the gaps and the skewness in the phenotype distribution did not inflate the Type I error rate in this study.

e. Quality control and meta-analysis

Each results file going into the meta-analysis contained the following information: SNP ID, coded allele (allele to which regression coefficient refers), non-coded allele, strand, beta (regression coefficient), standard error, *p*-value, allele frequency for the coded allele, *N* (sample size) and information (imputation quality score). The SNPs with a Minor Allele Frequency (MAF) < 1% and an imputation quality score < 40% were excluded. For some files, these quality-control filters were slightly adjusted to more stringent levels (Table S5). Summary file-specific Quantile-Quantile plots were visually inspected. After quality control the genomic control (GC) inflation factor λ *(28)* was calculated for each summary file (Table S5). The meta-analysis was performed using METAL *(29)*, with sample-size weighting and single GC. All studies provided GWA summary results for sex-specific analyses, and we also performed sex-specific meta-analyses.

To calculate standardized regression coefficients from the METAL output, we used the formula

$$\hat{\beta}_j \approx z_j \cdot \frac{\hat{\sigma}_y}{\sqrt{N_j \cdot 2 \cdot MAF_j \cdot (1 - MAF_j)}}$$

for SNP *j* with minor allele frequency MAF_j, sample size N_j, METAL *z*-statistic z_j, and standard deviation of the phenotype $\hat{\sigma}_y$ (equal to 1 for *EduYears* after standardizing the phenotype, and equal to $\sqrt{p_j \cdot (1 - p_j)}$ for *College*, where p_j is the proportion of cases in the sample (given in Table S3)). This formula is an approximation for N_j large and SNP *j* in Hardy-Weinberg equilibrium. To derive it, substitute the estimated standard error

$$SE\left(\hat{\beta}_j\right) \equiv \left(\frac{\frac{1}{N_j} \sum \left(y_{ij} - x_{ij} \hat{\beta}_j \right)^2}{\sum x_{ij}^{\ 2}} \right)^{\frac{1}{2}} = \left(\frac{1}{N_j} \frac{\hat{\sigma}_y^2 - \hat{\sigma}_{x,j}^2 \hat{\beta}_j^{\ 2}}{\hat{\sigma}_{x,j}^2} \right)^{\frac{1}{2}}$$

into the definition of the z-statistic $z_j \equiv \hat{\beta}_j / SE(\hat{\beta}_j)$. Squaring and solving for $\hat{\beta}_j{}^2$ yields

$$\hat{\beta}_j{}^2 = \frac{z_j{}^2 \hat{\sigma}_y^2}{N_j \hat{\sigma}_{x,j}^2 \left(1 + z_j{}^2 / N_j\right)}.$$

Assuming Hardy-Weinberg equilibrium,

$$\hat{\sigma}_{x,j}^2 = 2 \cdot MAF_j \cdot (1 - MAF_j).$$

Using the definition of the *z*-statistic,

$$\frac{z_j{}^2}{N_j} = \frac{\hat{\beta}_j{}^2}{\sqrt{N_j}\left[\sqrt{N_j} Var(\hat{\beta}_j)\right]} \to 0$$

almost surely because the term in brackets converges to a constant (by the Central Limit Theorem). The formula follows from taking the square-root of the expression for $\hat{\beta}_j{}^2$.

Again assuming Hardy-Weinberg equilibrium, we approximate the explained variance R^2 for SNP *j* by

$$R_j^2 \approx \frac{2 \cdot MAF_j \cdot (1 - MAF_j) \cdot \hat{\beta}_j^2}{\hat{\sigma}_y^2}.$$

For *EduYears*, to convert standardized regression coefficients to regression coefficients in units of years, we multiply each standardized coefficient by the standard deviation of *EduYears* (given in Table S3). For *College*, to generate the regression coefficient $\hat{\beta}_j^{College}$ we divide each standardized coefficient by $p_j(1 - p_j)$.

The odd ratio for SNP *j* for *College* is $OR_j = \exp\left(\hat{\beta}_j^{College}\right)$.

We calculate the marginal effect of SNP *j* as

$$ME_j = \frac{\frac{p_j}{1 - p_j} OR_j}{1 + \frac{p_j}{1 - p_j} OR_j} - p_j.$$

To understand this formula, note that $\frac{p_j}{1 - p_j}$ is the baseline odds that *College* is equal to 1, and hence

$$\frac{p_j}{1 - p_j} OR_j$$

is the odds for an individual with one additional risk allele. Converting these odds to a probability, the expression

$$\frac{p_j}{1 - p_j} OR_j \Big/ \left(1 + \frac{p_j}{1 - p_j} OR_j\right)$$

is the probability that *College* is equal to 1 for an individual with one additional risk allele. The marginal effect is the difference between that probability and the baseline probability. Finally, for each SNP displayed in the result tables, we used the SNP annotation database SCAN *(30)* to identify which gene it belongs to.

f. Discovery-stage genome-wide association meta-analysis

In the discovery stage, the GWA summary statistics were combined from 42 genome-wide association (GWA) studies in a meta-analysis of 101,069 individuals (40,564 males and 60,505 females) for *EduYears* and of 95,427 individuals (38,307 males and 57,120 females) for *College*. In the *EduYears* analysis, 59.9% of the individuals were female and 96.0% of the individuals were aged >30 years; see Table S4. In the *College* analysis, 59.9% of the individuals were female and 95.8% of the individuals were aged >30; see Table S4). After quality control, a total of 2,515,021 autosomal SNPs were meta-analyzed across 72 input files for *EduYears*. For *College* 2,510,674 autosomal SNPs were meta-analyzed across 65 input files. Only SNPs with an availability of ≥80% in the total sample were selected, resulting in 2,299,174 SNPs for *EduYears* and 2,309,290 SNPs for *College*. No additional genome-wide significant results emerged when the availability filter was not applied. After single GC, the overall genomic control inflation factor λ was 1.155 for *EduYears* and 1.154 for *College*. The λ_{1000} genomic control inflation factors *(31)* were 1.002 and 1.005, respectively, for the number of included individuals *(32)* (assuming that all study samples are controls for *EduYears*). The genomic control inflation factors are relatively high but comparable to those in other large GWAS studies on complex traits; see, for example, *(15)*. SNPs with p-values < 10^{-6} in the discovery stage were brought forward for further analysis in the replication stage. Using the clumping command in PLINK *(27)*, we selected SNPs with the strongest independent signals. The HapMap 2 CEU genotypes were used as reference panel; the physical threshold for clumping was 1000 kB, and the R^2 threshold for clumping was 0.01.

All study-specific GWAS results were quality controlled, crosschecked, and meta-analyzed using single genomic control and a sample-size weighting scheme at three independent analysis centers.

g. Replication-stage genome-wide association meta-analysis

The replication stage included 12 studies, comprising 25,490 individuals (11,936 males and 13,554 females) for both *EduYears* and *College*. A total of 53.2% of the individuals across studies were females, and 99.89% of the individuals were aged >30 years, see Table S4. Cohorts in the replication stage provided summary GWA statistics similar to those of the discovery-stage cohorts. The quality control procedures and meta-analysis techniques were identical to those of the discovery stage. The results are reported in Table 1.

h. Combined-stage genome-wide association meta-analysis

We conducted an overall meta-analysis, combining data from the discovery and replication stages. The GWA summary statistics were combined from all 54 (42 + 12) genome-wide association (GWA) studies of 126,559 individuals (52,500 males and 74,059 females) for *EduYears* and 120,917 individuals (50,243 males and 70,674 females) for *College*. In the *EduYears* analysis 58.5% of the individuals were female and 96.81% were aged >30 years; see Table S4. In the *College* analysis, 58.4% of the individuals were female and 96.66% were aged >30 years; see Table S4. Using the same quality control filters and meta-analysis techniques as in the discovery stage, we obtained a total of 2,521,321 and 2,518,942 autosomal SNPs meta-analyzed across 98 and 91 input

files for *EduYears* and for *College*, respectively. Filtering 80% SNP availability generated 2,310,444 SNPs for *EduYears* and 2,321,8963 SNPs for *College*. No additional genome-wide significant results were obtained when the availability filter was not applied. After single GC, the overall genomic control inflation factor λ *(28)* was 1.207 for *EduYears* and 1.206 for *College*. The λ_{1000} genomic control inflation factors *(31)* were 1.001 and 1.005, respectively. All replicated SNPs obtained genome-wide significance in the combined meta-analysis (Table 1). Using the clumping command in PLINK *(27)* (1000 kb, R^2 0.01), we identified 4 and 3 genome-wide significant loci for *EduYears* and *College*, respectively, in the combined meta-analysis (Table S6, S7). Three of these newly genome-wide significant SNPs (rs1487441, rs11584700 and rs4851264) are in linkage disequilibrium with the replicated SNPs. The remaining four are located in different loci and hence warrant further investigation: rs7309, a 3'UTR variant in TANK; rs11687170, close to GBX2; rs1056667, a 3'UTR variant in BTN2A1; and rs13401104 in LOC100128572. Future work should test these additional loci for replication.

QQ-plots of the meta-analysis results are provided in Figures S3 and S4. Manhattan plots summarizing the meta-analyses are displayed in Figures S5 and S6. Forest plots for all genome-wide significant SNPs show that the results are not driven by a few outlier cohorts or cohorts from a specific region (Figures S7-S15). Furthermore, these plots show that the pooled results are not driven by only one of the sexes. The effects of the identified SNPs are also broadly consistent across the two phenotype definitions (Tables S8-S9).

2. The heritability of educational attainment

Since *(33)*, a number of studies have estimated the heritability of educational attainment by contrasting the resemblance of monozygotic and dizygotic twins. Virtually without exception, these studies find that monozygotic twins are appreciably more similar than dizygotic twins on years of educational attainment. Table S10, constructed from the sources compiled by Amelia Branigan, Kenneth J. McCallum and Jeremy Freese *(34)*, lists findings from published studies of American, Western European and Australian samples of twins and the heritability implied by the twin correlations under the assumptions of the standard ACE model. We omit correlations obtained from unpublished sources.

a. Estimates from multiple sibling types

To explore the robustness of the twin-based heritability estimates to the inclusion of other types of siblings, we gathered a large sample of Swedish brothers and data on their educational attainment and cognitive function. This sample, which we refer to as the Brothers Sample and whose construction is described below, contains seven different types of siblings: monozygotic twins (MZ), dizygotic twins (DZ), full siblings reared together (FRT), full siblings reared apart (FRA), half siblings reared together (HRT), half siblings reared apart (HRA) and adoptees (ADO). These correlations are previously unpublished and were part of the dissertation research of one of the authors of this paper *(35)*.

Statistics Sweden maintains a comprehensive database called the Multi-Generation Registry. The registry includes all individuals born after 1931 who were also residents in Sweden at some point since 1961. For individuals born in Sweden in the 1960s, the registry generally contains high-quality information about their biological parents. The registry also records whether an individual was adopted. The structure of the registry thus makes identification of various sibling types straightforward.

To construct the *Brothers Sample*, we used data from the Multi-Generation Registry to identify all Swedish males born between 1950 and 1969, as well as their full brothers and half-brothers (regardless of birth year). We classified brothers with the same biological parents as full siblings and brothers who share only one biological parent as half-siblings. We next assigned to each pair of siblings a rearing status using the quinquennial census data, which records whether or not two brothers are domiciled in the same household. Such census data are available for 1960, 1965, 1970, 1975, 1980 and 1985. Brothers who resided in the same household in every census where both were 18 years of age or younger are classified as reared together. We refer to brothers who share neither biological parent but lived in the same household in every census as adoptees.

We removed brothers born in the same year from the sample (since an overwhelming majority of these individuals are twins whose zygosity we are unable to infer from the administrative records). Brothers who never resided in the same household were classified as reared apart. We discarded ambiguous cases; that is, siblings who were domiciled in the same household in some censuses but not others. The final sample of non-twin brothers was restricted to brother pairs where both were born between 1950 and 1970. With the exception of the adoptees, we also restrict the final sample to siblings who are at most five years apart in age. We then supplemented these data with a sample of twins with known zygosity, also born between 1950 and 1970, using data from the Swedish Twin Registry. The Swedish Twin Registry's data contains information on Swedish twin births since 1886 and onward, and it has been described in detail elsewhere *(36)*.

Creating all possible pairings of relatives from this sample produces: 1,409 pairs of monozygotic twins, 1,922 pairs of dizygotic twins, 206,518 pairs of full siblings reared together, 1,362 pairs of full siblings reared apart, 6,445 pairs of half-siblings reared together, 14,713 pairs of half-siblings reared apart and 858 pairs of adoptees. There are a total of 207,738 pairs with complete data on educational attainment and 154,951 pairs with complete data on cognitive function. The smaller number of pairs with complete data on cognitive function may give the impression that missing data is potentially a serious problem. However, the main reason for the smaller sample is that the conscription records have only been digitized for men born after 1951. Therefore, sibling pairs where one sibling is born before 1951 will be incomplete. For most birth years, over 95% of the men in the Brothers Sample are successfully matched to the conscription records. See *(35)* for a more detailed analysis of the sample.

We matched the *Brothers Sample* to administrative records with information about educational attainment and cognitive function. To measure years of education, we use data drawn from Statistics Sweden's administrative records. To measure cognitive function, we use data from the National Service Administration (which maintains the military conscription records). During the period that we study, Swedish men were required by law to participate in military conscription and underwent a comprehensive drafting procedure that involved taking a battery of mental tests. Cognitive function is measured using data from the Swedish Enlistment Battery *(17)*, a test similar to the U.S. Armed Forces Qualifying Test.

Table S11 reports cross-sibling pairwise correlations of educational attainment and cognitive function of the siblings in our sample. The diagonal entries represent the cross-sibling correlation for a particular trait, whereas the off-diagonal entries represent the cross-trait correlations between siblings. Siblings reared together always exhibit greater similarity than siblings reared apart, suggesting that differences in common environmental factors account for a substantial portion of variance across individuals. Consistent with a broad consensus in

behavior genetics *(37)*, the sibling correlations also suggest that genetic factors account for a larger fraction of variance than common environmental factors.

Finally, we use our data to estimate three highly stylized behavior-genetic models. Model 1 is simply the conventional ACE model. Models 2 and 3 make use of the additional moment conditions described below, to identify richer models which relax some of the restrictions of the ACE model. The three models are estimated by nonlinear least squares by solving

$$\hat{\Theta} = \arg\min \sum_{i=1}^{N} \left[\left(y_{i1} y_{i2} \right) - f_i(\Theta) \right]^2 ,$$

where *i* indexes the pair of brothers and $f_i(\Theta)$ is a moment condition that varies by sibling type. The variables are standardized so that they have mean zero and standard deviation one. In the baseline regressions, standard errors are clustered at the level of 1970 household.

The moment conditions for Model 1 are as follows: $h^2 + c^2$ for MZ pairs, $\frac{1}{2}h^2 + c^2$ for DZ pairs and full siblings reared together, $\frac{1}{2}h^2$ for full siblings reared apart, $\frac{1}{4}h^2 + c^2$ for half siblings reared together, $\frac{1}{4}h^2$ for half siblings reared apart, and c^2 for adoptees reared together. The ACE estimates for educational attainment (*N* = 216,091) are $\hat{h}^2 = 0.552$ (s.e. 0.027) and $\hat{c}^2 = 0.164$ (s.e. 0.014). The estimates are shown graphically in Figure S16. Despite its strong assumptions, a simple ACE model appears to fit the data surprisingly well.

Model 2 relaxes the assumptions of the ACE model in two ways. First, the degrees of genetic relatedness of full siblings (ρ_{FS}) and half-siblings (ρ_{HS}) are estimated rather than fixed at 0.5 and 0.25 respectively. Second, the model estimates separate c^2 coefficients for non-twin siblings who were reared together and twin siblings. Formally, we denote the amount of shared environmental variation in twins c_T^2 and then estimate the fraction of variation (λ) that is shared by non-twin siblings. The moment conditions are: $h^2 + c_T^2$ for MZ pairs, $\rho_{FS} h^2 + c_T^2$ for DZ pairs and full siblings reared together, $\rho_{FS} h^2$ for full siblings reared apart, $\rho_{HS} h^2 + \lambda c_T^2$ for half siblings reared together, $\rho_{HS} h^2$ for half siblings reared apart, and λc_T^2 for adoptees reared together. Notice that this model subsumes the ACE model as a special case with $\lambda = 1$, $\rho_{FS} = 0.5$ and $\rho_{HS} = 0.25$ The estimates from this model (*N* = 207,738) are $\hat{h}^2 = 0.494$ (s.e. 0.045) and $\hat{c}_T^2 = 0.211$ (s.e. = 0.033), $\hat{\lambda} = 0.705$ (s.e. = 0.099), $\hat{\rho}_{FS} = 0.591$ (s.e. = 0.052), $\hat{\rho}_{HS} = 0.247$ (s.e. = 0.028).

Model 3 allows the degree of environmental resemblance to vary more flexibly across sibling types, while maintaining standard assumptions about the genetic relatedness of full siblings (0.5), half siblings (0.25) and adoptees (0). In this model, the common environmental components are allowed to vary flexibly across MZ twins, DZ twins and all other co-reared non-twin siblings. We call the MZ environmental twin covariance c_{MZ}^2 and parameterize the two other environmental covariance terms as a scalar multiple of c_{MZ}^2. The moment conditions are: $h^2 + c_{MZ}^2$ for MZ pairs, $\frac{1}{2}h^2 + \lambda_T c_{MZ}^2$ for DZ pairs; $\frac{1}{2}h^2 + \lambda c_{MZ}^2$ for full siblings reared together, $\frac{1}{2}h^2$ for full siblings reared apart, $\frac{1}{4}h^2 + \lambda c_{MZ}^2$ for half siblings reared together, $\frac{1}{4}h^2$ for half siblings

reared apart, and λc^2_{MZ} for adoptees reared together. The parameter estimates for educational attainment (N = 207,738) are $\hat{h}^2 = 0.556$ (s.e. = 0.030) and $\hat{c}^2_{MZ} = 0.149$ (s.e. = 0.043), $\hat{\lambda}_1 = 1.51$ (s.e. = 0.422), $\hat{\lambda}_2 = 1.09$ (s.e. = 0.258).

Considered in their entirety, the results reinforce the conclusion that EA is a moderately heritable trait.

b. Estimating the variance in educational attainment explained by all SNPs

In addition to the analyses above, we also used the method developed by *(38)* to estimate the share of variance in *EduYears* and *College* that can be explained by all SNPs. This method provides a lower-bound estimate of narrow heritability, and the output can be interpreted as the fraction of variance that would be explained by the linear, additive effects of all the genotyped SNPs if these effects were observed without error.

There were 3,526 individuals from the QIMR cohort and 6,770 individuals from the STR cohort. We imputed the SNPs to the HapMap3 CEU panel and retained 1,121,675 SNPs after quality controls. We used the software GCTA to estimate the genetic relatedness between all the individuals and removed one of each pair of samples with estimated genetic relatedness > 0.025. We then estimated the variance explained by all the HapMap 3 SNPs by GREML analysis for *EduYear* and *College* using GCTA.

The results (see Table S12) suggest that ≈20% of the variance in educational attainment in the two samples can be attributed to genetic differences that are captured by the current SNP microarrays. The explanatory power of a linear polygenic score estimated in the same data will be lower because the coefficients used for constructing the score are estimated with error. This explains the difference between the estimates reported here and the performance of our polygenic scores in Figure 2.

3. Health and education

The health-education gradient is one of the most robustly documented and well-studied empirical relationships in social science *(39)* (see *(40)* for a review of the basic findings and an evaluation of the mechanisms at play). Researchers studying this relationship usually distinguish between three basic mechanisms that may explain the relationships. First, poor health in early-life (which may be partly due to genetic factors) may inhibit individuals from acquiring more education. Second, other factors, including heritable individual differences, could affect both schooling and health. Third, increased education may improve health, for example through effects on health-related behaviors.

There is evidence, mostly from twin and family studies, of some genetic overlap between (i) education and health-related behaviors (such as smoking or drinking; see *(41)*), and (ii) education and health outcomes (see, e.g. *(42, 43)*). Other research indicates that education has a causal impact on health. For example, *(44)* uses policy variation across states in compulsory schooling laws to instrument for educational attainment. These results imply that the causal effect of education on health is actually larger than the cross-sectional correlation. Other papers that have also taken quasi-experimental approaches include *(45)* and *(46)*.

To assess the genetic overlap between health and educational attainment due to common genetic variants, we estimated a bivariate GCTA model using the STR data. One of the STR questionnaires asks the respondents: "How would you rate your general health condition?" There are five possible responses, ranging from poor to

excellent. We assigned a value of 1 to the lowest category (poor), 2 to the second lowest category (not so good), and so on. The estimated genetic correlation between this health variable (measured on a continuous scale) and *EduYears* is 13.2% (standard error 23%). We also employed a binary-response model to estimate the genetic correlation between a dichotomized health variable and *College*, obtaining an estimate of 33% (standard error 33%). The estimates are imprecise but consistent with the hypothesis of positive genetic overlap (Table S13).

4. Exploring possible explanations for very small effect sizes

The effect sizes we find are much smaller than those found for other replicated SNP association results for complex physical traits such as body height *(15)*, BMI *(18)*, or metabolite profiles *(47)*. In this section, we explore three possible explanations:

A. Measurement error attenuates the estimated effect;

B. The genetic effect is conditional on specific environmental circumstances, and hence a meta-analysis approach that averages across different environments partially masks the genetic effect;

C. "Biologically distal" phenotypes such as years of education have smaller effect sizes than more "biologically proximal" phenotypes such as body height.

These factors are not mutually exclusive and may reinforce each other. Exploration of their relative importance may help to guide future research efforts.

To explore **A.**, we focus on analyses that use *EduYears* as the dependent variable (since the variable *College* used in the other analyses is measured similarly across studies). As a proxy for measurement quality of *EduYears* in a study, we use the number of distinct ISCED categories in the data.

The estimates of the effects of genetic variants are reported in the main text as unstandardized regression coefficients. In theory, to the extent that *EduYears* can be treated as a continuous variable, classical measurement error in *EduYears* should not attenuate the unstandardized regression coefficients (only reduce their precision).

Figure S17 plots the unstandardized coefficient against the number of categories that were available to respondents when they were asked about their educational attainment. For each of the three replicated SNPs, we ran weighted least squares regressions of the *EduYears* coefficients on the number of categories, with weights proportional to the sample size. Consistent with our expectation, we find no evidence that effects are weaker in cohorts with coarser measures. In all three regressions, we cannot reject the null hypothesis that there is no relationship (the smallest p-value is 0.515).

In contrast, measurement error in *EduYears* is expected to attenuate the standardized regression coefficients by increasing the standard deviation of *EduYears* (similarly attenuating the R^2 of the SNPs). To explore this possibility, we run analogous regressions and generate analogous plots using the standardized regression coefficients. Figure S18 plots the standardized coefficient against the number of categories that were available to respondents when they were asked about their educational attainment. The figure indicates that there is no significant attenuation (the smallest p-value is 0.359).

To get a sense for the amount of measurement error in our *EduYears* measure that may be due to errors in self report, we can exploit the fact that within the STR, data are available for both self report of highest educational attainment (a multiple-choice question with ten categories, plus additional open-ended questions about years of educational attainment) and a registry-based measure from government records. The correlation between the survey-based measure and the registry measure (after both are converted to the ISCED categorization and then to U.S. years-of-education equivalents) is 0.81, which suggests to us that the survey responses contain little measurement error relative to the registry-based measure. (Note that under some mild assumptions, 0.81 is an estimate of a lower bound of the reliability of the survey responses because the registry data could also contain errors.)

Regarding **B.**, note that in order to achieve a sample of $N > 100{,}000$, current data availability necessitated pooling GWAS results from different parts of the world and from individuals who completed their education at different points of time under vastly different circumstances. Therefore, the effects we identified are likely to be the ones that are most robust across various environments. Nevertheless, the meta-analysis results may also mask gene-environment interactions.

To explore one possible source of gene-environment interaction, we examined how the estimated genetic effects of our three replicated genome-wide-significant SNPs vary with birth cohort (an idea suggested to us by Steven Lehrer and Nicholas Christakis). Birth cohort is a proxy for a number of institutional changes in Western countries in the 20th century that were designed to raise the general level of educational attainment and to reduce inequalities in opportunity. These policies resulted in substantial increases in the rates of secondary and tertiary education.

We use *EduYears* as the dependent variable (also for the SNPs that we found to be significant with the *College* measure) because the large increases over time in the overall frequency of college completion makes it harder to interpret changes over time in genetic effects on the odds of college completion.

Figure S19 plots the study-specific unstandardized regression coefficient against the average birth cohort of participants in the study. The size of each data point is proportional to the study in question. We fit a regression line to these points by weighted least squares. There is no evidence that the effects of any of the three SNPs vary by birth cohort. In all three regressions, we cannot reject the null hypothesis that there is no relationship (the smallest p-value is 0.684).

To explore **C.**, we derived a theoretical framework (see section 7 below) within which a distal phenotype is caused by endophenotypes, which in turn are caused by SNPs. We explain why the fact that the polygenic score for educational attainment is *more* predictive of cognitive function than educational attainment is consistent with cognitive function being an endophenotype for education. We similarly conjecture (but cannot yet show) that personality and health traits may also be endophenotypes. The theoretical framework makes clear that SNPs are likely to be more weakly associated with a more distal phenotype such as educational attainment than with an endophenotype.

Considered jointly, our results suggest that the small effect sizes we find are not due to measurement error. We do not find evidence that the genetic effects interact with birth cohort. While we cannot exclude the possibility that stronger effects of individual SNPs on educational attainment exist that are conditional on other aspects of

the environment, we note that no single SNP reaches genome-wide significance in any particular cohort included in the meta-analysis, putting an upper bound on the effect sizes that can be expected within specific environments. There are strong theoretical reasons to expect that biologically-distal phenotypes will have weaker relationships with individual SNPs than more biologically-proximal phenotype do, and our empirical findings overall are most consistent with this explanation.

5. Biological annotation

In this section, we report the results from a series of bioinformatics analyses designed to explore possible biological mechanisms that may underlie the associations between the identified loci and educational outcomes.

We began by identifying all functional SNPs in LD with the seven SNPs that reached genome-wide significance in the combined analyses (see subsection *i*). These seven comprise the three original SNPs that reached genome-wide significance in the discovery stage and were subsequently replicated (see Table 1) as well as an additional four SNPs that reached genome-wide significance in the combined analyses and were not in linkage disequilibrium (LD) with the original three SNPs (see Tables S6 and S7).

Next, we examined whether any of these seven variants are associated with changes in gene expression levels in blood or brain tissue (see subsections *ii* and *iii*).

We subsequently turned to analyses that take as their input a larger set of SNPs than those meeting the stringent criterion for genome-wide significance. We conducted gene-based tests of association *(48)* (see subsection *iv*); pathway analyses *(49)* (see subsection *v*); and a recently developed method *(50)* (see subsection *vi*) that tests for enrichment of active chromatin in 34 different types of tissues in the regions implicated by the combined-stage GWAS meta-analyses.

From these primary biological follow-up analyses, we identify a set of "candidate" genes. These genes were used as inputs to a functional network analysis that uses gene co-expression data from multiple sources to predict a particular gene's likely functions *(51)* (see subsection *vii*). We also conduct a structured search of the existing genetics literature to explore what is currently known about phenotypes associated with the genes identified by our analyses (see subsection *viii*).

a. Characterizing genome-wide significant SNPs

i. Functional annotation

To identify coding or regulatory variants in close LD ($r^2 > 0.8$) with any of the seven signals that were either replicated or reached genome-wide significance in the combined analyses, we used the online tool HaploReg (http://www.broadinstitute.org/mammals/haploreg/haploreg.php). We observed 2 missense variants close to rs1056667 in gene *BTN1A1* and 3 missense variants close to rs11584700 in gene *LRRN2*, which is highly expressed in the brain and regulates axon guidance in model animals *(52)*. The complete set of results from the functional annotation lookup is given in Table S14. Four of the seven SNPs were not in close LD with any coding or regulatory variants and are therefore not listed in the table.

As a complementary analysis, we used the ENCODE custom tracks on the UCSC Genome browser (http://genome.ucsc.edu) to screen the implicated regions for overlap with DNAse I hypersensitivity sites and

open chromatin. We also used the online resource RegulomeDB (http://regulome.stanford.edu) to identify variants in the proximity of our seven SNPs with known functional annotation. However, we found no compelling evidence of enrichment.

ii. Gene expression eQTL analyses in brain tissue

We examined the association between each of the seven SNPs that either replicated or reached genome-wide significance in the combined analyses and the expression in brain tissue of nearby genes (within 1.2 Mb of the signal). We used data from two independent eQTL resources: SNPexpress *(53)* and Myers et al. *(54)*. The SNPexpress dataset contains expression (Affymetrix Human ST 1.o Exon array) and genotype (Illumina HumanHap 550K v1/3) data for 94 individuals. The Myers et al. dataset contains expression (Illumina Human Refseq-8 Expression BeadChip) and genotype (Affymetrix GeneChip Human Mapping 500K Array) data for 188 neurologically normal controls (data were also available for 176 Alzheimer's cases, but we did not include these data).

The two datasets were independently quality-controlled using identical filters and were analyzed separately. We removed SNPs with low minor allele frequency (MAF < 0.01), SNPs not in Hardy-Weinberg (HWE, $p < 1\times10^{-6}$), and SNPs with a low call rate (< 95%). To control for the effects of ancestry, we merged the data with the HapMap 3 ethnicity reference panels and conducted a multidimensional scaling (MDS) analysis using PLINK *(27)*, restricting the analyses to the SNPs present in both the target and HapMap reference datasets. We subsequently imputed the datasets to the HapMap 2 CEU reference panels using MACH (http://www.sph.umich.edu/csg/abecasis/MACH/).

Using MACH2QTL (http://www.sph.umich.edu/csg/abecasis/MACH/), we performed transcript-wide eQTL analyses for the seven SNPs. In these analyses, we controlled for sex, age, age^2, post-mortem interval, post-mortem interval^2, dummy variables for three sites, and ancestry (using the first four dimensions of the multidimensional scaling analysis). To adjust for hybridization, the analyses of the Myers et al. *(54)* dataset also included controls for date and brain region.

None of the observed *p*-values survived correction for multiple testing: the lowest nominal *p*-value observed in these analyses was $p \approx 1\times10^{-4}$. We note, however, that because the number of brain tissue samples available was small, our power to detect small effects on expression in brain tissue was limited.

iii. Gene expression eQTL analyses in blood tissue

We also examined the association between each of the seven SNPs that either replicated or reached genome-wide significance in the combined analyses and the expression in blood tissue of nearby genes (within 1.0 Mb of the signal). We performed this cis-eQTL mapping on three samples of unrelated individuals: 1,240 individuals from Fehrmann with expression data from the Illumina HT12v3 chip; 229 individuals from Fehrmann with expression data from the HT8v2 chip *(55, 56)*; and 891 individuals from the Estonian Biobank with expression data from the HT12v3 chip *(56)*. The gene expression data were obtained from total RNA in whole blood samples. As with the brain eQTL analyses, the genotype data were first filtered for MAF (> 0.01), HWE ($p \geq 1\times10^{-6}$), and call rate (95%) before imputation using the HapMap 2 CEU reference panel.

To avoid hybridization artifacts, we harmonized the data by aligning the gene expression probes to the human genome build 18 (Ensembl build 54) using BLAT, SOAPAlign v2, and BWA and excluding any probe that mapped to multiple genomic locations or contained more than two mismatches. Each expression dataset was normalized as follows: (1) quantile normalized, (2) $\log_2$ transformed, and (3) standardized to have a mean of zero and variance equal to one. We used the software MixupMapper *(57)* to identify and remove sample mix-ups. To correct for possible population structure, we residualized the gene expression data on the first four multi-dimensional scaling components obtained from the genotypic data. We residualized the resulting variable on 40 PCs (derived from the variance-covariance matrix of the genotypic data) that did not show any significant evidence of association with the genotypes (and might therefore have a biological interpretation).

To map the cis-eQTLs, we correlated the imputed genotypes with the transformed gene expression data. These analyses were conducted separately for each of the three samples and completed for each gene for which the midpoint of the probe was within 1.0 Mb of the SNP. The *Z*-statistics from the three samples were meta-analyzed, weighting each statistic by the sample size of the dataset. To test for statistical significance while correcting for multiple hypothesis testing (using a false discovery rate of 5%), we generated a distribution for the meta-analyzed Z-statistic under the null hypothesis by simulation: for each of 100 simulation runs, we permuted the sample labels and re-ran the meta-analysis. To determine whether the educational-attainment SNPs have independent cis-eQTL effects in a given loci, we performed conditional analysis as follows. We first determined which SNP showed the strongest cis-eQTL effect for each of the probes associated with the educational-attainment SNPs. Then, we adjusted the gene expression data for these effects using linear regression, and repeated the cis-eQTL analysis on the educational-attainment SNPs.

The analyses revealed several strong cis-regulatory signals for nearby genes (Table S15). Three of the educational-attainment-associated SNP eQTL effects—*BTN2A1*, *HMGN4*, and *MDM4*—remained significant even after removing the effect of the most significant SNP for the specific gene, thus suggesting an extra regulatory mechanism tagged by the GWAS signal.

b. Analyses aggregating effects across multiple SNP signals

iv. Gene-based tests

We used meta-analysis results from the combined-stage GWAS as input for VEGAS *(48)* to test for association at the level of the gene. In total, 17,661 genes were tested for *EduYears* and 17,676 for *College*. The 25 most strongly associated genes for both phenotypes are listed in Tables S16 and S17. After Bonferroni correction, 17 genes for *EduYears* and 7 for *College* are associated (p-value $\leq 2\times10^{-6}$). Of the 25 top genes for *EduYears* and *College*, 6 genes appear in both lists. Several of these genes (*TUFM*, *ATP2A1*, *ATXN2L*, and *SH2B1*) are located in adjacent positions on chromosome 16.

v. Pathway analyses

Pathway analysis typically entails two steps: first, identify genomic regions of interest (e.g., based on low p-values); and second, test whether these regions include genes that define known biological pathways more than expected by chance. To determine the genomic regions, we began by selecting a set of index SNPs that reached $p < 1\times10^{-5}$ in the combined-stage meta-analysis. A region surrounding each index SNP was extended to nominally associated SNPs ($p < 0.05$) within 250kb of the index SNP that were in moderate LD with the index

SNP ($r^2 > 0.5$). LD among SNPs was estimated from the HapMap 2 CEU reference panel using PLINK *(27)*. Any resulting regions that overlapped were subsequently merged, and only regions overlapping known genes were tested. In total, we identified 33 regions overlapping known genes for each of *EduYears* and *College*.

Next, we used INRICH *(58)* to test the identified genomic regions for overlap with 3,440 pathways listed in the Gene Ontology (GO) database that included between 5 and 200 genes *(59)*. Pathways showing suggestive enrichment (empirical $p < 0.05$) before multiple-testing corrections are listed in Table S18. None of these pathways demonstrated significant overlap with low *p*-value genomic regions in either the *EduYears* or *College* meta-analysis results after adjustment for multiple testing.

vi. Analyses of cell-type specificity

We employed a recently published method (http://www.broadinstitute.org/mpg/epigwas/, *(50)*) that tests for cell-type-specific enrichment of active chromatin, measured through H3K4me3 chromatin marks *(60)* in regions surrounding IndexSNPs identified by the combined-stage GWAS analysis. A recent paper shows that H3K4me3 chromatin marks are the most cell-type-specific marks in terms of co-localization with previously published GWAS loci *(50)*. The idea behind the method is that variants related to a particular phenotype may affect cell-type-specific gene expression by changing regulatory elements in cell types relevant to that phenotype. Hence, overlap between associated variants and chromatin marks should occur preferentially in the relevant cell type(s). Our analysis tested for enrichment of these chromatin marks in 34 different tissues.

We constructed the set of IndexSNPs by first identifying all SNPs that reached $p < 1\times10^{-5}$ in the combined-stage meta-analysis for each phenotype. We next pruned this set of SNPs to a final list of IndexSNPs, in which no two SNPs were in LD greater than $r^2 = 0.5$. For each IndexSNP, a locus region was defined, bounded on either side of the IndexSNP by the most distant SNP within 250 kb of the IndexSNP that was in LD ($r^2 > 0.8$) with the IndexSNP. For each SNP within each locus, regulatory activity scores were calculated as the height of the nearest H3K4me3 mark divided by distance from the SNP to the H3K4me3 mark. The SNP with the highest score within each IndexSNP locus region in a given tissue was designated the BestSNP, which served as the score representing that locus.

Cell-type-specificity scores per locus were estimated by normalizing BestSNP scores so that the sum of scores for a given locus across all cell types equaled 1. Cell-type-specificity scores per tissue were defined by summing normalized BestSNP scores across all loci within a given tissue. 10,000 sets of SNPs (matched to the IndexSNP regions having the same total number of H3K4me3 peaks) were sampled (from among background SNPs provided with the software) to estimate null distributions of cell-type-specificity scores per locus and per tissue. *P*-values for cell-type-specificity scores summed across all BestSNPs for each tissue (the observed per-tissue score) were estimated as the proportion of random SNP sets with a per-tissue score exceeding the observed per-tissue score. We identified loci with BestSNP cell-type-specificity scores falling at or above the 95th percentile of the corresponding null distribution as demonstrating greater than expected specificity within a given cell type *(50)*.

Figure S20 shows *p*-values for the cell-type-specific overlap of H3K4me3 marks and IndexSNP regions for each cell type. Four cell/tissue types showed significant overlap at nominal $p \leq 0.05$: for *EduYears*, anterior caudate ($p = 0.0089$), CD4+ naive primary cells ($p = 0.032$), hippocampus middle ($p = 0.05$) and muscle satellite

cultured cells (p = 0.0236); and for *College*, anterior caudate (p = 0.0007). Additionally, for *College* the mid-frontal lobe showed marginal enrichment at p = 0.0502. Only the anterior caudate results for *College* survive correction for multiple hypothesis testing.

The results from the analysis of overlap between H3K4me4 chromatin marks and education-associated SNP regions suggest gene expression regulatory function for those loci in specific cell types. In particular, we note that anterior caudate tissue appears enriched for both *EduYears* and *College* phenotypes, although only the latter survives multiple testing correction. Figure S21 shows cell-type-specificity scores per locus in the four nominally significant tissues and 95th-percentile threshold (dashed red line). Loci above the threshold were identified as showing greater than chance specificity within that particular cell type. Table S19 identifies these enriched loci, along with distance to the nearest chromatin mark.

c. Functional pathways and phenotypic associations of implicated genes

Table S20 provides an index listing every gene identified by any of the initial biological follow-up analyses (functional annotation, blood eQTL analyses, and the gene-based tests). The functional annotation column identifies genes with functional SNPs (missense, synonymous, or 3'UTR variants) in high LD ($r^2 > 0.8$) with one of the seven independent loci that were either replicated or significant in combined analyses (Tables 1, S6, S7; functional annotation results fully detailed in Table S14). The blood eQTL column lists genes showing a significant cis-eQTL signal within 1.0Mb of one of the seven loci (see Table S15 for blood eQTL details). The third column summarizes all genes that were significant in gene-based tests after correction for multiple testing (regardless of genomic location with regard to individually significant SNPs; full results are reported in Tables S16 for *EduYears*, and S17 for *College*). In the last column, we provide a map between the identified genes and their locations relative to the seven SNPs that were either replicated or reached genome-wide significance in the combined-sample analysis; that is, genes that were within 1.0 Mb of a significant SNP are labeled with the SNP identifier as well as the distance between the SNP and the nearest edge of the gene (or the location of the SNP within the gene, if appropriate). The complete list of genes presented in Table S20 was used as the input for the analyses that follow: examinations of likely gene functions as well as of previously-reported phenotypic associations.

vii. Gene function prediction using a large co-expression framework

We used a recently developed method to gain insight into the putative functions of all the genes listed in Table S20. This method takes as its input a list of genes and infers the probable functions of the genes by pooling published data on 80,000 gene expression profiles from humans, animals and cell lines. The method is described in a recent paper *(51)*, which also reports evidence that a prediction coming out of the framework was validated by subsequent wet lab experiments.

Gene-function prediction is based on the idea that genes with shared expression profiles are likely to have related biological functions. For example, if 50 genes are known to play a role in apoptosis, then a gene with unknown function that is strongly co-expressed with these 50 genes is likely to be part of apoptotic pathways as well. The method of *(51)* uses data on co-expression profiles to predict the likely functions of as-yet uncharacterized genes and refine our understanding of the function of other genes. The overall workflow of the

method has been graphically visualized at http://www.genenetwork.nl/genenetwork/ (described at the "method" link).

Table S21 lists all pathways associated with implicated genes after applying a false discovery rate criterion of < 0.05. Four genes queried (*BSN*, *GBX2*, *LRRN2*, *PIK3C2B*) tended to occur within neuronal pathways. Specifically, these genes were associated with axonal (*BSN*, *PIK3C2B*), dendritic (*BSN*, *LRRN2*), neuronal cell body (*LRRN2*), neuron fate (*GBX2*), and synaptic terms (*BSN*, *LRRN2*), as well as pathways related to learning and long-term memory (*BSN*) and glutamate receptor activity (*LRRN2*). In addition, several genes identified through gene-based tests (*PIK3C2B*, *IP6K3*, *ITPR3*, *TET2*) were implicated in muscular contractions and neuron-muscle junctions. Although some of these genes have previously been associated with these functional annotations (such as *BSN* with synaptic terms), others are novel associations detected through the applied gene co-expression analysis (such as *BSN* with long-term memory).

viii. Existing phenotypic asssociations for plausible candidate regions

Previous phenotypic associations were identified in early March 2013 from human and animal web databases (NHGRI's Catalog of Published Genome-Wide Association Studies, http://www.genome.gov/gwastudies; Mouse Genome Informatics, http://www.informatics.jax.org/; The Zebrafish Model Organism Database, http://zfin.org/). All databases were queried for implicated genes listed in Table S20, including alternate/previous gene symbols. From the human GWAS database, findings were considered relevant if a reported significant locus mapped within an implicated gene, or between an implicated gene and another gene, in the current build (b37). From the animal model databases, findings were considered relevant if they suggested neurological or central nervous system alterations caused by polymorphisms, mutations, or knockout models in an implicated gene. Table S22 lists, for each of the genes in Table S20, previously reported associations identified from the human GWAS, zebrafish, or mouse-model databases. This review is not intended as a comprehensive analysis of all phenotypes associated with these genes in humans or model organisms. Rather, we seek to provide an overview of previous findings, highlighting key results that suggest potential mechanisms of genetic influences on educational attainment for future study.

Several notable patterns emerge from previously reported phenotypic associations for genes identified in Table S20. Both the *MDM4-LRRN2* region on chromosome 1 (identified as potential candidates through blood eQTL and functional SNP characterization of the top associated locus tagged by rs11584700) and *STK24* on chromosome 13 (associated with *EduYears* in gene-based tests) have previously reported associations with cognitive phenotypes in humans *(61*, *62)*. In addition, the *GBX2* gene has been robustly demonstrated to affect neural development in both zebrafish *(63)* and mouse models *(64*, *65)*. These previous findings identify these genes as particularly interesting regions for future investigations of cognition-related phenotypes.

The remaining implicated regions show previous associations with basic health and disease phenotypes, primarily body size (including *TET2*, *ITPR3*, and the *ATXN2L-TUFM-SH2B1-ATP2A1* region) and inflammation (including *AFF3*, *BSN-APEH-MST1*, and *ATXN2L-TUFM-SH2B1-ATP2A1*). Further, the signals on chromosome 6p22-21 surround the Major Histocompatibility Complex (MHC), a dense region of genes, many of which are known to affect immune function and have been implicated in a range of psychiatric disorders including schizophrenia *(66)*. This connection suggests candidate regions for investigation of

pleiotropic, causal, or interactive genetic effects that may help inform our understanding of the etiology of the relationship between education and health.

d. Summary and discussion of findings from biological analyses

The supplementary analyses detailed in the previous sections suggest that the meta-analyses of educational attainment phenotypes identify several biologically plausible genomic loci that warrant future investigation. From Table S20 summarizing results of multiple forms of analysis, we note that two loci in particular, marked by rs11584700 at chromosome 1q32 and rs1056667 near the Major Histocompatibility Complex (MHC) on chromosome 6, show converging lines of evidence for association with educational attainment as well as plausible biological function. The region on 1q32 marked by replicated *College*-associated SNP rs11584700 was highlighted by each of the analyses listed in Table S20, showing LD with missense and 3'-UTR variants in *LRRN2*, a blood cis-eQTL signal located in *MDM4*, and significant gene-based association of *PIK3C2B* for the *College* phenotype. Among these genes, the *MDM4-LRRN2* locus has been associated with cognitive performance in humans *(61)*, and *MDM4* is known to be involved in central nervous system development in mouse models *(67)*. The genome-wide significant *EduYears* SNP rs1056667 is located near the gene-rich MHC on chromosome 6, a region that has been robustly shown to affect immune function *(66)*. Genes located in this region—including *LRRC16A*, *HMGN4*, four genes from the histone cluster 1 family, and five genes from the butyrophilin family—appeared in results from each of the analyses listed in Table S20.

Beyond the more standard methods of functional annotation, eQTL analysis, and gene-based tests, perhaps the most compelling biological evidence emerged from two novel, powerful methods for identifying potential biological mechanisms underlying the GWAS findings for educational attainment. Within the combined meta-analysis results, loci tagged by SNPs meeting $p < 1\times10^{-5}$ in the combined meta-analyses showed cell-type-specific overlap with chromatin marks, suggesting cell-type-specific gene expression regulation within the anterior caudate (for both *EduYears* and *College*). The caudate nucleus is located in the basal ganglia, and is strongly implicated in goal-directed behavior *(68)*. Co-expression-based gene-function prediction analysis identified several specific genes previously identified in Table S20 as likely involved in learning, long-term memory, and neuronal function or development pathways (including *GBX2*, *LRRN2*, and *PIK3C2B*, which are located near genome-wide significant loci, as well as *BSN*, which was identified as associated with *EduYears* in the gene-based tests). These genes have several previously reported associations with neural development or cognition-related phenotypes. *LRRN2* has been associated with cognitive performance in humans *(61)*. *GBX2* is known to be involved in anterior hindbrain development in both zebrafish and mouse models *(63, 64, 65)*, as well as being involved in striatal cholinergic interneuron development in mice *(65)*. In addition, *BSN* may influence glutamatergic synapse function in mice *(69)*.

Several of the implicated genes summarized in Table S20 suggest mechanisms potentially related to the well-established health-education gradient. In particular, human GWAS associations have been reported for *BSN* with inflammatory bowel disease (IBD; *(70, 71, 72)*). *MST1 (73, 74, 75, 76)*, *APEH (77)*, and *ATXN2L (78)* also have previously reported associations with various forms of IBD in humans, and *AFF3* has been associated with rheumatoid arthritis *(79)*, suggesting potential pleiotropic or mediation effects between genes related to inflammation and educational attainment. For example, experiencing IBD symptoms may have direct adverse

effects on educational outcomes, such as school attendance and performance *(80*, *81)*, as well as indirect effects mediated by psychosocial adversity, such as increased anxiety or depressive symptoms or family environment stress *(81)*. However, the *BSN* variant often associated with increased risk of IBD, the A allele of rs9858542, was marginally associated with increased levels of educational attainment within the combined-sample meta-analyses ($p_{\text{EduYears}} = 4.1\times10^{-7}$, $p_{\text{College}} = 9.5\times10^{-5}$). This counterintuitive positive association of rs9858542-A to both IBD and educational attainment, coupled with the previously reported negative phenotypic relationship between IBD and education, illustrates the complex relationships that may exist within the apparent genetic overlap between educational attainment and health outcomes. The association suggests *BSN* (along with other education-associated genomic regions previously reported for health phenotypes) as an interesting target for follow-up studies of pleiotropic or mediation effects involved in the etiology of the health-education gradient.

6. Prediction using linear polygenic scores

To investigate how much of the variance in educational attainment is captured by a linear polygenic score (*PGS*) in independent samples, we calculated the effect sizes of SNP *j* (γ_j) using

$$\gamma_j = \hat{\sigma}_{x,j} \cdot z_j ,$$

where (under the assumption of Hardy-Weinberg equilibrium)

$$\hat{\sigma}_{x,j} = \sqrt{2 \cdot MAF_j \cdot \left(1 - MAF_j\right)} .$$

There were 6,654 unrelated European Americans from the ARIC cohort with the SNP data imputed to the HapMap2 CEU panel *(82)*. The QIMR (N = 3,526) and STR (N = 6,770) cohorts were used as the independent validation samples and excluded from the meta-analysis, such that the results of the meta-analysis were based on (All Cohorts minus QIMR) and (All Cohorts minus STR), respectively. The SNP data of these two cohorts were also imputed to the HapMap2 CEU panel.

For the trait *College* (respectively, *EduYears*), in addition to using all available SNPs, we selected 3 (5), 113 (127) and 3,506 (3,369) SNPs that were significant at $p < 5\times10^{-8}$, $p < 5\times10^{-5}$ and $p < 5\times10^{-3}$. We selected the SNPs by a multiple-SNPs association analysis method *(83)* using the summary statistics from the meta-analysis and linkage disequilibrium (LD) between SNPs estimated from the ARIC (Atherosclerosis Risk in Communities Study) cohort. This method implements a step-wise model selection procedure to select all the top associated SNPs, taking LD into account. Therefore, the top associated SNPs selected by this approach are either independently (SNPs are in no LD) or jointly (SNPs are in LD but still have significant effects when fitted together in the model) associated with the phenotype. The advantage of using this approach is that we can identify multiple association signals at a locus, and we do not have to set an arbitrary threshold r^2 value for LD pruning. The effects of all the selected SNPs were re-estimated in a multiple SNP model *(83)*, which yields more accurate estimates for individual SNP effects, similar in spirit to the improvement one gets from adding correlated explanatory variables to a multivariate regression model. The analysis was performed using the GCTA software *(84)*. We then used the PLINK *(27)* profile scoring approach to create a *PGS* from these

selected SNPs for *College* and *EduYears* in a set of unrelated individuals of the QIMR (N = 3,526) and STR (N = 6,770) cohorts. The *PGS* for the i-th individual was calculated as

$$\hat{g}_i = \sum x_{ij}\hat{b}_j \ ,$$

where x_{ij} is the number of copies of the effect allele for SNP j, and $\hat{b}_j$ is the estimated SNP effect from the multiple-SNP analysis. In both QIMR and STR studies, the observed phenotypes of *EduYears* were adjusted for age, sex, and age×sex interaction, and standardized to z-scores. The prediction R^2 was calculated from a linear regression of the z-score for *EduYears* on the *PGS*.

In addition to examining how much of the variance in educational attainment is captured by the *PGS*, we also examined how much of the variance in cognitive function in the STR study *(17)* (N = 1,419) is captured by the *same PGS* (constructed for *College* and *EduYears*). The measures of cognitive function were also adjusted for age, sex, and age×sex interaction, and standardized to z-scores. The basic idea is that if cognitive function is an endophenotype for educational attainment, then the SNPs that are correlated with educational attainment may also be correlated with the cognitive function measure (see section 7 below). The estimated $R^2 \approx 2.5\%$ for cognitive function is slightly larger than the share of variance in educational attainment captured by the score in the STR sample. One possible interpretation is that some of the SNPs used to construct the score matter for education through their stronger, more direct effects on cognitive function (see section 7 below). A mediation analysis (Table S24) provides tentative evidence consistent with this interpretation.

Furthermore, we selected in the QIMR (respectively, STR) cohort the 572 full-sib pairs (2,774 DZ twins) from 572 (2,774) independent families for which we had both phenotype and genotype data available. In these samples, we regressed the difference in observed phenotypes (z-scores) on the difference in *PGS* between the full-sibs. This within-family analysis is unconfounded by possible population stratification. For the within-family analysis in the STR cohorts, only 399 of the 2,774 DZ twins had data on cognitive function available. The PGS for educational attainment constructed using all SNPs remains significant in these within-family analyses (Tables S23 and S25).

a. Projection of how much variance in educational attainment will be explained by a linear polygenic score as a function of the sample size used for estimating the score

(85) derived an approximation of the correlation between the polygenic score estimated in a discovery sample ($\hat{g}$) and its true value (g), which is the value it would attain if estimated in an infinite sample:

$$r^2\left(g,\hat{g}\right) \approx \frac{\lambda h^2}{\lambda h^2 + 1},$$

where h^2 the proportion of additive phenotypic variance captured by all SNPs, and λ^2 is the ratio of the number of individuals in the discovery sample (N) to the number of loci that contribute to the heritability (M), i.e. $\lambda = N/M$. In practice we do not know the number of loci that contribute to educational attainment. However, to make a prediction without knowing the number of loci we can use the effective number of independent segments that are segregating in the population, which is a function of effective population size (N_e) *(86)*. The resulting prediction of variance explained by a predictor is then for the case where all SNPs are used in a best linear

unbiased prediction (BLUP) analysis of a discovery sample with size N that uses individual-level genotype and phenotype data *(87)*.

The proportion of the phenotypic variance explained by a predictor based upon a discovery sample size of N is,

$$\begin{aligned} R^2\left(y,\hat{g}\right) &= R^2\left(g+e,\hat{g}\right) \\ &= \frac{\mathrm{cov}^2\left(g,\hat{g}\right)}{\mathrm{var}(\hat{g})\,\mathrm{var}(y)} \\ &= r^2(g,\hat{g})\frac{\mathrm{var}(g)}{\mathrm{var}(y)} \\ &= r^2(g,\hat{g})h^2 \\ &= \frac{\frac{N}{M}h^4}{\frac{N}{M}h^2+1} \end{aligned}$$

From *(86)*, $M \approx 2N_e kL/\log(2N_e L)$ where k is the number of chromosomes and L = average length of a chromosome. For human data, $k = 22$, $L \approx 1.6$ and $N_e \approx 10{,}000$, which gives an estimate of $M \approx 70{,}000$. In the case of educational attainment, $h^2 \approx 0.2$ (*(3)* and Table S12).

Using these parameters we predict the prediction accuracy for a range of values of N in Table S26 below. The predicted R^2 of 4% for a sample size of N = 100,000 is higher than what we observe with the real data (approximately 2-3%), which could be due to:

- the approximations used for calculating the effective number of loci;
- violation of an assumption underlying approximation *(85)*, the assumption being that the SNPs with a non-zero estimated effect included in the polygenic score are exactly the M SNPs related to educational attainment;
- the use of the estimate of h^2 of 0.2, which is from empirical estimates with a non-trivial standard error; or
- the fact that we performed a prediction analysis on summary statistics, which is less efficient than an analysis on individual-level genotype data.

The theoretical results imply that a GWAS of educational attainment on 1 million individuals will generate a polygenic score that can explain 15% of variance in educational attainment in a new sample. In the future, with denser SNP arrays or whole-sequence data, more phenotypic variance is likely to be accounted for by the genotypes, i.e., h^2 from the included genetic variants will be larger. For example, a value of h^2 of 0.4 results in R^2 values of 0.15, 0.30 and 0.34, for N of 100,000, 500,000 and 1 million, respectively.

7. Identifying genetic associations using biologically distal phenotypes

In this section, we sketch a simple formal framework for considering in general terms how genetic associations with a distal phenotype, such as educational attainment, may be informative regarding genetic associations with mediating traits (endophenotypes) that are more proximal to the direct effects of the genes.

Let Y denote the value of an individual's biologically-distal phenotype, for example, educational attainment. (To avoid cluttering notation, we suppress indexing variables by individual.) We assume that the phenotype is determined by a simple linear function of K genetically-influenced endophenotypes:

$$(1)\quad Y = \sum_{k=1}^{K} \gamma_k M_k + \varepsilon_Y \ ,$$

where M_k is the value of the individual's k^{th} endophenotype, γ_k is the effect of M_k on Y, and ε_Y is a random variable with mean zero that we assume is independent of the M_k's. For educational attainment, these mediating variables include cognitive function, personality traits such as perseverance, early-life health conditions, and many others. The error term ε_Y captures all other factors, including exogenous environmental factors that affect Y. Without loss of generality, we assume each $\gamma_k > 0$. We normalize Y and each M_k so that they have mean zero and variance one (hence regression coefficients are equal to partial correlation coefficients).

Let $j = 1,\ldots,J$ index the SNPs that are causally related to at least one of the mediating factors. We assume that each of the k endophenotypes is a simple linear function of the individual's genotype and determined by:

$$(2)\quad M_k = \sum_{j=1}^{J} \beta_{kj} X_j + \varepsilon_k \ ,$$

where X_j is the individual's genotype at SNP j (again, normalized to have mean zero and variance one), β_{kj} is the effect of X_j on M_k (which could be positive or negative, or possibly equal to 0 for particular SNP-j/endophenotype-k combinations), and ε_k is a random variable with mean zero that we assume is independent of the X_j's. The error term ε_k captures all other factors, including exogenous environmental factors that affect M_k.

The distal phenotype Y can be expressed as a function of the SNP genotypes by substituting equation (2) into equation (1):

$$(3)\quad Y = \sum_{j=1}^{J}\left(\sum_{k=1}^{K} \gamma_k \beta_{kj} \right) X_j + \left(\varepsilon_Y + \sum_{k=1}^{K} \gamma_k \varepsilon_k \right) = \sum_{j=1}^{J} \delta_j X_j + u_Y \ ,$$

where $\delta_j \equiv \sum_{k=1}^{K} \gamma_k \beta_{kj}$ is the effect of SNP j on the distal phenotype j (aggregated over all the mediating pathways), and $u_Y \equiv \left(\varepsilon_Y + \sum_{k=1}^{K} \gamma_k \varepsilon_k \right)$ is a mean-zero composite error term that is independent of the X_j's.

A GWAS of the distal phenotype Y estimates the δ_j's in equation (3).

An immediate implication of our theoretical framework is that if $\delta_j \neq 0$, then $\beta_{kj} \neq 0$ for at least one $k \in \{1,2,\ldots,K\}$. Hence if the GWAS of the distal phenotype Y credibly identifies a SNP, then that SNP can serve as a plausible "candidate gene" for the mediating traits.[1]

But if one is interested in identifying SNPs associated with a mediating trait, why not directly run a GWAS of the trait of interest?

[1] The converse does not hold: it is possible that $\beta_{k,j} \neq 0$ for some k, but $\delta_j = 0$. Thus, a GWAS on the distal phenotype may not identify all the SNPs related to an endophenotype of interest.

To understand the relevant tradeoffs within the context of our framework, denote by N_{distal} the sample size available for the distal GWAS and by N_k the sample size available for a direct GWAS of some endophenotype k. We will argue that if the distal phenotype is available in much larger samples, then a GWAS of Y may be better powered than a GWAS of M_k to identify SNPs associated with M_k.

The statistical power of a GWAS to detect the association of a phenotype with a SNP j is determined by the sample size of the study (N) and the population R^2 from the regression of the phenotype on SNP j ($R^2_{\text{phenotype,SNP}}$). Specifically, power is a function of the non-centrality-parameter (NCP) of a χ^2 test of association, which is approximately equal to $NR^2_{\text{phenotype,SNP}}$. Therefore, the GWAS on the distal phenotype in a sample of size N_{distal} will have approximately the same power as a GWAS on endophenotype k in a sample of size N_k if $N_{\text{distal}} R^2_{Y,j} = N_k R^2_{k,j}$. Since $R^2_{k,j} = \beta^2_{k,j}$ and $R^2_{Y,j} = \delta^2_j = \left(\sum_{k=1}^{K} \gamma_k \beta_{k,j} \right)^2$, the condition for equal power is $N_{\text{distal}} = N_k \dfrac{\beta^2_{k,j}}{\delta^2_j}$.

In theory, it is possible that δ^2_j exceeds $\beta^2_{k,j}$—that is, the SNP is *more* strongly associated with the distal phenotype than with the endophenotype—so that the GWAS on the distal phenotype has greater statistical power than the GWAS on the endophenotype even if $N_{\text{distal}} = N_k$. This situation can only arise if SNP j is not only associated with endophenotype k, but even more strongly associated with a weighted sum of the other endophenotypes $\sum_{k' \neq k} \gamma_{k'} M_{k'}$. We suspect that in practice, most SNPs detected in a GWAS study of a distal outcome do not have this property.

Another possibility is that a SNP operates mainly through one endophenotype. If SNP j affects the distal phenotype through exactly one channel, say endophenotype k, then $\delta^2_j = \gamma^2_k \beta^2_{k,j}$. This quantity is necessarily smaller than $\beta^2_{k,j}$, and hence the GWAS on the endophenotype has greater statistical power than the GWAS on the distal phenotype if $N_{\text{distal}} = N_k$. In the analysis that follows, we assume that $\delta^2_j = \gamma^2_k \beta^2_{k,j}$ in order to quantify the tradeoff between sample size and biological proximity.

Notice, however, that if this simplifying assumption does not hold, the calculations we report below will often be conservative in the sense of understating the benefits of the GWAS on the distal phenotype. To understand why, suppose that SNP j impacts the distal phenotype through multiple endophenotypes. We define a SNP as "mono-directional" for endophenotype k if its correlation with the distal phenotype has the same sign as its correlation with the weighted average of the other endophenotypes, $\sum_{k' \neq k} \gamma_{k'} M_{k'}$; and we define a SNP as "bi-directional" for endophenotype k otherwise. If a SNP is mono-directional, then $\delta^2_j > \gamma^2_k \beta^2_{k,j}$, making the SNP easier to detect than our calculations assume. In contrast, if the SNP is bi-directional, then it will typically be the case that $\delta^2_j < \gamma^2_k \beta^2_{k,j}$, and the calculations below are likely to overstate the benefits of the GWAS on the distal

outcome.[2] However, the SNPs most likely to be discovered by a GWAS on the distal phenotype are those that have the strongest association with the distal phenotype. Therefore, we expect that in practice, the SNPs discovered by a GWAS on the distal phenotype will generally be the mono-directional SNPs.

Assuming that $\delta_j^2 = \gamma_k^2 \beta_{k,j}^2$, the condition for equal power simplifies to $N_{\text{distal}} = N_k / \gamma_k^2$. To be concrete, suppose that the correlation between the distal outcome and the endophenotype is $\gamma_k = 0.5$; for comparison, the correlation between *EduYears* and cognitive function is 0.46 in the STR data analyzed in the prediction exercise of section 6 above. Then a GWAS on the distal phenotype with 100,000 individuals has the same power as a GWAS on the endophenotype with 25,000 individuals. If obtaining high-quality measurement of the endophenotype is costly or time-consuming, it may be more difficult to obtain a sample size of N_k for the endophenotype than a sample size of N_{distal} for the distal phenotype.

Figure S22 illustrates explicit power calculations. Each of the three curves graphs a locus of effect-size/sample-size pairs that gives a fixed level of statistical power—80%, 50%, and 16%, respectively—to detect an association with a SNP at $p = 5\times10^{-8}$. Effect size is measured in terms of the R^2 from a population regression of the phenotype on the SNP.

Our findings imply that for the SNPs with the largest associations with educational attainment, the R^2 from a population regression of educational attainment on SNP j is approximately 0.02% (that is, 0.0002). The black curve shows that our sample size of approximately $N_{\text{distal}} = 100{,}000$ individuals had about 16% power to detect a SNP of this effect size. Assuming as above that the SNP affects educational attainment only through its effect on a single endophenotype k and that $\gamma_k = 0.5$, the R^2 from a population regression of the endophenotype on SNP j is $\beta_{k,j}^2 = 0.0002/0.5^2 = 0.0008$. Consistent with the analytical derivation above, the black curve shows that for 16% power to detect this effect, a sample size of approximately $N_k = 25{,}000$ individuals is needed. The red and green curves numerically illustrate the tradeoff at other levels of statistical power.

We conclude this section by analyzing how well a polygenic score constructed to predict the distal outcome can be expected to predict the endophenotype.

The best possible polygenic score for the purpose of predicting endophenotype k is the deterministic component of the population regression equation (2): $\text{PGS}_k = \sum_{j=1}^{J} \beta_{k,j} X_j$. When estimated using GWAS results on endophenotype k in a sample of size N_k, the polygenic score is $\hat{\text{PGS}}_{k|N_k} = \sum_{j=1}^{J} \hat{\beta}_{k,j} X_j$. The polygenic score estimated using GWAS results on the distal phenotype in a sample of size N_{distal} is the estimated deterministic component of the population regression equation (3): $\hat{\text{PGS}}_{\text{distal}|N_{\text{distal}}} = \sum_{j=1}^{J} \hat{\delta}_j X_j$.

[2] In principle, even if a SNP is bi-directional for endophenotype k, it is possible that $\delta_j^2 > \beta_{k,j}^2$. This condition means that the SNP's effect on the distal trait is larger in magnitude than the SNP's effect on the endophenotype. However, for a bi-directional SNP, where the SNP's effect on the distal trait is in the opposite direction of the SNP's effect on the endophenotype of interest, we think this possibility is unlikely in most realistic settings.

Is $\hat{\text{PGS}}_{k|N_k}$ or $\hat{\text{PGS}}_{\text{distal}|N_{\text{distal}}}$ expected to be a better estimate of PGS_k? There are two countervailing effects. First and more obviously, the $\hat{\beta}_{k,j}$ weights used in constructing $\hat{\text{PGS}}_{k|N_k}$ are unbiased estimates of the optimal weights $\beta_{k,j}$, while in general the $\hat{\delta}_j$ weights used in constructing $\hat{\text{PGS}}_{\text{distal}|N_{\text{distal}}}$ are not unbiased estimates of the weights $\gamma_k\beta_{k,j}$. (The $\hat{\delta}_j$ weights are unbiased estimates of $\delta_j = \sum_{k=1}^{K}\gamma_k\beta_{k,j}$, but these reflect all the mediating pathways, not just the pathway involving endophenotype *k*.) This first effect favors $\hat{\text{PGS}}_{k|N_k}$ over $\hat{\text{PGS}}_{\text{distal}|N_{\text{distal}}}$.

However, there is a second effect. For reasons discussed above, if $N_{\text{distal}} > N_k$, then the $\hat{\delta}_j$'s may be estimated more precisely than the $\hat{\beta}_{k,j}$'s. If on average the $\hat{\delta}_j$'s are not too different from the $\beta_{k,j}$'s multiplied by the constant γ_k, then it may be the case that the $\hat{\delta}_j$'s end up as better estimates of the $\gamma_k\beta_{k,j}$'s than the $\hat{\beta}_{k,j}$'s are estimates of the $\beta_{k,j}$'s. Since γ_k is an irrelevant multiplicative constant—that is, $\sum_{j=1}^{J}\hat{\delta}_j X_j$ predicts exactly the same amount of variance in the endophenotype as $\sum_{j=1}^{J}\hat{\delta}_j X_j \big/ \gamma_k$ does—this second effect favors $\hat{\text{PGS}}_{\text{distal}|N_{\text{distal}}}$ over $\hat{\text{PGS}}_{k|N_k}$ (see *(85),(88)* for a derivation of the effect of sample size on the predictive power of an estimated linear polygenic score).

In the case of educational attainment and cognitive function, the second effect appears to dominate given currently available sample sizes. A recent paper on childhood intelligence *(20)* constructed a predictive score ($\hat{\text{PGS}}_{k|N_k}$) from GWAS findings of ~18,000 individuals and tested the score for association with childhood intelligence in three independent replication samples. The score explained 1.2%, 3.5% and 0.5% of variance in the three replication samples, corresponding to a sample-size weighted average of 1.08%. This is smaller than the approximately 2.5% of variance in cognitive function explained by the predictive score we constructed for educational attainment ($\hat{\text{PGS}}_{\text{distal}|N_{\text{distal}}}$).

In the context of our analysis, how is it possible that the polygenic score constructed to predict educational attainment does a better job predicting cognitive function than it does predicting educational attainment? There is a general reason why a polygenic score will tend to predict an endophenotype better than it predicts the distal phenotype: the error term in population regression equation (2) for endophenotype *k*, ε_k, has smaller variance than the composite error term in population regression equation (3) for the distal phenotype, $\varepsilon_Y + \sum_{k'=1}^{K}\gamma_{k'}\varepsilon_{k'}$.

The magnitude of the difference in predictive power is easiest to derive analytically in the extreme case where $\delta_j = \sum_{k'=1}^{K}\gamma_{k'}\beta_{k',j} \approx \gamma_k\beta_{k,j}$. This approximation would hold, for example, if $\gamma_k >> \gamma_{k'}$ for all $k' \neq k$ (the

mediating pathway k is especially important relative to other pathways, for example if k is the only mediating pathway) or if $\sum_{k' \neq k} \gamma_{k'} \beta_{k',j} \approx 0$ (the other mediating pathways cancel each other out). If $\delta_j = \gamma_k \beta_{k,j}$, then the best possible polygenic score for the distal outcome is identical to the best possible polygenic score for the endophenotype. However, the population R^2 of the polygenic score for predicting the endophenotype will be $\beta_{k,j}^2$, which is larger than the population R^2 of the polygenic score for predicting the distal phenotype, $\left(\gamma_k \beta_{k,j}\right)^2$.

8. Using a polygenic score as a control variable in a randomized experiment

In this section, we calibrate the gains in statistical power that may be afforded by using a polygenic score as a control variable in the context of a simple randomized experiment.

For concreteness, consider an experiment designed to estimate the effect of providing financial incentives for school attendance on educational attainment. Although our specific example is hypothetical, there are important real-world experiments testing how incentives matter for student achievement, e.g. *(89)*. Similarly, there are experiments testing how student achievement is affected by pre-school programs (e.g., *(90)*) or pre-birth interventions (e.g., *(91)*). In these and other examples, the intervention is expensive, and obtaining adequate statistical power from a relatively small sample size is a paramount challenge.

Let Y denote the level of an individual's educational attainment. (To avoid cluttering notation, we suppress indexing variables by individual.) Let N_X denote the number of experimental participants. Suppose proportion p of the participants are randomly assigned to the treatment group in which financial incentives are provided, and proportion 1–p of the participants are randomly assigned to the control group. The treatment effect, τ, is estimated by running the regression:

$$(1)\quad Y = \alpha + \sum_{j=1}^{J} \beta_j X_j + \tau I + \varepsilon,$$

where the X_j's are the values of J variables correlated with Y (such as sex, personality traits, and parental income), $I \in \{0,1\}$ is an indicator variable for assignment to control or treatment group, and the mean-zero error term ε captures all other factors that affect Y. We denote the variance of ε by σ^2.

Because I is randomly assigned, it is independent of the X_j's and of ε. Therefore, the treatment effect coefficient, $\hat{\tau}_X$, is an unbiased estimate of the true treatment effect τ, whether or not the X_j's are included in the regression. The standard error of $\hat{\tau}_X$, however, will tend to be smaller the stronger the predictive power of the X_j's for the outcome Y. To be precise, let R_X^2 denote the R^2 from the population regression of Y on X_1, X_2,…,X_J. The standard error of $\hat{\tau}_X$ is expected to be approximately equal to $\sqrt{\frac{\sigma^2}{N_X p(1-p)}\left(1 - R_X^2\right)}$.

Now suppose a polygenic score for educational attainment, PGS, is available for each individual in the experiment. Because the cost of genotyping is falling very rapidly, we anticipate that in a few years, it will be very inexpensive to collect genotypic data on experimental participants. Moreover, as long as the experimental participants are still alive, it is possible to collect their genotypic data and use a PGS as a control variable for re-analyzing experiments that may have been conducted many years in the past.

Now, suppose an otherwise-identical experiment with $N_{X\cup\text{PGS}}$ experimental participants is run, and the treatment effect, τ, is estimated by running the regression:

(2) $Y = \alpha + \sum_{j=1}^{J}\beta_j X_j + \gamma(\text{PGS}) + \tau I + u$.

Because I is randomly assigned, it is independent of the X_j's, of PGS, and of ε. As before, the treatment effect coefficient, $\hat{\tau}_{X\cup\text{PGS}}$, is an unbiased estimate of the true treatment effect τ. Now, however, the standard error of $\hat{\tau}_{X\cup\text{PGS}}$ is expected to be approximately equal to $\sqrt{\frac{\sigma^2}{N_{X\cup\text{PGS}}p(1-p)}\left(1-R^2_{X\cup\text{PGS}}\right)}$, where $R^2_{X\cup\text{PGS}}$ is the R^2 from the population regression of Y on X_1, X_2,…,X_J and PGS.

If PGS has predictive power for Y conditional on the X_j's, then $R^2_{X\cup\text{PGS}}$ is larger than R^2_X. Hence estimating regression (2) is expected to generate a smaller standard error—i.e., have greater statistical power—than estimating regression (1). To quantify the gain in statistical power, we solve for how much smaller the experimental sample size needs to be when regression (2) is used instead of regression (1) to generate the same anticipated standard error:

(3) $\frac{N_{X\cup\text{PGS}}}{N_X} = \frac{1-R^2_{X\cup\text{PGS}}}{1-R^2_X}$.

Table S27 calibrates this reduction in required sample size for a range of values of R^2_X and $R^2_{X\cup\text{PGS}}$.

The left and right panels examine the gain in power for experiments where the other control variables, the X_j's, jointly explain 10% and 20% of the variance in educational attainment, respectively.

The first value for $R^2_{X\cup\text{PGS}}$, which is 2 percentage points higher than R^2_X, corresponds to the joint explanatory power of the control variables when the PGS we have estimated in this paper is added the X_j's, assuming that the variance it captures does not overlap with the variance captured by the X_j's. In both panels, a 2% smaller sample size is required for any given level of statistical power. This calculation shows that at presently attainable levels of predictive power, the value of the score is almost certainly too low to result in cost-savings that pass the cost-benefit test.

The second and third values for $R^2_{X\cup\text{PGS}}$ are 12 and 15 percentage points higher than R^2_X, respectively. We explore these values because 12% and 15% correspond to the projected explanatory power that a polygenic score for educational attainment would attain if estimated in discovery samples of 500,000 or 1,000,000 individuals, respectively (see section 6 above). The left panel shows that when the other control variables have an R^2 of 10%, the respective reductions in sample size are 13% and 17%. These reductions can represent quite a

substantial savings in experimental cost in those instances where the intervention is very costly. For an example of the potential costs, *(92)* estimated the total undiscounted initial program cost of the Abecedarian program for five years as $76,939 (in 2010 dollars) per child. And *(93)* estimate that the two-year Perry preschool program cost a total of $19,208.61 (in 2010 dollars) per child. On the other hand, if the intervention is cheap, it is conceivable that even at these higher levels of explanatory power, the cost of genotyping dominates the cost saving obtained from a marginal reduction in sample size.

In interpreting the estimates we report here, it is important to remember that the predictive power of the score may vary across ethnic groups. If our score—which is estimated using a sample of Caucasians—has lower predictive power in non-Caucasians, then the benefits of using it as a control variable in a study of non-Caucasians will be smaller.

9. Data on cognitive function in STR

During the study period all Swedish men were required by law to participate in military conscription at or around the age of 18. The men in the STR sample enlisted at a point in time when exemptions from military duty were rare and typically only granted to men who could document a serious handicap that would make it impossible to complete training. The conscription procedure involved several medical and psychological examinations. We use data on performance on the Swedish Enlistment Battery, a test similar to the US Armed Forces Qualifying Test. See *(17)* for a detailed description. Most of the recruits took four subtests (logical, verbal, spatial and technical) which, for most of the study period, were graded on a scale from 0 to 40. Data are available for men born after 1936. To construct the final score, the four raw scores are summed, percentile-rank transformed, and convoluted with the inverse of the standard normal distribution. This procedure ensures that the final test scores are normally distributed. The construction of the final score is performed separately for each birth year in order to take into account small, occasional, year-to-year changes in the test.

10. Supplementary Figures

Figure S1. *EduYears* distribution in Rotterdam Study I

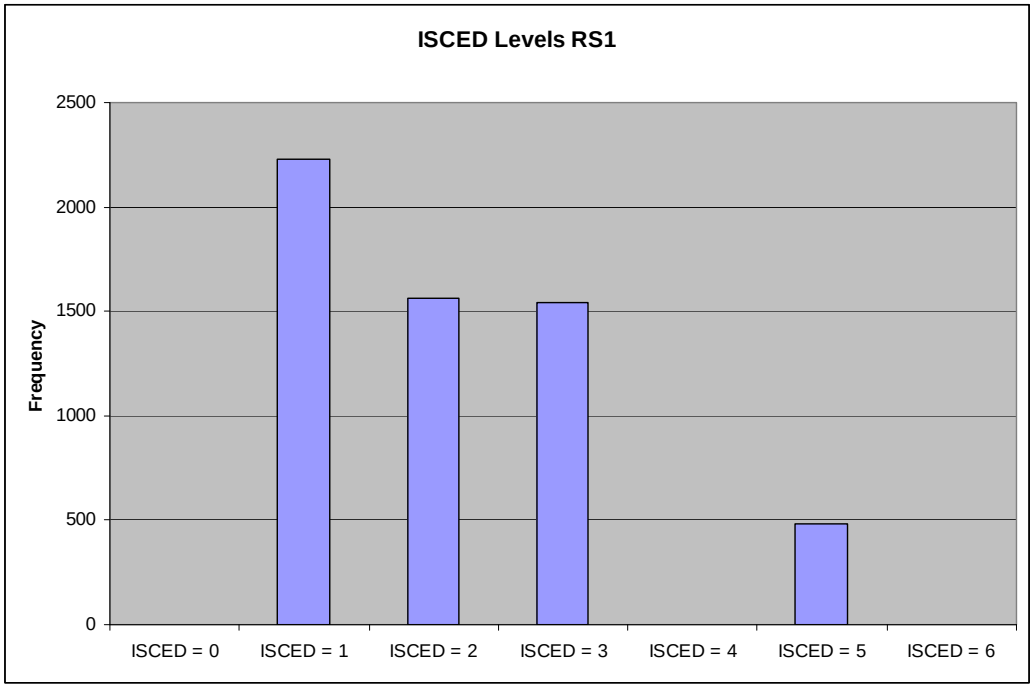
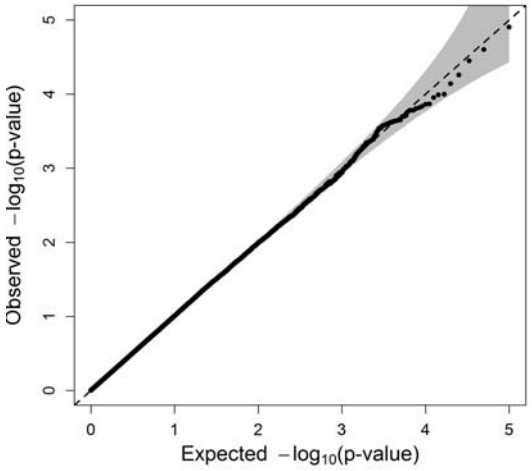
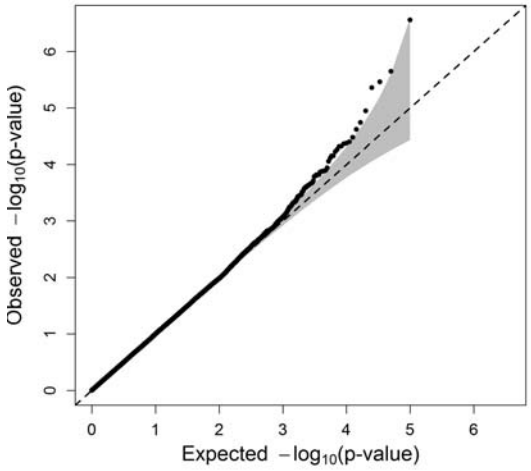


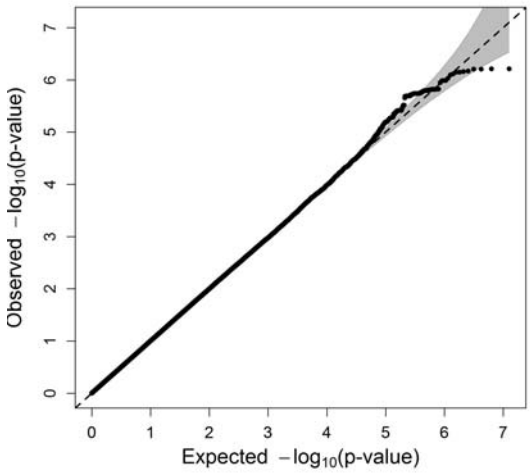
Figure S2. *P*-value distribution from simulation experiments. Panels A and B provide the Q-Q plots for the common and the rare SNP in simulation experiment 1, respectively. Panel C gives the Q-Q plot of simulation experiment 2. The gray shaded areas in the Q–Q plots represent the 95% confidence bands around the *p*-values under the null hypothesis.



A



B



C

Figure S3. Quantile-quantile plots of SNPs for *EduYears* in single genomic control meta-analysis. The gray shaded areas in the Q–Q plots represent the 95% confidence bands around the *p*-values under the null hypothesis.

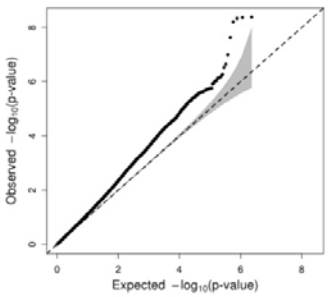
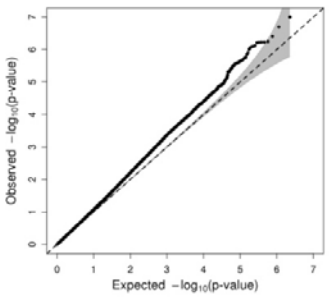
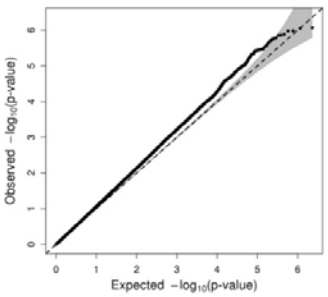
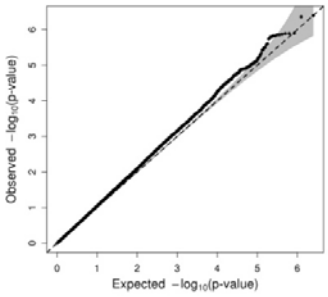
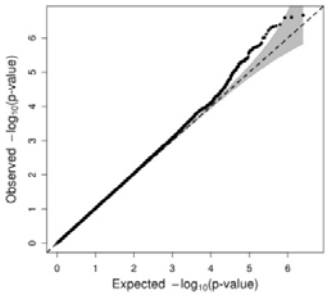
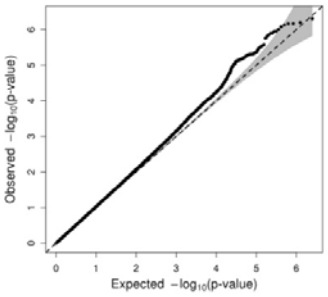
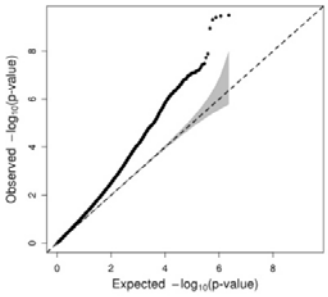
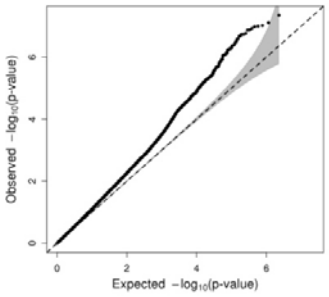
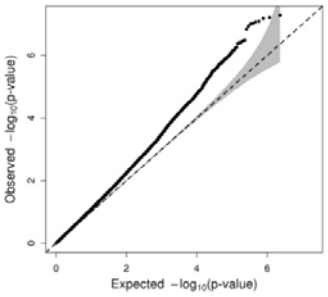
	Pooled	Males	Females
Discovery Stage			
Replication Stage			
Combined Stage			

Figure S4. Quantile-quantile plots of SNPs for *College* in single genomic control meta-analysis. The gray shaded areas in the Q–Q plots represent the 95% confidence bands around the *p*-values under the null hypothesis.

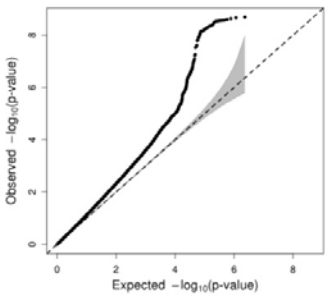
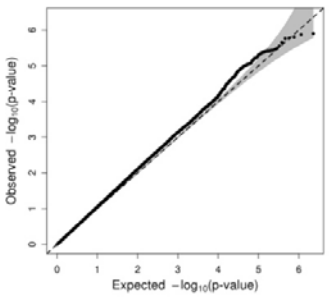
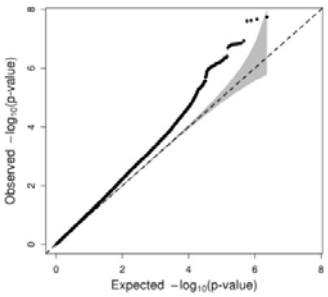
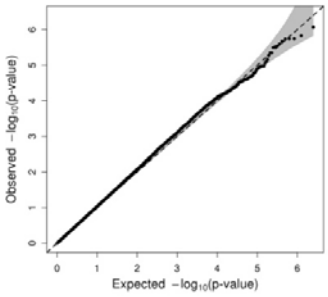
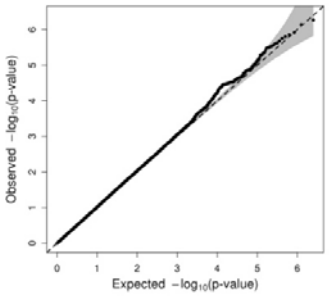
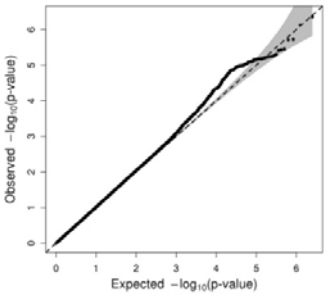
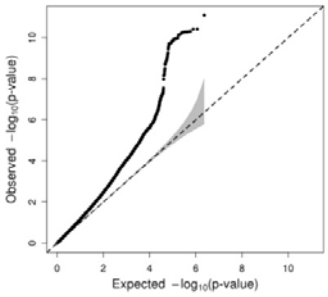
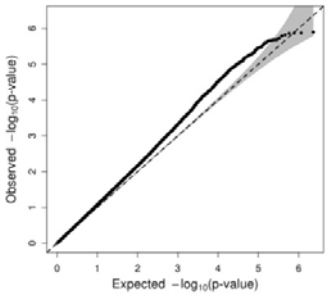
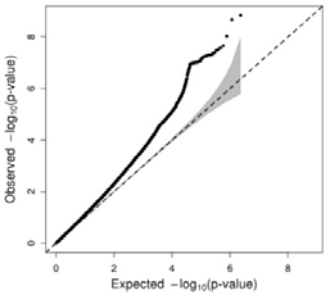
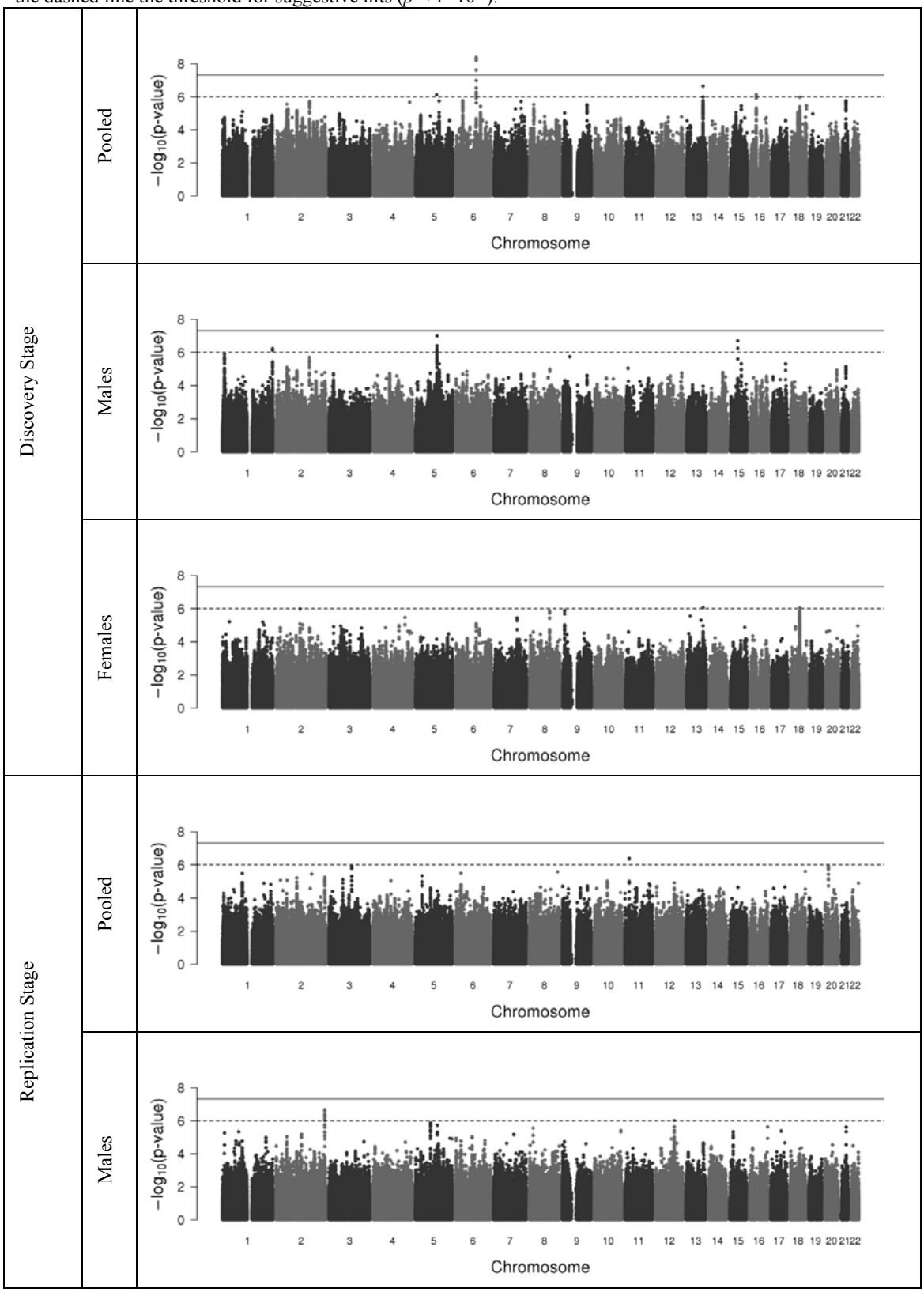
	Pooled	Males	Females
Discovery Stage			
Replication Stage			
Combined Stage			

Figure S5. Manhattan plots of SNPs for *EduYears* in single genomic control meta-analysis. SNPs are plotted on the x-axis according to their position on each chromosome against association with *EduYears* on the y-axis (shown as $-\log_{10}$ *p*-value). The solid line indicates the threshold for genome-wide significance ($p < 5\times10^{-8}$) and the dashed line the threshold for suggestive hits ($p < 1\times10^{-6}$).



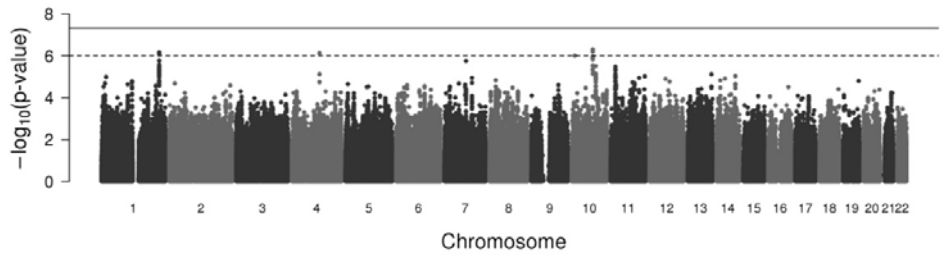
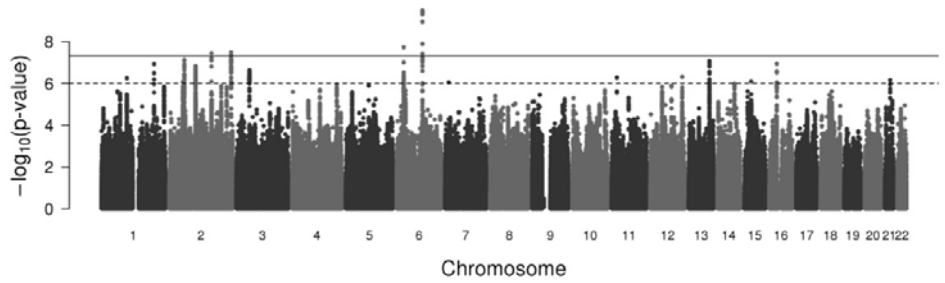
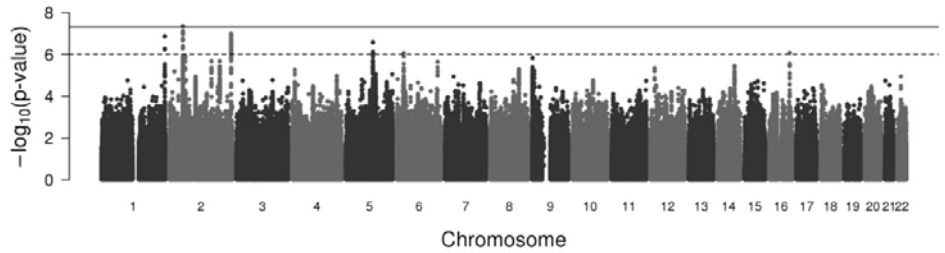
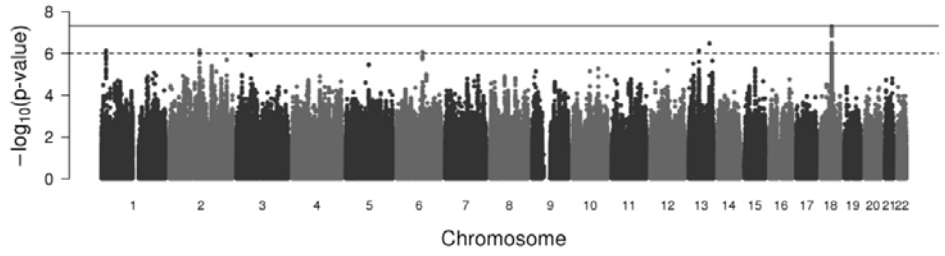
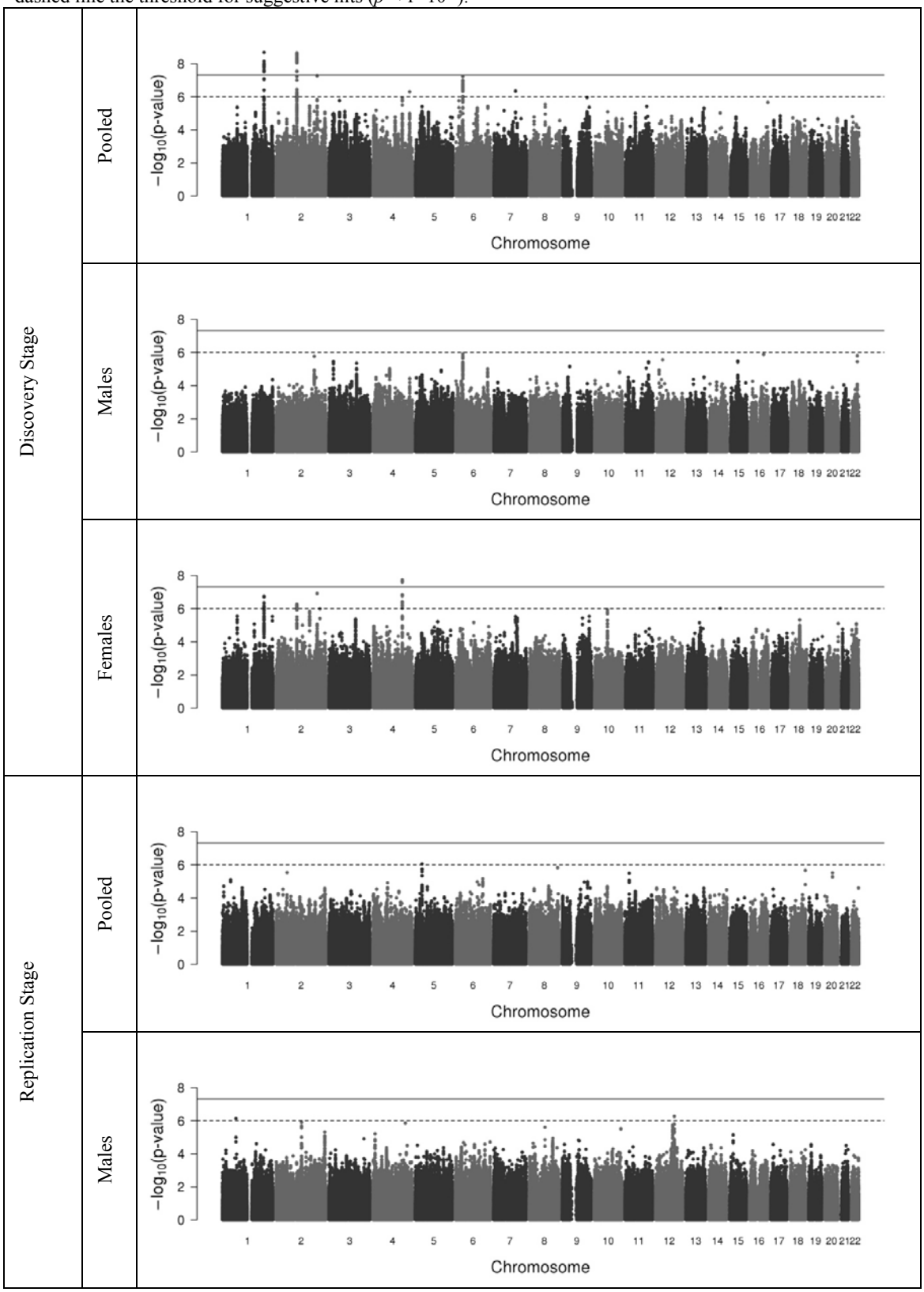
	Females	
Combined Stage	Pooled	
	Males	
	Females	

Figure S6. Manhattan plots of SNPs for *College* in single genomic control meta-analysis. SNPs are plotted on the *x*-axis according to their position on each chromosome against association with *College* on the *y*-axis (shown as $-\log_{10}$ *p*-value). The solid line indicates the threshold for genome-wide significance ($p < 5\times10^{-8}$) and the dashed line the threshold for suggestive hits ($p < 1\times10^{-6}$).



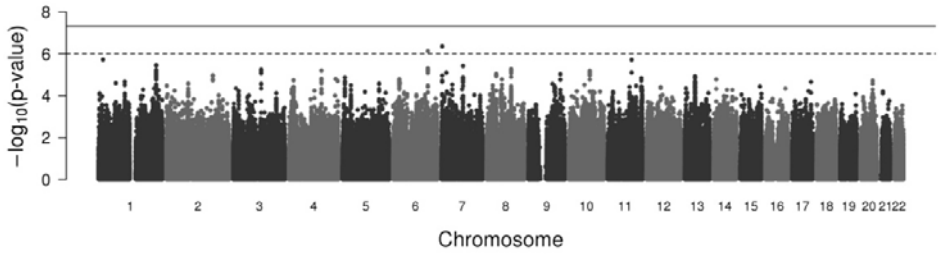
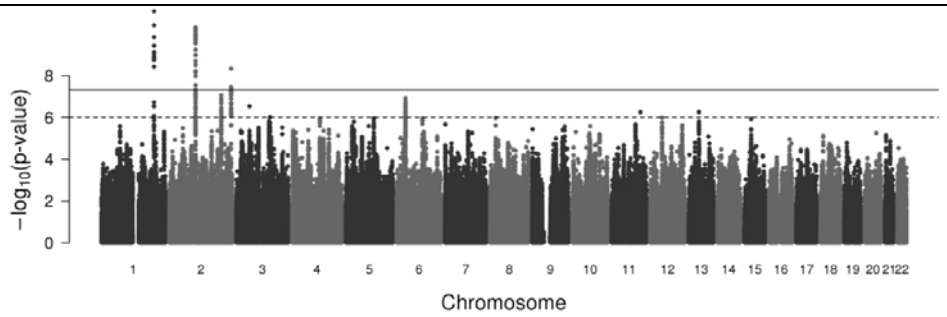
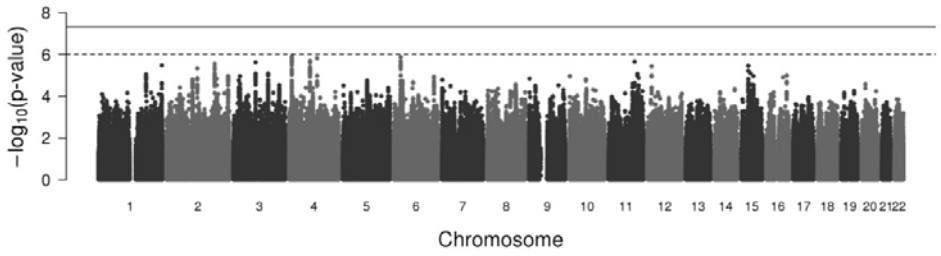
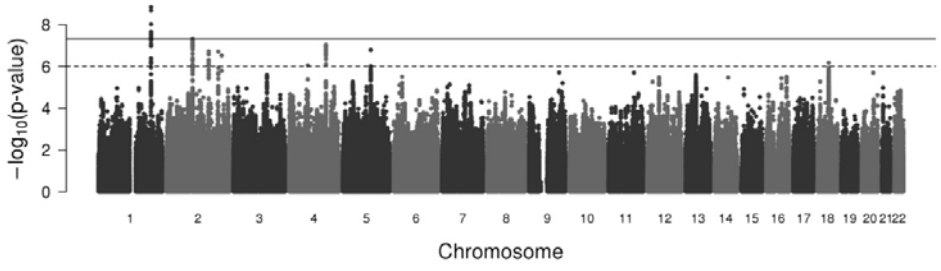
	Females	
Combined Stage	Pooled	
	Males	
	Females	

Figure S7. Forest plot for rs9320913, a genome-wide significant SNP for *EduYears* in the combined-stage GWAS. The gray lines represent the 95% confidence interval of the effect size estimate. The black rectangles are proportional to the square-root of the sample size.

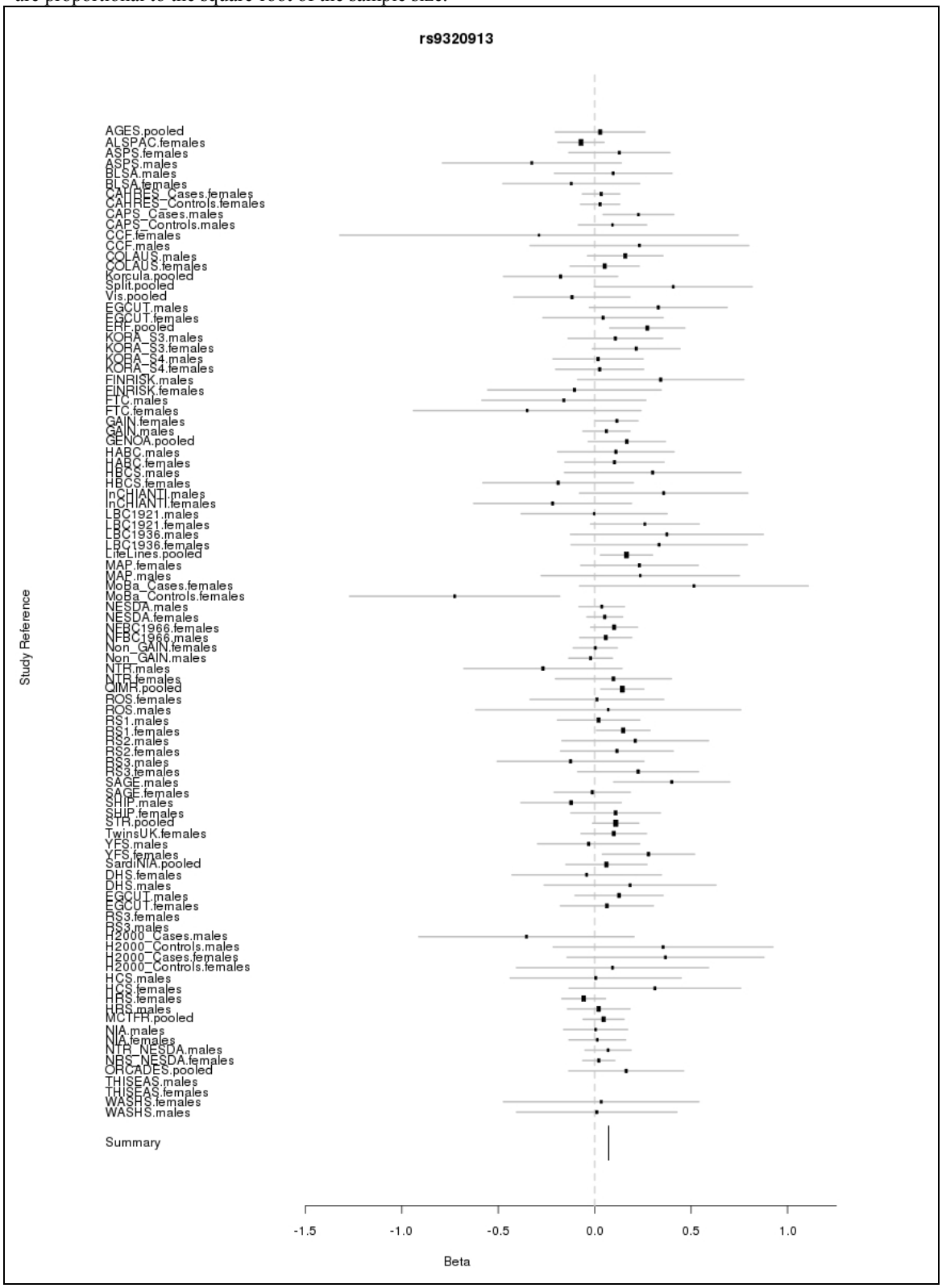


Figure S8. Forest plot for rs7309 that is genome-wide significant for *EduYears* in the combined discovery + replication stage. The gray lines represent the 95% confidence interval of the effect size estimate. The black rectangles are proportional to the square-root of the sample size.

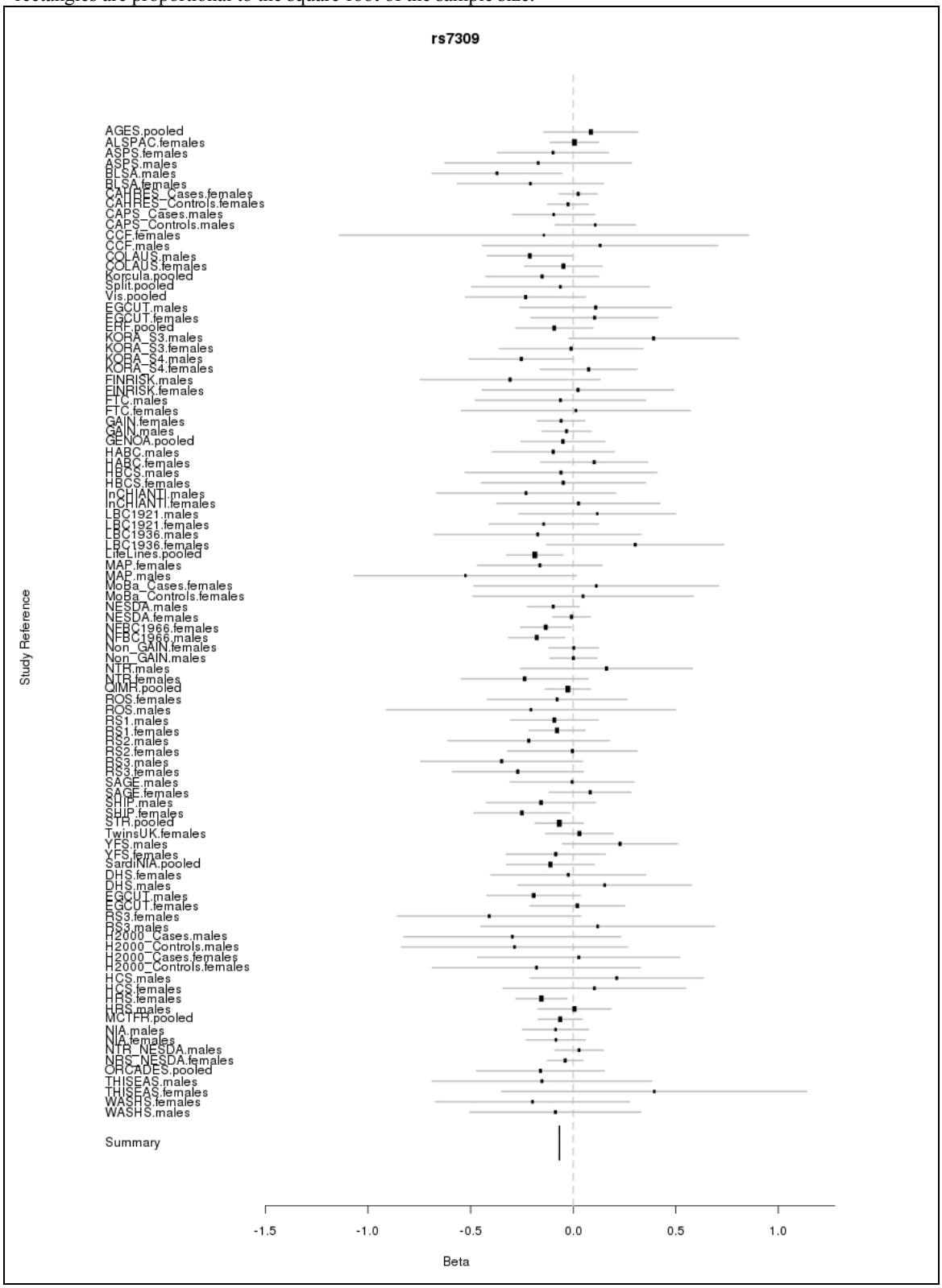


Figure S9. Forest plot for rs11687170 that is genome-wide significant for *EduYears* in the combined discovery + replication stage. The gray lines represent the 95% confidence interval of the effect size estimate. The black rectangles are proportional to the square-root of the sample size.

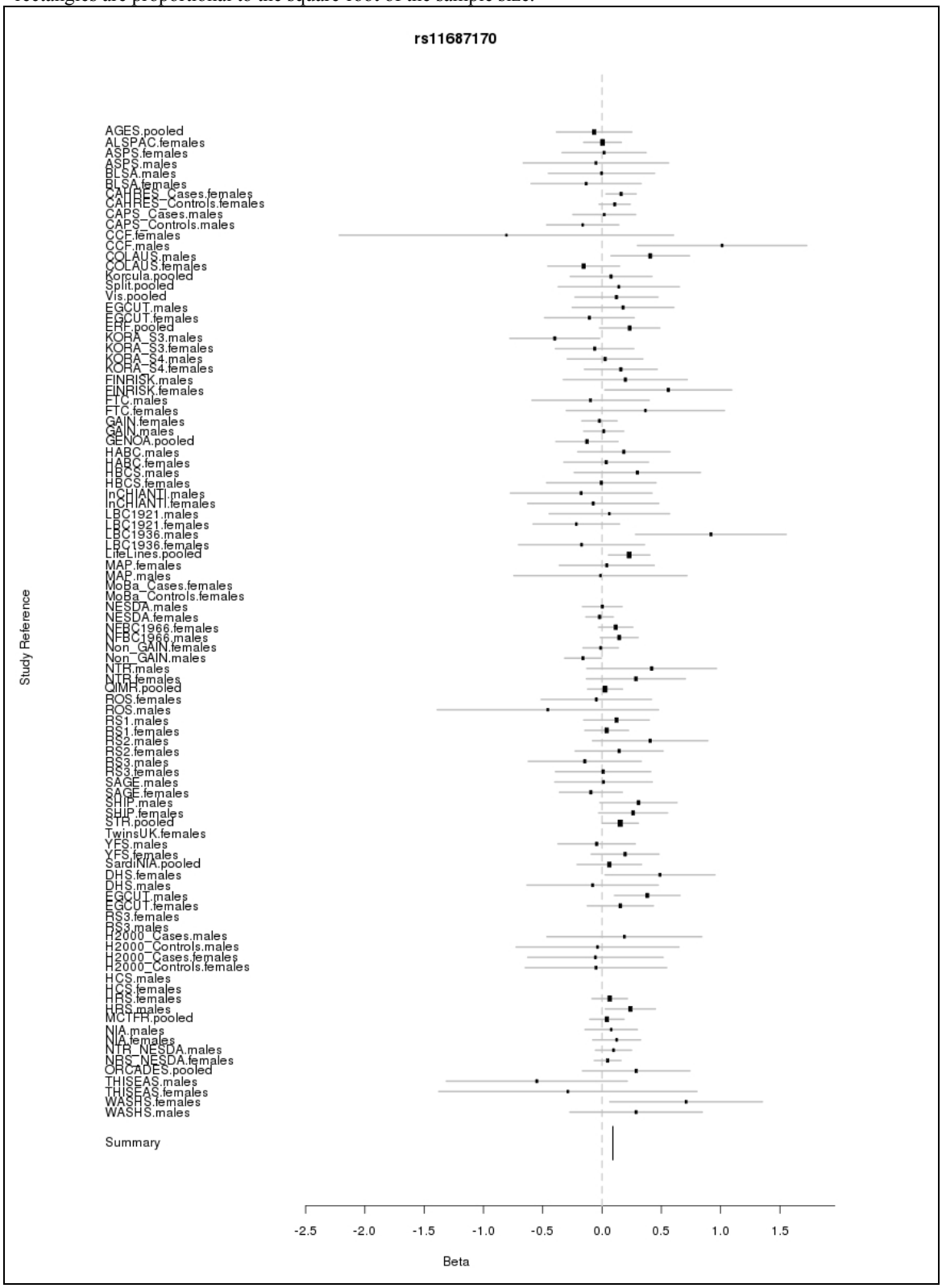


Figure S10. Forest plot for rs1056667 that is genome-wide significant for *EduYears* in the combined discovery + replication stage. The gray lines represent the 95% confidence interval of the effect size estimate. The black rectangles are proportional to the square-root of the sample size.

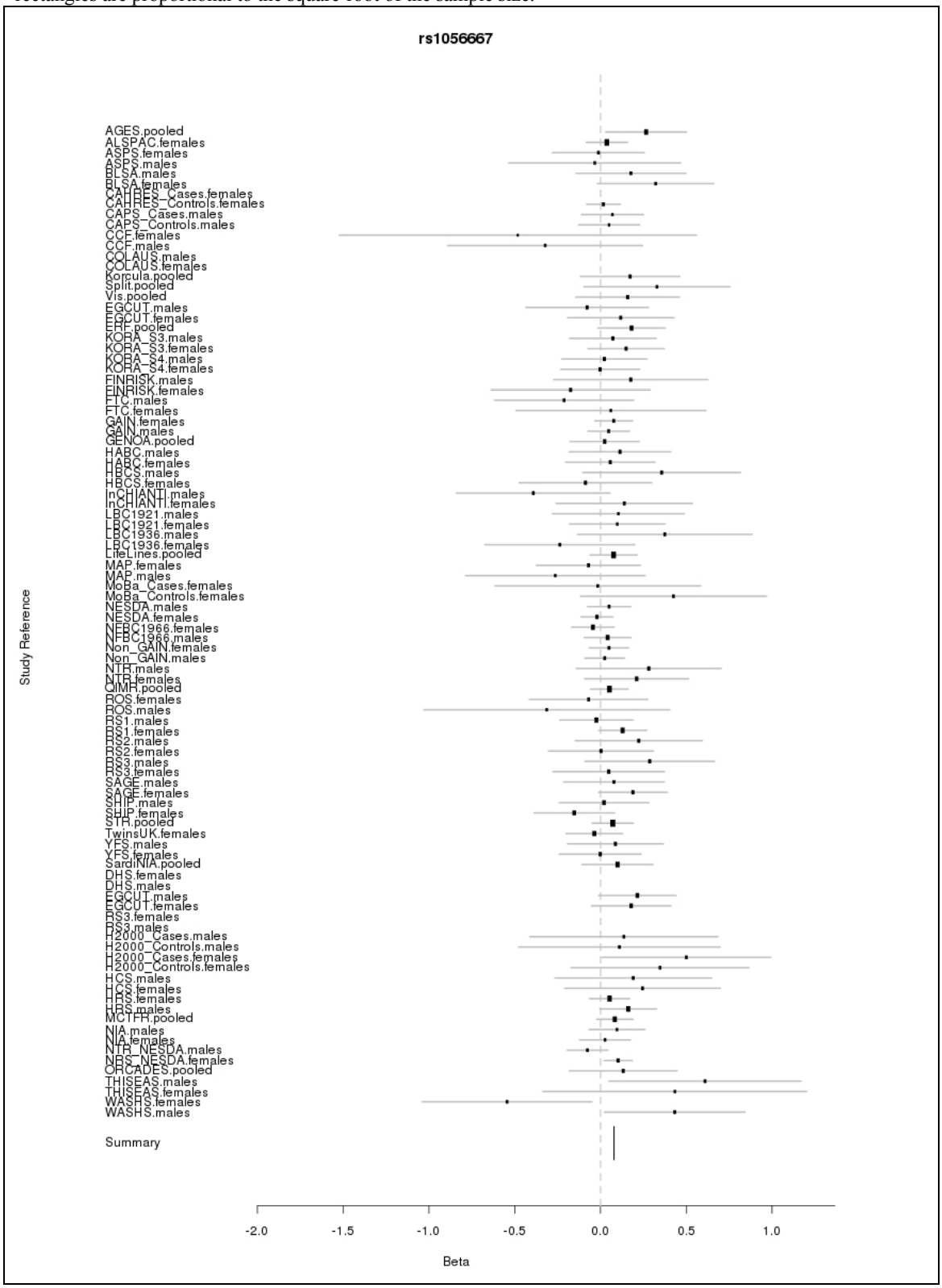


Figure S11. Forest plot for rs1487441 that is genome-wide significant for *EduYears* in the combined discovery + replication stage. The gray lines represent the 95% confidence interval of the effect size estimate. The black rectangles are proportional to the square-root of the sample size.

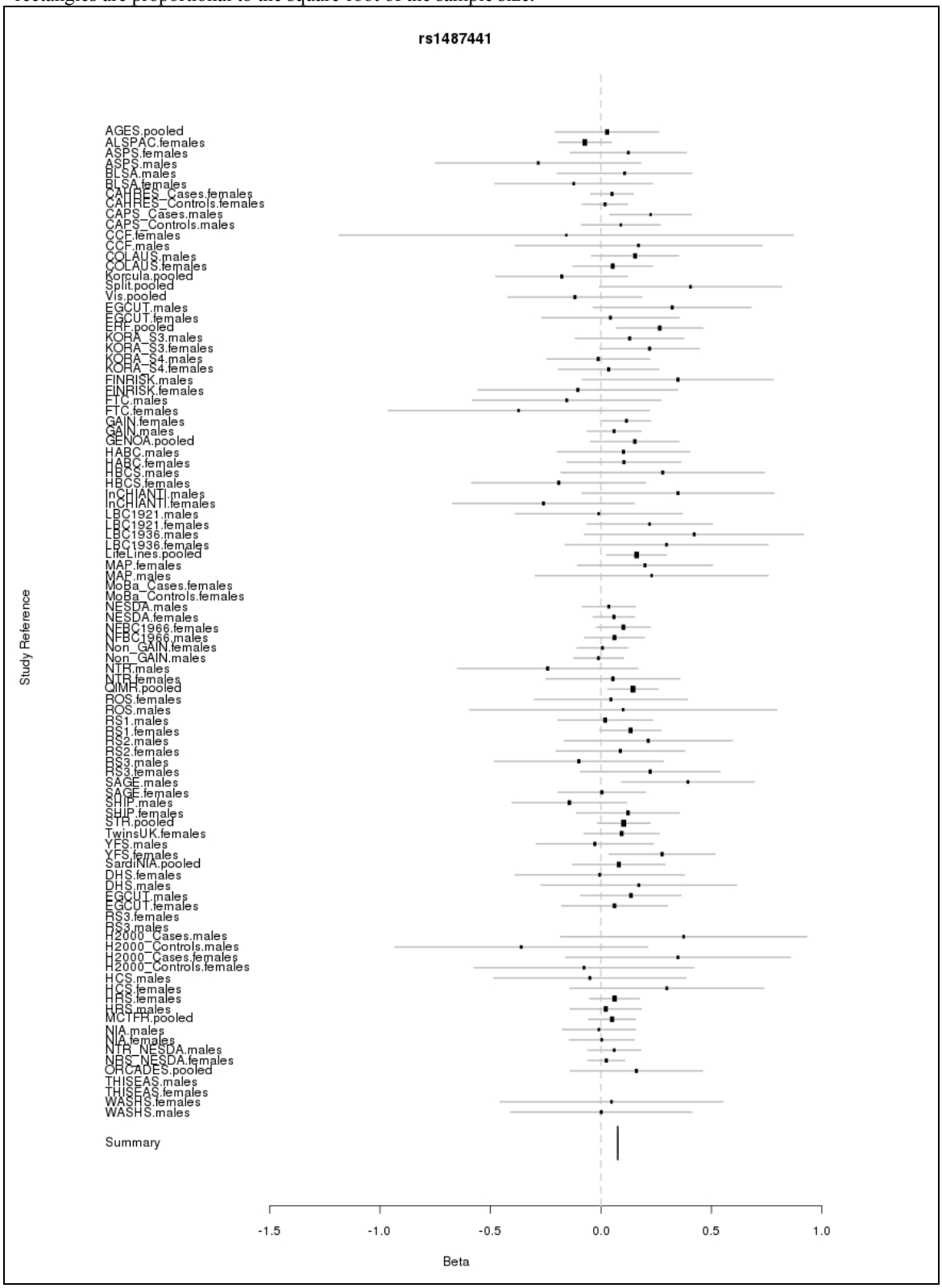


Figure S12. Forest plot for rs11584700 that is genome-wide significant for *College* in the combined discovery + replication stage. The gray lines represent the 95% confidence interval of the effect size estimate. The black rectangles are proportional to the square-root of the sample size.

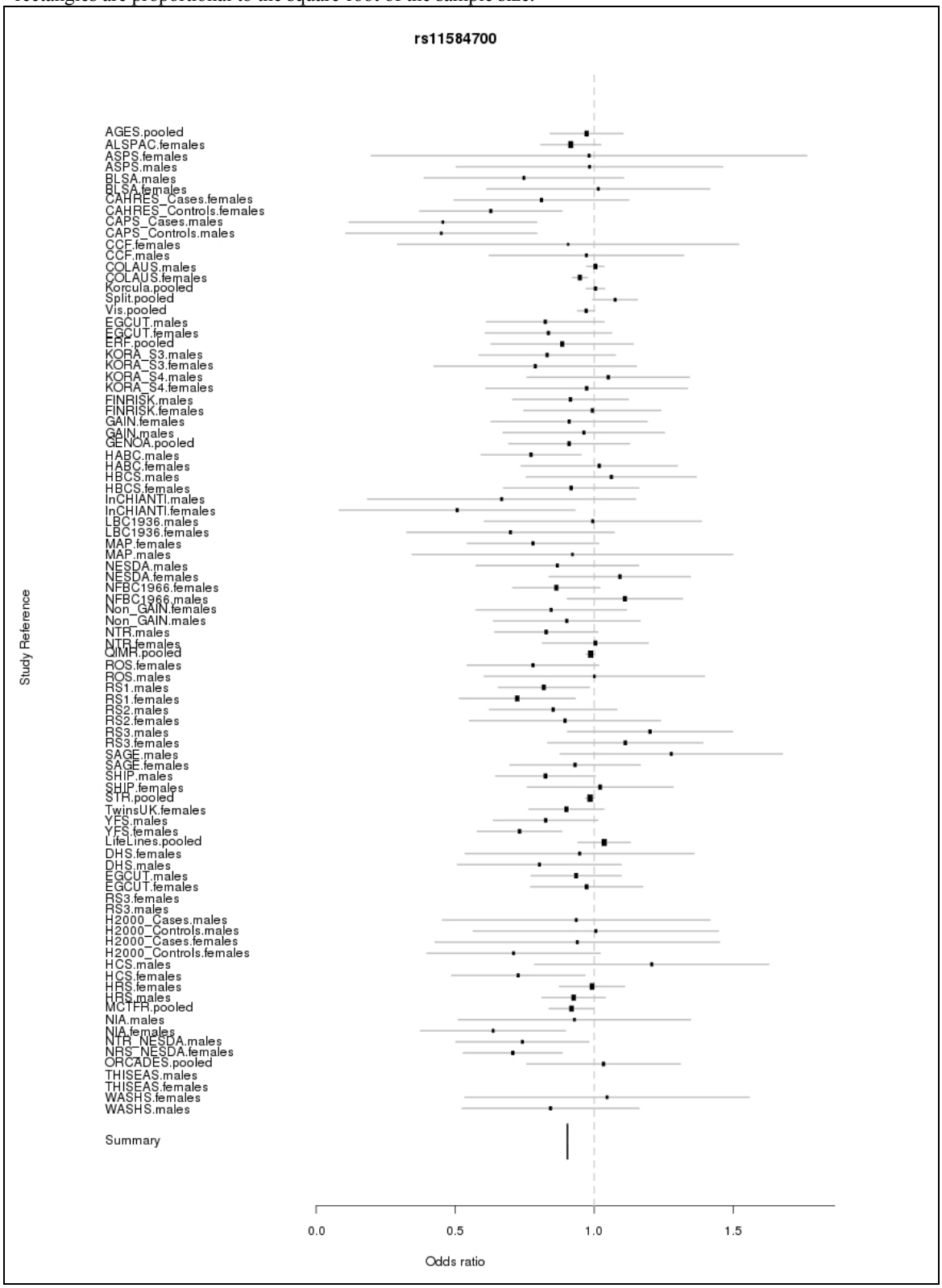


Figure S13. Forest plot for rs4851266 that is genome-wide significant for *College* in the combined discovery + replication stage. The gray lines represent the 95% confidence interval of the effect size estimate. The black rectangles are proportional to the square-root of the sample size.

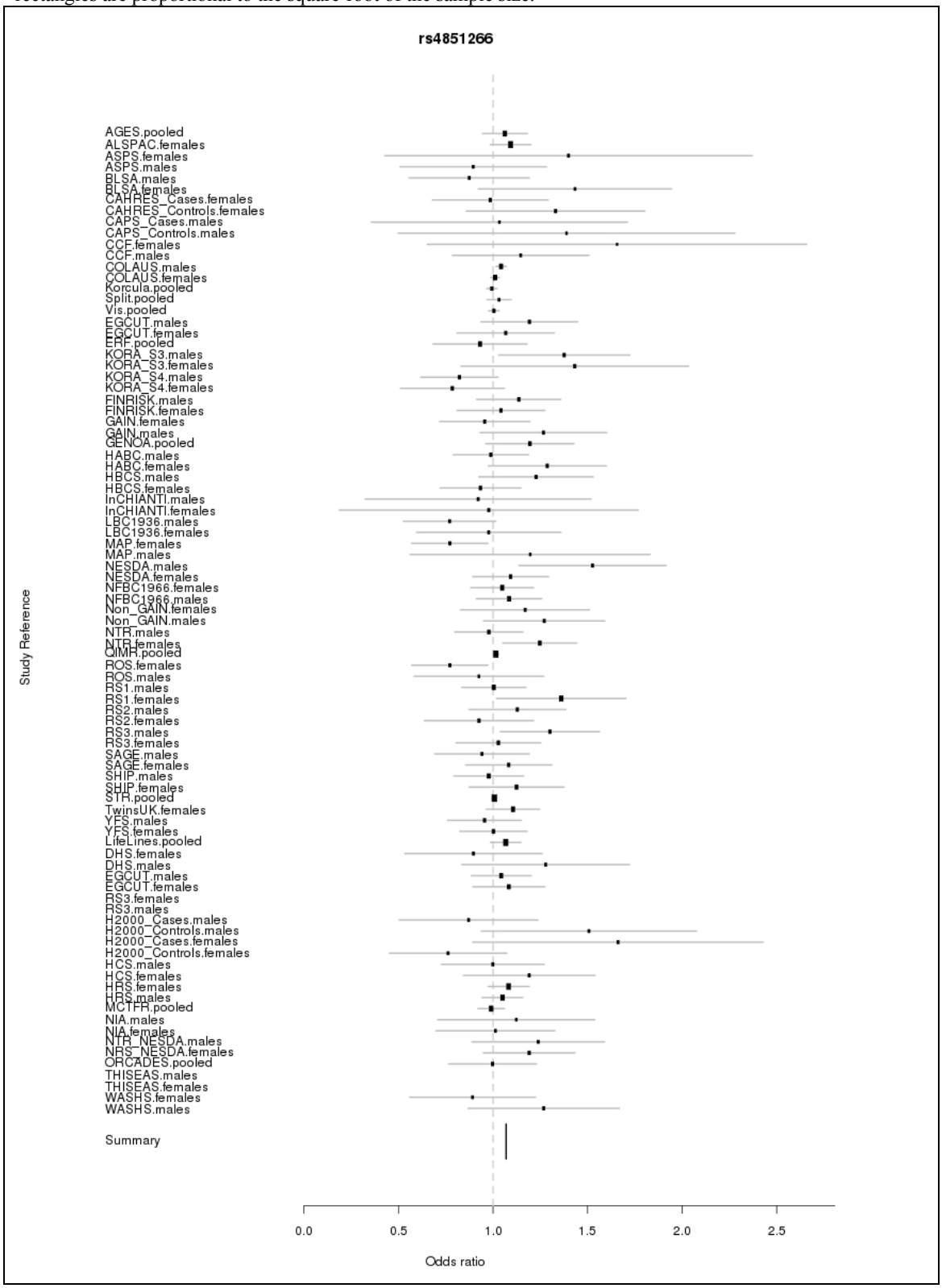


Figure S14. Forest plot for rs4851264 that is genome-wide significant for *College* in the combined discovery + replication stage. The gray lines represent the 95% confidence interval of the effect size estimate. The black rectangles are proportional to the square-root of the sample size.

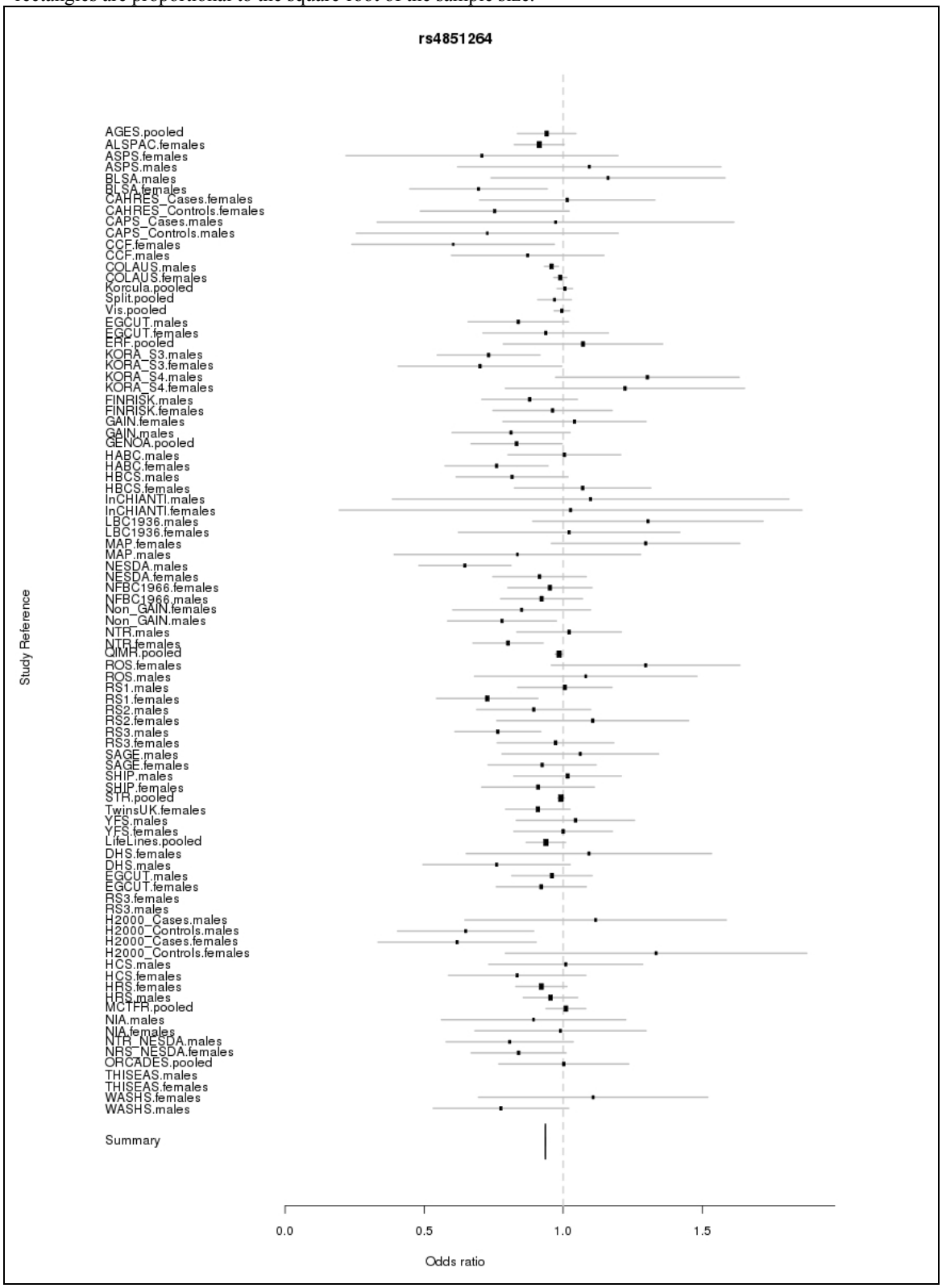


Figure S15. Forest plot for rs13401104 that is genome-wide significant for *College* in the combined discovery + replication stage. The gray lines represent the 95% confidence interval of the effect size estimate. The black rectangles are proportional to the square-root of the sample size.

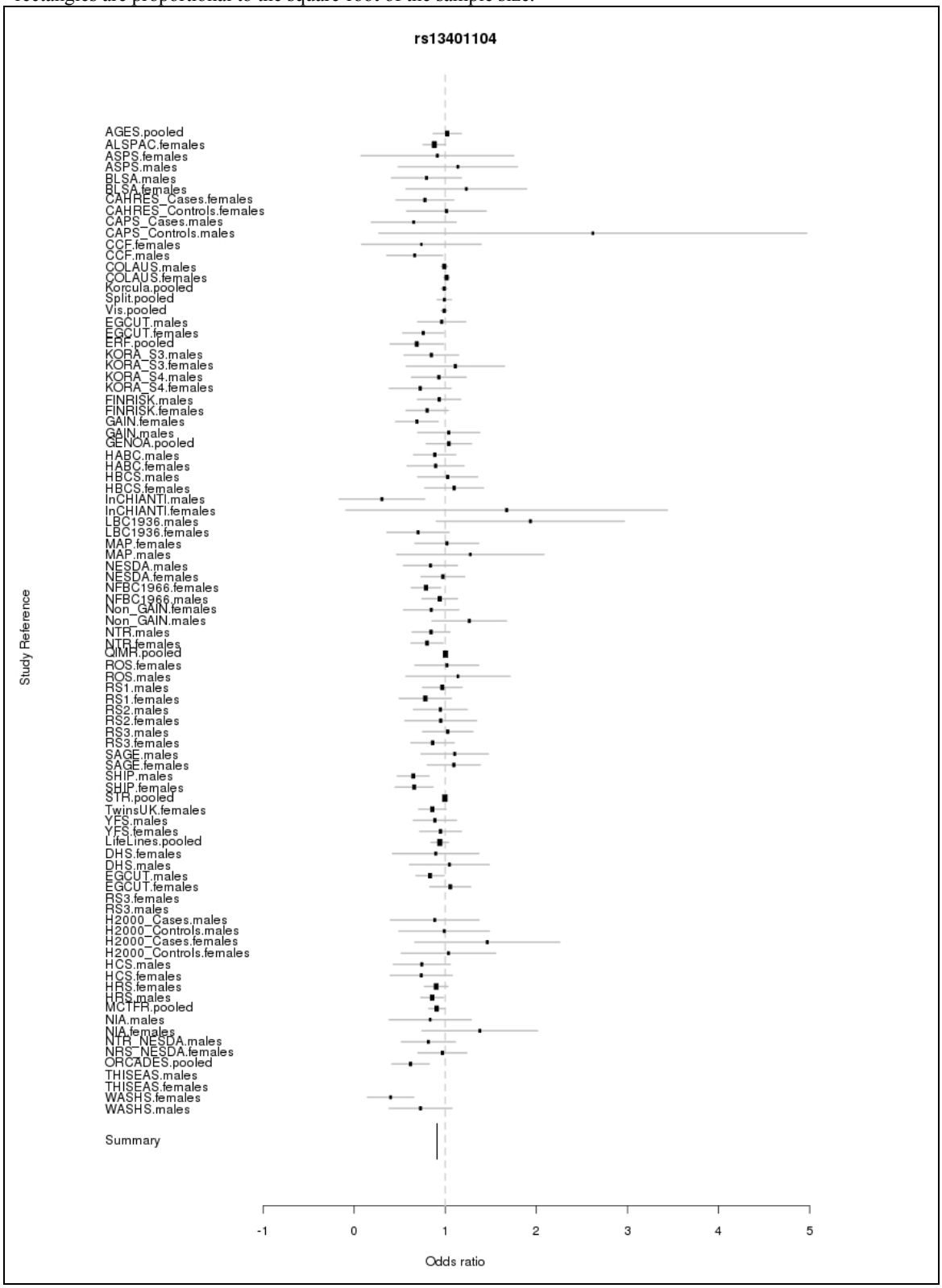


Figure S16. Comparison of empirical correlations with predicted correlations from ACE model. Predicted correlations were obtained when fitting the data to a simple ACE model and the sample correlations in the Swedish *Brothers Sample*.

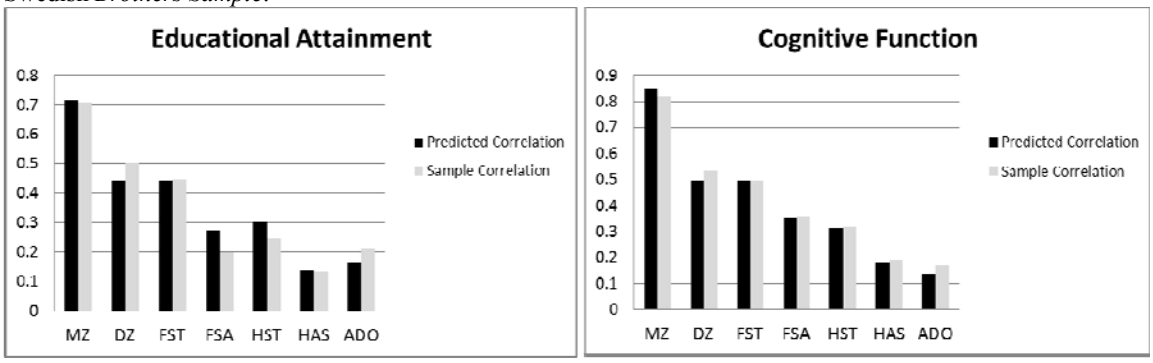


Figure S17. Plots of study-level unstandardized regression coefficients (β) against number of survey categories for SNP rs9320913 (panel A), rs11584700 (panel B), and rs4851266 (panel C). Each circle represents a study and is scaled proportional to the sample size of that study. The dashed lines indicate the sample-size-weighted average β, while the solid line is the weighted OLS regression line with weights proportional to the sample size. The stars represent the weighted-average β for each number of survey categories. The *p*-values of the regressions are 0.515 (panel A), 0.711 (panel B), and 0.665 (panel C). Notice that to facilitate comparisons and interpretation, we report the coefficients from the regression with the *EduYears* measure even in those instances where the genome-wide significant SNP was detected using the *College* measure.

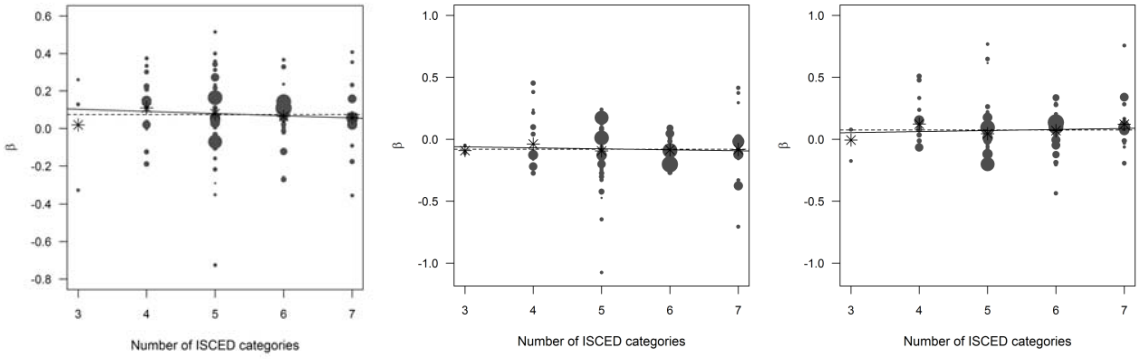


Figure S18. Plots of study-level standardized regression coefficients (β) against number of survey categories for SNP rs9320913 (panel A), rs11584700 (panel B), and rs4851266 (panel C). Each circle represents a study and is scaled proportional to the sample size of that study. The dashed lines indicate the sample-size-weighted average β, while the solid line is the weighted OLS regression line with weights proportional to the sample size. The stars represent the weighted average β for each number of survey categories. The *p*-values of the regressions are 0.359 (panel A), 0.924 (panel B), and 0.753 (panel C). Notice that to facilitate comparisons and interpretation, we report the coefficients from the regression with the *EduYears* measure even in those instances where the genome-wide significant SNP was detected using the *College* measure.

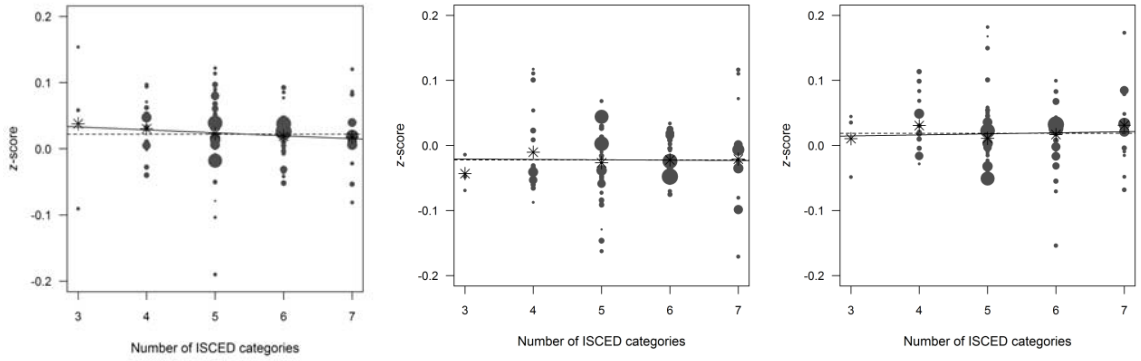


Figure S19. Plots of study-level *EduYears* regression coefficients (β) versus average birth year for SNP rs9320913 (panel A), rs11584700 (panel B), and rs4851266 (panel C). Each circle represents a study and is scaled proportionally to the sample size of that study. The dashed lines indicate the sample-size-weighted average β, while the solid line is the weighted OLS regression line with weights proportional to the sample size. The *p*-values of the regressions are 0.684 (panel A), 0.824 (panel B), and 0.829 (panel C). Notice that to facilitate comparisons and interpretation, we report the coefficients from the regression with the *EduYears* measure even in those instances where the genome-wide significant SNP was detected using the *College* measure.

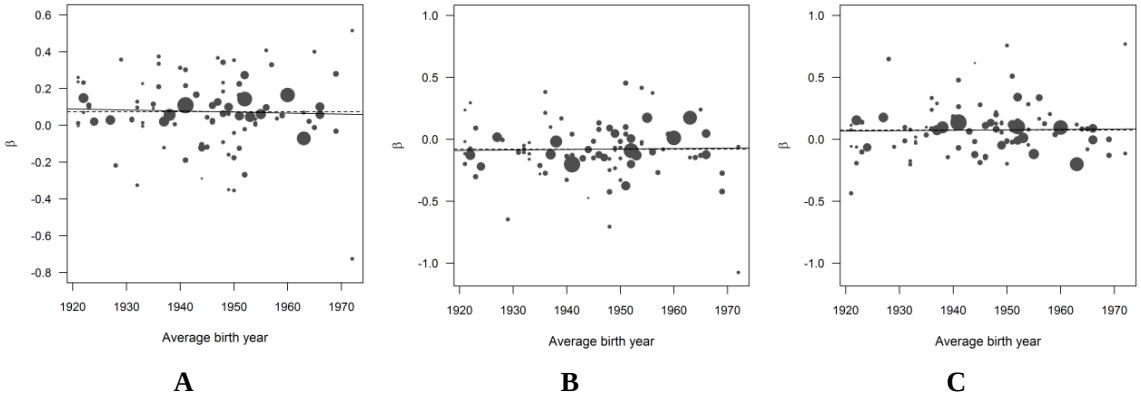


Figure S20. Cell-type-specific overlap between H3K4me3 marks and loci tagged by SNPs meeting $p < 1\times10^{-5}$ for each tissue/cell type. Tissue/cell types with *p*-values to the right of the dashed line are enriched at nominal $p \leq 0.05$.

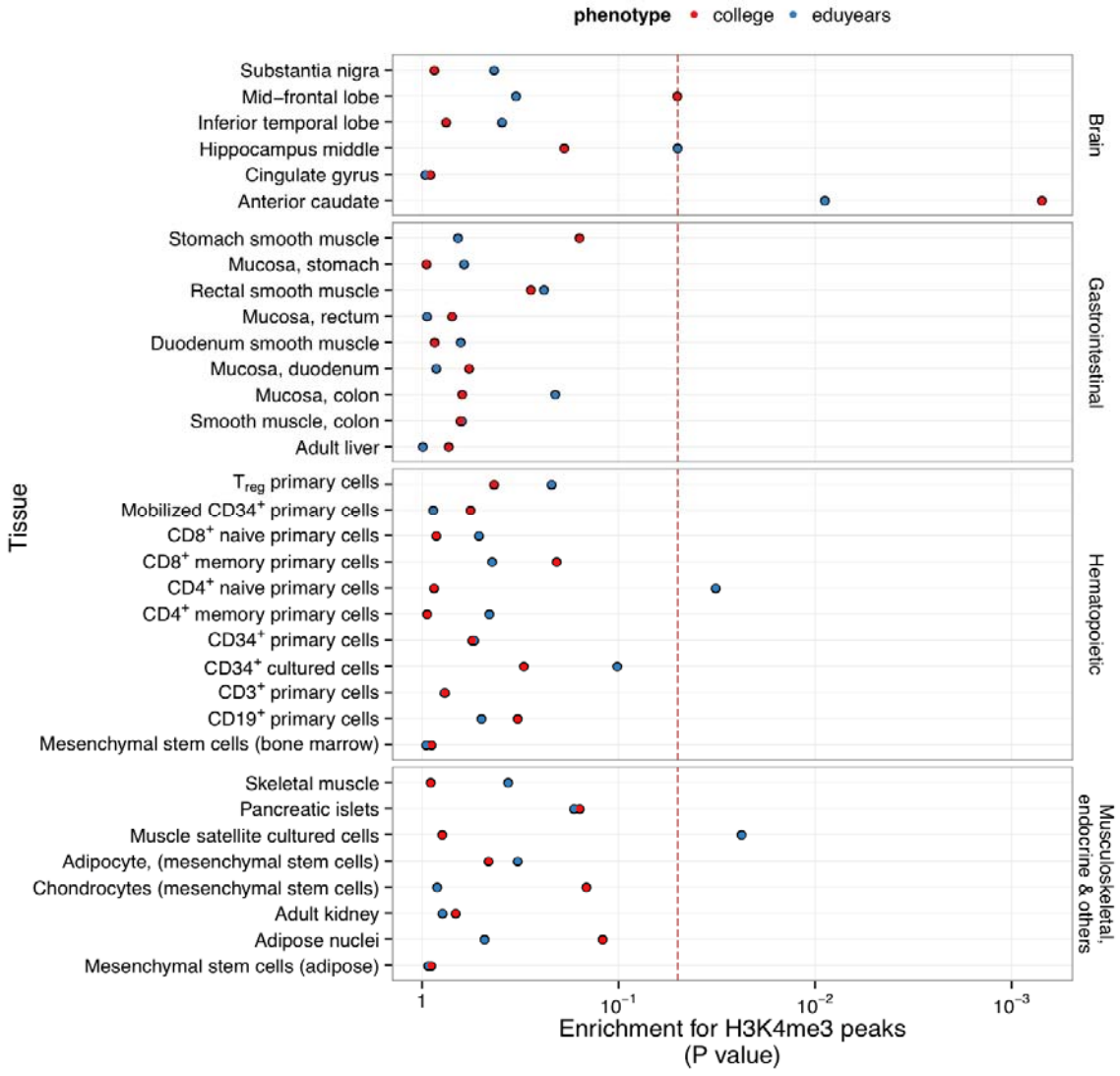


Figure S21. Distribution of cell-type specificity scores for loci tagged by SNPs meeting $p < 1 \times 10^{-5}$ within cell types significantly enriched for overlap between H3K4me3 marks and associated loci (Figure S20). Loci above the dashed line (identified and described in Table S19) show specificity to that cell type at nominal $p < 0.05$.

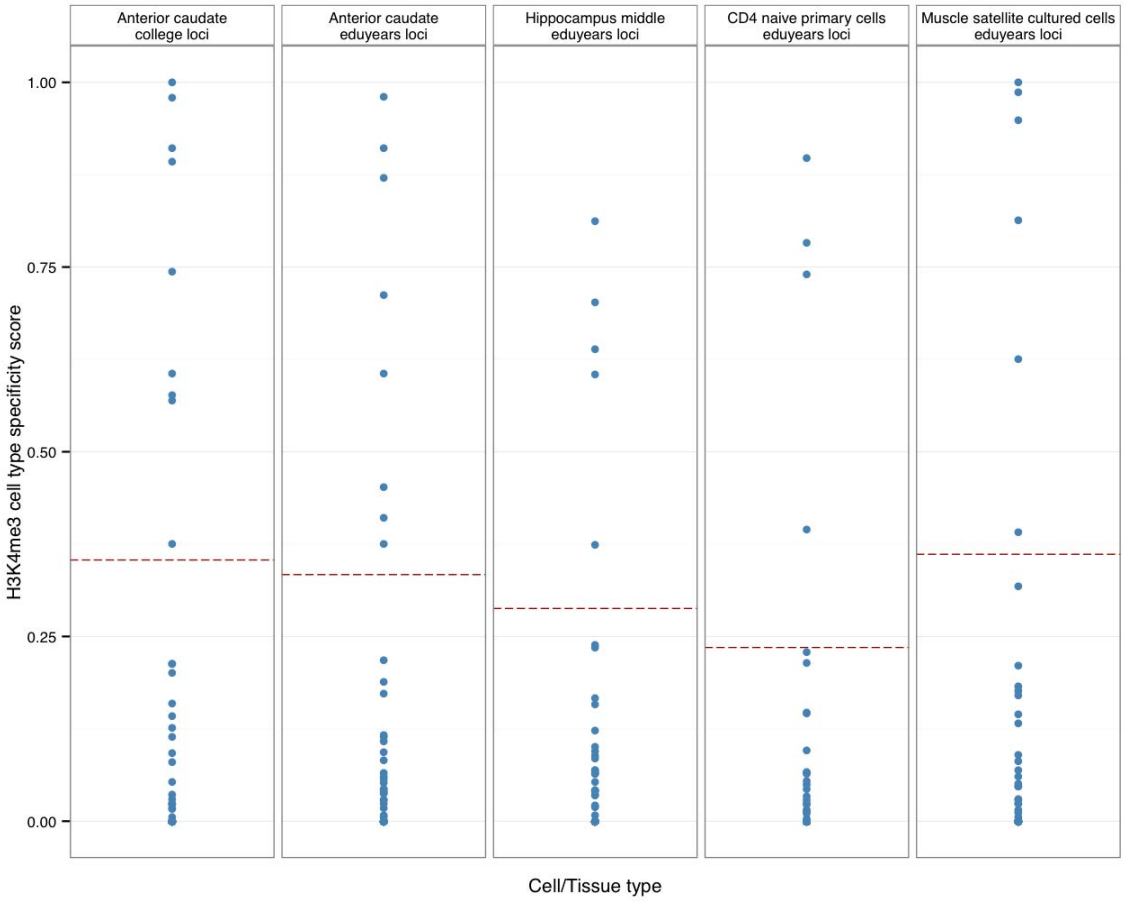
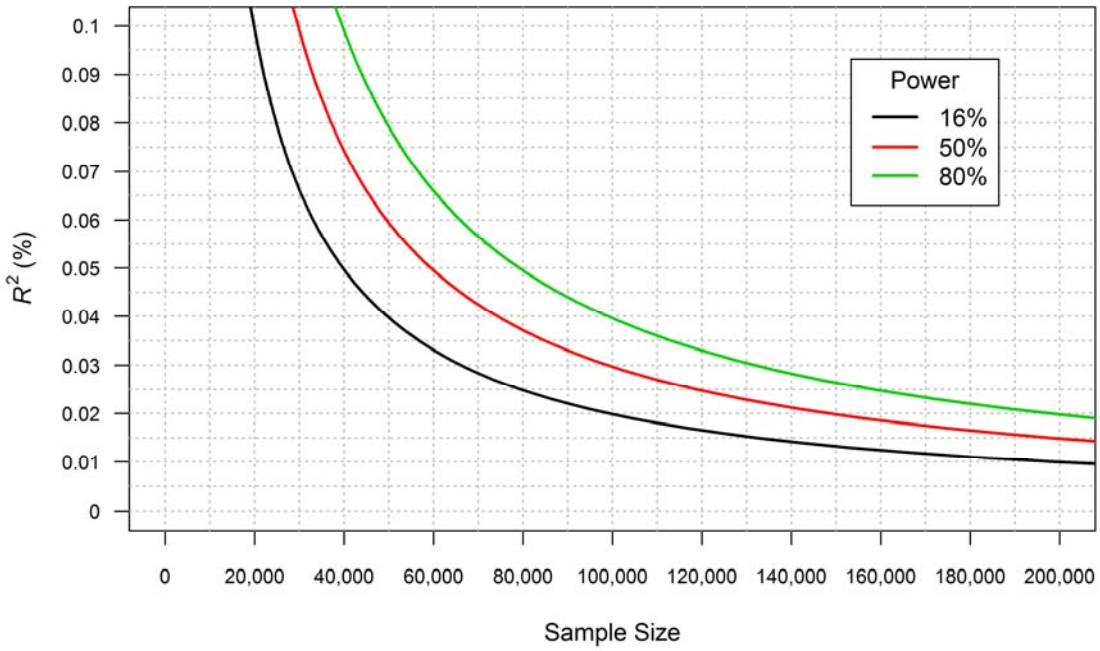


Figure S22. An illustration of the tradeoff between a GWAS on a distal phenotype in a larger sample against a GWAS on the endophenotype in a smaller sample. The *y*-axis shows effect sizes in terms of the population R^2 from the regression of the phenotype on the single SNP, ranging from 0 to 0.1% in increments of 0.01% (one-hundredth of one percent). The *x*-axis is the sample size. In these calculations, we assume that the only source of correlation between the SNP and *Y* is the effect of the SNP on the endophenotype of interest. Each curve graphs the locus of effect-size/sample-size pairs that gives a given level of power to detect the association at $p=5\times10^{-8}$.



11. Supplementary Tables

Table S1. Study design, numbers of individuals and sample quality control for GWAS cohorts. "Call rate" refers to the genotyping success rate, i.e., the minimum percentage of successfully genotyped SNPs.

Study				Sample QC			
Short name	Full name	Study design	Total sample size (*N*)	Call rate	Other exclusions	Sample in analysis (*N*)	References
Discovery Stage							
AGES	Age, Gene/Environment Susceptibility–Reykjavik Study	Population-based	3,219	≥97%	1) Mismatch to previous genotypes 2) Missing EA phenotype	3,212	*(94)*
ALSPAC	Avon Longitudinal Study of Parents and Children	Prospective pregnancy cohort	8,340	≥95%	1) IBD above 10% 2) Inconclusive X chromosome heterozygosity 3) Do not cluster with CEU HapMap on IBS plot 4) High autosomal heterozygosity 5) Missing EA phenotype	6,919	*(95)*
ASPS	Austrian Stroke Prevention Study	Population-based	922	≥97%	1) Mismatch between called and phenotypic gender 2) Other sample failures 3) Missing EA phenotype	848	*(96)* *(97)*
BLSA	Baltimore Longitudinal Study of Aging	Community-dwelling	848	≥98.5%	1) Sex mismatch 2) Missing EA phenotype	821	*(98)*
CAHRES-Cases	Cancer Hormone Replacement Epidemiology in Sweden	Case-control	1,321	≥96%	1) Missing EA phenotype	788	*(99)*
CAHRES-Controls	Cancer Hormone Replacement Epidemiology in Sweden	Case-control	1,524	≥96%	1) Missing EA phenotype	709	As CAHRES-Cases
CAPS-Cases	Cancer Prostate Sweden	Case-control	3,030	≥95%	1) Missing EA phenotype	240	*(100)* *(101)* *(102)*
CAPS-Controls	Cancer Prostate Sweden	Case-control	1,960	≥95%	1) Missing EA phenotype	219	As CAPS-Cases
CCF	Cleveland Clinic Foundation	Clinically selected (Lone Atrial Fibrillation)	495	≥97%	1) High heterozygosity (FDR<1%) 2) Sex mismatch 3) High IBS (IBS>=0.95); 4) PC outlier (more than 6 SDs away)	485	-

					4) Missing EA phenotype		
CoLaus	Etude Cohorte Lausannoise	Population-based	6,189	≥90%	1)PCA outliers removed 2)Related individuals 3) Missing EA phenotype	5,410	*(103)*
Cr_Kor	Croatia Korcula	Population-based Isolate	969	≥97%	1)Duplicate samples 2) Missing EA phenotype	843	*(104)*
Cr_Spl	Croatia Split	Population-based Isolate	535	≥97%	1)Duplicate samples 2) Missing EA phenotype	417	As Croatia Korcula
Cr_Vis	Croatia Vis	Population-based Isolate	1,026	≥95%	1)Duplicate samples 2) Missing EA phenotype	864	As Croatia Korcula
EGCUT	Estonian Genome Center, University of Tartu	Population-based	1,537	≥95%	1) Gender mismatch 2) Duplicates and/or 1st or 2nd degree relatives 3) Missing EA phenotype	1,537	*(56)*
ERF	Erasmus Rucphen Family study	Family-based	3,485	≥95%	1) Failing IBS checks 2) Sex chromosome checks 3) Ethnic outliers removed 4) Missing EA phenotype	2,380	*(106)* *(107)*
FINRISK	The National FINRISK Study	Population-based	38,031	≥95%	1) Excess heterozygosity 2) Relatedness and/or failed gender check 3) Missing EA phenotype	1,837	*(108)*
FTC	Finnish Twin Cohort	Family-based	1,387	≥95%	1) Gender discrepancy; 2) Heterozygosity check (threshold for inclusion - 0.03<F< 0.05) 3) Only one individual per family was included in the analysis 4) Missing EA phenotype	729	*(109)*
GAIN	Genetic Association Information Network Schizophrenia-Controls	Case-control	1,442	≥97%	1) Exclude individuals with population heterozygosity rate ± 3 s.d. 2) Exclude duplicates and individuals with IBD > 0.185. 3) Ethnic outliers (based Hapmap CEU) using IBS distances > 3 s.d. 4) Schizophrenia cases were excluded 5) Missing EA phenotype	1,164	*(110)*
GENOA	Genetic Epidemiology Network of Arteriopathy	Family-based	1,509	≥95%	1) MAF<0.01 2) SNPs not in HapMap	1,439	*(111)*

					3) Outliers (±6 SDs) on first 10 PCs from EIGENSTRAT 4) Missing EA phenotype		
HABC	Health ABC	Population based	1,663	≥97%	1) Sample failure 2) Genotypic sex mismatch 3) First degree relatives 4) Missing EA phenotype	1,659	*(112)*
HBCS	Helsinki Birth Cohort Study	Birth-cohort	1,728	≥95%	1) Gender discrepancy 2) Missing EA phenotype	1,717	*(113)*
InCHIANTI	Invecchiare in Chinati	Population-based	1,210	≥97%	1) Sex mismatch 2) Heterogeneity > 0.3 3) Missing EA phenotype	1,164	*(114)*
KORA S3	Kooperative Gesundheitsforschung in der Region Augsburg	Population-based	1,644	>93%	1) Gender discrepancy 2) Missing EA phenotype	1,595	*(115)*
KORA S4	Kooperative Gesundheitsforschung in der Region Augsburg	Population-based	1,814	>93%	1) Gender discrepancy 2) Missing EA phenotype	1,809	As KORA S3
LifeLines	The LifeLines Cohort Study	Population-based	8,132	≥95%	1) Sex mistmatch, 2) Duplicates / Cryptic relationships 3) None caucasians, ethnic cluster outlier (>4SD) 4) Excess/lack of heterozygosity (>3SD) 5) Missing EA phenotype	7,493	*(116)*
LBC1921	Lothian Birth Cohort 1921	Population-based birth-cohort	517	≥95%	1) Gender discrepancy 2) Relatedness (PiHAT > 0.25) 3) Non-caucasian descent 4) Missing EA phenotypes	515	*(117)*
LBC1936	Lothian Birth Cohort 1936	Population-based birth-cohort	1005	≥95%	1) Gender discrepancy 2) Relatedness (PiHAT > 0.25) 3)Non-caucasian descent 4) Missing EA phenotypes	1,003	*(118)*
MoBa-Cases	Mother and Child Cohort of NIPH-Cases	Population-based, nested case-control study	951	≥98%	1) Missing EA phenotypes	354	*(119)* *(120)*
MoBa-Controls	Mother and Child Cohort of NIPH-Controls	Population-based, nested case-control study	970	≥98%	1) Missing EA phenotypes	405	As MoBa-Cases
NESDA	Netherlands Study of Depression and Anxiety	Case-control	1,847	≥90%	1) Uncertain linkage between genotype and phenotype 2) Samples with evidence of contamination 3) Samples that failed	1,517	*(121)*

					genotyping 4) Miscellaneous failures (for example, data consistent with the presence of XO and XXY sex chromosome status). 5) Missing EA phenotypes		
NFBC1966	Northern Finland Birth Cohort 1966	Population-based	12,138	≥95%	1) Low mean heterozygosity [exclude if <0.29 & MDS outliers 2) Duplicates: concordance with other DNA>0.99 3) Contaminated samples: IBS pairwise with most other samples >0.99 4) IBS pairwise sharing>0.20 5) Withdrew consent 6) Gender mismatch: genotypic gender different from phenotypic 7) Missing EA phenotypes	5,371	*(122)* *(123)*
nonGAIN	Non-Genetic Association Information Network Schizophrenia-Controls	Case-control	1,364	≥97%	1) Exclude individuals with population heterozygosity rate ± 3 s.d. 2) Exclude duplicates and individuals with IBD > 0.185 3) Ethnic outliers (Hapmap CEU) using IBS distances > 3 s.d. 4) Missing EA phenotypes	1,109	*(110)*
NTR	Netherlands Twin Register	Family-based	5,856	≥90%	1) Extreme autosomal heterozygosity (F > -.10) 2) Gender discrepancy 3) Discrepancy between genetic and reported relatedness 4) Only unrelated individuals were chosen 5) Missing EA phenotypes	2,650	-
QIMR	Queensland Institute of Medical Research	Family based	10,506	≥95%	1) Individuals age < 30 years 2) Missing EA phenotypes	7,985	*(124)*

RS-I	Rotterdam Study Baseline	Population-based	7,983	≥97.5%	1) Gender mismatch with typed Xlinked markers 2) excess autosomal heterozygosity > 0.336~FDR>0.1% 3) duplicates and/or 1st or 2nd degree relatives using IBS probabilities >97% from PLINK 4) ethnic outliers using IBS distances > 3SD from PLINK 5) Missing EA phenotypes	5,806	*(125)* *(126)*
RS-II	Rotterdam Study Extension of Baseline	Population-based	3,011	≥97.5%	1) Gender mismatch with typed Xlinkedmarkers 2) excess autosomal heterozygosity (F<-0.055) 3) duplicates and/or 1st degree relatives using IBD PiHAT >40% from PLINK 4) ethnic outliers IBS distances > 4SD mean HapMAP CEU cluster from PLINK 5) Missing EA phenotypes	1,641	As RS-I
RS-III	Rotterdam Study Young	Population-based	3,932	≥97.5%	1) Gender mismatch with typed Xlinked markers 2) excess autosomal heterozygosity (F<-0.055) 3) duplicates and/or 1st degree relatives using IBD PiHAT >40% from PLINK 4) ethnic outliers IBS distances > 4SD mean HapMAP CEU cluster from PLINK 5) Missing EA phenotypes	2,014	As RS-I
RUSH-MAP	Rush University Medical Center - Memory and Aging Project	Epidemiological cohort	1,565	≥95%	1) Genotype-derived sex discordant with reported sex 2) Failed heterozygosity test 3) Missing EA phenotypes	888	*(127)*
RUSH-ROS	Rush University Medical Center - Religious Orders Study	Epidemiological cohort	1,170	≥95%	1) Genotype-derived sex discordant with reported sex 2) Failed heterozygosity test 3) Missing EA phenotypes	810	As RUSH-MAP
SAGE	Study of Addiction: Genetics and	Case-control	3,829	≥98%	1) Batch effects, chromosomal	1,321	*(128)*

	Environment				anomalies, Mendelian errors, sex-check 2) minor allele frequency > 1% 3)HWE P > E-4 4) sample misidentification, relatedness, other misspecifications 5) Missing EA phenotypes		
SardiNIA	SardiNIA Study of Aging	Family-based	6,148	≥95%	1) Sex-check 2) Other sample failures 3) Missing EA phenotypes	3,639	*(129)*
SHIP	Study of Health in Pomerania	Population-based	4,308	≥92%	1) Gender mismatch with typed X-linked markers 2) duplicates by estimated IBD 3) Missing EA phenotypes	3,556	*(130)*
STR	Swedish Twin Registry	Family-based	9,836	≥97%	1) Sex-check (heterozygosity of X-chomosomes) 2) deviations in heterozygosity of more then 5 SD from the population mean 3) Cryptically relatedness check; 4) Missing EA phenotypes	9,553	*(131)*
TwinsUK	St Thomas' UK Adult Twin Registry	Family-based	5,654	≥98%	1) sample call rate <98% 2) heterozygosity across all SNPs ≥2 SD from the sample mean 3) evidence of non-European ancestry as assessed by principle component analysis (PCA) comparison with HapMap3 populations 4) observed pairwise identity by descent (IBD) probabilities suggestive of sample identity errors 5) Missing EA phenotypes	2,619	*(132)*
YFS	The Cardiovascular Risk in Young Finns Study	Population-based	2,442	≥95%	1) Missing gender 2) Related individuals and duplicates 5) Missing EA phenotypes	2,029	*(133)*
Replication Stage							
DHS	Dortmund Health Study	Epidemiological	1,312	≥95%	1) PCA failed	953	*(134)*

		cohort			2) No DNA 3) Aged < 30 years 4) Missing EA phenotypes		*(135)*
EGCUT	Estonian Genome Center, University of Tartu	Population based	3,755	≥95%	1) Gender mismatch 2) Duplicates and/or 1st or 2nd degree relatives 3) Missing EA phenotypes	3,755	As EGCUT (Discovery stage)
H2000-Cases	Health 2000	Population-based random sample (case-control subcohort selected on basis of metabolic syndrome)	857	≥95%	1) Excess heterozygosity 2) Relatedness and/or failed gender check 3) Missing EA phenotypes	852	*(136)*
H2000-Controls	Health 2000	Population-based random sample (case-control subcohort selected on basis of metabolic syndrome)	868	≥95%	1) Excess heterozygosity 2) Relatedness and/or failed gender check 3) Missing EA phenotypes	864	As H2000-Cases
HCS	Hunter Community Study	Population based	1,230	≥95%	1) Gender mismatch 2) Duplicates and/or 1st degree relatives 3) Pan-European ancestry evident in principal components 4) Missing EA phenotypes	1,094	*(137)*
HRS	Health and Retirment Study	Population based	12,507	≥98%	1) Gender discrepancy 2) Ethnic outliers 3) Duplicates and/or relatives with KC>1/32 4) Missing EA phenotypes	8,626	*(138)*
MCTFR	Minnesota Center For Twin and Family Research	Family-based	7,278	≥99%	1) >5000 uncalled SNPs 2) low GenCall score 3) extreme hetero- or homozygosity 4) sample mix-up or unable to confirm known genetic relationships 5) Missing EA phenotypes	3,830	*(139)*
NIA	National Institute of Aging	Family based	1,072	≥97%	1) Removed non-caucasians and non-controls 2) Sex-check (heterozygosity of X-chomosomes) 3) Removed missing phenotypes Pruned family data to sample of unrelateds (retained all founders	622	*(140)*

					from each family; retained one non-founder from each family with no founders in sample) 4) Missing EA phenotypes		
NTR	Netherlands Twin Register	Family based	3,124	≥95%	1) Extreme autosomal heterozygosity (F > -.10) 2) Gender discrepancy 3) Discrepancy between genetic and reported relatedness 4) Only unrelated individuals were chosen 5) Missing EA phenotypes	1,317	*(141)*
ORCADES	The Orkney Complex Disease Study	Population-based	895	≥97%	1) Ethnic outliers 2) Duplicates 3) Gender mismatch 4) Excess IBS incompatible with pedigree 5) Missing EA phenotypes	810	*(142)*
RS-III	Rotterdam Study Young (additionallty genotyped individuals)	Population-based	976	≥97.5%	1) Gender mismatch with typed Xlinked markers 2) excess autosomal heterozygosity (F<-0.055) 3) duplicates and/or 1st degree relatives using IBD PiHAT >40% from PLINK 4) ethnic outliers IBS distances > 4SD mean HaMAP CEU cluster from PLINK 5) Missing EA phenotypes	976	As RS-I
THISEAS	The Hellenic study of Interactions between SNPs & Eating in Atherosclerosis Susceptibility	Case- control	1,075	≥95%	1) Gender mismatch 2) Heterozygosity 3) Ethnic outliers 4) Missing EA phenotypes	831	*(143)*
WASHS	Western Australia Sleep Health Study	Clinically selected (sleep problems; BMI<30 or BMI>40)	1,301	≥97%	1) Sex mismatch using sexcheck in PLINK 2) Relatedness (IBD > 0.1875 using PLINK) 3) Heterozygosity (h > 4sd using PLINK) 4) PCA outliers removed by eye after evaluating cluster plot comparing to HapMap CEU r3 5) Missing EA phenotypes	960	*(144)*

Table S2. ISCED classification scheme

ISCED Levels	Definition	US years of schooling (EduYears)	College
0	Pre-primary education	1	0
1	Primary education or first stage of basic education	7	0
2	Lower secondary or second stage of basic education	10	0
3	(Upper) secondary education	13	0
4	Post-secondary non-tertiary education	15	0
5	First stage of tertiary education (not leading directly to an advanced research qualification)	19	1
6	Second stage of tertiary education (leading to an advanced research qualification, e.g. a Ph.D.)	22	1

Table S3. Study-specific educational attainment measure and phenotype distribution.

Study	Educational attainment measure	ISCED transformation		*N* per ISCED category							*N*		
				0	1	2	3	4	5	6	Total	Non-College	College
Discovery Stage													
AGES	What is the highest level or year of school that you completed? 0) Did not go to school 1) Elementary school 2) High school 3) Industiral College,midwife,nurse´s aid, art/music education 4) Farmer´s College 5)-House keeping 6) Seamanship 7)Junior College 8) Business school 9) Teacher´s College/Nursing school 10) University / Technical College	0) ISCED 0 1) ISCED 1 2) ISCED 2 3) ISCED 2 4) ISCED 2 5) ISCED 2 6) ISCED 2 7) ISCED 4 8) ISCED 4 9) ISCED 4 10) ISCED 5	Females Males Pooled	2 0 2	550 217 767	449 80 529	0 0 0	425 635 1060	437 417 854	0 0 0	1,863 1,349 3,212	932 1,426 2,358	437 417 854
ALSPAC	What is the highest level of school that you completed? 1) None 2) CSE only 3) O-levels 4) A-levels 5) University degree and vocational training	1) ISCED 0 2) ISCED 2 3) ISCED 3 4) ISCED 4 5) ISCED 5	Females Males Pooled	274 0 274	0 0 0	1246 0 1246	2525 0 2525	1800 0 1800	1074 0 1074	0 0 0	6,919 0 6,919	5,845 0 5,845	1,074 0 1,074
ASPS	Please state your highest completed education: 1) Compulsory schooling not 2) Compulsory schooling	1) ISCED 0 or 1 (not in study) 2) ISCED 2 3) ISCED 3 4) ISCED 4 5) ISCED 5	Females Males Pooled	0 0 0	0 0 0	352 217 569	111 89 200	0 0 0	19 60 79	0 0 0	482 366 848	306 463 769	60 19 79

	3) High school 4) College 5) University degree												
BLSA	How many years of education do you have? 1) 7 2) 8 3) 9 4) 10 5) 11 6) 12 (High School) 7) 13 8) 14 (AA of AS) 9) 15 10) 16 (College) 11) 17 12) 18 (Some master) 13) 19 14) 20 15) 21 16) 22	1) ISCED 1 2) ISCED 1 3) ISCED 2 4) ISCED 2 5) ISCED 2 6) ISCED 3 7) ISCED 3 8) ISCED 4 9) ISCED 4 10) ISCED 5 11) ISCED 5 12) ISCED 5 13) ISCED 6 14) ISCED 6 15) ISCED 6 16) ISCED 6 Because BLSA is a US sample of highly educated for EduYears analysis actual years of schooling were used.	Females Males Pooled	0 0 0	1 2 3	4 8 12	55 30 85	44 35 79	219 248 467	48 127 175	371 450 821	104 75 179	267 375 642
CAHRES-Cases	What is your education? 1) Elementary school 2) 9 year compulsory school 3) Junior secondary school/girl school 4) Gymnasium 5) High school/University 6) Other 7) Other, what	If 6): ISCED 5 Else if (3) = 1 or 4)=1) and 6) = 1: ISCED 4 Else if 3) = 1 or 4) = 1 or 6) = 1: ISCED 3 Else if 2) = 1 or 1) = 1: ISCED 2	Females Males Pooled	0 0 0	0 0 0	360 0 360	272 0 272	61 0 61	95 0 95	0 0 0	788 0 788	693 0 693	95 0 95
CAHRES-Controls	What is your education? 1) Elementary school 2) 9 year compulsory school 3) Junior secondary	If 6): ISCED 5 Else if (3) = 1 or 4)=1) and 6) = 1: ISCED 4 Else if 3) = 1 or 4) = 1 or 6) = 1: ISCED 3	Females Males Pooled	0 0 0	0 0 0	331 0 331	240 0 240	53 0 53	85 0 85	0 0 0	709 0 709	624 0 624	85 0 85

	school/girl school 4) Gymnasium 5) High school/University 6) Other 7) Other, what	Else if 2) = 1 or 1) = 1: ISCED 2											
CAPS-Cases	What schools have you attended? 1) Elementary school 2) Vocational School 3) Lowever secondary school 4) 2-year upper secondary school 5) 3-4 upper secondary school 6) College/university degree 7) Other training	If 6) = 1: ISCED 5 Else if (4) = 1 or 5) = 1) and 7) = 1: ISCED 4 Else if 4) =1 or 5) = 1 or 7) = 1: ISCED 3 Else if 1) = 1 or 3) = 1: ISCED 2 If 2) individuals always have at least elementary school and/or something else, which then is set to the level of education.	Females	0	0	0	0	0	0	0	0	0	0
			Males	0	0	147	62	3	28	0	240	212	28
			Pooled	0	0	147	62	3	28	0	240	212	28
CAPS-Controls	What schools have you attended? 1) Elementary school 2) Vocational School 3) Lowever secondary school 4) 2-year upper secondary school 5) 3-4 upper secondary school 6) College/university degree 7) Other training	If 6) = 1: ISCED 5 Else if (4) = 1 or 5) = 1) and 7) = 1: ISCED 4 Else if 4) =1 or 5) = 1 or 7) = 1: ISCED 3 Else if 1) = 1 or 3) = 1: ISCED 2 If 2) individuals always have at least elementary school and/or something else, which then is set to the level of education.	Females	0	0	0	0	0	0	0	0	0	0
			Males	0	0	133	53	8	25	0	219	194	25
			Pooled	0	0	133	53	8	25	0	219	194	25
CCF	Highest level of education? 1) Grade school 2) High school 3) 2-year college	1) ISCED 2 2) ISCED 3 3) ISCED 4 4) ISCED 5 5) ISCED 6	Females	0	0	3	28	25	31	31	118	56	62
			Males	0	0	12	73	44	106	132	367	129	238
			Pooled	0	0	15	101	69	137	163	485	185	300

	4) 4-year college 5) Graduate level												
CoLaus	What is the highest level of school that you completed? 1) Scolarité obligatoire 2) Apprentissage 3) Baccalauréat, maturité 4) Maîtrise, diplôme supérieur (technicum, etc..) 5) Université, hautes écoles How many years of education do you have?	If 1) not completed: ISCED 0 If 1) and <7 years of education: ISCED 1 If 1) and 7+ years of education: ISCED 2 If 2) or 3) and <14 years of education: ISCED 3 If 2) or 3) and 14+ years of education: ISCED 4 If 4) or 5) and <20 years of education: ISCED 5 If 4) or 5) and 20+ years of education: ISCED 5	Females Males Pooled	3 2 5	121 130 251	565 306 871	1,056 793 1,849	322 389 711	668 727 1,395	128 200 328	2,863 2,547 5,410	2,067 1,620 3,687	796 927 1,723
Cr_Kor	How many years of school have you completed?	0: ISCED 0 1-7: ISCED 1 8-10: ISCED 2 11-13: ISCED 3 14-15: ISCED 4 16-19: ISCED 5 20+: ISCED 6.	Females Males Pooled	3 0 3	85 22 107	99 48 147	264 150 414	39 39 78	49 43 92	2 0 2	541 302 843	490 259 749	51 43 94
Cr_Spl	How many years of school have you completed?	0: ISCED 0 1-7: ISCED 1 8-10: ISCED 2 11-13: ISCED 3 14-15: ISCED 4 16-19: ISCED 5 20+: ISCED 6.	Females Males Pooled	1 0 1	16 3 19	13 7 20	108 91 199	39 23 62	63 44 107	2 7 9	242 175 417	177 124 301	65 51 116
Cr_Vis	How many years of school have you completed?	0: ISCED 0 1-7: ISCED 1 8-10: ISCED 2 11-13: ISCED 3 14-15: ISCED 4 16-19: ISCED 5 20+: ISCED 6.	Females Males Pooled	0 0 0	155 60 215	138 60 198	156 175 331	21 31 52	29 36 65	0 3 3	499 365 864	470 326 796	29 39 68
EGCUT	How many years of schooling do you have	If 0-3: ISCED 0 If 4 or Primary	Females Males	0 0	11 32	90 113	447 386	27 13	212 159	24 23	811 726	575 544	236 182

	and whati is the highest level school graduated? 1) Elementary school 2) Secondary school 3)Professional secondary school 4) High school 5) Professional high school 6) Professional higher education 7) Lower university degree (BcS and McS) 8) Higer university degree (PhD or MD)	School: ISCED 1 If 9 or Secondary: ISCED 2 If 12 or High School: or Professional Secondary: SCED 3 If 13-15 : ISCED 4 If 16-20 or lower university degree: ISCED 5 If 21+ or higher university degree (PhD, MD) : ISCED 6	Pooled	0	43	203	833	40	371	47	1,537	1,119	418
ERF	What is your highest completed education? 1) primary education 2) primary education, plus a higher not completed education 3) lower vocational education 4) lower secondary education 5) intermediate vocational education 6) general secondary education 7) higher vocational education 8) university	1) ISCED 1 2) ISCED 1 3) ISCED 2 4) ISCED 2 5) ISCED 4 6) ISCED 3 7) ISCED 5 8) ISCED 5	Females	0	466	560	148	97	36	0	1,307	1,271	36
			Males	0	376	404	62	141	90	0	1,073	983	90
			Pooled	0	842	964	210	238	126	0	2,380	2,254	126
FINRISK	In 1997 questionnaire: What is your highest level of education? 1) Primary school 2) Middle school 3) Vocational school 4) High school/college level education 5) Academical degree	In 1997 questionnaire: 1) ISCED 1 2) ISCED 2 3) ISCED 3 4) ISCED 3 / ISCED 4 5) ISCED 5 / ISCED 6	Females	0	216	87	243	5	299	0	850	551	299
			Males	0	198	80	319	1	375	0	973	598	375
			Pooled	0	414	167	562	6	674	0	1,823	1,149	674

	In 2002 and 2007 questionnaire: What is your highest level of education? 1) Primary school/Basic education 2) Middle school 3) Vocational school 4) High school 5) College-level education 6) Polytechnical degree 7) Academical degree	In 2002 and 2007 questionnaire: 1) ISCED 1/ ISCED 2 2) ISCED 2 3) ISCED 3 4) ISCED 3 5) ISCED 4 6) ISCED 5 7) ISCED 5 / ISCED 6												
FTC	Q1. What is your basic education? 1) Less than primary school 2) Primary school 3) Less than junior high-school 4) Junior high-school 5) Some high-school studies 6) Senior high-school Q2. What professional training have you completed (after the basic education)? 1) None 2) Lower level vocational school 3) Advanced vocational school 4) High school 5) Polytechnic 6) University	Q2=1 & Q1=1,2: ISCED 1 Q2=1 & Q1=3,4,5: ISCED 2 Q2=1 & Q1=6: ISCED 3 Q2=2 & Q1=1,2,3,4,5,6: ISCED 3 Q2 = 3 & Q1 = 1,2,3,4,5: ISCED 3 Q2 = 4 & Q1 = 6: ISCED 3 Q2 = 3 & Q1 = 6: ISCED 4 Q2 = 5 & Q1 = 2,3,4,5,6: ISCED 5 Q2 = 6 & Q1 = 2,3,4,5,6: ISCED 5	Females	0	48	13	167	17	25	0	270	245	25
			Males	0	90	14	313	9	33	0	459	426	33
			Pooled	0	138	27	480	26	58	0	729	671	58
GAIN	What is your highest degree received? 1) Less than high	1) ISCED 1 2) ISCED 2 3) ISCED 3	Females	0	8	51	178	173	175	19	604	410	194
			Males	0	12	30	146	153	197	22	560	341	219
			Pooled	0	20	81	324	326	372	41	1,164	751	413

	school 2) Some high school, no diploma 3) Graduated from high school, Diploma or equivalent (GED) 4) Some college, no degree 5) Associate degree (for example: AA, AS) 6) Bachelor's degree 7) Master's degree 8) Professional degree (for example: MD, DDS, LLB, JD) 9) Doctorate degree	4) ISCED 4 5) ISCED 4 6) ISCED 5 7) ISCED 5 8) ISCED 6 9) ISCED 6											
GENOA	What is the highest level of education that you have completed? Precollege Years/Grade: 0-12 or GED Technical/Trade School Years: 1-3, ≥4 College/University Years: 1-3, ≥4 Professional/Graduate School Years: 1-3, ≥4	Precollege Years/Grade: 0: ISCED 0 1-6: ISCED 1 7-9: ISCED 2 10-12 or GED: ISCED 3 Technical/Trade School Years: 1-3, ≥4: ISCED 15 College/University Years: 1-2: ISCED 4 3, ≥4: ISCED 5 Professional/Graduate School Years: 1-3: ISCED 5 ≥4: ISCED 6	Females Males Pooled	0 0 0	0 0 0	22 50 72	357 268 625	257 195 452	132 99 231	26 33 59	794 645 1,439	636 513 1,149	158 132 290
HABC	What is the highest grade or year of school that you completed? 0) No formal education 1-12) Grade 12 13)	0-1) ISCED 0 2-7) ISCED 1 8-10) ISCED 2 11-13) ISCED 3 14-15) ISCED 4 16-19) ISCED 5 no ISCED 6 defined, because no	Females Males Pooled	0 0 0	5 14 19	39 64 103	315 234 549	238 218 456	185 347 532	0 0 0	782 877 1,659	597 530 1,127	185 347 532

	Vocational/tradeschoo l without high school or the GED 14) Vocational/trade school after high school 15) Some college/Associate degree 16) College graduate (4 or 5 year program) 17) Master's degree (or post-graduate training) 18) Doctoral degree (PhD, MD, EdD, DVM, DDS, JD, etc.)	observations with this level in Health ABC cohort.											
HBCS	Highest educational attainment derived from the registry (Statistics Finland). EA is categorized in four classes: 1) folk school, elementary school or less 2) Learning profession 3) high school 4) college (upper or lower academic degree)	1) ISCED 1 2) ISCED 2 3) ISCED 3 4) ISCED 5	Females Males Pooled	0 0 0	374 201 575	171 163 334	246 192 438	0 0 0	191 179 370	0 0 0	982 735 1,717	791 556 1,347	191 179 370
InCHIANTI	What is your highest attained degree? 1) No schooling 2) Elementary 3) Lower secondary education 4) High School 5) University degree	1) ISCED 0 2) ISCED 1 3) ISCED 2 4) ISCED 3 5) ISCED 5	Females Males Pooled	218 65 283	291 283 574	74 95 169	50 53 103	0 0 0	14 21 35	0 0 0	647 517 1,164	633 496 1,129	14 21 35
KORA S3	ISCED classification derived on combination of these two questions:	Q2=1 & Q1=1: ISCED 0 Q2= 1 & Q1=2,4: ISCED 2	Females Males Pooled	4 1 5	0 0 0	203 50 253	461 402 863	77 188 265	56 153 209	0 0 0	801 794 1,595	745 641 1,386	56 153 209

	Q1: What is your highest completed education? 1) kein Abschluss (No degree) 2) Hauptschule, Volksschule (Primary School) 3) Berufsschule / Lehre (Vocational School) 4) Mittlere Reife, Realschule (Secondary School) 5) Fachschule, Techniker-, Meisterschule (Technical School) 6) Abitur, Fachabitur (Tertiary School) 7) Universität (University) Q2: What is your highest vocational education? 1) kein Abschluss (No degree) 2) Berufsschule (Lehre) (Vocational School) 3) Fachschule, Techniker-/Meisterschule (Technical School) 4) Ingenieur-Schule, Polytechnikum (Engineer/Polytechnic School) 5) Fachhochschule, Universität (University)	Q2=1 & Q1=3,6: ISCED 3 Q2=1 & Q1=5: ISCED 4 Q2=1 & Q1=7: ISCED 5 Q2=2 $ Q1=1-4: ISCED 3 Q2=2 $ Q1=5-6: ISCED 4 Q2=2 $ Q1=7: ISCED 5 Q2=3 $ Q1=1-6: ISCED 4 Q2=3 $ Q1=7: ISCED 5 Q2=4 $ Q1=1-7: ISCED 5 Q2=5 $ Q1=1-7: ISCED 5											
KORA S4	ISCED classification derived on	Q2=1 & Q1=1,2: ISCED 2	Females	8	0	171	497	165	88	0	929	841	88
			Males	3	0	38	422	233	184	0	880	696	184

	combination of these two questions: Q1: What is your highest completed education? 1) Hauptschule, Volksschule (Primary School) 2) Mittlere Reife, Realschule (Secondary School) 3) Abitur, Fachabitur, Fachhochschulreife (Tertiary School) 4) Hochschule, Fachhochschule, Universität (University) 5) Sonstiger Abschluss (Other qualification) 6) Kein Abschluss (No degree) Q2: What is your highest vocational education? 1) kein Abschluss (No degree) 2) Berufsschule (Lehre) (Vocational School) 3) Fachschule, Techniker-/Meisterschule (Technical School) 4) Ingenieur-Schule, Polytechnikum (Engineer/Polytechnic School) 5) Sonstiger Abschluss (Other qualification)	Q2=1 & Q1=3: ISCED 3 Q2=1 & Q1=4: ISCED 5 Q2=1 & Q1=6: ISCED 0 Q2=2 & Q1=1,2,6: ISCED 3 Q2=2 & Q1=3: ISCED 4 Q2=2 & Q1=4: ISCED 5 Q2=3 & Q1=1,2,3,6: ISCED 4 Q2=3 & Q1=4: ISCED 5 Q2=4 & Q1=1,2,3,4,6: ISCED 5 Other combinations excluded.	Pooled	11	0	209	919	398	272	0	1,809	1,537	272
LifeLines	What is your highest	1) ISCED 0	Females	26	153	1,461	1,577	0	1,043	0	4,260	3,217	1,043

completed education? 1) No Education (not finished elementary school) 2) Lower education (elementary school) 3) Lower or preparatory applied education (e.g. lower technical school, lower vocational education in business and administration , preparatory middle-level applied education) 4) Middle general continued education(e.g. further extended primary education, (further) extended primary education ,middle-level applied education-short , preparatory middle-level applied education theoretical) 5) Middle-level applied education(e.g. middle-level applied education-long, middle level applied/technical training, upper vocational education in business and administration) 6) Higher general and preparatory education(e.g. higher general continued education, preparatory scientific	2) ISCED 1 3) ISCED 2 4) ISCED 2 5) ISCED 3 6) ISCED 3 7) ISCED 5 8) ISCED 5	Males	22	97	1,059	1,119	0	936	0	3,233	2,297	936
		Pooled	48	250	2,520	2,696	0	1,979	0	7,493	5,514	1,979

	education, higher commoner's school) 7) Higher professional education or pre university education(e.g. higher professional education, higher level applied/technical training, higher vocational education in business and administration) 8) Scientific education (university)													
LBC1921	How many years did you spend in full-time education? 0) 7 1) 8-10 2) 11-17 3) 18+	0) ISCED 1 1) ISCED 2 2) ISCED 3 3) ISCED 5	Females Males Pooled	0 0 0	0 2 2	187 114 301	111 93 204	0 0 0	3 5 8	0 0 0	301 214 515	298 209 507	3 5 8
LBC1936	What is the highest qualification you have achieved? 0) No qualification 1) O-level or equivalent 2) A-level or equivalent 3) Semi-professional/professional quealifications 4) Degree	0) ISCED 1 1) ISCED 3 2) ISCED 3 3) ISCED 4 4) ISCED 5	Females Males Pooled	0 0 0	78 96 174	0 0 0	295 265 560	64 57 121	58 90 148	0 0 0	495 508 1,003	437 418 855	58 90 148
MoBa-Cases	What is your highest completed education? 1) 9-yr secondary school 2) 1-2 yr high school 3) Technical high school 4) 3 yr high school general studies, junior	1) ISCED 1 2) ISCED 2 3) ISCED 3 4) ISCED 3 5) ISCED 5 6) ISCED 6	Females Males Pooled	0 0 0	13 0 13	19 0 19	90 0 90	0 0 0	150 0 150	82 0 82	354 0 354	122 0 122	232 0 232

	5) Regional technical school, 4-yr university bachelor degree 6) University, technical college												
MoBa-Controls	What is your highest completed education? 1) 9-yr secondary school 2) 1-2 yr high school 3) Technical high school 4) 3 yr high school general studies, junior 5) Regional technical school, 4-yr university bachelor degree 6) University, technical college	1) ISCED 1 2) ISCED 2 3) ISCED 3 4) ISCED 3 5) ISCED 5 6) ISCED 6	Females Males Pooled	0 0 0	8 0 8	15 0 15	92 0 92	0 0 0	195 0 195	95 0 95	405 0 405	115 0 115	290 0 290
NESDA	1) No degree or some years of primary school 2) Primary education 3) Secondary Special Education 4) VBO/LBO (housekeeping-, vodational-, technical school or internal professional training), MBO-short 5) Leerlingwezen, ULO 6) MAVO, MULO, VMBO 7) MBO-lang, or internal professional training on MBO-level 8) HAVO, VWO, Gymnasium, HBS, MMS 9) HBO or internal	1) ISCED 0 2) ISCED 1 3) ISCED 2 4) ISCED 2 5) ISCED 2 6) ISCED 2 7) ISCED 3 8) ISCED 3 9) ISCED 5 10) ISCED 5 11) Mainly ISCED 5, some added manually to ISCED 4 12) Missing 13) Missing	Females Males Pooled	12 9 21	61 28 89	278 134 412	268 171 439	3 2 5	371 180 551	0 0 0	993 524 1,517	622 344 966	371 180 551

	professional training on HBO-level 10) Scientific education, university 11) Else, namely: (propedeuse University, first year nursary etc.) 12) Don't know 13) Not applicable												
NFBC1966	What is your education? 1) Primary education (less than 9) 2) Lower secondary education (9) 3) Lower levels of upper secondary education (10-11) 4) Upper level of upper secondary education (12) 5) Lowest level of tertiary education (13-14) 6) Lower-degree level of tertirary education (15) 7) Higher-degree level of tertiary education (16) 8) Doctorate or equivalent level of tertiary education	1) ISCED 1 2) ISCED 2 3) ISCED 3 4) ISCED 3 5) ISCED 4 6) ISCED 5 7) ISCED 5 8) ISCED 6	Females Males Pooled	0 0 0	10 16 26	117 164 281	1,899 1,726 3,625	371 282 653	391 372 763	11 12 23	2,799 2,572 5,371	2,397 2,188 4,585	402 384 786
nonGAIN	What is your highest degree received? 1) Less than high school 2) Some high school, no diploma 3) Graduated from high school, Diploma or equivalent (GED)	1) ISCED 1 2) ISCED 2 3) ISCED 3 4) ISCED 4 5) ISCED 4 6) ISCED 5 7) ISCED 5 8) ISCED 6 9) ISCED 6	Females Males Pooled	0 0 0	7 6 13	27 21 48	163 127 290	168 177 345	148 223 371	13 29 42	526 583 1,109	365 331 696	161 252 413

	4) Some college, no degree 5) Associate degree (for example: AA, AS) 6) Bachelor's degree 7) Master's degree 8) Professional degree (for example: MD, DDS, LLB, JD) 9) Doctorate degree												
NTR	What is your highest finished education with a diploma? 0) No education finished 1) Primary school only 2) Lower vocational education (LB0) 3) General Secundary education (LAVO, MAVO) 4) Higher secundary education (HAVO, VWO) 5) Intermediate vocational education (MBO) 6) Higher vocational education (HBO) 7) University 8) PhD	0) ISCED 0 1) ISCED 1 2) ISCED 2 3) ISCED 2 4) ISCED 3 5) ISCED 3 6) ISCED 5 7) ISCED 5 8) ISCED 6	Females Males Pooled	38 32 70	256 134 390	248 82 330	451 282 733	0 0 0	571 468 1,039	30 58 88	1,594 1,056 2,650	993 530 1,523	601 526 1,127
QIMR	Three different educational scales were used Scale 1: 1) Less than 7 years' schooling 2) 8-10 years' schooling 3) 8-10 years' schooling and	Scale 1: 1) ISCED 1 2) ISCED 2 3) ISCED 3 4) ISCED 3 5) ISCED 4 6) ISCED 4 7) ISCED 5 8) ISCED 5bis (20 years of schooling)	Females Males Pooled	0 0 0	91 69 160	1,225 565 1,790	1,239 989 2,228	1,116 912 2,028	503 500 1,003	370 406 776	4,544 3,441 7,985	3,671 2,535 6,206	873 906 1,779

	apprenticeship or diploma 4) 11-12 years' schooling 5) 11-12 years' schooling and apprenticeship or diploma 6) Technical/Teacher's College 7) University first degree 8) University post graduate training	Scale 2: 1) ISCED 1 2) ISCED 2 3) ISCED 3 4) ISCED 4 5) ISCED 4 6) ISCED 5 7) ISCED 5bis (20 years of schooling)
	Scale 2: 1) Less than 7 years' schooling 2) 8-10 years' schooling 3) 11-12 years' schooling 4) Apprenticeship, diploma, etc. 5) Technical/Teacher's College 6) University first degree 7) University post graduate training	Scale 3: 1) ISCED 1 2) ISCED 2 3) ISCED 3 4) ISCED 3 5) ISCED 4 6) ISCED 5 7) ISCED 5bis (20 years of schooling)
	Scale 3: 1) Primary 2) Secondary Junior (SC) 3) Secondary Senior (HSC) 4) Apprenticeship, diploma, etc. 5) Tertiary undergraduate 6) Tertiary graduate 7) University post	

	graduate training												
RS-I	What is your highest attained education? 1) Primary education 2) Primary education, plus a higher not completed education 3) Lower vocational education 4) Lower secundary education 5) Intermediate vocational education 6) General secundary education 7) Higher vocational education 8) University	1) ISCED 1 2) ISCED 1 3) ISCED 2 4) ISCED 2 5) ISCED 3 6) ISCED 3 7) ISCED 5 8) ISCED 5	Females Males Pooled	0 0 0	1,598 627 2,225	1,007 555 1,562	675 865 1,540	0 0 0	135 344 479	0 0 0	3,415 2,391 5,806	3,280 2,047 5,327	135 344 479
RS-II	What is your highest attained education? 0) Primary education 1) lower vocational education 2) intermediate general education 3) Intermediate vocational education 4) General secundary education 5) Higher vocational education, first fase 6) Higher vocational education, second fase Only individuals were included that answered yes to the question: Did you finish this education with a degree? 0) No 1) Yes	0) ISCED 1 1) ISCED 2 2) ISCED 2 3) ISCED 3 4) ISCED 3 5) ISCED 5 6) ISCED 5	Females Males Pooled	0 0 0	92 45 137	467 184 651	199 326 525	0 0 0	101 227 328	0 0 0	859 782 1,641	758 555 1,313	101 227 328
RS-III	What is the highest	0) ISCED 1	Females	0	125	499	256	0	250	0	1,130	880	250

	education that you finished with a degree? 0) Primary education 1) Lower vocational education 2) intermediate secundary education 3) Intermediate vocational education 4) General secondary education 5) Higher education (HBO) 6) Higher education (University)	1) ISCED 2 2) ISCED 2 3) ISCED 3 4) ISCED 3 5) ISCED 5 6) ISCED 5	Males Pooled	0 0	76 201	206 705	310 566	0 0	292 542	0 0	884 2,014	592 1,472	292 542
RUSH-MAP	What is the highest grade or year of regular school you completed? 0-30 years	If ≥22: ISCED 6 If ≥19 ISCED 5 If ≥ 15 ISCED 4 If ≥13 ISCED 3 If ≥10 ISCED 2 If ≥7 ISCED 1 If ≥1 ISCED 0	Females Males Pooled	1 0 1	14 7 21	209 46 255	133 37 170	245 121 366	37 28 65	4 6 10	643 245 888	602 211 813	41 34 75
RUSH-ROS	What is the highest grade or year of regular school you completed? 0-30 years	If ≥22: ISCED 6 If ≥19 ISCED 5 If ≥ 15 ISCED 4 If ≥13 ISCED 3 If ≥10 ISCED 2 If ≥7 ISCED 1 If ≥1 ISCED 0	Females Males Pooled	0 3 3	4 6 10	23 21 44	15 12 27	299 73 372	153 108 261	38 55 93	532 278 810	341 115 456	191 163 354
SAGE	What is the highest grade in school you completed? Grade 1-12 (listed by respondent as 1-12); Technical school/1 year of college=13; 2 years of college=14; 3 years of college=15; 4 years of colege=16; graduate/doctorate=17	If ≥16 ISCED 5 If 14-15, ISCED 4 If 11-13 ISCED 3 If 8-10 ISCED 2 If 2-7 ISCED 1 If 1 ISCED 0	Females Males Pooled	2 0 2	3 4 7	26 19 45	217 146 363	174 83 257	423 224 647	0 0 0	845 476 1,321	422 252 674	423 224 647
SardiNIA	What is your highest degree?	1) ISCED 0 2) ISCED 1	Females Males	69 37	472 415	836 780	487 267	0 0	191 85	0 0	2,055 1,584	1,864 1,499	191 85

	1) Iliterate 2) 5th grade 3) 8th grade 4) High school 5) University degree	3) ISCED 2 4) ISCED 3 5) ISCED 5	Pooled	106	887	1,616	754	0	276	0	3,639	3,363	276
SHIP	Measure based on two questions: Q1: What is your highest general education? 1) Noch Schüler(in) ohne Abschluß 2) Schulabgang ohne Abschluß 3) Volks- oder Hauptschulabschluß 4) Mittlere Reife, Realschulabschluß, Fachschulreife 5) Abschluß polytechnische Oberschule 6) Fachhochschulreife, fachgebundene Hochschulreife, Fachoberschule 7) Abitur, allgemeine Hochschulreife, EOS mit Facharbeiterabschluß 8) Fachhochschulreife, Facharbeiter mit Abitur 9) Anderer Abschluß (auch: keine Angabe!) Q2: What kind of vocational training do you have? 1) Noch in beruflicher Ausbildung oder Student	If Q1=1,2,9: ISCED 0 If Q1=3 & Q2≠4,5,6,7: ISCED 1 If Q1=4,5 & Q2=3 & Q2≠4,5,6,7: ISCED 2 If Q1=3 & Q2=4: ISCED 2 If Q1=3,4,5,6,7,8 & Q2=4 & Q2≠3,5,6,7: ISCED 3 If Q1=6,7,8: ISCED 3 If Q1=3,4,5,6,7,8 & Q2=5 & Q2≠3,4,6,7,8: ISCED 4 If Q1=3 & Q2=5: ISCED 4 If Q1=3,4,5,6,7,8 & Q2=6,7 & Q2≠4,5,8: ISCED 5 If Q1=3 & Q2=6,7: ISCED 5	Females Males Pooled	41 60 101	261 93 354	129 156 285	899 1,039 1,938	277 145 422	187 269 456	0 0 0	1,794 1,762 3,556	1,607 1,493 3,100	187 269 456

	2) Nicht in beruflicher Ausbildung, bisher kein Ausbildungsabschluß 3) Beruflich-betriebliche Anlernzeit, aber keine Lehre; Teilfacharbeiterabschluß 4) Lehre mit Abschlußprüfung, beruflich-betriebliche Ausbildung 5) Fach- oder Berufsfachschulabschluß, z. B. Handelsschule, Fachakademie 6) Abschluß Fachhochschule, Ingenieurschule, Polytechnikum 7) Hochschulabschluß 8) Anderer beruflicher Abschluß												
STR	Data from Statistics Sweden which contains information on the ISCED level for the year 2005.	N.A.	Females	0	1,054	395	2,248	76	1,263	20	5,056	3,773	1,283
			Males	0	1,130	277	1,939	169	921	61	4,497	3,515	982
			Pooled	0	2,184	672	4,187	245	2,184	81	9,553	7,288	2,265
TwinsUK	At what age did you finish full time education? At what age did you finish or stop full-time education? At what age did you finish continuous full-time education? Please indicate all the qualifications you have:	A combination of age when finished education and highest qualification were used to determine ISCED If 0-10: ISCED 0 If 11-16: ISCED 1/2 If 16-18: ISCED 2/3 If 18-21: ISCED 4/5 If 21-23: ISCED 5 If >23: ISCED 6	Females	0	2	838	1,332	66	361	20	2,619	2,238	381
			Males	0	0	0	0	0	0	0	0	0	0
			Pooled	0	2	838	1,332	66	361	20	2,619	2.238	381

	University Higher vocational Teaching Nursing A-level Middle vocational O-level 5+ Lower vocational O-level Clerical Other No qualification	Qualifications University: ISCED 5/6 Higher vocational: ISCED 3 Teaching: ISCED 4 Nursing: ISCED 4 A-level: ISCED 3 Middle vocational ISCED 2 Lower vocational: ISCED 2 O-level: ISCED 2 Clerical: ISCED 3 No qualification: ISCED 1										
YFS	Based on 2 questions: Q1: Highest degree of completed studies? 1) Vocational school 2) Occupational / vocational college 3) University of applied sciences 4) University studies (no final degree) 5) Lower university degree (Bachelors degree) 6) Higher university degree (Masters) 7) Licentiate degree 8) Doctoral degree Q2: What is your basic education? 1) Comprehensive school (9 years) 1) Previous form of comprehensive school (up to 8 years) 2) High-school (after comprehensive school	Q1: 1) ISCED 3 2) ISCED 4 3) ISCED 5 4) ISCED 5 5) ISCED 5 6) ISCED 5 7) ISCED 6 8) ISCED 6 Q2: 1) ISCED 2 2) ISCED 2 3) ISCED 3	Females Males Pooled	0 0 0	0 0 0	28 45 73	307 379 686	319 172 491	423 301 724	37 18 55	1,114 915 2,029	654 596 1,250	460 319 779

	3 years)											
Discovery stage total			Females	702	6,649	13,075	20,877	7,063	11,139	1000	60,505	48,366	12,139
			Males	234	4,491	6,547	14,405	4,551	9,144	1,192	40,564	30,228	10,336
			Pooled	936	11,140	19,622	35,282	11,614	20,283	2,192	101,069	78,594	22,475
Replication Stage (in-silico GWA studies)													
DHS	Two questions, the first question is concerned with primary and secondary education. The other question is concerned with universities, vocational full-time schools and the like. Combination of these two gives: 1) School not completed and no vocational training 2) Lower secondary school without vocational traing and year of birth < 1954 3) Lower secondary school and birthyear ≥ 1954 or intermediate school certificate, but no vocational education 4) Lower secondary school plus vocational training 5) upper secondary school without vocational training 6) upper secondary school and vocational	1) ISCED 1 2) ISCED 1 3) ISCED 2 4) ISCED 3 5) ISCED 3 6) ISCED 4 7) ISCED 5	Females	0	67	17	306	45	66	0	501	435	66
			Males	0	25	9	284	33	101	0	452	351	101
			Pooled	0	92	26	590	78	167	0	953	786	167

	training 7) university or university of applied science												
EGCUT	How many years of schooling do you have and whati is the highest level school graduated? 1.Elementary school 2. Secondary school 3.Professional secondary school 4. High school 5. Professional high school 6. Professional higher education 7. Lower university degree (BcS and McS) 8. Higer university degree (PhD or MD)	If 0-3: ISCED 0 If 4 or Primary School: ISCED 1 If 9 or Secondary: ISCED 2 If 12 or High School: or Professional Secondary: SCED 3 If 13-15: ISCED 4 If 16-20 or lower university degree: ISCED 5 If 21+ or higher university degree (PhD, MD) : ISCED 6	Females Males Pooled	16 12 28	122 142 264	313 488 801	900 996 1896	0 0 0	312 403 715	11 40 51	1674 2081 3755	1,351 1,638 2,989	323 443 766
H2000-Cases	Q1: What is your basic education level? 1)Less than grammar school 2) Grammar school 3)Two additional years of education after grammar school in “jatkokoulu” (could be roughly translated to civics school) 4)Part of middle school or comprehensive school (less than 9 years) 5)Middle school 6) Comprehensive school	Q1: 1) ISCED 0 2) ISCED 1 3) ISCED 1 4) ISCED 1 5) ISCED 2 6) ISCED 2 7) ISCED 2 8) ISCED 3 Q2: 1)ISCED level depends on basic level education (Q1) 2)ISCED level depends on basic level education (Q1) 3) ISCED 3	Females Males Pooled	3 3 6	133 112 245	50 23 73	63 107 170	139 121 260	43 54 97	0 1 1	431 421 852	388 366 754	43 55 98

	7)Part of high school or high school examination 8)Matriculation examination Q2: What is your highest level of education or degree after basic education? 1)No vocational education 2) Vocational course or course at work 3)Vocational school or apprenticeship 4)Vocational college (ie. technical school) 5) College degree 6) Vocational programme for specialist vocational qualifications 7)University of applied sciences 8) Lower university degree (Bachelors degree) 9) Higher university degree (Masters degree) 10) Licentiate degree 11)Doctoral degree	4) ISCED 4 5) ISCED 4 6) ISCED 4 7) ISCED 5 8) ISCED 5 9) ISCED 5 10) ISCED 6 11) ISCED 6											
H2000-Controls	Q1: What is your basic education level? 1)Less than grammar school 2) Grammar school 3)Two additional years of education after grammar school in “jatkokoulu” (could	Q1: 1) ISCED 0 2) ISCED 1 3) ISCED 1 4) ISCED 1 5) ISCED 2 6) ISCED 2 7) ISCED 2 8) ISCED 3	Females Males Pooled	2 5 7	110 116 226	33 20 53	88 98 186	148 116 264	60 59 119	4 5 9	445 419 864	381 355 736	64 64 128

	be roughly translated to civics school) 4)Part of middle school or comprehensive school (less than 9 years) 5)Middle school 6) Comprehensive school 7)Part of high school or high school examination 8)Matriculation examination Q2: What is your highest level of education or degree after basic education? 1)No vocational education 2) Vocational course or course at work 3)Vocational school or apprenticeship 4)Vocational college (ie. technical school) 5) College degree 6) 7)University of applied sciences 8) Lower university degree (Bachelors degree) 9) Higher university degree (Masters degree) 10) Licentiate degree 11)Doctoral degree	Q2: 1)ISCED level depends on basic level education (Q1) 2)ISCED level depends on basic level education (Q1) 3) ISCED 3 4) ISCED 4 5) ISCED 4 6) ISCED 4 7) ISCED 5 8) ISCED 5 9) ISCED 5 10) ISCED 6 11) ISCED 6											
HCS	What is your highest level of education? 1) Primary schooling	1) ISCED 1 2) ISCED 2 3) ISCED 3	Females Males Pooled	0 0 0	12 18 30	136 102 238	188 98 286	79 192 271	118 151 269	0 0 0	533 561 1,094	415 410 825	118 151 269

	only 2) Secondary schooling not completed. 3) Secondary schooling completed 4) Trade qualification or technical college 5) University or other tertiary study	4) ISCED 4 5) ISCED 5											
HRS	1) What is the highest grade of school or year of college you completed? Highest Degree Attained Used for GED and higher: 2) How many years of schooling do you have?	If Degree =0 & Schoolyears = 0: ISCED 0 If Degree =0 & Schoolyears = 1-6: ISCED 1 If Degree =0 & Schoolyears = 7-9: ISCED 2 If Degree =0 & Schoolyears = 10-12: ISCED 2 If Degree=1 (GED): ISCED 3 If Degree=2 (HS): ISCED 3 If Degree=3 (2 year college): ISCED 4 If Degree=4 (College: ISCED 5 If Degree=5 (Masters): ISCED 5 If Degree=6 (Prof): ISCED 6	Females Males Pooled	5 5 10	24 36 60	624 427 1051	3,084 1,869 4,953	278 157 435	979 946 1,925	42 150 192	5,036 3,590 8,626	4,015 2,494 6,509	,1021 1,096 2,117
MCTFR	What's your highest obtained degree? 1) Less than high school diploma 2) High school diploma, General Educational Development, or high school equivalent	1) ISCED 2 2) ISCED 3 3) ISCED 4 4) ISCED 5 5) ISCED 5/6 (US years=20)	Females Males Pooled	0 0 0	0 0 0	24 27 51	862 768 1630	97 95 192	925 658 1583	153 221 374	2,061 1,769 3,830	983 890 1,873	1,078 879 1,957

	3) Technical degree, business certificate, or business college 4) Bachelor's degree, associate degree, or some college 5) Professional degree (such as master's degree, JD, MD, PhD)												
NIA	How many years of schooling do you have?	If 0-5: ISCED 0 If 6-8: ISCED 1 If 9-11: ISCED 2 If 12-13: ISCED 3 If 14-15: ISCED 4 If 16-20: ISCED 5 If 21+: ISCED 6	Females	1	6	8	115	56	162	2	350	186	164
			Males	0	5	6	79	31	143	8	272	121	151
			Pooled	1	11	14	194	87	305	10	622	307	315
NTR	What is your highest finished education with a diploma? 0) No education finished 1) Primary school only 2) Lower vocational education (LB0) 3) General Secundary education (LAVO, MAVO) 4) Higher secundary education (HAVO, VWO) 5) Intermediate vocational education (MBO) 6) Higher vocational education (HBO) 7) University 8) PhD	0) ISCED 0 1) ISCED 1 2) ISCED 2 3) ISCED 2 4) ISCED 3 5) ISCED 3 6) ISCED 5 7) ISCED 5 8) ISCED 6	Females	8	74	86	297	0	371	27	863	465	398
			Males	4	29	38	139	0	229	15	454	210	244
			Pooled	12	103	124	436	0	600	42	1317	675	642

ORCADES	Q1: What is the highest educational qualification you have obtained? 1) O grades, standard grades, CSE, leaving cert or equivalent 2) Highers, A levels or equivalent 3) Certificates or diplomas, eg City & Guilds, SCOTVEC, SVQ, HNC, HND, etc 4) Bachelor or Master's degree 5) Doctorate or other higher degree 6) Professional qualification, eg accountancy 7) Other, please specify 8) None of these Q2: How old were you when you left the school? As people start school at 5 in Scotland years of education can be derived.	1) ISCED 2 2) ISCED 3 3) ISCED 4 4) ISCED 5 5) ISCED 6 6) ISCED 5 7) ISCED 2 if Age < 17, ISCED 3 if Age >= 17 8) ISCED 2 if Age < 17, ISCED 3 if Age >= 17 9) ISCED 2 if Age < 17, ISCED 3 if Age >= 17	Females	0	0	228	56	100	53	2	439	384	55
			Males	0	0	205	31	94	38	3	371	330	41
			Pooled	0	0	433	87	194	91	5	810	714	96
RS-III	What is the highest education that you finished with a degree? 0) Primary education 1) Lower vocational education 2) intermediate secundary education 3) Intermediate vocational education 4) General secondary education	0) ISCED 1 1) ISCED 2 2) ISCED 2 3) ISCED 3 4) ISCED 3 5) ISCED 5 6) ISCED 5	Females	0	69	256	115	0	112	0	552	440	112
			Males	0	42	110	128	0	144	0	424	280	144
			Pooled	0	111	366	243	0	256	0	976	720	256

	5) Higher education (HBO) 6) Higher education (University)												
THISEAS	1) Primary school- 6 years of education 2) Secondary education (high school & lyceum) - 12 years of education 3) Technological school - 14 to 15 years of education 4) Higher education (university)- equal to or more than 16 years of education How many years of education do you have?	If 0-5: ISCED 0 If 6-8: ISCED 1 If 9-11: ISCED 2 If 12: ISCED 3 If 13-15: ISCED 4 If 16-23: ISCED 5 (master) If 21+: ISCED 6 (doctorate)	Females Males Pooled	10 6 16	81 132 213	21 44 65	71 146 217	28 67 95	67 151 218	1 6 7	279 552 831	211 395 606	68 157 225
WASHS	What is the highest level of education you have completed? 1) Did not go to school 2) Primary school (please specify highest Grade completed) 3) Secondary school (please specify highest Years compelted) 4) Tertiary instituion (please specify Degree obtained) 5) Other educational institstution (please specify)	1) ISCED 0 2) ISCED 1 3) ISCED 2 / 3: In some cases, the total number of years of school, including primary, was given, and in other cases, the number of years spent only in secondary school was given. These situations were reviewed by a trained researcher familiar with the Australian school system and the ISCED coding was determined accordingly. 4) ISCED 5 / 6: researcher evaluated the Degree	Females Males Pooled	0 2 2	10 28 38	161 200 361	74 127 201	70 116 186	75 93 168	0 4 4	390 570 960	315 473 788	75 97 172

	5) Evaluated by researcher											
Replication-stage total		Females	45	708	1,957	6,219	1,040	3,343	242	13,554	9,969	3,585
		Males	37	685	1,699	4,870	1,022	3,170	453	11,936	8,313	3,623
		Pooled	82	1,393	3,656	11,089	2,062	6,513	695	25,490	18,282	7,208
Combined-stage total		Females	747	7,357	15,032	27,096	8,103	14,482	1,242	74,059	58,335	15,724
		Males	271	5,176	8,246	19,275	5,573	12,314	1,645	52,500	38,541	13,959
		Pooled	1,018	12,533	23,278	46,371	13,676	26,796	2,887	126,559	96,876	29,683

Table S4. Study-specific age and birth-year statistics.

		N	Age				Birth year			
Study			mean	SD	min	max	mean	SD	min	max
Discovery Stage										
AGES	Females	1,863	76.33	5.54	66	95	1926.86	5.61	1908	1936
	Males	1,349	76.52	5.32	67	94	1926.70	5.38	1910	1935
	Pooled	3,212	76.41	5.45	66	95	1926.80	5.52	1908	1936
ALSPAC	Females	6,919	28.69	4.66	15	44	1962.95	4.68	1948	1977
	Males	0	-	-	-	-	-	-	-	-
	Pooled	6,919	28.69	4.66	15	44	1962.95	4.68	1948	1977
ASPS	Females	482	65.95	8.12	50	85	1931.98	6.23	1909	1945
	Males	366	64.93	7.77	46	82	1931.96	6.17	1913	1949
	Pooled	848	65.51	7.98	46	85	1931.97	6.21	1909	1949
BLSA	Females	371	69.18	16.71	30	101	1936.80	16.66	1904	1977
	Males	450	73.66	14.24	31	96	1931.51	14.94	1902	1977
	Pooled	821	71.63	15.55	30	101	1933.90	15.95	1902	1977
CAHRES-Cases	Females	788	78.74	6.26	66	92	1931.41	6.26	1919	1944
	Males	0	-	-	-	-	-	-	-	-
	Pooled	788	78.74	6.26	66	92	1931.41	6.26	1919	1944
CAHRES-Controls	Females	709	79.05	6.32	66	91	1931.12	6.31	1919	1944
	Males	0	-	-	-	-	-	-	-	-
	Pooled	709	79.05	6.32	66	91	1931.12	6.31	1919	1944
CAPS-Cases	Females	0	-	-	-	-	-	-	-	-
	Males	240	68.80	7.73	50	81	1933.00	7.83	1921	1953
	Pooled	240	68.80	7.73	50	81	1933.00	7.83	1921	1953
CAPS-Controls	Females	0	-	-	-	-	-	-	-	-
	Males	219	66.67	7.42	49	80	1935.43	7.50	1922	1954
	Pooled	219	66.67	7.42	49	80	1935.43	7.50	1922	1954
CCF	Females	118	62.47	10.01	30	83	1944.25	9.97	1923	1978
	Males	367	58.39	9.56	31	84	1948.34	9.62	1923	1976
	Pooled	485	59.39	9.82	30	84	1947.34	9.86	1923	1978
CoLaus	Females	2,863	53.88	10.72	35	75	1950.64	10.82	1928	1970
	Males	2,547	52.92	10.77	34	75	1951.62	10.84	1928	1970
	Pooled	5,410	53.43	10.75	34	75	1951.10	10.84	1928	1970
Cr_Kor	Females	541	56.52	12.58	30	98	1950.48	12.58	1909	1977
	Males	302	59.04	12.73	30	90	1947.96	12.73	1917	1977
	Pooled	843	57.42	12.68	30	98	1949.58	12.68	1909	1977
Cr_Spl	Females	242	53.51	11.07	30	79	1955.72	11.07	1930	1979
	Males	175	52.54	12.76	30	85	1956.37	12.76	1924	1979
	Pooled	417	53.10	11.74	30	85	1955.90	11.74	1924	1979
Cr_Vis	Females	499	58.48	14.41	30	93	1944.52	14.41	1910	1973
	Males	365	57.82	13.01	30	88	1945.18	13.01	1915	1973

	Pooled	864	58.20	13.83	30	93	1944.80	13.83	1910	1973
EGCUT	Females	811	60.60	18.63	30	103	1945.25	18.60	1905	1979
	Males	726	48.24	13.65	30	90	1957.24	14.29	1913	1979
	Pooled	1,537	55.94	18.63	30	103	1949.75	18.59	1905	1979
ERF	Females	1,307	51.59	12.35	30	86	1951.73	12.49	1914	1974
	Males	1,073	51.98	12.23	30	88	1951.41	12.43	1915	1974
	Pooled	2,380	51.77	12.29	30	89	1951.59	12.46	1914	1974
FINRISK	Females	850	57.86	11.05	30	74	1944.48	12.45	1923	1977
	Males	973	54.90	11.54	30	74	1947.60	11.85	1923	1977
	Pooled	1,823	46.28	11.40	30	74	1946.14	12.23	1923	1977
FTC	Females	270	55.06	4.47	41	64	1948.70	5.07	1938	1961
	Males	459	54.77	4.49	45	67	1948.95	4.80	1937	1957
	Pooled	729	54.88	4.48	41	67	1948.86	4.90	1937	1961
GAIN	Females	604	54.07	13.84	30	89	1951.93	13.84	1917	1976
	Males	560	56.99	14.50	30	90	1949.01	14.50	1916	1976
	Pooled	1,164	55.48	14.23	30	90	1950.52	14.23	1916	1976
GENOA	Females	794	55.08	10.85	25	83	1942.92	11.14	1914	1974
	Males	645	55.63	10.78	27	90	1942.19	10.98	1908	1972
	Pooled	1,439	55.33	10.82	24	89	1942.59	11.08	1908	1974
HABC	Females	782	73.63	2.79	69	80	1923.31	2.83	1917	1928
	Males	877	73.90	2.88	69	80	1923.06	2.92	1917	1928
	Pooled	1,659	73.77	2.84	69	80	1923.18	2.88	1917	1928
HBCS	Females	982	61.53	3.04	56	69	1940.75	3.04	1934	1944
	Males	735	61.40	2.74	57	69	1940.99	2.67	1934	1944
	Pooled	1,717	61.47	2.92	56	69	1940.85	2.82	1934	1944
InCHIANTI	Females	647	70.70	13.46	30	102	1927.75	13.44	1896	1969
	Males	517	69.01	13.02	30	97	1929.46	13.04	1902	1970
	Pooled	1,164	69.95	13.29	30	102	1928.51	13.29	1896	1970
KORA S3	Females	801	53.03	8.98	30	69	1940.97	8.98	1925	1964
	Males	794	53.54	9.40	30	69	1940.46	9.40	1925	1964
	Pooled	1,595	53.28	9.20	30	69	1940.72	9.20	1925	1964
KORA S4	Females	929	53.65	8.75	32	74	1946.35	8.75	1926	1968
	Males	880	54.23	8.88	31	72	1945.77	8.88	1928	1969
	Pooled	1,809	53.93	8.82	31	74	1946.07	8.82	1926	1969
LBC1921	Females	301	79.10	0.57	77	80	1921.00	0.00	1921	1921
	Males	214	79.11	0.59	77	80	1921.00	0.00	1921	1921
	Pooled	515	79.10	0.58	77	80	1921.00	0.00	1921	1921
LBC1936	Females	495	69.61	0.84	67	71	1936.00	0.00	1936	1936
	Males	508	69.59	0.84	67	71	1936.00	0.00	1936	1936
	Pooled	1,003	69.60	0.84	67	71	1936.00	0.00	1936	1936
LifeLines	Females	4,260	48.35	10.16	30	89	1960.18	9.91	1920	1980
	Males	3,233	48.86	10.51	30	87	1959.69	10.27	1922	1980
	Pooled	7,493	48.57	10.31	30	89	1959.97	10.07	1920	1980

MoBa-Cases	Females	354	31.80	1.38	30	34	1971.60	2.20	1966	1976
	Males	0	-	-	-	-	-	-	-	-
	Pooled	354	31.80	1.38	30	34	1971.60	2.20	1966	1976
MoBa-Controls	Females	405	31.80	1.37	30	34	1971.70	2.10	1966	1976
	Males	0	-	-	-	-	-	-	-	-
	Pooled	405	31.80	1.37	20	34	1971.70	2.10	1966	1976
NESDA	Females	993	46.22	9.45	30	65	1958.94	9.52	1939	1976
	Males	524	47.59	9.11	30	64	1957.54	9.12	1940	1976
	Pooled	1,517	46.69	9.35	30	65	1959.45	9.40	1939	1976
NFBC1966	Females	2,799	31.00	0.00	31	31	1966.00	0.00	1966	1966
	Males	2,572	31.00	0.00	31	31	1966.00	0.00	1966	1966
	Pooled	5,371	31.00	0.00	31	31	1966.00	0.00	1966	1966
nonGAIN	Females	526	52.11	14.02	30	90	1953.89	14.02	1916	1976
	Males	583	53.84	13.69	30	87	1952.16	13.69	1913	1976
	Pooled	1,109	53.02	13.88	30	90	1952.98	13.87	1916	1976
NTR	Females	1,594	50.16	12.08	30	91	1955.82	12.41	1917	1979
	Males	1,056	53.25	12.86	30	81	1952.32	12.58	1923	1980
	Pooled	2,650	51.39	12.49	30	91	1954.43	12.59	1917	1980
QIMR	Females	4,544	44.82	10.19	30	101	1951.06	11.68	1900	1975
	Males	3,441	45.12	9.98	30	101	1952.89	10.52	1900	1975
	Pooled	7,985	44.95	10.09	30	101	1951.85	11.24	1900	1975
RS-I	Females	3,415	69.97	9.42	55	99	1921.60	9.56	1893	1938
	Males	2,391	68.05	8.09	55	95	1923.66	8.33	1983	1938
	Pooled	5,806	69.18	8.95	55	99	1922.45	9.13	1893	1938
RS-II	Females	859	64.54	7.66	55	96	1935.04	8.41	1906	1944
	Males	782	65.36	8.57	55	95	1935.82	7.51	1907	1944
	Pooled	1,641	64.97	8.15	55	95	1935.41	8.00	1906	1944
RS-III	Females	1130	56.21	6.08	45	97	1950.50	60.4	1910	1960
	Males	884	55.99	5.51	45	84	1950.70	5.41	1922	1960
	Pooled	2,014	56.11	5.84	45	97	1950.60	5.77	1910	1960
RUSH-MAP	Females	643	80.99	6.90	55	101	1921.61	7.28	1901	1948
	Males	245	81.38	5.99	64	95	1920.98	6.68	1906	1939
	Pooled	888	81.10	6.66	55	101	1921.44	7.12	1901	1948
RUSH-ROS	Females	532	76.28	7.36	60	95	1921.22	9.12	1901	1946
	Males	278	74.59	7.21	64	102	1921.76	8.18	1896	1940
	Pooled	810	75.70	7.35	60	102	1921.41	8.81	1896	1946
SAGE	Females	845	38.59	5.82	30	65	1965.28	5.95	1938	1975
	Males	476	38.89	5.44	30	63	1964.91	5.58	1940	1975
	Pooled	1,321	38.70	5.68	30	65	1965.15	5.82	1938	1975
SardiNIA	Females	2,055	51.96	14.00	30	101	1955.52	14.22	1900	1980
	Males	1,584	53.91	14.49	30	94	1953.34	14.72	1909	1980
	Pooled	3,639	52.81	14.25	30	101	1954.57	14.48	1900	1980
SHIP	Females	1,794	52.22	13.95	30	81	1946.04	13.08	1918	1971

	Males	1,762	54.34	14.14	30	81	1943.97	14.42	1918	1971
	Pooled	3,556	53.27	14.08	30	81	1945.02	14.13	1918	1971
STR	Females	5,056	63.41	8.81	47	89	1941.59	8.81	1916	1958
	Males	4,497	64.28	8.64	47	89	1940.72	8.64	1916	1958
	Pooled	9,553	63.82	8.74	47	89	1941.18	8.74	1916	1958
TwinsUK	Females	2,619	51.03	10.72	30	80	1949.39	11.14	1919	1978
	Males	0	-	-	-	-	-	-	-	-
	Pooled	2,619	51.03	10.72	30	80	1949.39	11.14	1919	1978
YFS	Females	1,114	37.73	4.98	30	45	1969.27	4.98	1962	1977
	Males	915	37.70	5.04	30	45	1969.30	5.04	1962	1977
	Pooled	2,029	37.72	5.01	30	45	1969.28	5.01	1962	1977
Replication Stage (in-silico GWA studies)										
DHS	Females	501	53.08	12.50	30	74	1950.13	12.47	1929	1974
	Males	452	55.49	12.30	30	74	1947.69	12.35	1929	1974
	Pooled	953	54.22	12.46	30	74	1948.97	12.46	1929	1974
EGCUT	Females	1674	58.72	15.71	30	99	1948.24	15.75	1910	1979
	Males	2081	59.97	15.79	30	100	1947.30	15.81	1910	1980
	Pooled	3755	59.87	15.76	30	100	1947.68	15.78	1910	1980
H2000-Cases	Females	431	52.37	11.72	30	75	1947.14	11.75	1924	1970
	Males	421	49.25	10.45	30	75	1950.26	10.44	1924	1970
	Pooled	852	50.83	11.21	30	75	1948.68	1.22	1924	1970
H2000-Controls	Females	445	51.98	11.59	30	75	1947.54	11.59	1924	1970
	Males	419	49.26	10.39	30	75	1950.25	10.37	1925	1970
	Pooled	864	50.66	11.10	30	75	1948.86	11.09	1924	1970
HCS	Females	533	65.69	7.15	55	86	1940.08	7.34	1921	1951
	Males	561	66.55	7.80	55	85	1939.28	7.90	1920	1951
	Pooled	1,094	66.13	7.50	55	86	1939.67	7.63	1920	1951
HRS	Females	5,036	68.33	10.82	33	101	1938.07	10.81	1905	1974
	Males	3,590	68.99	9.83	37	107	1937.40	9.83	1900	1970
	Pooled	8,626	68.61	10.42	33	107	1937.79	10.42	1900	1974
MCTFR	Females	2061	42.85	5.30	30	60	1954.39	6.41	1934	1974
	Males	1769	44.91	5.67	30	65	1952.23	6.72	1926	1972
	Pooled	3830	43.80	5.57	30	65	1953.40	6.64	1926	1974
NIA	Females	350	75.78	9.15	42	103	1932.61	10.20	1903	1958
	Males	272	76.59	7.97	52	103	1933.25	10.50	1903	1962
	Pooled	622	76.10	8.71	42	103	1932.89	10.33	1903	1962
NTR	Females	863	42.27	10.83	30	88	1964.41	11.31	1926	1980
	Males	454	43.06	11.67	30	78	1963.32	12.31	1928	1979
	Pooled	1317	42.54	11.13	30	88	1964.03	11.67	1926	1980
ORCADES	Females	439	54.48	14.25	25	91	1951.68	14.27	1914	1979
	Males	371	56.42	13.98	27	90	1949.71	14.05	1915	1979
	Pooled	810	55.37	14.15	25	92	1950.8	14.19	1914	1979
RS-III	Females	540	58.67	7.90	46	87	1948.52	7.78	1921	1960

	Males	414	59.72	8.32	45	89	1947.67	8.10	1918	1960
	Pooled	976	59.26	8.15	45	89	1948.04	7.97	1918	1960
THISEAS	Females	279	57.58	13.41	30	87	1949.84	13.70	1909	1979
	Males	552	56.95	11.67	31	89	1950.02	11.62	1920	1978
	Pooled	831	57.16	12.28	30	89	1949.96	12.35	1909	1979
WASHS	Females	390	53.33	12.38	30	92	1954.35	12.36	1915	1980
	Males	570	53.10	12.55	30	83	1954.32	12.45	1925	1980
	Pooled	960	53.19	12.47	30	92	1954.33	12.41	1915	1980

Table S5. Information on genotyping methods, quality control of SNPs, imputation, and statistical analyses. "Call rate" refers to the genotyping success rate, i.e., the minimum percentage of successfully genotyped SNPs. "SNPs in analysis after QC" includes the removal of non-HapMap SNPs and technical artifacts, such as SNPs with missing effect size, standard error, etc.; in other words, it is the number of HapMap SNPs that could be handled by METAL.

	Genotyping						Imputation			Association analysis				
		Inclusion criteria					Inclusion criteria							
Study	Platform	Genotyping calling algorithm	MAF	Call rate	*p* for HWE	SNPs that met QC criteria	Imputation software	MAF	Imputation quality	Sample	SNPs in analysis after QC	λ	Analysis software	Additional covariates
Discovery stage														
AGES	Illumina Human370CNV	BeadStudio	≥1%	≥97%	$\geq 10^{-6}$	326,034	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			ProbABEL	
										Females	2,385,826	1.038		
										Males	2,385,826	1.017		
										Pooled	2,385,826	1.054		
										College				
										Females	2,385,826	1.015		
										Males	2,385,826	1.023		
										Pooled	2,385,826	1.015		
ALSPAC	Illumina Human550 quad array	Illumina	≥1%	≥95%	$\geq 10^{-7}$	526,688	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			MACH2DAT MACH2QTL	
										Females	2,437,714	1.045		
										Males	-	-		
										Pooled	-	-		
										College				
										Females	2,437,714	1.048		
										Males	-	-		
										Pooled	-	-		
ASPS	Illumina Human610-Quad BeadChip	Illumina	≥1%	≥98%	$\geq 10^{-6}$	550,635	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			GenABEL	
										Females	2,433,424	1.032		
										Males	2,433,424	1.030		
										Pooled				
										College				
										Females	2,432,966	1.092		
										Males	2,433,424	1.039		
										Pooled	-	-		
BLSA	Illumina 550K	Beadstudio	≥1%	≥99%	$\geq 10^{-5}$	514,027	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			Merlinoffline/ ProbABEL	
										Females	2,441,287	0.995		
										Males	2,441,287	1.024		
										Pooled	-	-		
										College				
										Females	2,438,930	1.014		
										Males	2,437,102	1.038		
										Pooled	-	-		

CAHRES-Cases	Illumina HumanHap300+ 240S	BeadStudio	≥3%	≥90%	$\geq 10^{-6}$	510,578	IMPUTE	≥2.5%	$R^2 \geq 0.4$	*EduYears* Females Males Pooled *College* Females Males Pooled	 2,309,251 - - 2,309,251 - -	 0.995 - - 1.023 - -	SNPTEST	For age only ((Birth-year – 1900) $/10)^3$ included
CAHRES-Controls	Illumina HumanHap300+ 240S	BeadStudio	≥3%	≥90%	$\geq 10^{-6}$	512,223	IMPUTE	≥2.5%	$R^2 \geq 0.4$	*EduYears* Females Males Pooled *College* Females Males Pooled	 2,334,910 - - 2,334,910 - -	 1.020 - - 1.037 - -	SNPTEST	For age only ((Birth-year – 1900) $/10)^3$ included
CAPS-Cases	Affymetrix GeneChip Human 500K	BRLMM	≥1%	≥95%	$\geq 10^{-6}$	330,124	IMPUTE	≥5%	$R^2 \geq 0.4$	*EduYears* Females Males Pooled *College* Females Males Pooled	 - 2,101,503 - - 2,101,503 -	 - 1.021 - - 1.060 -	SNPTEST	For age only ((Birth-year – 1900) $/10)^3$ included
CAPS-Controls	Affymetrix GeneChip Human 500K	BRLMM	≥1%	≥95%	$\geq 10^{-6}$	330,124	IMPUTE	≥5%	$R^2 \geq 0.4$	*EduYears* Females Males Pooled *College* Females Males Pooled	 - 2,101,359 - - 2,01,359 -	 - 1.017 - - 1.105 -	SNPTEST	For age only ((Birth-year – 1900) $/10)^3$ included
CCF	Illumina Hap550 v1 or v3 and Hap610 v1	GenCall	≥1%	≥97%	FDR < 0.20	479,618	MACH	≥1%	$R^2 \geq 0.4$	*EduYears* Females Males Pooled *College* Females Males Pooled	 2,416,880 2,428,591 - 2,415232 2,428,498 -	 0.994 1.016 - 1.061 1.028 -	ProbABEL, R	
CoLaus	Affymetrix GeneChip Human	BRLMM	≥1%	≥90%	$\geq 10^{-6}$	390,631	IMPUTE	≥1%	$R^2 \geq 0.4$	*EduYears* Females Males	 2,353,219 2,353,775	 1.037 1.017	Matlab	

	Mapping 500K									Pooled	-	-	
										College			
										Females	2,353,219	1.031	
										Males	2,353,775	1.023	
										Pooled	-	-	
Cr_Kor	Illumina Hap370CNV	GenomeStudio	≥1%	≥98%	$\geq 10^{-6}$	307,625	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			ProbABEL
										Females	2,341,221	1.016	
										Males	2,337,497	1.019	
										Pooled	2,342,048	1.010	
										College			
										Females	2,341,221	1.022	
										Males	2,337,497	1.019	
										Pooled	2,342,048	1.030	
Cr_Spl	Illumina Hap370CNV	GenomeStudio	≥1%	≥98%	$\geq 10^{-6}$	321,456	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			ProbABEL
										Females	2,385,908	1.007	
										Males	2,383,368	1.023	
										Pooled	2,387,759	1.005	
										College			
										Females	2,385,908	1.006	
										Males	2,383,368	1.027	
										Pooled	2,387,759	1.016	
Cr_Vis	Illumina Hap300v1	BeadStudio	≥1%	≥98%	$\geq 10^{-6}$	285,491	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			ProbABEL
										Females	2,380,057	1.001	
										Males	2,376,145	1.003	
										Pooled	2,379,760	1.002	
										College			
										Females	2,380,057	1.028	
										Males	2,376,145	1.029	
										Pooled	2,379,760	1.008	
EGCUT	370CNV	GenomeStudio	≥1%	≥95%	$\geq 10^{-6}$	311,028	IMPUTE	≥1%	$R^2 \geq 0.4$	*EduYears*			SNPTEST
										Females	2,340,499	1.009	
										Males	2,340,416	1.028	
										Pooled	-	-	
										College			
										Females	2,338,956	1.018	
										Males	2,339,313	1.006	
										Pooled	-	-	
ERF	Illumina 6K, 318K, 370K, Affymetrix 250K, Illumina610K	GenCall & BRLMM	≥1%	≥98%	$\geq 10^{-6}$	487,573	MACH 1.0.16	≥1%	$R^2 \geq 0.4$	*EduYears*			ProbABEL
										Females	2,394,464	1.021	
										Males	2,394,464	1.022	
										Pooled	2,394,464	1.042	
										College			
										Females	2,394,100	0.999	

										Males	2,3944,12	1.132	
										Pooled	2,394,455	1.155	
FINRISK	Illumina Human610-Quad BeadChip	Illuminus	≥5%	≥95%	$\geq 10^{-7}$	554,988	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			ProbABEL
										Females	2,415,737	1.009	
										Males	2,415,737	1.008	
										Pooled	-	-	
										College			
										Females	2,415,737	1.020	
										Males	2,415,737	1.019	
										Pooled	-	-	
FTC	Illumina Human670-QuadCustom	Illuminus	≥1%	≥95%	$\geq 10^{-6}$	549,060	IMPUTE	≥1%	$R^2 \geq 0.4$	*EduYears*			SNPTEST
										Females	2,407,305	0.993	
										Males	2,409,575	1.003	
										Pooled	-	-	
										College			
										Females	-	-	
										Males	-	-	
										Pooled	-	-	
GAIN	Affymetrix _SNP6_build36.1	Birdseed	≥5%	≥95%	$\geq 10^{-7}$	649,668	IMPUTE	≥1%	$R^2 \geq 0.4$	*EduYears*			SNPTEST
										Females	2,409,360	1.002	
										Males	2,408,859[3]	1.004	
										Pooled	-	-	
										College			
										Females	2,409,360	1.013	
										Males	2,408,860	1.025	
										Pooled	-	-	
GENOA	Affymetrix 6.0 and Illumina 1M-Duo BeadChip	Birdseed & Genome Studio	≥1%	≥95%	NA	Affymetrix: 596,941; Illumina: 804,154	MACH	2.5%	$R^2 \geq 0.4$	*EduYears*			R
										Females	2,275,983	0.989	
										Males	2,275,983	0.993	
										Pooled	2,275,983	0.988	
										College			
										Females	2,275,975[4]	1.126	
										Males	2,275,982[5]	1.132	
										Pooled	2,275,983	1.244	
HABC	Illumina Human 1M -Duo	Beadstudio	≥1%	≥97%	$\geq 10^{-6}$	914,263	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			R
										Females	2,460,895	1.024	
										Males	2,460,895	1.006	
										Pooled	-	-	
										College			

[3] This number includes the removal of rs1259286, p-value = 6.28^{-13}, R^2 = 0.4005.
[4] This number includes the removal of rs10161503, rs11112192, rs1240825, rs17046226, rs17273464, rs2966996, rs770931 and rs9810816, all with Beta > abs(10^{14}).
[5] This number includes the removal of rs17331350, Beta = -6.93^{14}, MAF = 0.028.

										Females	2,460,895	1.025		
										Males	2,460,895	1.025		
										Pooled	-	-		
HBCS	Modified Illumina Infinum 610K Quad	BeadStudio	≥5%	≥95%	$\geq 10^{-6}$	509,947	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			PLINK (directly genotyped), ProbABEL (imputed)	
										Females	2,417,111	1.009		
										Males	2,417,111	1.020		
										Pooled				
										College				
										Females	2,416,556	1.020		
										Males	2,416,556	1.026		
										Pooled	-	-		
InCHIANTI	Illumina 550K	Beadstudio	≥1%	≥99%	$\geq 10^{-7}$	498,838	MACH	≥5%	$R^2 \geq 0.4$	*EduYears*			Merlinoffline/ ProABEL	Study site
										Females	2,168,258	1.005		
										Males	2,168,258	1.019		
										Pooled				
										College				
										Females	2,165,506	0.972		
										Males	2,165,543	1.019		
										Pooled	-	-		
KORA S3	Affymetrix 500k	BRLMM	≥1%	≥95%	$\geq 10^{-6}$	379,392	IMPUTE	≥2,5%	$R^2 \geq 0.4$	*EduYears*			QUICKTEST	World War 2 dummy (born between 1919 and 1937)
										Females	2,276,751	1.018		
										Males	2,276,573	1.007		
										Pooled	-	-		
										College				
										Females	2,275,449	1.034		
										Males	2,276,573	1.004		
										Pooled	-	-		
KORA S4	Affymetrix 6.0	Birdseed2	None	None	None	909,622	IMPUTE	≥2,5%	$R^2 \geq 0.4$	*EduYears*			QUICKTEST	World War 2 dummy (born between 1919 and 1937)
										Females	2,338,017[6]	1.016		
										Males	2,339,187	1.012		
										Pooled	-	-		
										College				
										Females	2,337,953	1.031		
										Males	2,339,189	1.017		
										Pooled	-	-		
LifeLines	Illumina CytoSNP v 2.0-300K	GenomeStudio	≥1%	≥95%	$\geq 10^{-4}$	254,374	BEAGLE	≥1%	$R^2 \geq 0.4$	*EduYears*			PLINK (dosage module)	First 10 PC's instead of 4 PC's
										Females	2,024,591	1.034		
										Males	2,025,047	1.034		
										Pooled	2,024,909	1.079		
										College				
										Females	2,024,591	1.033		

[6] This number includes the removal of rs12123886: *p*-value 9.28^{-14}, MAF 0.030, $R^2 = 0.48$.

										Males	2,025,047	1.024		
										Pooled	2,024,909	1.054		
LBC1921	Illumina 610 quad v1	GenomeStudio	≥1%	≥98%	$\geq 10^{-3}$	535,709	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			MACH2QTL	Age in days instead of years due to cohort setup
										Females	2,432,460	1.010		
										Males	2,432,460	1.003		
										Pooled	-	-		
										College				
										Females	-	-		
										Males	-	-		
										Pooled	-	-		
LBC1936	Illumina 610 quad v1	GenomeStudio	≥1%	≥98%	$\geq 10^{-3}$	535,709	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			MACH2QTL	Age in days instead of years due to cohort setup
										Females	2,433,592	1.011		
										Males	2,433,592	1.020		
										Pooled	-	-		
										College				
										Females	2,428,839	1.018		
										Males	2,431,922	1.024		
										Pooled	-	-		
MoBa-Cases	Illumina 660W quad	GenCall	≥0.5%	≥95%	$\geq 10^{-4}$	453,126	PLINK	≥1%	$R^2 \geq 0.4$	*EduYears*			PLINK	
										Females	1,855,625[7]	1.028		
										Males	-	-		
										Pooled	-	-		
										College				
										Females	-	-		
										Males	-	-		
										Pooled	-	-		
MoBa-Controls	Illumina 660W quad	GenCall	≥0.5%	≥95%	$\geq 10^{-4}$	453,126	PLINK	≥1%	$R^2 \geq 0.4$	*EduYears*			PLINK	
										Females	1,852,751[8]	0.995		
										Males	-	-		
										Pooled	-	-		
										College				
										Females	-	-		
										Males	-	-		
										Pooled	-	-		
NESDA	Of 1517 subjects 1433 were genotyped on Perlegen	Perlegen proprietary algorithm	≥1%	≥95%	$\geq 10^{-6}$	435,291	IMPUTE	≥1%	$R^2 \geq 0.4$	*EduYears*			SNPTEST	
										Females	2,366,555	1.005		
										Males	2,364,096	1.006		
										Pooled	-	-		

[7] Results for MoBa-Cases were additionally filtered on callrate ≥95%
[8] Results for MoBa-Controls were additionally filtered on callrate ≥95%.

	600k and 84 on Affymetrix 6.0.									*College*				
										Females	2,366,444	1.021		
										Males	2,364,124	1.029		
										Pooled	-	-		
NFBC1966	Illumina HumanCNV-370DUO Analysis BeadChip	Beadstudio	≥1%	95%, for MAF< 5% call rate ≥99%	≥5.7× 10^{-7}	324,896	IMPUTE	≥2,5%	R^2≥0.4	*EduYears*			SNPTEST	
										Females	2,290,644[9]	1.009		
										Males	2,290,602	1.022		
										Pooled	-	-		
										College				
										Females	2,290,544	0.996		
										Males	2,290,273	1.027		
										Pooled	-	-		
nonGAIN	Affymetrix_SNP6_build36.1	Birdseed	≥5%	≥95%	≥10^{-7}	598,153	IMPUTE	≥1%	R^2≥0.4	*EduYears*			SNPTEST	
										Females	2,401,283	1.011		
										Males	2,401,549	0.993		
										Pooled	-	-		
										College				
										Females	2,401,283	1.021		
										Males	2,401,549	1.017		
										Pooled	-	-		
NTR	Perlegen 600k, Illumina 660k, Illumina 370k, Affymetrix 6, Illumina 1m	Perlegen proprietary, Illumina Genome Studio, Affymetrix Genotyping Console	≥1%	≥95%	≥10^{-6}	Per individual : min. = 311,567. max.: 932,824. mean: 481,415.13	IMPUTE	≥1%	R^2≥0.4	*EduYears*			SNPTEST	Dummy for Mammoet Law (born before 1956)
										Females	2,302,982	1.016		
										Males	2,301,882	1.013		
										Pooled				
										College				
										Females	2,303,244	1.012		
										Males	2,302,434	1.027		
										Pooled	-	-		
QIMR	Illumina 610, 370 ,317	BeadStudio	≥1%	≥95%	≥10^{-7}	269,840	MACH	≥1%	R^2≥0.4	*EduYears*			MERLIN - offline	
										Females	2,398,497	1.004		
										Males	2,398,497	1.016		
										Pooled	2,398,497	1.021		
										College				
										Females	2,398,497	0.991		
										Males	2,398,497	1.006		
										Pooled	2,398,497	1.009		
RS-I	Illumina HumanHap 550 V.3	BeadStudio Genecall	≥1%	≥98%	≥10^{-6}	512,349	MACH	≥1%	R^2≥0.4	*EduYears*			MACH2DAT	
										Females	2,433,150	1.025		
										Males	2,433,150	1.002		

[9] This number includes the removal of rs2152709: *p*-value 5.90^{-36}, MAF 0.027, R^2 = 0.54.

										Pooled	-	-		
										College				
										Females	2,432,894	1.027		
										Males	2,433,150	1.012		
										Pooled	-	-		
RS-II	Illumina HumanHap 550 V.3 DUO; Illumina HumanHap 610 QUAD	Genomestudio Genecall	≥1%	≥97.5 %	$\geq 10^{-6}$	466,389	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			MACH2DAT	
										Females	2,432,613	1.015		
										Males	2,432,613	1.004		
										Pooled	-	-		
										College				
										Females	2,431,307	1.021		
										Males	2,432,575	1.021		
										Pooled	-	-		
RS-III	Illumina HumanHap 610 QUAD	Genomestudio Genecall	≥1%	≥97.5 %	$\geq 10^{-6}$	514,073	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			MACH2DAT	Dummy for Mammoet Law (born before 1956)
										Females	2,436,797	1.005		
										Males	2,436,797	1.015		
										Pooled	-	-		
										College				
										Females	2,436,796	1.014		
										Males	2,436,797	1.021		
										Pooled	-	-		
RUSH-MAP	Affymetrix 6.0	Birdsuite, Broad Institute	≥1%	≥95%	$\geq 10^{-6}$	645,349	MACH	≥2,5%	$R^2 \geq 0.4$	*EduYears*			PLINK	
										Females	2,322,227	0.998		
										Males	2,316,021	0.992		
										Pooled	-	-		
										College				
										Females	2,319,944	1.024		
										Males	2,311,832	1.033		
										Pooled	-	-		
RUSH-ROS	Affymetrix 6.0	Birdsuite, Broad Institute	≥1%	≥95%	$\geq 10^{-6}$	645,349	MACH	≥2,5%	$R^2 \geq 0.4$	*EduYears*			PLINK	
										Females	2,319,944	1.011		
										Males	2,319,384	1.026		
										Pooled	-	-		
										College				
										Females	2,319,944	1.024		
										Males	2,319,380	1.026		
										Pooled	-	-		
SAGE	Illumina 1M	BeadStudio	≥1%	≥98%	$\geq 10^{-4}$	948,658	IMPUTE	≥1%	$R^2 \geq 0.4$	*EduYears*			PLINK	In College analysis only three PC's included. Dummy
										Females	2,429,089	1.028		
										Males	2,426,685	1.023		
										Pooled	-	-		
										College				
										Females	2,429,508	1.022		

										Males	2,427,206	1.048		Cocaine-study versus not.
SardiNIA	Affymetrix 10k, 500k, 1M	BRLMM	≥5%	≥95%	$\geq 10^{-7}$	765,419	MACH	≥1%	$R^2 \geq 0.4$	Pooled EduYears	-	-	Merlin	
										Females	2,134,355	1.026		
										Males	2,134,355	1.053		
										Pooled	2,134,355	1.071		
										College				
										Females	-	-		
										Males	-	-		
										Pooled	-	-		
SHIP	Affymetrix Human SNP Array 6.0	Birdseed2	≥0%	>92%	≥0%	869,224	IMPUTE	≥1%	$R^2 \geq 0.4$	*EduYears*			QUICKTEST	World War 2 dummy (for people aged between 6 and 30 in 1939-1945)
										Females	2,430,317	0.994		
										Males	2,430,785	1.018		
										Pooled	-	-		
										College				
										Females	2,430,191	1.009		
										Males	2,430,763	1.020		
										Pooled	-	-		
STR	Illumina HumanOmniExpress-12v1_A	GenomeStudio	≥1%	≥97%	$\geq 10^{-7}$	644,556	IMPUTE	≥1%	$R^2 \geq 0.4$	*EduYears*			Merlin-offline	
										Females	2,440,323	1.014		
										Males	2,440,323	1.017		
										Pooled	2,440,323	1.027		
										College				
										Females	2,440,323	1.012		
										Males	2,440,323	1.012		
										Pooled	2,440,323	1.021		
TwinsUK	HumanHap300 and HumanHap610	Illiminus	≥5%	≥97%	$\geq 10^{-6}$	557,427	IMPUTE	≥1%	$R^2 \geq 0.4$	*EduYears*			SNPTEST	
										Females	2,326,644	1.016		
										Males	-	-		
										Pooled	-	-		
										College				
										Females	2,326,684	1.004		
										Males	-	-		
										Pooled	-	-		
YFS	Illumina custom made BeadChip Human 670K-Quad	Illiminus	≥ 1%	≥ 95%	$\geq 10^{-6}$	546,674	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			PLINK	
										Females	2,409,746	1.002		
										Males	2,409,712	1.005		
										Pooled				

										College				
										Females	2,409,746	1.002		
										Males	2,409,697	1.003		
										Pooled	-	-		
Replication Stage														
DHS	Illumina, HumanOmni2.5 -4v1_D	Genome-Studio	≥5%	≥95%	$\geq 10^{-6}$	1.480.368	MACH	≥5%	$R^2 \geq 0.4$	*EduYears*			PLINK	For age only ((birthyear - 1900)/10)^3) included
										Females	2,140,522	0.987		
										Males	2,139,604	0.992		
										Pooled	-	-		
										College				
										Females	2,140,513	1.016		
										Males	2,139,604	1.007		
										Pooled	-	-		
EGCUT	Illumina OmniExpress	Genome-Studio	≥1%	≥95%	$\geq 10^{-6}$	615,574	IMPUTE	≥1%	$R^2 \geq 0.4$	*EduYears*			SNPTEST	
										Females	2,370,624	1.023		
										Males	2,371,992	1.033		
										Pooled	-	-		
										College				
										Females	2,411,856	1.014		
										Males	2,412,486	1.021		
										Pooled	-	-		
H2000-Cases	Illumina Human610-Quad BeadChip	Illuminus	≥5%	≥95%	$\geq 10^{-6}$	555,418	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			ProbABEL	
										Females	2,462,032	1.011		
										Males	2,461,943	1.001		
										Pooled	-	-		
										College				
										Females	2,456,973	0.987		
										Males	2,458,679	1.005		
										Pooled	-	-		
H2000-Controls	Illumina Human610-Quad BeadChip	Illuminus	≥5%	≥95%	$\geq 10^{-6}$	555,418	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			ProbABEL	
										Females	2,462,169	1.002		
										Males	2,461,678	1.009		
										Pooled	-	-		
										College				
										Females	2,459,673	1.004		
										Males	2,459,363	1.008		
										Pooled	-	-		
HCS	Illumina 610K-Quad	GeneomeStudio GeneCall	≥1%	≥95%	$\geq 10^{-6}$	551,551	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			MACH2DAT, MACH2QTL	
										Females	2,395,501	1.003		
										Males	2,395,501	1.008		
										Pooled	-	-		
										College				
										Females	2,395,590	1.017		

										Males	2,395,771	1.015		
										Pooled	-	-		
HRS	Illumina Omni2.5 Beadchip	GenomeStudio	≥1%	≥98%	$\geq 10^{-4}$	551,936	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			PLINK	
										Females	2,441,592	1.024		
										Males	2,441,232	1.012		
										Pooled	-	-		
										College				
										Females	2,441,592	1.019		
										Males	2,441,232	1.005		
										Pooled	-	-		
MCTFR	Illumina 660W Quad	None; Illumina calls	≥1%	≥99%	$\geq 10^{-6}$	527,829	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			R	
										Females	2,443,350	1.022		
										Males	2,443,258	1.029		
										Pooled	2,443,493	1.020		
										College				
										Females	2,443,350	1.019		
										Males	2,443,258	1.029		
										Pooled	2,443,493	1.021		
NIA	Illumina Human610-Quadv1_B	Illuminus26	≥5%	≥95%	$\geq 10^{-7}$	532,255	IMPUTE	≥2.5%	$R^2 \geq 0.4$	*EduYears*			SNPTEST	
										Females	2,342,357	1.011		
										Males	2,339,767	1.018		
										Pooled	-	-		
										College				
										Females	2,342,358	1.026		
										Males	2,339,767	1.053		
										Pooled	-	-		
NTR	Affymetrix 6	Affymetrix Genotyping Console	≥1%	≥95%	$\geq 10^{-5}$	666,284	BEAGLE was used for phasing, Minimac to impute	≥1%	$R^2 \geq 0.4$	*EduYears*			SNPTEST	Dummy for Mammoet Law (born before 1956)
										Females	2,358,404	0.996		
										Males	2,355,914	1.000		
										Pooled	-	-		
										College				
										Females	2,357,570	0.996		
										Males	2,353,439	1.032		
										Pooled	-	-		
ORCADES	Illumina Hap300	Beadstudio	≥5%	≥98%	$\geq 10^{-6}$	298,785	MACH	≥1%	$R^2 \geq 0.4$	*EduYears*			ProbABEL	
										Females	2,381,185	1.022		
										Males	2,380,826	1.012		
										Pooled	2,382,691	1.003		
										College				
										Females	2,381,184	1.089		
										Males	2,380,824	1.077		
										Pooled	2,382,691	1.107		
RS-III	Illumina	Genomestudio	≥1%	≥97.5	$\geq 10^{-6}$	513,329	N.A.	N.A.	N.A.	*EduYears*			MACH2DAT	Dummy

	HumanHap 610 QUAD	Genecall		%						Females Males Pooled *College* Females Males Pooled	491,225 491,255 - 489,588 489,613 -	0.996 1.011 - 1.011 1.044 -		for Mammoet Law (born before 1956)
THISEAS	Illumina OmniExpress	Illuminus	NA	NA	NA	733,202	N.A.	≥1%	N.A.	*EduYears* Females Males Pooled *College* Females Males Pooled	 595,330 595,339 - 595,331 595,340 -	 1.008 1.028 - 1.011 1.016 -	PLINK	
WASHS	Illumina HumanOmni2.5-8	BeadStudio	≥1%	≥95%	≥5.7×10^{-7}	1,463,846	MACH/minimac	≥1%	$R^2\geq 0.4$	*EduYears* Females Males Pooled *College* Females Males Pooled	 2,455,025 2,455,115 - 2,455,026 2,455,116 -	 1.000 0.986 - 1.022 1.007 -	R	

Table S6. *EduYears* association results for the 4 additional independent loci that reached genome-wide significance ($p < 5\times10^{-8}$) in the combined meta-analysis. I^2 represents the % heterogeneity of effect size between the discovery stage studies. p_{het} is the heterogeneity *p*-value. SNPs are ordered according to ascending *p*-value in the combined stage. The *p*-value in the Replication stage meta-analysis is from a one-sided test.

						Discovery stage				Replication stage		Combined stage			Combined stage – sex-specific			
SNP	Chr	Position (bp)	Nearest gene	Effective allele	Freq[a]	Beta	*P*-value[b]	I^2	P_{het}	Beta	*P*-value[b]	Beta	*P*-value[b]	P_{het}	Beta (Males)	*P*-value[b] (Males)	Beta (Females)	*P*-value[b] (Females)
rs1487441	6	98660615	LOC100129158	A	0.480	0.106	4.36×10^{-9}	6.1	0.333	0.078	1.05×10^{-2}	0.101	3.22×10^{-10}	0.688	0.093	2.51×10^{-4}	0.102	8.50×10^{-7}
rs1056667	6	26618543	BTN1A1	T	0.538	0.074	7.41×10^{-5}	0.0	0.952	0.159	1.59×10^{-6}	0.093	1.86×10^{-8}	0.762	0.128	8.53×10^{-7}	0.066	1.95×10^{-3}
rs11687170	2	236722883	GBX2	T	0.770	0.093	1.68×10^{-5}	9.3	0.262	0.163	7.83×10^{-4}	0.107	3.25×10^{-8}	0.278	0.170	1.00×10^{-7}	0.065	7.35×10^{-3}
rs7309	2	161800886	TANK	A	0.476	-0.085	2.22×10^{-6}	0.0	0.725	-0.093	2.16×10^{-3}	-0.088	3.60×10^{-8}	0.867	-0.115	4.40×10^{-6}	-0.071	5.67×10^{-4}

[a]Frequency in combined stage meta-analysis.

[b]All *p*-values are based on the sample-size weighted meta-analysis (fixed effects).

Table S7. *College* association results for the 3 additional independent loci that reached genome-wide significance ($p < 5\times10^{-8}$) in the combined stage meta-analysis. I^2 represents the % heterogeneity of effect size between the discovery stage studies. p_{het} is the heterogeneity *p*-value. SNPs are ordered according to ascending *p*-value in the combined stage. The *p*-value in the Replication stage meta-analysis is from a one-sided test.

						Discovery stage				Replication stage		Combined stage			Combined stage – sex-specific			
SNP	Chr	Position (bp)	Nearest gene	Effective allele	Freq[a]	OR	*P*-value[b]	I^2	P_{het}	OR	*P*-value[b]	OR	*P*-value[b]	P_{het}	OR (Males)	*P*-value[b] (Males)	OR (Females)	*P*-value[b] (Females)
rs11584700	1	202843606	LRRN2	A	0.780	0.921	2.07×10^{-9}	13.8	0.179	0.912	4.86×10^{-4}	0.919	8.24×10^{-12}	0.221	0.934	6.11×10^{-4}	0.911	2.12×10^{-9}
rs4851264	2	100176620	LOC150577	T	0.605	0.952	2.52×10^{-9}	26.2	0.031	0.952	2.75×10^{-3}	0.952	4.93×10^{-11}	0.046	0.949	1.96×10^{-5}	0.950	4.87×10^{-8}
rs13401104	2	236770257	LOC100128572	A	0.180	0.926	8.37×10^{-6}	6.5	0.330	0.866	1.34×10^{-5}	0.913	4.60×10^{-9}	0.199	0.901	2.99×10^{-5}	0.920	1.72×10^{-5}

[a]Frequency in combined stage meta-analysis.
[b]All *p*-values are based on the sample-size weighted meta-analysis (fixed effects).

Table S8. Comparison of *EduYears* associated SNPs (Table 1 and Supplementary Table 6) in *College* analysis (Combined stage).

	EduYears		*College*	
SNP	Beta	*P*-value	OR	*P*-value
rs1487441	0.101	3.22×10^{-10}	1.035	1.16×10^{-6}
rs9320913	0.101	3.50×10^{-10}	1.035	1.28×10^{-6}
rs1056667	0.093	1.86×10^{-8}	1.029	6.22×10^{-5}
rs11687170	0.107	3.25×10^{-8}	1.075	5.39×10^{-7}
rs7309	-0.088	3.60×10^{-8}	0.971	2.58×10^{-5}
rs3783006	0.088	8.45×10^{-8}	1.031	1.82×10^{-5}
rs8049439	0.086	1.15×10^{-7}	1.029	8.38×10^{-5}
rs13188378	-0.097	1.37×10^{-4}	0.908	6.75×10^{-2}

Table S9. Comparison of College associated SNPs (Table 1 and Supplementary Table 7) in *EduYears* analysis (Combined stage).

	College		*EduYears*	
SNP	OR	*P*-value	Beta	*P*-value
rs11584700	0.919	8.24×10^{-12}	-0.095	3.25×10^{-7}
rs4851264	0.952	4.93×10^{-11}	-0.083	4.17×10^{-7}
rs4851266	1.050	5.33×10^{-11}	0.082	5.61×10^{-7}
rs13401104	0.913	4.60×10^{-9}	-0.107	4.74×10^{-8}
rs2054125	1.376	2.12×10^{-7}	0.105	7.12×10^{-5}
rs3227	1.037	3.24×10^{-7}	0.074	7.58×10^{-6}
rs4073894	1.062	5.55×10^{-6}	0.080	1.88×10^{-5}
rs12640626	1.034	7.48×10^{-6}	0.070	1.31×10^{-5}

Table S10. Previously published twin study findings on the heritability of educational attainment

Country	Gender	Cohort	rMZ	NMZ	rDZ	NDZ	Falconer h^2
Australia *(145)*	Male	1893-1950	0.7	216	0.53	94	0.34
	Female		0.77	520	0.55	299	0.44
Australia *(145)*	Male	1951-1965	0.74	226	0.47	161	0.54
	Female		0.75	479	0.49	290	0.52
Australia *(146)*	Male	1964-1971	0.674	282	0.532	164	0.284
	Female		0.705	320	0.319	158	0.772
Finland *(147)*	Male	1936-1955	0.83	1506	0.58	3504	0.5
	Female		0.86	2028	0.62	3870	0.48
Norway *(148)*	Male	1915-1939	0.86	259	0.77	313	0.18
	Female		0.89	405	0.75	425	0.28
Norway *(148)*	Male	1940-1949	0.82	253	0.48	284	0.68
	Female		0.85	342	0.68	400	0.34
Norway *(148)*	Male	1950-1960	0.85	370	0.47	463	0.76
	Female		0.89	518	0.66	576	0.46
Sweden *(149)*	Mixed	1926-1958	0.76	2492	0.55	3368	0.42
United States *(150)*	Male	1939-1957	0.76	1019	0.54	907	0.44
United States *(151)*	Male	1936-1955	0.65	512	0.42	772	0.46
	Female		0.72	758	0.57	1154	0.3
United States *(152)*	Male	1917-1927	0.764	1234	0.545	1167	0.438

Notes: When correlations for multiple cohorts are available in a country, we order them chronologically. Gender is equal to mixed if the estimated correlation coefficients were obtained from a mixed-sex sample. Cohort gives the range of the birth years of the twins used to compute the correlations. rMZ and NMZ are, respectively, the sample correlation in monozygotic twins' years of educational attainment and the number of pairs of twins used to compute the correlation. rDZ and NMZ are defined analogously for the dizygotic twins. Falconer h^2 is calculated as $2\times(\text{rMZ} - \text{rDZ})$. This list was compiled by Amelia Branigan, Kenneth J. McCallum and Jeremy Freese *(34)*.

Table S11. Cross-Sib Correlations in Swedish *Brothers Sample*

	Education	Cognitive Function	Education	Cognitive Function
Twins	MZ		DZ	
Education	0.709		0.502	
Cognitive Function	0.512	0.822	0.383	0.534
Full Brothers	Together		Apart	
Education	0.445		0.198	
Cognitive Function	0.364	0.497	0.198	0.359
Half Brothers	Together		Apart	
Education	0.246		0.134	
Cognitive Function	0.208	0.320	0.133	0.191
Adoptees				
Education	0.213			
Cognitive Function	0.149	0.170		

This table reports cross-sib correlations for educational attainment and cognitive function in seven sibling types. Education is years of education residualized on a third order age polynomial. Cognitive function is measured using data from the Swedish Enlistment Battery, a test similar to the US Armed Forces Qualifying Test. Most of the recruits took four subtests (logical, verbal, spatial and technical) which, for most of the study period, were graded on a scale from 0 to 40. To construct the final score, the four raw scores are summed, percentile-rank transformed, and convoluted with the inverse of the standard normal distribution. This procedure ensures that the final test scores are normally distributed. The construction of the final score is performed separately for each birth year in order to take into account small, occasional, year-to-year changes in the test.

Table S12. Estimated variance explained by all SNPs for EduYears and College. The GCTA analysis for *College* was performed on the observed 0-1 scale and transformed to the underlying scale assuming a threshold model. Notice that the GCTA estimate is based on a sample that overlaps with the sample used by *(3)*.

Cohort	Phenotype	*N*	h^2_G	*SE*	*LRT*	*p*-value
QIMR	*EduYears*	2281	0.356	0.138	7.03	4.0×10^{-3}
	College	2281	0.599	0.277	4.90	1.0×10^{-2}
STR	*EduYears*	5678	0.228	0.058	17.04	2.0×10^{-5}
	College	5678	0.213	0.106	4.44	2.0×10^{-2}
QIMR+STR	*EduYears*	7959	0.224	0.042	31.38	1.0×10^{-8}
	College	7959	0.254	0.078	12.33	2.0×10^{-4}

Table S13. Estimating the genetic correlation between Educational Attainment and health status by whole-genome bivariate analysis using genome-wide SNP data. For *College* and dichotomous health data, the whole-genome bivariate analysis was performed on the observed 0/1 scale.

	Univariate GCTA			Bivariate GCTA			
Phenotype	*N*	h_g^2	SE	r_g	SE	LRT ($r_g = 0$)	*p*-value
EduYears	5,650	0.166	0.060	0.132	0.227	0.338	0.3
Health	5,650	0.210	0.061				
College (dichotomous)	5,650	0.125	0.058	0.333	0.328	1.058	0.2
Health (dichotomous)	5,650	0.130	0.060				

Table S14. SNP functional annotation. Genome-wide significant SNPs are listed as headers, with SNPs in strong LD reported in the rows beneath.

SNP	Chr.	Position	LD with GWAS SNP	Reference allele	Other allele	Minor Allele Frequency	Gene	dbSNP functional annotation	Amino Acid change
					rs1056667				
rs1321479	6	26501897	0.93	T	C	0.45	*BTN1A1*	synonymous	
rs3736781	6	26505362	0.93	G	A	0.45	*BTN1A1*	missense	A[Ala] > T[Thr]
rs3736782	6	26505403	0.93	C	A	0.45	*BTN1A1*	synonymous	
rs9393728	6	26509330	0.93	C	G	0.45	*BTN1A1*	missense	D[Asp] > E[Glu]
rs4871	6	26545632	0.93	G	A	0.45	*HMGN4*	synonymous	
rs4573	6	26546808	0.87	T	C	0.43	*HMGN4*	3'-UTR	
					rs11584700				
rs3789045	1	204586812	0.87	C	T	0.23	*LRRN2*	3'-UTR	
rs11588857	1	204587047	0.87	G	A	0.23	*LRRN2*	missense	P[Pro] > S[Ser]
rs3747631	1	204587569	0.87	G	C	0.23	*LRRN2*	missense	L[Leu] > V[Val]
rs3789044	1	204589101	0.87	G	A	0.23	*LRRN2*	missense	P[Pro] > L[Leu]
					rs7309				
rs7309	2	162092640	1	G	A	0.52	*TANK*	3'-UTR	

Table S15. Gene expression blood eQTL analysis results. gSNP – variant associated with educational attainment; eSNP – variant identified as having the strongest cis-effect on a given gene; FDR – false discovery rate; LD – linkage disequilibrium; * denotes that the probe is not annotated**.**

			Single SNP Estimates							Conditioned on eSNP		
GWAS lead-SNP (gSNP)	cis-affected gene	Probe ID	gSNP allele assessed	gSNP *P-value*	gSNP-FDR	eSNP	eSNP allele assessed	eSNP *P-value*	eSNP-FDR	LD between gSNP-eSNP	gSNP *P-value*	gSNP-FDR
rs4851266	*AFF3*	650753	T	2.46×10^{-10}	<< 0.05	rs6749757	A	1.70×10^{-29}	<< 0.05	0.31	7.14×10^{-1}	1.00
rs1056667	*BTN2A1*	2570477	C	9.44×10^{-8}	<< 0.05	rs2273193	C	7.85×10^{-40}	<< 0.05	0.01	1.85×10^{-4}	2.50×10^{-3}
rs1056667	*BTN2A1*	1110093	C	6.23×10^{-4}	0.012	rs2393664	T	9.77×10^{-5}	<< 0.05	0.13	6.95×10^{-2}	1.00
rs1056667	*BTN2A2*	5420709	C	2.97×10^{-3}	0.046	rs3799378	G	1.08×10^{-9}	<< 0.05	0.24	7.91×10^{-1}	1.00
rs1056667	*BTN3A1*	3130600	C	3.62×10^{-7}	<< 0.05	rs7744254	C	1.95×10^{-32}	<< 0.05	0.11	2.30×10^{-1}	1.00
rs1056667	*BTN3A2*	4610674	C	6.12×10^{-35}	<< 0.05	rs3799378	G	9.81×10^{-198}	<< 0.05	0.24	2.41×10^{-2}	0.89
rs1056667	*HIST1H2AC HIST1H2BD HIST1H4A*	290730	C	4.20×10^{-5}	<< 0.05	rs1009181	C	9.81×10^{-198}	<< 0.05	0.05	9.28×10^{-2}	1.00
rs1056667	*HIST1H2AC HIST1H2BD HIST1H4A*	6200669	C	8.41×10^{-4}	0.013	rs1009181	C	9.81×10^{-198}	<< 0.05	0.05	6.49×10^{-2}	1.00
rs1056667	*HIST1H2BK*	6110630	C	5.58×10^{-10}	<< 0.05	rs10946899	A	1.83×10^{-33}	<< 0.05	0.24	8.06×10^{-1}	1.00
rs1056667	*HMGN4*	5270689	C	2.76×10^{-82}	<< 0.05	rs9379886	T	3.33×10^{-84}	<< 0.05	0.55	6.09×10^{-6}	<< 0.05
rs1056667	*LRRC16A*	6450022	C	1.87×10^{-3}	0.031	rs9366619	C	4.48×10^{-31}	<< 0.05	0.00	1.56×10^{-3}	0.078
rs11584700	*MDM4*	5420471	G	1.63×10^{-9}	<< 0.05	rs7556371	G	4.16×10^{-29}	<< 0.05	0.05	1.08×10^{-4}	<< 0.05
rs1056667	*	290273	C	5.90×10^{-26}	<< 0.05	rs2093169	T	5.46×10^{-125}	<< 0.05	0.19	3.90×10^{-1}	1.00
rs1056667	*	3390050	C	4.30×10^{-6}	<< 0.05	rs6456762	T	4.88×10^{-16}	<< 0.05	0.04	5.04×10^{-3}	0.23
rs7309	*TANK*	2230113	A	1.74×10^{-8}	<< 0.05	rs17705608	G	1.88×10^{-9}	<< 0.05	0.71	5.94×10^{-1}	1.00

Table S16. Gene-based *p*-values for the top 25 genes associated with *EduYears* in the combined-stage meta-analysis (using VEGAS).

Chr.	Gene	Number of SNPs	Start position	Stop position	*p*-value (Pooled)	*p*-value (Males)	*p*-value (Females)	*p*-value (Pooled *College*)
2	*GBX2*	70	236739045	236741391	$<1.00\times10^{-6}$	$<1.00\times10^{-6}$	9.77×10^{-3}	3.90×10^{-5}
13	*STK24*	280	97900455	98027397	$<1.00\times10^{-6}$	5.96×10^{-3}	3.50×10^{-5}	$<1.00\times10^{-6}$
16	*LOC440350-4*	1	28677525	28690630	$<1.00\times10^{-6}$	7.28×10^{-4}	4.40×10^{-4}	4.00×10^{-6}
16	*TUFM*	26	28761232	28765230	$<1.00\times10^{-6}$	4.91×10^{-4}	4.46×10^{-4}	5.00×10^{-6}
2	*ASB18*	78	236768253	236837727	1.00×10^{-6}	$<1.00\times10^{-6}$	1.32×10^{-2}	1.51×10^{-4}
3	*APEH*	51	49686438	49695938	1.00×10^{-6}	2.06×10^{-4}	1.58×10^{-3}	7.10×10^{-5}
3	*NICN1*	33	49434769	49441761	1.00×10^{-6}	6.43×10^{-4}	1.06×10^{-3}	1.12×10^{-4}
3	*RNF123*	52	49701993	49733966	1.00×10^{-6}	3.60×10^{-4}	2.24×10^{-3}	1.04×10^{-4}
6	*BTN1A1*	104	26609473	26618631	1.00×10^{-6}	2.80×10^{-5}	1.09×10^{-2}	4.57×10^{-4}
6	*HMGN4*	79	26646550	26655143	1.00×10^{-6}	3.80×10^{-5}	7.22×10^{-3}	5.46×10^{-4}
6	*IHPK3*	201	33797420	33822660	1.00×10^{-6}	5.16×10^{-4}	2.53×10^{-4}	2.00×10^{-6}
10	*C10orf88*	82	124680408	124703909	1.00×10^{-6}	2.03×10^{-3}	3.62×10^{-4}	3.60×10^{-5}
16	*ATP2A1*	28	28797309	28823331	1.00×10^{-6}	6.09×10^{-4}	7.13×10^{-4}	1.00×10^{-5}
16	*ATXN2L*	23	28741914	28756059	1.00×10^{-6}	5.09×10^{-4}	5.01×10^{-4}	8.00×10^{-6}
16	*SH2B1*	31	28782814	28793027	1.00×10^{-6}	5.28×10^{-4}	6.07×10^{-4}	6.00×10^{-6}
3	*BSN*	86	49566925	49683986	2.00×10^{-6}	1.46×10^{-4}	1.45×10^{-3}	3.80×10^{-5}
3	*MST1*	47	49696391	49701099	2.00×10^{-6}	3.31×10^{-4}	1.82×10^{-3}	9.10×10^{-5}
3	*TCTA*	34	49424642	49428913	3.00×10^{-6}	7.47×10^{-4}	1.23×10^{-3}	1.36×10^{-4}
6	*C6orf125*	177	33773323	33787482	3.00×10^{-6}	7.79×10^{-4}	1.16×10^{-3}	4.00×10^{-6}
3	*AMT*	33	49429214	49435016	4.00×10^{-6}	6.47×10^{-4}	1.35×10^{-3}	1.11×10^{-4}
10	*FAM24A*	51	124660206	124662617	4.00×10^{-6}	1.87×10^{-3}	1.75×10^{-3}	6.80×10^{-5}
18	*KATNAL2*	109	42780784	42881663	4.00×10^{-6}	5.41×10^{-1}	$<1.00\times10^{-6}$	2.10×10^{-5}
1	*CEP170*	70	241354352	241485331	6.00×10^{-6}	1.09×10^{-3}	4.59×10^{-3}	6.50×10^{-5}
4	*TET2*	130	106287391	106420407	6.00×10^{-6}	1.16×10^{-2}	1.60×10^{-4}	5.00×10^{-4}
16	*RABEP2*	27	28823242	28844033	6.00×10^{-6}	1.30×10^{-3}	1.68×10^{-3}	3.30×10^{-5}

Table S17. Gene-based *p*-values for the top 25 genes associated with *College* in the combined stage meta-analysis (using VEGAS).

Chr.	Gene	Number of SNPs	Start position	Stop position	*p*-value (Pooled)	*p*-value (Males)	*p*-value (Females)	*p*-value (Pooled *EduYears*)
1	*PIK3C2B*	109	202658380	202726097	$<1.00\times10^{-6}$	7.09×10^{-2}	<1.00×10-6	8.88×10^{-3}
2	*ASB18*	78	236768253	236837727	$<1.00\times10^{-6}$	1.22×10^{-4}	2.36×10-4	1.00×10^{-6}
2	*GBX2*	70	236739045	236741391	$<1.00\times10^{-6}$	1.30×10^{-5}	2.62×10-4	$<1.00\times10^{-6}$
6	*C6orf125*	178	33773323	33787482	$<1.00\times10^{-6}$	8.90×10^{-5}	2.72×10-3	3.00×10^{-6}
4	*TET2*	130	106287391	106420407	1.00×10^{-6}	1.58×10^{-3}	7.94×10-4	6.00×10^{-6}
6	*IHPK3*	202	33797420	33822660	1.00×10^{-6}	4.70×10^{-5}	2.26×10-3	1.00×10^{-6}
6	*ITPR3*	228	33697138	33772326	1.00×10^{-6}	2.00×10^{-4}	3.25×10-3	1.60×10^{-5}
3	*CCDC14*	113	125114963	125162945	4.00×10^{-6}	6.12×10^{-3}	3.82×10-4	6.62×10^{-3}
10	*PSD*	39	104152365	104168891	4.00×10^{-6}	1.54×10^{-3}	2.63×10-3	6.90×10^{-5}
10	*NFKB2*	28	104144218	104152271	6.00×10^{-6}	1.14×10^{-3}	2.11×10-3	6.30×10^{-5}
4	*C4orf44*	85	3220564	3235638	7.00×10^{-6}	4.61×10^{-2}	8.80×10-5	1.49×10^{-4}
10	*ELOVL3*	34	103976132	103979334	7.00×10^{-6}	1.02×10^{-3}	2.35×10-3	4.00×10^{-5}
10	*GBF1*	77	103995298	104132639	7.00×10^{-6}	7.45×10^{-4}	1.30×10-3	2.80×10^{-5}
12	*PITPNM2*	57	122033979	122160928	7.00×10^{-6}	1.52×10^{-2}	1.15×10-4	2.10×10^{-5}
3	*MST1R*	45	49899439	49916310	8.00×10^{-6}	4.12×10^{-3}	1.15×10-3	1.06×10^{-3}
3	*ROPN1*	75	125170568	125192889	8.00×10^{-6}	2.68×10^{-2}	1.93×10-4	8.31×10^{-3}
1	*PPP1R15B*	93	202639114	202647567	9.00×10^{-6}	1.63×10^{-1}	8.00×10-6	2.47×10^{-2}
12	*ARL6IP4*	14	122030832	122033413	9.00×10^{-6}	1.90×10^{-2}	4.00×10-5	5.80×10^{-5}
3	*TRAIP*	50	49841031	49868996	1.00×10^{-5}	2.91×10^{-3}	1.61×10-3	5.98×10^{-4}
3	*UBA7*	39	49817641	49826395	1.10×10^{-5}	1.94×10^{-3}	3.19×10-3	3.78×10^{-4}
12	*OGFOD2*	15	122025306	122030541	1.10×10^{-5}	2.19×10^{-2}	2.60×10-5	8.50×10^{-5}
3	*IHPK1*	54	49736731	49798977	1.20×10^{-5}	7.80×10^{-4}	8.11×10-3	6.80×10^{-5}
3	*AMIGO3*	39	49729968	49732127	1.40×10^{-5}	4.46×10^{-4}	1.12×10-2	1.90×10^{-5}
3	*RNF123*	52	49701993	49733966	1.40×10^{-5}	3.67×10^{-4}	1.28×10-2	1.00×10^{-6}
10	*PITX3*	36	103979935	103991221	1.40×10^{-5}	8.78×10^{-4}	1.69×10-3	3.10×10^{-5}

Table S18. Pathway-based *p*-values for pathways showing suggestive overlap ($p < 0.05$) with genomic regions meeting *p*-value $<1\times10^{-5}$ in the combined-stage GWAS meta-analysis for (A) *EduYears* and (B) *College*. Size refers to the number of genomic intervals defining the pathway, while Overlap indicates the number of LD-independent intervals defined by SNPs meeting $p<1\times10^{-5}$ in the combined discovery and replication GWAS meta-analysis that overlap with genomic intervals defining the pathway. *P* lists the empirical *p*-value, using 1×10^{6} permutations, and Corrected *P* provides the *p*-value adjusted for multiple testing, using 1×10^{4} permutations.

A. *EduYears*

			EduYears			*College*		
Pathway	GO ID	Size	Overlap	*P*	Corrected *P*	Overlap	*P*	Corrected *P*
focal adhesion	GO:0005925	98	3	0.022	1.00	0	1.000	1.00
purine base metabolic process	GO:0006144	33	2	0.024	1.00	0	1.000	1.00
I-kappaB kinase/NF-kappaB cascade	GO:0007249	20	2	0.005	0.93	0	1.000	1.00
sulfotransferase activity	GO:0008146	36	2	0.033	1.00	0	1.000	1.00
oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen	GO:0016702	44	2	0.021	1.00	3	0.001	0.41
vinculin binding	GO:0017166	8	2	0.004	0.90	0	1.000	1.00
lamellipodium	GO:0030027	72	4	0.005	0.93	1	1.000	1.00
lamellipodium assembly	GO:0030032	15	2	0.030	1.00	0	1.000	1.00
endocytic vesicle	GO:0030139	27	2	0.013	0.99	2	0.007	0.94
integral to Golgi membrane	GO:0030173	43	2	0.014	0.99	0	1.000	1.00
heat shock protein binding	GO:0031072	61	2	0.038	1.00	0	1.000	1.00
platelet dense tubular network membrane	GO:0031095	8	2	0.024	1.00	1	1.000	1.00
sarcoplasmic reticulum membrane	GO:0033017	17	2	0.029	1.00	0	1.000	1.00
phosphoinositide binding	GO:0035091	56	2	0.040	1.00	2	0.028	1.00
focal adhesion assembly	GO:0048041	14	2	0.002	0.80	0	1.000	1.00

B. *College*

			College			*EduYears*		
Pathway	GO ID	Size	Overlap	*P*	Corrected *P*	Overlap	*P*	Corrected *P*
spliceosome assembly	GO:0000245	19	2	0.003	0.85	1	1.000	1.00
calcium channel activity	GO:0005262	46	2	0.045	1.00	1	1.000	1.00
Notch signaling pathway	GO:0007219	52	2	0.042	1.00	0	1.000	1.00
locomotory behavior	GO:0007626	57	2	0.038	1.00	0	1.000	1.00
methyltransferase activity	GO:0008168	104	3	0.017	0.99	2	0.148	1.00
oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen	GO:0016702	44	3	0.001	0.41	2	0.021	1.00
triglyceride biosynthetic process	GO:0019432	31	2	0.015	0.99	0	1.000	1.00
endocytic vesicle	GO:0030139	27	2	0.007	0.94	2	0.013	0.99
phosphoinositide binding	GO:0035091	56	2	0.028	1.00	2	0.040	1.00
neuron development	GO:0048666	28	2	0.010	0.98	1	1.000	1.00
positive regulation of NF-kappaB transcription factor activity	GO:0051092	69	2	0.037	1.00	0	1.000	1.00
response to calcium ion	GO:0051592	47	2	0.015	0.99	1	1.000	1.00

Table S19. Loci with cell-type specificity scores ≥ 95th percentile (see Figure S21). IndexSNPs were identified as $p < 1\times10^{-5}$ (pruned to remove SNPs correlated at $r^2 > 0.5$) from either the *EduYears* or *College* combined stage meta-analyses. BestSNP provides the identity of the SNP in LD with the IndexSNP that displays the highest cell-type specificity score (Score), calculated as the height of the nearest H3K4me3 peak divided by the distance (Distance, in bp) between the H3K4me3 peak and BestSNP and subsequently normalized such that sum of scores for a given locus across the 34 tissues equals one.

Tissue (Phenotype)	IndexSNP	BestSNP	Score	Distance
Brain Anterior Caudate (*EduYears*)				
	rs6882046	rs6882046	0.45	834
	rs6742801	rs67003507	0.91	37
	rs9320913	6:98566506	0.38	57
	rs791903	rs4711343	0.61	5
	rs2955259	rs13110775	0.41	267
	rs12433424	rs1449108	0.87	2480
	rs7333699	rs12853561	0.98	0
	rs8034147	rs34480933	0.71	41
Brain Anterior Caudate (*College*)				
	rs6742801	rs67003507	0.91	37
	rs3121417	rs3121417	0.98	434
	rs9320913	6:98566506	0.38	57
	rs791903	rs4711343	0.61	5
	rs12640626	rs12640626	0.74	323
	rs11802889	rs61817490	0.57	866
	rs2540989	rs1206419	1.00	2340
	rs9563527	rs11842899	0.58	327
	rs4365358	rs4321256	0.89	52
Brain Hippocampus Middle (*EduYears*)				
	rs6882046	rs6882046	0.37	832
	rs7713243	rs324888	0.60	50
	rs9320913	6:98566506	0.81	31
	rs2955259	rs3797042	0.64	180
	rs2930734	rs2930726	0.70	216
CD4 Naïve Primary Cells (*EduYears*)				
	rs2137835	rs2137835	0.74	6
	rs10176262	rs4851368	0.90	283
	rs1892700	rs16990773	0.40	72
	rs6984449	rs4292704	0.78	100
Muscle Satellite Cultured Cells (*EduYears*)				
	rs3789044	rs12043569	0.81	815
	rs889956	2:57388609	1.00	367
	rs1056667	rs6918360	0.63	1
	rs652049	rs530614	0.39	136
	rs1391439	rs2047409	0.99	5
	rs11248332	rs10794575	0.95	44

Table S20. Implicated gene loci demonstrating promising eQTL (Table S15) or functional SNP annotation (Table S14) of top loci or association in gene-based tests (Tables S16, S17). The last column gives the distance (in kilobases) from nearby (< 1.0 Mb) independent top associated SNPs (replicated SNPs: rs9320913, rs11584700, rs4851266; or additional independent SNPs meeting $p<5\times10^{-8}$ in the combined meta-analysis: rs7309, rs11687170, rs1056667, rs13401104). All other columns list the Table location for details of evidence.

Location	Gene	Functional annotation	Blood eQTL	Gene-based tests	Distance to replicated or significant SNP marker
1q32	*PIK3C2B*			Table S17	rs11584700, 117.5kb
1q32	*MDM4*		Table S15		rs11584700, 57.3kb
1q32	*LRRN2*	Table S14			rs11584700, 9.3kb
2q11-q12	*AFF3*		Table S15		rs4851266, 59.4kb
2q24-q31	*TANK*	Table S14	Table S15		rs7309, 3'UTR
2q37	*GBX2*			Tables S16, S17	rs11687170, 16.2kb; rs13401104, 28.9kb
2q37	*ASB18*			Tables S16, S17	rs13401104, intronic; rs11687170, 45.4kb
3p21	*NICN1*			Table S16	
3p21	*BSN*			Table S16	
3p21	*APEH*			Table S16	
3p21	*MST1*			Table S16	
3p24	*RNF123*			Table S16	
4q24	*TET2*			Table S17	
6p22	*LRRC16A*		Table S15		rs1056667, 889.8kb
6p22	*HIST1H4A*		Table S15		rs1056667, 488.3kb
6p22	*HIST1H2AC*		Table S15		rs1056667, 385.6kb
6p21	*HIST1H2BD*		Table S15		rs1056667, 339.0kb
6p22	*BTN3A2*		Table S15		rs1056667, 132.0kb
6p22	*BTN2A2*		Table S15		rs1056667, 115.5kb
6p22	*BTN3A1*		Table S15		rs1056667, 95.1kb
6p22	*BTN2A1*		Table S15		rs1056667, 40.7kb
6p22	*BTN1A1*	Table S14		Table S16	rs1056667, 3'UTR
6p21	*HMGN4*	Table S14	Table S15	Table S16	rs1056667, 28.0kb
6p21	*HIST1H2BK*		Table S15		rs1056667, 595.5kb
6p21	*ITPR3*			Table S17	
6p21	*MNF1 (C6orf125)*			Table S17	
6p21	*IP6K3 (IHPK3)*			Tables S16, S17	
10q26	*C10orf88*			Table S16	
13q31-q32	*STK24*			Table S16	
16p11	*NPIPL1 (LOC440350)*			Table S16	
16p11	*ATXN2L*			Table S16	
16p11	*TUFM*			Table S16	
16p11	*SH2B1*			Table S16	
16p12	*ATP2A1*			Table S16	

Table S21. Results of gene function prediction analysis in 80,000 gene expression profiles of identified genes (Table S20). Pathway terms originate from several databases: (1) Gene Ontology Biological Processes, (2) Gene Ontology Molecular Function, (3) Gene Ontology Cellular Component, (4) Reactome, and (5) KEGG. Terms directly related to neuronal or central nervous system function are marked with an asterisk. *P*-values refer to the correlation between the Gene principal component profile and the Term principal component profile, uncorrected for multiple testing; all reported terms meet False discovery rate < 0.05. The Annotated column indicates if the gene has previously been listed as a member of that term (Y) or not (N). Results are sorted alphabetically by gene name.

Gene		Term	*P*-value	Annotated
AFF3	[1]	cartilage condensation	2.3×10^{-6}	N
APEH	[1]	cofactor metabolic process	3.6×10^{-15}	N
APEH	[1]	porphyrin-containing compound biosynthetic process	9.5×10^{-15}	N
APEH	[1]	tetrapyrrole biosynthetic process	9.5×10^{-15}	N
APEH	[1]	heme biosynthetic process	1.1×10^{-14}	N
APEH	[1]	porphyrin-containing compound metabolic process	2.4×10^{-13}	N
APEH	[1]	tetrapyrrole metabolic process	2.4×10^{-13}	N
APEH	[1]	aerobic respiration	2.4×10^{-13}	N
APEH	[3]	mitochondrial matrix	1.0×10^{-12}	N
APEH	[4]	Mitochondrial tRNA aminoacylation	1.4×10^{-12}	N
APEH	[1]	cofactor biosynthetic process	1.7×10^{-12}	N
APEH	[1]	heme metabolic process	2.6×10^{-12}	N
APEH	[1]	tricarboxylic acid cycle	5.9×10^{-12}	N
APEH	[2]	coenzyme binding	7.7×10^{-12}	N
APEH	[1]	tRNA aminoacylation for protein translation	8.5×10^{-12}	N
APEH	[2]	aminoacyl-tRNA ligase activity	9.0×10^{-12}	N
APEH	[2]	ligase activity, forming aminoacyl-tRNA and related compounds	9.0×10^{-12}	N
APEH	[2]	ligase activity, forming carbon-oxygen bonds	9.0×10^{-12}	N
APEH	[1]	acetyl-CoA catabolic process	1.2×10^{-11}	N
APEH	[4]	Metabolism of porphyrins	1.3×10^{-11}	N
APEH	[2]	cofactor binding	3.8×10^{-11}	N
APEH	[1]	amino acid activation	9.5×10^{-11}	N
APEH	[1]	tRNA aminoacylation	9.5×10^{-11}	N
APEH	[1]	coenzyme metabolic process	2.0×10^{-10}	N
APEH	[5]	Aminoacyl-tRNA biosynthesis	2.1×10^{-10}	N
APEH	[1]	heterocycle biosynthetic process	2.6×10^{-10}	N
APEH	[1]	fatty acid beta-oxidation using acyl-CoA oxidase	3.1×10^{-10}	N
APEH	[5]	Citrate cycle (TCA cycle)	4.7×10^{-10}	N
APEH	[4]	Citric acid cycle (TCA cycle)	4.8×10^{-10}	N
APEH	[5]	Valine, leucine and isoleucine biosynthesis	1.4×10^{-9}	N
APEH	[5]	Porphyrin and chlorophyll metabolism	1.6×10^{-9}	N
APEH	[4]	tRNA Aminoacylation	1.9×10^{-9}	N
APEH	[1]	coenzyme catabolic process	2.1×10^{-9}	N
APEH	[4]	Pyruvate metabolism and Citric Acid (TCA) cycle	2.8×10^{-9}	N
APEH	[3]	mitochondrial inner membrane	4.3×10^{-9}	N
APEH	[3]	organelle inner membrane	5.3×10^{-9}	N
APEH	[1]	fatty acid beta-oxidation	7.2×10^{-9}	N
APEH	[1]	nicotinamide nucleotide metabolic process	9.2×10^{-9}	N
APEH	[2]	oxidoreductase activity, acting on the CH-CH group of donors	1.5×10^{-8}	N
APEH	[2]	small protein activating enzyme activity	3.9×10^{-8}	N
APEH	[4]	Metabolism of carbohydrates	4.0×10^{-8}	N
APEH	[3]	mitochondrial envelope	1.9×10^{-7}	N
APEH	[2]	oxidoreductase activity, acting on the aldehyde or oxo group of donors, disulfide as ...	2.8×10^{-7}	N
APEH	[2]	ATPase activity	3.1×10^{-7}	N
APEH	[3]	mitochondrial membrane	3.5×10^{-7}	N
APEH	[2]	acyl-CoA dehydrogenase activity	4.3×10^{-7}	N
APEH	[2]	ATPase activity, coupled	5.1×10^{-7}	N
APEH	[4]	Mitochondrial Fatty Acid Beta-Oxidation	5.9×10^{-7}	N
APEH	[2]	lyase activity	6.6×10^{-7}	N
APEH	[3]	signalosome	7.9×10^{-7}	N
APEH	[4]	Peroxisomal lipid metabolism	8.1×10^{-7}	N
APEH	[4]	Purine metabolism	1.0×10^{-6}	N
APEH	[3]	microbody lumen	1.9×10^{-6}	N

APEH	3	peroxisomal matrix	1.9×10^{-6}	N
APEH	5	Galactose metabolism	3.4×10^{-6}	N
APEH	5	Valine, leucine and isoleucine degradation	1.4×10^{-5}	N
APEH	5	Butanoate metabolism	1.4×10^{-5}	N
APEH	5	One carbon pool by folate	2.8×10^{-5}	N
APEH	5	Fatty acid metabolism	3.3×10^{-5}	N
APEH	5	Peroxisome	4.5×10^{-4}	N
APEH	5	Non-homologous end-joining	7.6×10^{-4}	N
ATP2A1	1	actin-mediated cell contraction	4.7×10^{-236}	N
ATP2A1	1	actin-myosin filament sliding	2.2×10^{-233}	N
ATP2A1	1	muscle filament sliding	2.2×10^{-233}	N
ATP2A1	1	skeletal muscle contraction	6.4×10^{-230}	Y
ATP2A1	2	structural constituent of muscle	3.2×10^{-222}	N
ATP2A1	1	actin filament-based movement	3.6×10^{-214}	N
ATP2A1	4	Striated Muscle Contraction	2.0×10^{-201}	N
ATP2A1	3	contractile fiber part	5.1×10^{-191}	Y
ATP2A1	3	contractile fiber	1.2×10^{-189}	Y
ATP2A1	3	myofibril	3.0×10^{-187}	Y
ATP2A1	3	myosin filament	1.4×10^{-186}	N
ATP2A1	3	sarcomere	1.3×10^{-185}	Y
ATP2A1	3	striated muscle thin filament	1.6×10^{-182}	N
ATP2A1	1	multicellular organismal movement	9.9×10^{-181}	Y
ATP2A1	1	musculoskeletal movement	9.9×10^{-181}	Y
ATP2A1	3	muscle myosin complex	9.0×10^{-178}	N
ATP2A1	3	myosin II complex	1.3×10^{-165}	N
ATP2A1	4	Muscle contraction	1.8×10^{-153}	N
ATP2A1	3	myosin complex	9.8×10^{-143}	N
ATP2A1	1	striated muscle contraction	1.4×10^{-141}	Y
ATP2A1	1	muscle contraction	3.3×10^{-126}	Y
ATP2A1	1	muscle system process	4.2×10^{-121}	Y
ATP2A1	3	I band	1.1×10^{-108}	Y
ATP2A1	2	tropomyosin binding	2.2×10^{-100}	N
ATP2A1	3	actin cytoskeleton	3.0×10^{-98}	N
ATP2A1	2	titin binding	1.5×10^{-90}	N
ATP2A1	3	sarcoplasmic reticulum	8.7×10^{-90}	Y
ATP2A1	3	sarcoplasm	5.0×10^{-86}	Y
ATP2A1	3	A band	1.3×10^{-82}	Y
ATP2A1	2	actin binding	3.0×10^{-70}	N
ATP2A1	3	Z disc	6.6×10^{-69}	N
ATP2A1	3	pseudopodium	7.0×10^{-66}	N
ATP2A1	1	actin filament-based process	5.8×10^{-63}	N
ATP2A1	3	sarcoplasmic reticulum membrane	2.5×10^{-57}	Y
ATP2A1	2	microfilament motor activity	7.3×10^{-57}	N
ATP2A1	1	regulation of striated muscle contraction	2.5×10^{-56}	Y
ATP2A1	1	muscle organ development	1.2×10^{-51}	N
ATP2A1	1	striated muscle adaptation	5.1×10^{-50}	N
ATP2A1	1	muscle cell fate commitment	5.6×10^{-49}	N
ATP2A1	1	muscle structure development	5.2×10^{-47}	N
ATP2A1	1	myofibril assembly	3.8×10^{-44}	N
ATP2A1	1	regulation of muscle contraction	7.9×10^{-44}	Y
ATP2A1	1	skeletal muscle tissue development	4.9×10^{-43}	N
ATP2A1	1	skeletal muscle organ development	1.9×10^{-41}	N
ATP2A1	5	Tight junction	1.4×10^{-40}	N
ATP2A1	5	Viral myocarditis	3.1×10^{-39}	N
ATP2A1	2	myosin binding	7.8×10^{-35}	N
ATP2A1	2	calmodulin binding	6.7×10^{-33}	N
ATP2A1	5	Cardiac muscle contraction	7.2×10^{-26}	N
ATP2A1	5	Hypertrophic cardiomyopathy (HCM)	3.0×10^{-25}	N
ATP2A1	5	Dilated cardiomyopathy	2.4×10^{-20}	N
ATP2A1	5	Calcium signaling pathway	1.2×10^{-8}	Y
ATP2A1	5	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	7.1×10^{-8}	N
ATP2A1	5	Leukocyte transendothelial migration	2.7×10^{-6}	N
ATP2A1	5	Glycolysis / Gluconeogenesis	3.1×10^{-5}	N

ATP2A1		[5]	Insulin signaling pathway	8.4×10^{-5}	N
ATP2A1		[5]	Arginine and proline metabolism	3.1×10^{-4}	N
ATP2A1		[5]	Thyroid cancer	5.6×10^{-4}	N
ATXN2L		[1]	positive regulation of gene expression, epigenetic	9.9×10^{-13}	N
ATXN2L		[2]	transcription cofactor activity	6.5×10^{-11}	N
ATXN2L		[2]	transcription factor binding transcription factor activity	8.6×10^{-11}	N
ATXN2L		[2]	protein binding transcription factor activity	2.0×10^{-10}	N
ATXN2L		[2]	tau-protein kinase activity	7.0×10^{-10}	N
ATXN2L		[2]	transcription corepressor activity	1.5×10^{-9}	N
ATXN2L		[1]	chromatin disassembly	3.1×10^{-9}	N
ATXN2L		[1]	nucleosome disassembly	3.1×10^{-9}	N
ATXN2L		[1]	protein-DNA complex disassembly	3.1×10^{-9}	N
ATXN2L		[3]	npBAF complex	1.4×10^{-8}	N
ATXN2L		[3]	nuclear chromatin	4.9×10^{-8}	N
ATXN2L		[4]	EGFR downregulation	4.9×10^{-8}	N
ATXN2L		[3]	nBAF complex	3.1×10^{-7}	N
ATXN2L		[3]	chromatin remodeling complex	7.0×10^{-7}	N
ATXN2L		[3]	SWI/SNF-type complex	1.4×10^{-6}	N
ATXN2L		[3]	PRC1 complex	1.8×10^{-6}	N
ATXN2L		[5]	Valine, leucine and isoleucine biosynthesis	1.9×10^{-6}	N
ATXN2L		[3]	SWI/SNF complex	4.7×10^{-6}	N
ATXN2L		[3]	sex chromosome	5.8×10^{-6}	N
ATXN2L		[3]	histone methyltransferase complex	5.9×10^{-6}	N
ATXN2L		[3]	methyltransferase complex	5.9×10^{-6}	N
ATXN2L		[5]	Aminoacyl-tRNA biosynthesis	9.6×10^{-6}	N
ATXN2L		[5]	Vasopressin-regulated water reabsorption	2.4×10^{-5}	N
BSN	*	[3]	synapse part	8.0×10^{-36}	Y
BSN	*	[4]	Neuronal System	1.1×10^{-33}	N
BSN	*	[3]	synapse	6.4×10^{-31}	Y
BSN	*	[3]	synaptic vesicle membrane	6.2×10^{-30}	N
BSN	*	[3]	synaptic membrane	3.7×10^{-29}	N
BSN	*	[1]	neurotransmitter secretion	1.0×10^{-28}	N
BSN	*	[4]	Transmission across Chemical Synapses	2.0×10^{-28}	N
BSN	*	[4]	Ras activation uopn Ca2+ infux through NMDA receptor	2.5×10^{-27}	N
BSN	*	[4]	CREB phosphorylation through the activation of CaMKII	7.5×10^{-26}	N
BSN	*	[1]	synaptic vesicle exocytosis	6.4×10^{-25}	N
BSN	*	[3]	dendrite	8.4×10^{-25}	N
BSN	*	[3]	dendritic spine	9.8×10^{-25}	N
BSN	*	[3]	neuron spine	9.8×10^{-25}	N
BSN	*	[3]	dendritic spine head	3.0×10^{-24}	N
BSN	*	[3]	postsynaptic density	3.0×10^{-24}	N
BSN	*	[4]	Glutamate Neurotransmitter Release Cycle	3.9×10^{-24}	N
BSN	*	[1]	neurotransmitter transport	1.3×10^{-23}	N
BSN		[2]	voltage-gated cation channel activity	1.8×10^{-23}	N
BSN	*	[3]	main axon	3.2×10^{-23}	N
BSN	*	[3]	postsynaptic membrane	1.2×10^{-22}	N
BSN	*	[4]	Post NMDA receptor activation events	1.5×10^{-22}	N
BSN		[4]	Potassium Channels	2.0×10^{-22}	N
BSN	*	[4]	Unblocking of NMDA receptor, glutamate binding and activation	2.3×10^{-22}	N
BSN		[3]	cation channel complex	4.0×10^{-22}	N
BSN	*	[4]	Activation of NMDA receptor upon glutamate binding and postsynaptic events	4.8×10^{-22}	N
BSN	*	[3]	synaptic vesicle	6.3×10^{-22}	N
BSN		[2]	gated channel activity	9.1×10^{-22}	N
BSN	*	[3]	axon part	9.4×10^{-22}	N
BSN	*	[1]	regulation of neurotransmitter levels	1.3×10^{-21}	N
BSN		[3]	clathrin coated vesicle membrane	1.6×10^{-21}	N
BSN		[3]	ion channel complex	2.8×10^{-21}	N
BSN	*	[4]	Dopamine Neurotransmitter Release Cycle	3.5×10^{-21}	N
BSN	*	[4]	Serotonin Neurotransmitter Release Cycle	3.5×10^{-21}	N
BSN		[2]	voltage-gated channel activity	3.8×10^{-21}	N
BSN		[2]	voltage-gated ion channel activity	3.8×10^{-21}	N

BSN	*	4	Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell	5.9×10^{-21}	N
BSN	*	3	presynaptic membrane	7.1×10^{-21}	N
BSN	*	1	synaptic vesicle transport	1.0×10^{-20}	N
BSN		2	ion channel activity	1.4×10^{-20}	N
BSN		2	substrate-specific channel activity	4.8×10^{-20}	N
BSN		2	channel activity	4.5×10^{-19}	N
BSN		2	passive transmembrane transporter activity	4.5×10^{-19}	N
BSN	*	3	axon	1.5×10^{-18}	N
BSN	*	1	regulation of synaptic transmission	2.4×10^{-18}	N
BSN	*	1	regulation of alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionate selective glutama...	2.8×10^{-18}	N
BSN		4	Voltage gated Potassium channels	7.7×10^{-18}	N
BSN		2	cation channel activity	8.9×10^{-18}	N
BSN	*	4	CREB phosphorylation through the activation of Ras	3.2×10^{-17}	N
BSN		2	voltage-gated potassium channel activity	3.5×10^{-17}	N
BSN	*	1	neuron-neuron synaptic transmission	1.2×10^{-16}	N
BSN	*	4	Neurotransmitter Release Cycle	1.3×10^{-16}	N
BSN	*	1	regulation of synaptic plasticity	1.3×10^{-16}	N
BSN	*	1	regulation of neurological system process	2.8×10^{-16}	N
BSN	*	1	long-term memory	3.9×10^{-16}	N
BSN		3	voltage-gated calcium channel complex	5.1×10^{-16}	N
BSN	*	1	regulation of transmission of nerve impulse	6.7×10^{-16}	N
BSN		2	potassium ion transmembrane transporter activity	8.0×10^{-16}	N
BSN		3	calcium channel complex	1.6×10^{-15}	N
BSN		2	potassium channel activity	2.1×10^{-15}	N
BSN	*	1	regulation of neuronal synaptic plasticity	2.5×10^{-15}	N
BSN	*	4	GABA synthesis, release, reuptake and degradation	1.1×10^{-14}	N
BSN	*	4	Acetylcholine Neurotransmitter Release Cycle	1.4×10^{-14}	N
BSN	*	1	dendrite morphogenesis	3.2×10^{-14}	N
BSN		1	generation of a signal involved in cell-cell signaling	3.7×10^{-14}	N
BSN		1	signal release	3.7×10^{-14}	N
BSN	*	1	synaptic transmission, glutamatergic	4.5×10^{-14}	N
BSN		2	syntaxin-1 binding	6.4×10^{-14}	N
BSN		2	voltage-gated calcium channel activity	7.6×10^{-14}	N
BSN		2	ligand-gated channel activity	8.4×10^{-14}	N
BSN		2	ligand-gated ion channel activity	8.4×10^{-14}	N
BSN		4	Interaction between L1 and Ankyrins	2.0×10^{-13}	N
BSN	*	1	glutamate secretion	3.3×10^{-13}	N
BSN		1	membrane depolarization	3.9×10^{-13}	N
BSN		2	metal ion transmembrane transporter activity	4.3×10^{-13}	N
BSN	*	4	GABA receptor activation	5.6×10^{-13}	N
BSN	*	4	Glutamate Binding, Activation of AMPA Receptors and Synaptic Plasticity	7.8×10^{-13}	N
BSN	*	1	learning	2.2×10^{-12}	N
BSN	*	2	GABA receptor activity	1.1×10^{-11}	N
BSN		2	SNARE binding	1.2×10^{-11}	N
BSN		2	syntaxin binding	1.3×10^{-11}	N
BSN		5	Calcium signaling pathway	1.4×10^{-10}	N
BSN		5	Long-term potentiation	8.0×10^{-9}	N
BSN	*	5	Taste transduction	4.9×10^{-6}	N
BSN		5	Cardiac muscle contraction	3.7×10^{-5}	N
BSN		5	Amyotrophic lateral sclerosis (ALS)	7.0×10^{-5}	N
BSN	*	5	Neuroactive ligand-receptor interaction	5.6×10^{-4}	N
BTN family		4	Antigen Presentation: Folding, assembly and peptide loading of class I MHC	2.2×10^{-45}	N
BTN family		3	MHC class I protein complex	2.5×10^{-34}	N
BTN family		4	Interferon gamma signaling	9.0×10^{-34}	N
BTN family		1	cellular response to interferon-gamma	9.1×10^{-33}	N
BTN family		1	response to interferon-gamma	2.0×10^{-31}	N
BTN family		3	MHC protein complex	9.8×10^{-31}	N
BTN family		1	interferon-gamma-mediated signaling pathway	3.2×10^{-30}	N
BTN family		1	antigen processing and presentation of peptide antigen via MHC	9.0×10^{-30}	N

			class I		
BTN family		[1]	antigen processing and presentation of peptide antigen	3.2×10^{-29}	N
BTN family		[4]	ER-Phagosome pathway	4.0×10^{-28}	N
BTN family		[1]	antigen processing and presentation	1.7×10^{-27}	N
BTN family		[4]	Antigen processing-Cross presentation	4.5×10^{-27}	N
BTN family		[3]	ER to Golgi transport vesicle	6.9×10^{-25}	N
BTN family		[5]	Antigen processing and presentation	4.3×10^{-24}	N
BTN family		[3]	ER to Golgi transport vesicle membrane	3.4×10^{-23}	N
BTN family		[2]	MHC class I receptor activity	4.9×10^{-23}	N
BTN family		[4]	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	1.0×10^{-21}	N
BTN family		[4]	Adaptive Immune System	7.0×10-21	N
BTN family		[5]	Graft-versus-host disease	1.4×10^{-20}	N
BTN family		[5]	Viral myocarditis	2.1×10^{-20}	N
BTN family		[5]	Allograft rejection	8.8×10^{-20}	N
BTN family		[5]	Type I diabetes mellitus	1.3×10^{-18}	N
BTN family		[4]	Interferon Signaling	2.1×10^{-17}	N
BTN family		[2]	MHC class I protein binding	8.1×10^{-17}	N
BTN family		[5]	Autoimmune thyroid disease	1.2×10^{-16}	N
BTN family		[4]	Class I MHC mediated antigen processing & presentation	3.8×10^{-16}	N
BTN family		[2]	threonine-type endopeptidase activity	6.1×10^{-16}	N
BTN family		[2]	threonine-type peptidase activity	6.1×10^{-16}	N
BTN family		[3]	proteasome core complex	7.7×10^{-16}	N
BTN family		[3]	integral to endoplasmic reticulum membrane	1.2×10^{-15}	N
BTN family		[3]	intrinsic to endoplasmic reticulum membrane	5.6×10^{-14}	N
BTN family		[1]	response to type I interferon	7.8×10^{-14}	N
BTN family		[3]	transport vesicle membrane	1.2×10^{-13}	N
BTN family		[1]	cellular response to type I interferon	1.3×10^{-13}	N
BTN family		[1]	type I interferon-mediated signaling pathway	1.3×10^{-13}	N
BTN family		[4]	Interferon alpha/beta signaling	3.1×10^{-13}	N
BTN family		[1]	antigen processing and presentation of exogenous antigen	5.1×10^{-13}	N
BTN family		[4]	Cytokine Signaling in Immune system	1.7×10^{-12}	N
BTN family		[2]	MHC protein binding	8.3×10^{-12}	N
BTN family		[5]	Cell adhesion molecules (CAMs)	1.1×10^{-11}	N
BTN family		[4]	Negative regulators of RIG-I/MDA5 signaling	2.0×10^{-10}	N
BTN family		[3]	MHC class II protein complex	2.7×10^{-10}	N
BTN family		[3]	integral to organelle membrane	8.6×10^{-10}	N
BTN family		[4]	Translocation of ZAP-70 to Immunological synapse	1.4×10^{-9}	N
BTN family		[5]	Primary immunodeficiency	2.5×10^{-9}	N
BTN family		[4]	PD-1 signaling	2.5×10^{-9}	N
BTN family		[4]	Phosphorylation of CD3 and TCR zeta chains	6.1×10^{-9}	N
BTN family		[4]	Generation of second messenger molecules	1.0×10^{-8}	N
BTN family		[3]	intrinsic to organelle membrane	1.2×10^{-8}	N
BTN family		[5]	Natural killer cell mediated cytotoxicity	3.1×10^{-8}	N
BTN family		[4]	Downstream TCR signaling	4.3×10^{-8}	N
BTN family		[3]	transport vesicle	8.8×10^{-8}	N
BTN family		[5]	Proteasome	1.7×10^{-6}	N
BTN family		[5]	Endocytosis	5.0×10^{-5}	N
BTN family		[5]	Intestinal immune network for IgA production	5.8×10^{-5}	N
C10orf88		[2]	retinoic acid receptor binding	1.3×10^{-6}	N
C10orf88		[2]	retinoid X receptor binding	3.2×10-6	N
C10orf88		[4]	Mitotic Spindle Checkpoint	3.8×10^{-6}	N
C10orf88		[4]	Inactivation of APC/C via direct inhibition of the APC/C complex	4.4×10^{-6}	N
C10orf88		[4]	Inhibition of the proteolytic activity of APC/C required for the onset of anaphase by...	4.4×10^{-6}	N
C10orf88		[4]	APC-Cdc20 mediated degradation of Nek2A	7.8×10^{-6}	N
C10orf88		[5]	Basal transcription factors	4.6×10^{-4}	N
GBX2	*	[1]	cell differentiation in spinal cord	4.7×10^{-12}	N
GBX2		[1]	stem cell differentiation	1.2×10^{-10}	N
GBX2	*	[1]	dorsal spinal cord development	2.2×10^{-10}	N
GBX2	*	[1]	spinal cord development	2.6×10^{-10}	N
GBX2	*	[1]	spinal cord dorsal/ventral patterning	2.8×10^{-10}	N
GBX2	*	[1]	spinal cord patterning	7.3×10^{-10}	N

GBX2	*	1	nerve development	1.4×10^{-9}	N
GBX2	*	1	neural tube development	2.0×10^{-9}	Y
GBX2		1	regionalization	2.5×10^{-9}	Y
GBX2	*	1	neuron fate commitment	2.6×10^{-9}	N
GBX2	*	1	positive regulation of neuron differentiation	4.6×10^{-9}	N
GBX2		1	pattern specification process	5.0×10^{-9}	Y
GBX2	*	1	cranial nerve development	6.0×10^{-9}	N
GBX2	*	1	neuron fate specification	9.5×10^{-9}	N
GBX2		1	morphogenesis of embryonic epithelium	2.3×10^{-8}	N
GBX2	*	1	negative regulation of glial cell differentiation	2.5×10^{-8}	N
GBX2		1	cochlea morphogenesis	4.6×10^{-8}	N
GBX2	*	1	parasympathetic nervous system development	5.3×10^{-8}	N
GBX2	*	1	neuromuscular process	5.8×10^{-8}	N
GBX2		1	cell fate specification	5.9×10^{-8}	N
GBX2		5	Basal cell carcinoma	9.3×10^{-6}	N
GBX2		2	Notch binding	1.5×10^{-5}	N
GBX2		5	Renal cell carcinoma	5.2×10^{-5}	N
GBX2		5	Notch signaling pathway	8.2×10^{-5}	N
GBX2		5	Aldosterone-regulated sodium reabsorption	3.2×10^{-4}	N
GBX2		5	Proximal tubule bicarbonate reclamation	6.6×10^{-4}	N
HIST1H family		3	nucleosome	3.5×10^{-82}	Y
HIST1H family		1	regulation of gene silencing	2.5×10^{-80}	N
HIST1H family		1	nucleosome assembly	8.3×10^{-77}	Y
HIST1H family		3	protein-DNA complex	2.6×10^{-75}	Y
HIST1H family		1	chromatin assembly	1.6×10^{-74}	Y
HIST1H family		1	nucleosome organization	2.6×10^{-73}	Y
HIST1H family		1	protein-DNA complex assembly	7.3×10^{-73}	Y
HIST1H family		5	Systemic lupus erythematosus	5.9×10^{-72}	Y
HIST1H family		1	chromatin assembly or disassembly	1.6×10^{-71}	Y
HIST1H family		1	protein-DNA complex subunit organization	1.1×10^{-70}	Y
HIST1H family		1	DNA packaging	3.3×10^{-67}	Y
HIST1H family		1	DNA conformation change	5.5×10^{-65}	Y
HIST1H family		3	chromatin	6.8×10^{-60}	Y
HIST1H family		4	RNA Polymerase I Promoter Opening	1.2×10^{-55}	Y
HIST1H family		1	regulation of megakaryocyte differentiation	5.3×10^{-51}	Y
HIST1H family		1	cellular macromolecular complex assembly	1.5×10^{-48}	Y
HIST1H family		4	RNA Polymerase I Chain Elongation	3.2×10^{-48}	Y
HIST1H family		4	RNA Polymerase I Promoter Clearance	4.1×10^{-47}	Y
HIST1H family		4	RNA Polymerase I Transcription	2.2×10^{-46}	Y
HIST1H family		4	Meiotic Recombination	7.3×10^{-43}	Y
HIST1H family		4	Amyloids	1.7×10^{-42}	Y
HIST1H family		4	Packaging Of Telomere Ends	7.4×10^{-42}	Y
HIST1H family		4	Activation of DNA fragmentation factor	8.4×10^{-38}	Y
HIST1H family		4	Apoptosis induced DNA fragmentation	8.4×10^{-38}	N
HIST1H family		1	megakaryocyte differentiation	3.9×10^{-37}	Y
HIST1H family		1	chromatin organization	4.3×10^{-36}	Y
HIST1H family		1	CenH3-containing nucleosome assembly at centromere	5.0×10^{-36}	Y
HIST1H family		1	DNA replication-independent nucleosome assembly	5.0×10^{-36}	Y
HIST1H family		1	DNA replication-independent nucleosome organization	5.0×10^{-36}	Y
HIST1H family		4	Deposition of New CENPA-containing Nucleosomes at the Centromere	7.3×10^{-36}	Y
HIST1H family		4	Nucleosome assembly	7.3×10^{-36}	N
HIST1H family		4	Meiotic Synapsis	7.4×10^{-34}	Y
HIST1H family		1	chromatin remodeling at centromere	3.5×10^{-33}	Y
HIST1H family		4	Meiosis	8.7×10^{-33}	Y
HIST1H family		4	RNA Polymerase I, RNA Polymerase III, and Mitochondrial Transcription	9.0×10^{-33}	Y
HIST1H family		1	gene silencing	2.8×10^{-32}	N
HIST1H family		1	histone exchange	3.4×10-32	Y
HIST1H family		3	chromosomal part	3.6×10^{-32}	Y
HIST1H family		1	ATP-dependent chromatin remodeling	8.7×10^{-30}	Y
HIST1H family		4	Telomere Maintenance	1.6×10^{-28}	Y
HIST1H family		4	Chromosome Maintenance	4.7×10^{-19}	Y

HIST1H family		4	Transcription	$4.3\times10\text{-}^{17}$	Y
HIST1H family		4	Apoptotic execution phase	4.9×10^{-15}	N
HMGN4		5	Basal transcription factors	9.8×10^{-4}	N
IP6K3		1	muscle cell fate commitment	6.9×10^{-12}	N
IP6K3		1	striated muscle cell development	1.6×10^{-10}	N
IP6K3		1	skeletal muscle tissue development	4.4×10^{-10}	N
IP6K3		1	skeletal muscle organ development	6.0×10^{-10}	N
IP6K3		1	muscle system process	1.8×10^{-9}	N
IP6K3	*	1	neuromuscular junction development	1.9×10^{-9}	N
IP6K3		1	muscle organ development	2.3×10^{-9}	N
IP6K3		3	I band	3.2×10^{-9}	N
IP6K3		3	myofibril	3.9×10^{-9}	N
IP6K3		1	muscle structure development	4.5×10^{-9}	N
IP6K3		1	muscle contraction	4.8×10^{-9}	N
IP6K3		3	contractile fiber	7.0×10^{-9}	N
IP6K3		1	muscle fiber development	1.9×10^{-8}	N
IP6K3		2	structural constituent of muscle	1.9×10^{-8}	N
IP6K3		1	multicellular organismal movement	2.1×10^{-8}	N
IP6K3		1	muscle cell development	2.1×10^{-8}	N
IP6K3		1	musculoskeletal movement	2.1×10^{-8}	N
IP6K3		3	sarcomere	2.8×10^{-8}	N
IP6K3		3	contractile fiber part	3.7×10^{-8}	N
IP6K3		1	skeletal muscle contraction	4.7×10^{-8}	N
IP6K3		1	striated muscle contraction	1.0×10^{-7}	N
IP6K3	*	2	acetylcholine-activated cation-selective channel activity	1.6×10^{-7}	N
IP6K3		1	muscle cell differentiation	2.7×10^{-7}	N
IP6K3		2	titin binding	3.8×10^{-7}	N
IP6K3		3	sarcoplasm	5.1×10^{-7}	N
IP6K3		1	skeletal muscle fiber development	7.2×10^{-7}	N
IP6K3		3	acetylcholine-gated channel complex	7.3×10^{-7}	N
IP6K3		3	Z disc	8.2×10^{-7}	N
IP6K3		3	myosin filament	9.7×10^{-7}	N
IP6K3		1	striated muscle cell differentiation	1.4×10^{-6}	N
IP6K3		4	Acetylcholine Binding And Downstream Events	1.6×10^{-6}	N
IP6K3	*	4	Activation of Nicotinic Acetylcholine Receptors	1.6×10^{-6}	N
IP6K3	*	4	Postsynaptic nicotinic acetylcholine receptors	1.6×10^{-6}	N
IP6K3		3	sarcoplasmic reticulum	2.0×10^{-6}	N
IP6K3	*	4	Presynaptic nicotinic acetylcholine receptors	2.8×10^{-6}	N
IP6K3		5	Hypertrophic cardiomyopathy (HCM)	4.5×10^{-6}	N
IP6K3		1	striated muscle tissue development	4.7×10^{-6}	N
IP6K3		1	actin-mediated cell contraction	4.9×10^{-6}	N
IP6K3	*	3	neuromuscular junction	5.3×10^{-6}	N
IP6K3		4	Striated Muscle Contraction	6.3×10^{-6}	N
IP6K3	*	5	Cardiac muscle contraction	1.3×10^{-5}	N
IP6K3	*	2	acetylcholine binding	1.3×10^{-5}	N
IP6K3		3	sarcolemma	2.6×10^{-5}	N
IP6K3		4	Muscle contraction	3.2×10^{-5}	N
IP6K3	*	2	acetylcholine receptor activity	3.4×10^{-5}	N
IP6K3		3	actin cytoskeleton	4.1×10^{-5}	N
IP6K3		3	sarcoplasmic reticulum membrane	5.7×10^{-5}	N
IP6K3		3	myosin complex	9.8×10^{-5}	N
IP6K3	*	4	Highly calcium permeable postsynaptic nicotinic acetylcholine receptors	1.0×10^{-4}	N
IP6K3		3	pseudopodium	1.2×10^{-4}	N
IP6K3		5	Dilated cardiomyopathy	1.3×10^{-4}	N
ITPR3		3	lateral plasma membrane	8.4×10^{-11}	N
ITPR3		3	basal plasma membrane	2.5×10^{-10}	N
ITPR3		1	hemidesmosome assembly	2.9×10^{-9}	N
ITPR3		3	basal part of cell	3.6×10^{-9}	N
ITPR3		5	VEGF signaling pathway	1.5×10^{-8}	N
ITPR3		3	laminin complex	1.8×10^{-7}	N
ITPR3		2	protein kinase C activity	2.4×10^{-7}	N
ITPR3		4	Cell junction organization	3.3×10^{-7}	N

ITPR3	*	1	neural crest cell migration	4.1×10-7	N
ITPR3	*	5	Neuroactive ligand-receptor interaction	5.7×10^{-7}	N
ITPR3		4	Cell-Cell communication	8.3×10^{-7}	N
ITPR3		3	cell-cell junction	9.7×10^{-7}	N
ITPR3		5	Thyroid cancer	1.2×10^{-6}	N
ITPR3		3	basal lamina	3.3×10^{-6}	N
ITPR3		3	leading edge membrane	8.1×10^{-6}	N
ITPR3		3	lamellipodium membrane	8.8×10^{-6}	N
ITPR3		5	Small cell lung cancer	3.7×10^{-5}	N
ITPR3		5	Amino sugar and nucleotide sugar metabolism	6.7×10^{-5}	N
ITPR3		5	Glycosaminoglycan degradation	1.6×10^{-4}	N
ITPR3		5	Pathways in cancer	2.5×10^{-4}	N
ITPR3		5	ECM-receptor interaction	2.5×10^{-4}	N
LRRC16A		3	cell leading edge	1.1×10^{-6}	Y
LRRC16A		2	unfolded protein binding	8.4×10^{-6}	N
LRRC16A		3	filopodium	9.0×10^{-6}	N
LRRC16A		2	gamma-catenin binding	2.1×10^{-5}	N
LRRC16A		4	Eicosanoid ligand-binding receptors	3.5×10^{-5}	N
LRRC16A		5	Adherens junction	9.3×10^{-5}	N
LRRN2	*	3	dendrite	1.2×10^{-10}	N
LRRN2	*	3	dendritic spine head	3.3×10^{-10}	N
LRRN2	*	3	postsynaptic density	3.3×10^{-10}	N
LRRN2	*	3	dendritic spine	6.0×10^{-10}	N
LRRN2	*	3	neuron spine	6.0×10^{-10}	N
LRRN2	*	2	extracellular-glutamate-gated ion channel activity	9.9×10^{-10}	N
LRRN2	*	2	ionotropic glutamate receptor activity	1.5×10^{-9}	N
LRRN2	*	1	synapse organization	2.5×10^{-9}	N
LRRN2	*	3	ionotropic glutamate receptor complex	2.6×10^{-9}	N
LRRN2	*	1	regulation of synapse organization	3.8×10^{-9}	N
LRRN2	*	1	positive regulation of nervous system development	1.3×10^{-8}	N
LRRN2	*	1	positive regulation of synapse assembly	1.3×10^{-8}	N
LRRN2		3	outer membrane-bounded periplasmic space	2.5×10^{-8}	N
LRRN2		3	periplasmic space	2.5×10^{-8}	N
LRRN2	*	1	regulation of transmission of nerve impulse	2.6×10^{-8}	N
LRRN2	*	1	regulation of synapse structure and activity	3.6×10^{-8}	N
LRRN2	*	1	regulation of synapse assembly	4.2×10^{-8}	N
LRRN2		1	positive regulation of cellular component biogenesis	6.8×10^{-8}	N
LRRN2	*	3	excitatory synapse	8.7×10^{-8}	N
LRRN2	*	1	regulation of synaptic transmission	8.7×10^{-8}	N
LRRN2	*	2	glutamate receptor activity	9.7×10^{-8}	N
LRRN2	*	3	synapse part	2.5×10^{-7}	N
LRRN2		3	cell envelope	2.7×10^{-7}	N
LRRN2		3	external encapsulating structure part	2.7×10^{-7}	N
LRRN2	*	3	synapse	2.8×10^{-7}	N
LRRN2	*	1	synapse assembly	2.9×10^{-7}	N
LRRN2		1	regulation of glomerulus development	4.2×10^{-7}	N
LRRN2		3	external encapsulating structure	4.4×10^{-7}	N
LRRN2	*	3	neuronal cell body	4.6×10^{-7}	N
LRRN2	*	1	regulation of neurological system process	6.8×10^{-7}	N
LRRN2		3	cell body	7.3×10^{-7}	N
LRRN2	*	3	synaptic membrane	7.6×10^{-7}	N
LRRN2	*	3	postsynaptic membrane	2.0×10^{-6}	N
LRRN2		4	Potassium Channels	1.0×10^{-5}	N
LRRN2	*	3	dendritic shaft	2.4×10^{-5}	N
LRRN2	*	3	alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid selective glutamate receptor...	2.8×10^{-5}	N
MDM4		1	regulation of RNA splicing	1.3×10^{-7}	N
MDM4		1	protein alkylation	7.9×10^{-7}	N
MDM4		1	protein methylation	7.9×10^{-7}	N
MDM4		2	N-methyltransferase activity	1.2×10^{-6}	N
MDM4		2	histone-lysine N-methyltransferase activity	1.3×10^{-6}	N
MDM4		2	protein methyltransferase activity	3.9×10^{-6}	N
MDM4		2	lysine N-methyltransferase activity	8.0×10^{-6}	N

MDM4		2	protein-lysine N-methyltransferase activity	8.0×10^{-6}	N
MDM4		2	histone methyltransferase activity	8.7×10^{-6}	N
MDM4		2	S-adenosylmethionine-dependent methyltransferase activity	1.3×10^{-5}	N
MDM4		2	protein serine/threonine/tyrosine kinase activity	3.6×10^{-5}	N
MDM4		4	PI3K/AKT activation	9.3×10^{-5}	N
MDM4		3	heterogeneous nuclear ribonucleoprotein complex	1.3×10^{-4}	N
PIK3C2B	*	1	regulation of oligodendrocyte differentiation	1.1×10^{-7}	N
PIK3C2B		4	Nitric oxide stimulates guanylate cyclase	2.0×10^{-7}	N
PIK3C2B		2	Ras guanyl-nucleotide exchange factor activity	1.1×10^{-6}	N
PIK3C2B		2	guanyl-nucleotide exchange factor activity	2.6×10^{-6}	N
PIK3C2B		4	Platelet homeostasis	6.3×10^{-6}	N
PIK3C2B		5	B cell receptor signaling pathway	7.3×10^{-5}	N
PIK3C2B	*	5	Axon guidance	1.2×10^{-4}	N
RNF123		1	heme biosynthetic process	6.4×10^{-26}	N
RNF123		1	porphyrin-containing compound biosynthetic process	5.8×10^{-24}	N
RNF123		1	tetrapyrrole biosynthetic process	5.8×10^{-24}	N
RNF123		1	porphyrin-containing compound metabolic process	1.7×10^{-22}	N
RNF123		1	tetrapyrrole metabolic process	1.7×10^{-22}	N
RNF123		1	heme metabolic process	1.0×10^{-20}	N
RNF123		4	Metabolism of porphyrins	2.3×10^{-15}	N
RNF123		1	hemoglobin metabolic process	8.2×10^{-15}	N
RNF123		1	protein deubiquitination	4.1×10^{-13}	N
RNF123		1	protein modification by small protein removal	2.6×10^{-11}	N
RNF123		2	polyubiquitin binding	2.1×10^{-10}	N
RNF123		1	protein K48-linked deubiquitination	6.4×10^{-10}	N
RNF123		1	cofactor biosynthetic process	6.7×10^{-10}	N
RNF123		5	Porphyrin and chlorophyll metabolism	8.5×10^{-9}	N
RNF123		1	response to arsenic-containing substance	1.8×10^{-8}	N
RNF123		1	actin filament capping	2.2×10^{-8}	N
RNF123		1	pigment biosynthetic process	3.1×10^{-8}	N
RNF123		1	negative regulation of actin filament depolymerization	3.8×10^{-8}	N
RNF123		2	ubiquitin-specific protease activity	5.2×10^{-8}	N
RNF123		2	small conjugating protein binding	5.3×10^{-8}	N
RNF123		2	ubiquitin binding	1.1×10^{-7}	N
RNF123		2	small conjugating protein-specific protease activity	2.9×10^{-7}	N
RNF123		2	ferrous iron binding	3.1×10^{-7}	N
RNF123		2	protein serine/threonine/tyrosine kinase activity	3.9×10^{-7}	N
RNF123		3	CUL4 RING ubiquitin ligase complex	1.1×10^{-6}	N
RNF123		5	Valine, leucine and isoleucine biosynthesis	1.1×10^{-4}	N
RNF123		5	ABC transporters	1.1×10^{-4}	N
RNF123		5	SNARE interactions in vesicular transport	4.0×10^{-4}	N
RNF123		5	Non-small cell lung cancer	4.1×10^{-4}	N
STK24		2	Rho guanyl-nucleotide exchange factor activity	2.6×10^{-8}	N
STK24		4	G alpha (12/13) signalling events	1.8×10^{-7}	N
STK24		5	Adherens junction	1.8×10^{-7}	N
STK24		4	NRAGE signals death through JNK	8.7×10^{-7}	N
STK24		2	receptor signaling protein activity	2.3×10^{-6}	N
STK24		5	Thyroid cancer	2.2×10^{-4}	N
STK24		5	Regulation of actin cytoskeleton	4.2×10^{-4}	N
STK24		5	Renal cell carcinoma	5.8×10^{-4}	N
STK24		5	ErbB signaling pathway	7.7×10^{-4}	N
TANK		4	NOD1/2 Signaling Pathway	1.3×10^{-14}	N
TANK		4	Death Receptor Signalling	3.1×10^{-14}	N
TANK		4	Extrinsic Pathway for Apoptosis	3.1×10^{-14}	N
TANK		1	pattern recognition receptor signaling pathway	8.2×10^{-13}	N
TANK		1	toll-like receptor signaling pathway	8.9×10^{-13}	N
TANK		1	positive regulation of T cell mediated immunity	9.8×10^{-13}	N
TANK		1	innate immune response-activating signal transduction	1.3×10^{-12}	N
TANK		1	positive regulation of innate immune response	4.2×10^{-12}	N
TANK		1	positive regulation of leukocyte mediated immunity	6.6×10^{-12}	N
TANK		1	positive regulation of lymphocyte mediated immunity	6.6×10^{-12}	N
TANK		1	positive regulation of NF-kappaB transcription factor activity	7.5×10^{-12}	N

TANK	[1]	positive regulation of adaptive immune response based on somatic recombination of imm...	1.3×10^{-11}	N
TANK	[1]	activation of innate immune response	1.4×10^{-11}	N
TANK	[1]	toll-like receptor 4 signaling pathway	1.5×10^{-11}	N
TANK	[1]	alpha-beta T cell proliferation	1.7×10^{-11}	N
TANK	[1]	positive regulation of adaptive immune response	2.2×10^{-11}	N
TANK	[1]	positive regulation of interleukin-10 production	2.7×10^{-11}	N
TANK	[5]	Apoptosis	3.3×10^{-11}	N
TANK	[1]	positive regulation of defense response	6.6×10^{-11}	N
TANK	[1]	toll-like receptor 3 signaling pathway	7.2×10^{-11}	N
TANK	[4]	Regulation of IFNG signaling	9.5×10^{-11}	N
TANK	[1]	Toll signaling pathway	1.0×10^{-10}	N
TANK	[1]	MyD88-independent toll-like receptor signaling pathway	1.3×10^{-10}	N
TANK	[1]	regulation of innate immune response	1.9×10^{-10}	N
TANK	[1]	positive regulation of leukocyte proliferation	2.0×10^{-10}	N
TANK	[4]	Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signaling pa...	8.3×10^{-10}	N
TANK	[4]	MyD88-independent cascade initiated on plasma membrane	1.1×10^{-9}	N
TANK	[5]	Toll-like receptor signaling pathway	1.1×10^{-9}	N
TANK	[4]	Toll Like Receptor 3 (TLR3) Cascade	1.2×10^{-9}	N
TANK	[4]	TRIF mediated TLR3 signaling	1.2×10^{-9}	N
TANK	[5]	RIG-I-like receptor signaling pathway	1.3×10^{-9}	Y
TANK	[2]	tumor necrosis factor receptor binding	1.8×10^{-9}	N
TANK	[5]	NOD-like receptor signaling pathway	2.9×10^{-9}	N
TANK	[4]	Innate Immune System	3.8×10^{-9}	Y
TANK	[4]	Activated TLR4 signalling	6.6×10^{-9}	N
TANK	[4]	TAK1 activates NFkB by phosphorylation and activation of IKKs complex	9.6×10^{-9}	N
TANK	[4]	Toll Like Receptor 4 (TLR4) Cascade	1.4×10^{-8}	N
TANK	[4]	Toll Receptor Cascades	1.5×10^{-8}	N
TANK	[4]	TRAF6 mediated NF-kB activation	1.6×10^{-8}	N
TANK	[2]	tumor necrosis factor receptor superfamily binding	3.3×10^{-8}	N
TANK	[4]	NFkB and MAP kinases activation mediated by TLR4 signaling repertoire	4.2×10^{-8}	N
TANK	[4]	Interleukin-1 signaling	5.6×10^{-8}	N
TANK	[4]	TRAF6 Mediated Induction of proinflammatory cytokines	1.4×10^{-7}	N
TANK	[4]	MyD88 cascade initiated on plasma membrane	2.4×10^{-7}	N
TANK	[4]	Toll Like Receptor 10 (TLR10) Cascade	2.4×10^{-7}	N
TANK	[4]	Toll Like Receptor 5 (TLR5) Cascade	2.4×10^{-7}	N
TANK	[5]	Cytokine-cytokine receptor interaction	9.3×10^{-7}	N
TANK	[5]	Leishmania infection	3.3×10^{-6}	N
TANK	[5]	T cell receptor signaling pathway	4.7×10^{-6}	N
TANK	[5]	Jak-STAT signaling pathway	6.6×10^{-6}	N
TANK	[5]	Amyotrophic lateral sclerosis (ALS)	1.4×10^{-5}	N
TANK	[5]	Pancreatic cancer	1.7×10^{-4}	N
TANK	[5]	Small cell lung cancer	4.5×10^{-4}	N
TANK	[5]	Epithelial cell signaling in Helicobacter pylori infection	1.0×10^{-3}	N
TET2	[2]	thyroid hormone receptor binding	1.1×10^{-6}	N
TET2	[1]	positive regulation of gene expression, epigenetic	1.8×10^{-5}	N
TET2	[2]	kinase activator activity	4.0×10^{-5}	N
TET2	[4]	Transcriptional Regulation of White Adipocyte Differentiation	7.2×10^{-5}	N
TET2	* [4]	BMAL1:CLOCK/NPAS2 Activates Gene Expression	8.6×10^{-5}	N
TET2	[5]	Other glycan degradation	8.2×10^{-4}	N
TUFM	[3]	mitochondrial matrix	9.1×10^{-34}	Y
TUFM	[3]	mitochondrial inner membrane	2.5×10^{-23}	N
TUFM	[3]	organelle inner membrane	4.1×10^{-23}	N
TUFM	[2]	4 iron, 4 sulfur cluster binding	5.6×10^{-23}	N
TUFM	[3]	mitochondrial membrane	9.9×10^{-23}	N
TUFM	[3]	mitochondrial envelope	1.0×10^{-22}	N
TUFM	[3]	mitochondrial nucleoid	1.3×10^{-22}	Y
TUFM	[3]	nucleoid	2.3×10^{-21}	Y
TUFM	[1]	mitochondrion organization	2.6×10^{-17}	N
TUFM	[4]	Mitochondrial tRNA aminoacylation	1.5×10^{-16}	N

TUFM		[1]	aerobic respiration	3.4×10^{-15}	N
TUFM		[3]	mitochondrial membrane part	4.7×10^{-15}	N
TUFM		[2]	iron-sulfur cluster binding	9.0×10^{-15}	N
TUFM		[2]	metal cluster binding	9.0×10^{-15}	N
TUFM		[4]	Citric acid cycle (TCA cycle)	7.1×10^{-14}	N
TUFM		[1]	cellular respiration	7.5×10^{-14}	N
TUFM		[4]	The citric acid (TCA) cycle and respiratory electron transport	9.8×10^{-14}	N
TUFM		[1]	oxidative phosphorylation	4.7×10^{-13}	N
TUFM		[1]	respiratory electron transport chain	6.5×10^{-13}	N
TUFM		[1]	quinone cofactor metabolic process	1.1×10^{-12}	N
TUFM		[3]	respiratory chain	1.2×10^{-12}	N
TUFM		[4]	Pyruvate metabolism and Citric Acid (TCA) cycle	2.6×10^{-12}	N
TUFM		[3]	mitochondrial respiratory chain	2.9×10^{-12}	N
TUFM	*	[5]	Parkinson's disease	7.6×10^{-12}	N
TUFM		[3]	mitochondrial ribosome	8.5×10^{-12}	N
TUFM		[3]	organellar ribosome	8.5×10^{-12}	N
TUFM		[1]	mitochondrial translation	1.5×10^{-11}	N
TUFM		[1]	cofactor metabolic process	1.6×10^{-11}	N
TUFM		[1]	electron transport chain	2.2×10^{-11}	N
TUFM		[1]	ATP synthesis coupled electron transport	3.0×10^{-11}	N
TUFM		[1]	mitochondrial ATP synthesis coupled electron transport	3.0×10^{-11}	N
TUFM		[4]	Respiratory electron transport	3.4×10^{-11}	N
TUFM		[4]	Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat prod...	5.8×10^{-11}	N
TUFM		[1]	energy derivation by oxidation of organic compounds	6.1×10^{-11}	N
TUFM	*	[5]	Huntington's disease	1.0×10^{-10}	N
TUFM		[1]	mitochondrial RNA metabolic process	$1.7\times10^{-1}0$	N
TUFM		[3]	ribosome	1.7×10^{-10}	N
TUFM		[1]	branched chain family amino acid catabolic process	1.8×10^{-10}	N
TUFM		[1]	generation of precursor metabolites and energy	1.8×10^{-10}	N
TUFM		[5]	Valine, leucine and isoleucine degradation	2.1×10^{-10}	N
TUFM		[1]	tRNA metabolic process	2.5×10^{-10}	N
TUFM		[1]	cofactor biosynthetic process	3.5×10^{-10}	N
TUFM		[5]	Oxidative phosphorylation	3.7×10^{-10}	N
TUFM		[1]	tricarboxylic acid cycle	4.3×10^{-10}	N
TUFM		[4]	Mitochondrial Fatty Acid Beta-Oxidation	4.4×10^{-10}	N
TUFM		[3]	small ribosomal subunit	4.9×10^{-10}	N
TUFM		[3]	mitochondrial small ribosomal subunit	5.0×10^{-10}	N
TUFM		[3]	organellar small ribosomal subunit	5.0×10^{-10}	N
TUFM		[3]	integral to mitochondrial membrane	6.0×10^{-10}	N
TUFM		[1]	coenzyme metabolic process	7.4×10^{-10}	N
TUFM		[1]	acetyl-CoA catabolic process	7.6×10^{-10}	N
TUFM		[2]	unfolded protein binding	2.2×10^{-9}	N
TUFM		[5]	Citrate cycle (TCA cycle)	3.0×10^{-9}	N
TUFM		[2]	hydrogen ion transporting ATP synthase activity, rotational mechanism	5.4×10^{-9}	N
TUFM		[2]	cofactor binding	1.0×10^{-8}	N
TUFM	*	[5]	Alzheimer's disease	1.3×10^{-8}	N
TUFM		[3]	mitochondrial respiratory chain complex I	1.9×10^{-8}	N
TUFM		[3]	NADH dehydrogenase complex	1.9×10^{-8}	N
TUFM		[3]	respiratory chain complex I	1.9×10^{-8}	N
TUFM		[2]	oxidoreductase activity, acting on NADH or NADPH	1.9×10^{-8}	N
TUFM		[2]	aminoacyl-tRNA ligase activity	2.5×10^{-8}	N
TUFM		[2]	ligase activity, forming aminoacyl-tRNA and related compounds	2.5×10^{-8}	N
TUFM		[2]	ligase activity, forming carbon-oxygen bonds	2.5×10^{-8}	N
TUFM		[2]	NADH dehydrogenase (quinone) activity	2.7×10^{-8}	N
TUFM		[2]	NADH dehydrogenase (ubiquinone) activity	2.7×10^{-8}	N
TUFM		[2]	NADH dehydrogenase activity	2.7×10^{-8}	N
TUFM		[5]	Aminoacyl-tRNA biosynthesis	2.9×10^{-8}	N
TUFM		[2]	oxidoreductase activity, acting on the CH-CH group of donors	3.9×10^{-8}	N
TUFM		[2]	structural constituent of ribosome	6.5×10^{-8}	N
TUFM		[4]	tRNA Aminoacylation	7.3×10^{-8}	N
TUFM		[4]	Gluconeogenesis	7.5×10^{-8}	N

TUFM	[4]	RNA Polymerase III Transcription Initiation From Type 1 Promoter	1.6×10^{-7}	N
TUFM	[5]	Propanoate metabolism	1.7×10^{-7}	N
TUFM	[4]	RNA Polymerase III Transcription Initiation From Type 2 Promoter	2.1×10^{-7}	N
TUFM	[4]	Formation of ATP by chemiosmotic coupling	3.8×10^{-7}	N
TUFM	[4]	RNA Polymerase III Chain Elongation	4.3×10^{-7}	N
TUFM	[2]	NAD binding	4.3×10^{-7}	N
TUFM	[2]	coenzyme binding	5.0×10^{-7}	N
TUFM	[5]	Butanoate metabolism	6.0×10^{-7}	N
TUFM	[2]	modified amino acid binding	6.7×10^{-7}	N
TUFM	[2]	oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acce...	7.4×10^{-7}	N
TUFM	[2]	translation factor activity, nucleic acid binding	9.9×10^{-7}	Y
TUFM	[5]	beta-Alanine metabolism	1.1×10^{-6}	N
TUFM	[5]	Selenoamino acid metabolism	1.4×10^{-6}	N
TUFM	[4]	Branched-chain amino acid catabolism	1.8×10^{-6}	N
TUFM	[5]	Fatty acid metabolism	3.3×10^{-6}	N
TUFM	[5]	Pyruvate metabolism	4.0×10^{-6}	N
TUFM	[5]	RNA polymerase	1.7×10^{-5}	N
TUFM	[5]	Valine, leucine and isoleucine biosynthesis	1.8×10^{-5}	N
TUFM	[5]	Glycolysis / Gluconeogenesis	9.4×10^{-5}	N
TUFM	[5]	Cardiac muscle contraction	1.0×10^{-4}	N
TUFM	[5]	Lysine degradation	1.7×10^{-4}	N
TUFM	[5]	Oocyte meiosis	3.7×10^{-4}	N
TUFM	[5]	Glyoxylate and dicarboxylate metabolism	3.8×10^{-4}	N

Table S22. Candidate-gene regions (see Table S20) with previously reported associations in human GWAS and/or evidence of neurological or central nervous system function in mouse or zebrafish models.

MDM4, LRRN2

Central nervous system development (*MDM4*, mouse, *(67)*)

Cognitive performance *(61)*

Cavities (*LRRN2*, *(155)*)

AFF3

Type 1 diabetes nephropathy *(158)*

Rheumatoid arthritis *(79)*

TANK

Subcutaneous adipose tissue *(162)*

Treatment response for severe sepsis *(164)*

Amyotrophic lateral sclerosis *(165)*

GBX2, ASB18

Anterior hindbrain development (*GBX2*, zebrafish, *(63)*; *GBX2*, mouse, *(64)*)

Striatal cholinergic interneuron development (*GBX2*, mouse, *(65)*

Response to statin therapy (*ASB18*, *(169)*

BSN, APEH, MST1

Glutamatergic synapse function (*BSN*, mouse, *(69)*)

Crohn's disease (*BSN*, *(70*, *71)*; *MST1*, *(73*, *74)*

Ulcerative colitis (*BSN*, *(72)*; *APEH*, *(77)*; *MST1*, *(76)*)

Inflammatory bowel disease (*MST1*, *(75)*)

Primary sclerosing cholangitis (*MST1*, *(173)*)

TET2

Height *(15)*

Pulmonary function *(153)*

Tuberculosis *(154)*

Prostate cancer *(156)*

Immune response to anthrax vaccine *(157)*

LRRC16A, BTN1A1

Iron status biomarkers *(159)*

Mean platelet volume (*LRRC16A*, *(160)*

Platelet counts (*LRRC16A*, *(161)*)

Uric acid levels (*LRRC16A*, *(163)*)

ITPR3

Height *(15)*

Obesity *(166)*

Graves' disease *(167)*

IP6K3

Serum phosphorous levels *(168)*

STK24

Alzheimer's disease *(62)*

Longevity *(170)*

ATXN2L, TUFM, SH2B1, ATP2A1

Inflammatory bowel disease (early onset) (*ATXN2L*, *(78)*)

Body mass index (*SH2B1*, *(18*, *171*, *172)*)

Weight (*SH2B1*, *(171)*)

Table S23. Regression results of the polygenic scores (*PGSs*) on *College*, *EduYears* and *Cognitive function* in a set of unrelated individuals of the QIMR (N = 3,526) and STR (N = 6,770) cohorts using SNPs selected from the meta-analysis excluding the QIMR and STR cohorts. Results for cognitive function are based on a sample of 1,419 individuals from STR. The R^2's reported in this table are illustrated in Figure 2.

		Prediction in QIMR				Prediction in STR			
Phenotype (PGS)		$p_{SNPs} < 5\times10^{-8}$	$p_{SNPs} < 5\times10^{-5}$	$p_{SNPs} < 5\times10^{-3}$	All SNPs	$p_{SNPs} < 5\times10^{-8}$	$p_{SNPs} < 5\times10^{-5}$	$p_{SNPs} < 5\times10^{-3}$	All SNPs
EduYears (*College*)	R^2 (%)	0.023	0.210	1.180	2.910	0.170	0.230	0.720	1.800
	p-value	0.370	0.007	9.1×10^{-11}	1.4×10^{-24}	6.1×10^{-4}	6.9×10^{-5}	3.1×10^{-12}	1.2×10^{-28}
EduYears (*EduYears*)	R^2 (%)	0.005	0.560	1.020	2.820	0.110	0.370	0.610	1.880
	p-value	0.689	7.6×10^{-6}	1.7×10^{-9}	7.1×10^{-24}	6.5×10^{-3}	6.4×10^{-7}	1.4×10^{-10}	1.0×10^{-29}
Cognitive function (*College*)	R^2 (%)					0.000	0.160	0.380	2.380
	p-value					0.986	0.137	0.021	5.3×10^{-9}
Cognitive function (*EduYears*)	R^2 (%)					0.190	0.420	0.220	2.580
	p-value					0.103	0.015	0.077	1.2×10^{-9}

Table S24. Results of a mediation analysis on educational attainment using the polygenic scores (*PGSs*) from Supplementary Table 12 and a measure of cognitive function in a set of unrelated individuals in the STR sample (N = 1,419). All variables are standardized to z-scores. The effect sizes should be interpreted in standard-deviation units. The indirect effect measures the extent to which *EduYears* changes when the *PGS* is held fixed and cognitive function changes to the level it would have attained had the *PGS* increased by one unit *(174)*. Put another way, the indirect effect is the difference between the coefficient on the *PGS* with cognitive function as a covariate and the coefficient without it. When cognitive function is included as a covariate, the coefficient on *PGS* declines by ~2/3 and is no longer statistically distinguishable from zero. These findings are consistent with the hypothesis that cognitive function mediates the relationship between the *PGS* and educational attainment.

	PGS = polygenic score from GWAS for *College*			*PGS* = polygenic score from GWAS for *EduYears*		
	Est.	SE	*P*	Est.	SE	*P*
EduYears regressed on *PGS*						
Polygenic score	0.0974	0.0256	1.5×10^{-4}	0.1156	0.0254	5.8×10^{-6}
Cognitive function regressed on *PGS*						
Polygenic score	0.1464	0.0265	3.9×10^{-8}	0.1536	0.0263	6.6×10^{-9}
EduYears regressed on *PGS* + *Cognitive function*						
Polygenic score	0.0321	0.0230	1.6×10^{-1}	0.0475	0.0228	3.8×10^{-2}
Cognitive function	0.4464	0.0234	4.6×10^{-72}	0.4436	0.0234	3.0×10^{-71}
Indirect effect	0.0653	0.0123	1.3×10^{-7}	0.0681	0.0122	2.9×10^{-8}

Table S25. Within-family regression results of the polygenic scores (*PGSs*) on *College*, *EduYears* and *Cognitive function* in the QIMR and STR cohorts using SNPs selected from the meta-analysis excluding the QIMR and STR cohorts. Analyses for QIMR are based on 572 full-sib pairs from independent 572 families (QIMR), and analyses for STR are based on 2,774 DZ twins from 2,774 independent families. Results for cognitive function are based on a sample of 798 individuals from 399 independent families in STR.

		Prediction in QIMR				Prediction in STR				Prediction in QIMR + STR			
Phenotype (PGS)		$p_{SNPs} < 5\times10^{-8}$	$p_{SNPs} < 5\times10^{-5}$	$p_{SNPs} < 5\times10^{-3}$	All SNPs	$p_{SNPs} < 5\times10^{-8}$	$p_{SNPs} < 5\times10^{-5}$	$p_{SNPs} < 5\times10^{-3}$	All SNPs	$p_{SNPs} < 5\times10^{-8}$	$p_{SNPs} < 5\times10^{-5}$	$p_{SNPs} < 5\times10^{-3}$	All SNPs
EduYears (*College*)	R^2 (%)	0.110	0.037	0.210	0.100	0.055	0.000	0.230	0.370	0.017	0.003	0.220	0.310
	P	0.419	0.648	0.279	0.443	0.216	0.878	0.012	0.001	0.455	0.739	0.006	0.001
EduYears (*EduYears*)	R^2 (%)	0.34	0.096	0.81	0.034	0.01	0.01	0.04	0.25	0.002	0.001	0.110	0.190
	P	0.165	0.459	0.031	0.660	0.669	0.563	0.290	0.009	0.791	0.846	0.065	0.011
Cognitive function (*College*)	R^2 (%)					0.41	0.41	0.13	0.11				
	P					0.203	0.201	0.474	0.035				
Cognitive function (*EduYears*)	R^2 (%)					0.16	0.29	0.02	0.76				
	P					0.432	0.282	0.780	0.082				

Table S26. Theoretically-approximated prediction accuracy of a linear polygenic score for educational attainment, depending on sample size *N* to estimate the effects of individual SNPs using GWAS. $R^2_{y,\hat{g}}$ is the expected prediction accuracy and $r^2_{g,\hat{g}}$ is the correlation between the polygenic score estimated in a discovery sample ($\hat{g}$) and its true value (*g*).

N	$r^2_{g,\hat{g}}$	$R^2_{y,\hat{g}}$
100,000	0.22	0.04
500,000	0.59	0.12
1,000,000	0.74	0.15

Table S27. The reduction in required sample size from including PGS as a control variable.

	$R^2_X = 0.10$			$R^2_X = 0.20$		
	$R^2_{X\cup\text{PGS}} = 0.12$	$R^2_{X\cup\text{PGS}} = 0.22$	$R^2_{X\cup\text{PGS}} = 0.25$	$R^2_{X\cup\text{PGS}} = 0.22$	$R^2_{X\cup\text{PGS}} = 0.32$	$R^2_{X\cup\text{PGS}} = 0.35$
$\frac{N_{X\cup\text{PGS}}}{N_X}$	0.98	0.87	0.83	0.98	0.85	0.81

12. Supplementary Notes

Author contributions

Daniel Benjamin, David Cesarini, and Philipp Koellinger conceived and designed the study and organized the consortium. Cornelius Rietveld, Sarah Medland, Jaime Derringer, and Nico Martin performed the meta-analyses. Cornelius Rietveld conducted the gene-based tests. Peter Visscher contributed to the design of the study, statistical methods and interpretation of the post-GWAS analyses. Jian Yang and Peter Visscher conducted GREML and prediction analyses. Sarah Medland, Tõnu Esko, Harm-Jan Westra and Lude Franke conducted the expression analyses. Tõnu Esko and Lude Franke performed gene-function prediction analyses. Konstantin Shakhbazov performed cell-type-specificity analyses. Jaime Derringer performed the pathway analyses and summarized all biological follow-up results. Adriaan Hofman organized the work on phenotype harmonization. Philipp Koellinger, Daniel Benjamin, David Cesarini, and Sarah Medland wrote the first draft of the manuscript. Cornelius Rietveld prepared most of the tables and figures in the supplementary materials, with the help of Matthijs van der Loos and Jian Yang. Niels Rietveld, Daniel Benjamin, David Cesarini, Jaime Derringer, Philipp Koellinger and Peter Visscher all wrote substantial portions of the supplementary materials. Chris Chabris, Jan-Emmanuel De Neve, Jaime Derringer, Magnus Johannesson, David Laibson, Nick Martin, Michelle Meyer, Nicholas Timpson, Roy Thurik, André Uitterlinden, Cornelia van Duijn, and Peter Visscher critically reviewed and edited the manuscript.

The advisory board members of the SSGAC (Dalton Conley, George Davey Smith, Albert Hofman, Robert Krueger, David Laibson, Sarah Medland, Michelle Meyer, and Peter Visscher) helped to facilitate the establishment of the consortium and provided crucial advice and ideas throughout the project. Jonathan Beauchamp contributed to the early conceptualization of the study.

Cohort-specific contributions

Cohort	Author	Individual study design and management	Data collection	Genotyping	Genotype preparation	Phenotype preparation	Study data analysis
AGES	Albert V. Smith				X	X	X
AGES	Vilmundur Gudnason	X	X				
AGES	Gudny Eiriksdottir	X	X				
AGES	Tamara B. Harris	X					
AGES	Lenore J. Launer	X					
ALSPAC	Nicholas J. Timpson				X	X	X
ALSPAC	George Davey Smith	X					
ALSPAC	George McMahon				X		X
ALSPAC	Beate St Pourcain				X	X	
ALSPAC	Susan M. Ring		X	X			
ALSPAC	David M. Evans				X	X	
ALSPAC	Debbie A. Lawlor	X					
ASPS	Reinhold Schmidt	X	X				X
ASPS	Katja E. Petrovic		X				X
ASPS	Helena Schmidt			X	X	X	X
ASPS	Marisa Loitfelder						X
BLSA	Dena G. Hernandez			X			
BLSA/InCHIANTI	Toshiko Tanaka						X
BLSA/InCHIANTI	Luigi Ferrucci	X					
BLSA/InCHIANTI//SardiNIA	Antonio Terracciano	X (SardiNIA)				X	X
Cahres/Caps	Sara Hägg						X
Cahres/Caps	Erik Ingelsson						X
Cahres	Per Hall	X	X				
Caps	Henrik Grönberg	X	X				
Cahres	Jingmei Li				X	X	
CCF	Mina K. Chung	X	X	X	X	X	
CCF	John Barnard	X		X	X	X	X
CCF	David R. Van Wagoner	X		X	X		
CoLaus	Zoltán Kutalik						X
CoLaus	Pedro Marquea Vidal					X	X
CoLaus	François Bastardot		X			X	
CoLaus	Martin Preisig	X				X	
CoLaus	Peter Vollenweider	X				X	
CoLaus	Gérard Waeber	X				X	
CROATIA-Korcula	Igor Rudan	X	X				
CROATIA-Korcula	Harry Campbell	X					

CROATIA-Korcula	Veronique Vitart		X	X	X	X	X
CROATIA-Split	Ivana Kolcic		X			X	X
CROATIA-Vis	Caroline Hayward	X	X	X	X	X	X
CROATIA-Vis	Ozren Polasek	X	X			X	X
CROATIA-Vis	Alan F. Wright	X					
DHS	Klaus Berger	X	X				
DHS	Jürgen Wellmann				X	X	X
DHS	Peter Lichtner			X			
ERF	Carla A. Ibrahim-Verbaas					X	X
ERF	Najaf Amin				X		X
ERF	Ben A. Oostra	X	X	X			
ERF	Cornelia M. van Duijn	X	X	X			
EGCUT	Lili Milani			X			X
EGCUT	Tõnu Esko			X	X		X
EGCUT	Anu Realo					X	
EGCUT	Eva Reinmaa						X
EGCUT	Jüri Allik					X	
EGCUT	Krista Fischer					X	X
EGCUT	Andres Metspalu	X	X				
FINRISK	Marja-Liisa Nuotio						X
FINRISK	Kati Kristiansson						X
FINRISK	Erkki Vartiainen	X	X				
FINRISK	Markus Perola	X	X				
FTC	Jaakko Kaprio	X	X			X	
FTC	Samuli Ripatti	X		X	X		
FTC	Antti Latvala					X	
FTC	Antti-Pekka Sarin				X		X
GAIN/nonGAIN/NIA	Thais S. Rizzi				X		X
GAIN/nonGAIN/NIA	Danielle Posthuma					X	X
GENOA	Lawrence F. Bielak					X	X
GENOA	Patricia A. Peyser	X					
GENOA	Wei Zhao				X		X
GENOA	Mariza de Andrade			X			
GENOA	Sharon L.R. Kardia	X	X				
GENOA	Jennifer A. Smith					X	X
H2000	Niina Eklund					X	X
H2000	Ida Surakka				X		
H2000	Tomi E. Mäkinen	X				X	
H2000	Veikko Salomaa	X	X				
HABC	Melissa E. Garcia	X				X	
HABC	Kurt Lohman				X		X

HABC	Yongmei Liu			X	X		X
HABC	Tamara B. Harris	X	X			X	
HABC	Daniel S. Evans						x
HBCS	Jari Lahti		X		X	X	X
HBCS	Elisabeth Widen			X	X		
HBCS	Aarno. Palotie			X	X		
HBCS	Johan G. Eriksson	X	X				
HBCS	Katri Räikkönen	X	X			X	
HCS	Christopher J. Oldmeadow						X
HCS	Elizabeth G. Holliday				X		
HCS	Rodney J. Scott		X	X	X		
HCS	John R. Attia	X				X	
HRS	Jennifer A. Smith				X		X
HRS	Jessica D. Faul	X	X			X	
HRS	Sharon L.R. Kardia	X			X		
HRS	David R. Weir	X	X				
INCHIANTI	Stefania Bandinelli	X	X				
KORA	Eva Albrecht						X
KORA	Christian Gieger						X
KORA	Rolf Holle	X				X	
KORA	Christina Holzapfel						X
KORA	Thomas Illig			X	X		
KORA	Andreas Mielck					X	
KORA	H.- Erich Wichmann	X	X				
LifeLines	Martin F. Elderson,		X			X	
LifeLines	Judith M. Vonk		X			X	
LifeLines	Harold Snieder	X		X	X		
LifeLines	Behrooz Z. Alizadeh,			X	X	X	X
LifeLines	Ute Bültmann,					X	
LBC1921/1936	Gail Davies			X	X	X	X
LBC1921/1936	David C. Liewald			X	X		
LBC1921/1936	John M. Starr	X	X				
LBC1921/1936	Ian J. Deary	X	X			X	
MCTFR	Jaime Derringer					X	X
MCTFR	Robert F. Krueger					X	
MCTFR	Jeffrey A. Boatman						X
MCTFR	Robert M. Kirkpatrick					X	
MCTFR	Michael B. Miller				X		
MCTFR	Jae Hoon Sul				X		
MCTFR	Matt McGue	X	X				
MCTFR	William G. Iacono	X	X				

MCTFR	Aldo Rustichini					X	
MoBa	Bo Jacobsson			X	X	X	X
MoBa	Ronny Myhre			X	X	X	X
MoBa	Håkon Gjessing			X		X	X
MoBa	Astanand Jugessur					X	X
MoBa	Jennifer R. Harris					X	X
NESDA	Wouter J. Peyrot				X	X	X
NESDA	Brenda Penninx	X	X	X			
NFBC1966	Marika Kaakinen						X
NFBC1966	Marjo-Riitta Järvelin	X	X				
NFBC1966	Rauli Svento	X					
NIA	Christiaan de Leeuw						X
NTR	Abdel Abdellaoui						X
NTR	Jouke-Jan Hottenga				X		
NTR	Gonneke Willemsen					X	
NTR	Dorret I. Boomsma		X				
ORCADES	Peter K. Joshi				X	X	X
ORCADES	Nicholas D. Hastie	X					
ORCADES	James F. Wilson	X	X	X			
QIMR	Nicolas W. Martin	X				X	X
QIMR	Sang H. Lee						X
QIMR	Dale R. Nyholt			X	X		
QIMR	Pamela A. Madden P	X	X	X			
QIMR	Andrew C. Heath	X	X	X			
QIMR	Grant W. Montgomery		X	X			
QIMR	Nicholas G. Martin	X	X	X			
QIMR	Sarah E. Medland				X		X
RS	Patrick J.F. Groenen	X					
RS	Albert Hofman	X					
RS	Philipp D. Koellinger					X	
RS	Cornelius A. Rietveld					X	X
RS	Fernando Rivadeneira		X	X	X		
RS	A. Roy Thurik	X					
RS	André G. Uitterlinden		X	X	X		
RS	Henning W. Tiemeier	X					
RS	Frank J.A. van Rooij		X			X	
RS	Matthijs J.H.M. van der Loos						X
RUSH (MAP/ROS)	David A. Bennett	X	X				
RUSH (MAP/ROS)	Patricia A. Boyle		X			X	X
RUSH (MAP/ROS)	Phil L. De jager			X	X		
RUSH (MAP/ROS)	Lei Yu					X	X

SAGE	Laura J. Bierut	X	X	X	X	X	
SAGE	Arpana Agrawal				X	X	X
SAGE	Peng Lin				X		X
SAGE	John P. Rice	X	X	X	X		X
SardiNIA	David Schlessinger	X	X				
SardiNIA	Osorio Meirelles						X
SardiNIA	Marco Masala		X				
SardiNIA	Francesco Cucca	X	X				
SHIP	Sebastian E. Baumeister	X	X			X	X
SHIP	Alexander Teumer			X	X	X	
SHIP	Henry Völzke	X	X			X	X
SHIP	Wolfgang Hoffmann	X	X				
STR	Patrik K.E. Magnusson	X	X		X	X	
STR	Paul Lichtenstein	X	X				
STR	Magnus Johannesson	X	X				
STR	Matthijs J.H.M. van der Loos						X
STR	David Cesarini					X	
THISEAS	Stavroula Kanoni			X			X
THISEAS	Maria Dimitriou		X			X	
THISEAS	Panos Deloukas			X			
THISEAS	George V. Dedoussis	X					
TwinsUK	Lydia Quaye				X	X	X
TwinsUK	Lynn Cherkas	X	X				
TwinsUK	Juliette M. Harris	X	X				
YFS	Terho Lehtimäki	X		X			
YFS	Olli T. Raitakari	X	X				
YFS	Jorma Viikari	X					
YFS	Mika Kähönen	X					
WASHS	Kelly S. Benke				X	X	X
WASHS	Matthew Kowgier				X		
WASHS	Lyle J. Palmer	X	X	X			
WASHS	Sutapa Mukherjee	X	X			X	

13. Additional acknowledgements

New York University in Abu Dhabi provided generous funding that facilitated the completion of this paper.

17 participating cohorts are also members of the earlier GENTREPRENEUR consortium, which developed from a joint effort of the Erasmus School of Economics and the Erasmus Medical Centre of the Erasmus University Rotterdam starting in 2007 to link entrepreneurship phenotypes to genotypes using GWAS. While being the first large scale attempt to link a socio-economic outcome to genetic information using GWAS, the GENTREPRENEUR consortium did not identify any genome-wide significant loci that replicate *(175)*. Its setup and experience contributed much to the SSGAC and its working methods.

The results from a GWAS of educational attainment in the QIMR data have been separately reported *(124)*.

Group banner **The Lifelines Cohort Study**: Behrooz Z. Alizadeh (1), Rudolf A. de Boer (2), H. Marike Boezen (1), Marcel Bruinenberg (3), Lude Franke (4), Pim van der Harst (2), Hans L. Hillege (1,2), Melanie M. van der Klauw (5), Gerjan Navis (6), Johan Ormel (7), Dirkje S. Postma (8), Judith G.M. Rosmalen (7), Joris P. Slaets (9), Harold Snieder (1), Ronald P. Stolk (1), Bruce H.R. Wolffenbuttel (5), Cisca Wijmenga (4).

(1) Department of Epidemiology, University of Groningen, University Medical Center Groningen, The Netherlands,

(2) Department of Cardiology, University of Groningen, University Medical Center Groningen, The Netherlands,

(3) LifeLines Cohort Study, University of Groningen, University Medical Center Groningen, The Netherlands,

(4) Department of Genetics, University of Groningen, University Medical Center Groningen, The Netherlands,

(5) Department of Endocrinology, University of Groningen, University Medical Center Groningen, The Netherlands,

(6) Department of Internal Medicine, Division of Nephrology, University of Groningen, University Medical Center Groningen, The Netherlands,

(7) Interdisciplinary Center of Psychopathology of Emotion Regulation (ICPE), Department of Psychiatry, University of Groningen, University Medical Center Groningen, The Netherlands,

(8) Department of Pulmonology, University of Groningen, University Medical Center Groningen, The Netherlands,

(9) University Center for Geriatric Medicine, University of Groningen, University Medical Center Groningen, The Netherlands.

National Institute on Aging Intramural Research Program: The following authors are affiliated with the National Institute on Aging (NIA), National Institutes of Health (NIH), and are in part supported by the Intramural Research Program of the National Institute on Aging, NIH, Baltimore, MD: Luigi Ferrucci, Melissa E. Garcia, Tamara B. Harris, Dena G. Hernandez, Lenore J. Launer, David Schlessinger, Toshiko Tanaka and Antonio Terracciano.

Description of the ALSPAC sample

ALSPAC recruited 14,541 pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992. 14,541 is the initial number of pregnancies for which the mother enrolled in the ALSPAC study and had either returned at least one questionnaire or attended a "Children in Focus" clinic by 19/07/99. Of these initial pregnancies, there was a total of 14,676 fetuses, resulting in 14,062 live births and 13,988 children who were alive at 1 year of age.

When the oldest children were approximately 7 years of age, an attempt was made to bolster the initial sample with eligible cases who had failed to join the study originally. As a result, when considering variables collected from the age of seven onwards (and potentially abstracted from obstetric notes) there are data available for more than the 14,541 pregnancies mentioned above.

The number of new pregnancies not in the initial sample (known as Phase I enrolment) that are currently represented on the built files and reflecting enrolment status at the age of 18 is 706 (452 and 254 recruited during Phases II and III respectively), resulting in an additional 713 children being enrolled. The phases of enrolment are described in more detail in the cohort profile paper: http://ije.oxfordjournals.org/content/early/2012/04/14/ije.dys064.full.pdf+html.

The total sample size for analyses using any data collected after the age of seven is therefore 15,247 pregnancies, resulting in 15,458 fetuses. Of this total sample of 15,458 fetuses, 14,775 were live births and 14,701 were alive at 1 year of age.

A 10% sample of the ALSPAC cohort, known as the Children in Focus (CiF) group, attended clinics at the University of Bristol at various time intervals between 4 to 61 months of age. The CiF group were chosen at random from the last 6 months of ALSPAC births (1432 families attended at least one clinic). Excluded were those mothers who had moved out of the area or were lost to follow-up, and those partaking in another study of infant development in Avon.

Please note that the study website contains details of all the data that is available through a fully searchable data dictionary: http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/.

AGES (Age, Gene/Environment Susceptibility–Reykjavik Study) - The Age, Gene/Environment Susceptibility-Reykjavik Study is funded by NIH contract N01-AG-12100, the NIA Intramural Research Program, Hjartavernd (the Icelandic Heart Association), and the Althingi (the Icelandic Parliament). Genotyping was conducted at the NIA IRP Laboratory of Neurogenetics. Researchers interested in using the AGES data must obtain approval from the AGES study group. Researchers using the data are required to follow the terms of a research agreement between them and the AGES investigators. In accordance with Icelandic law, individual level data cannot be released to external investigators, only summary GWAS results. Investigators interested in collaboration can work on individual data at the Icelandic Heart Association site. For further information contact Prof. V. Gudnason (v.gudnason@hjarta.is).

ALSPAC (Avon Longitudinal Study of Parents and Children) - We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and the Wellcome Trust (Grant ref:

092731) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and Lenore J. Launer, Nicholas J. Timpson, George Davey Smith, George McMahon, Beate St Pourcain, Susan M. Ring, David M. Evans and Debbie A. Lawlor will serve as guarantors for the contents of this paper. George Davey Smith, David Evans, Debbie Lawlor and Nicholas Timpson are supported by the MRC as part of the Integrated Epidemiology Unit. Details of access procedures are described in our access policy (http://www.bristol.ac.uk/alspac/researchers/data-access/policy/). Genome Wide data is held by ALSPAC and, due to its potential for disclosure of identity, current ethical constraints require these data to be analysed only in Bristol. We are working towards secure remote access that will enable direct access to these data in the future. Individual SNP data can be released under the terms of a Data Transfer Agreement. For further information or to apply for access please contact the ALSPAC Executive (alspac-exec@bris.ac.uk).

ARIC (Atherosclerosis Risk in Communities Study) - The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research. ARIC genotype data have been deposited in the NIH GWAS repository (dbGaP). Researchers wishing to use the ARIC genetic data must first apply to dbGaP for access. The process to request access to any dbGaP study is done via the dbGaP authorized access system.

ASPS (Austrian Stroke Prevention Study) – The authors thank the staff and the participants of the ASPS for their valuable contributions. We thank Birgit Reinhart for her long-term administrative commitment and Ing Johann Semmler for technical assistance with the creation of the DNA bank. Researchers must obtain approval from the Steering Committee of the Austrian Stroke Prevention Study and from the Institutional Ethics Committee of the Medical University Graz, Austria. Researchers using the data are required to follow the terms of an Assistance Agreement containing a number of clauses designed to ensure protection of privacy and compliance with relevant laws. For further information, contact Reinhold Schmidt (reinhold.schmidt@medunigraz.at).

BLSA (Baltimore Longitudinal Study of Aging) - This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Institute on Aging. A portion of that support was through a R&D contract with MedStar Research Institute. Researchers interested in using BLSA data should know that individual level data cannot be released to external investigators, only summary GWAS results; and that they are required to follow the terms of a research agreement between them and BLSA investigators, submitting an IRB-approved protocol and specific plan to the Steering Committee for consideration (as specified at the website http://www.blsa.nih.gov).

CAHRES (Cancer Hormone Replacement Epidemiology in Sweden) - The CAHRES study was supported by funding from the Agency for Science, Technology and Research of Singapore (A*STAR), the United States National Institute of Health (NIH) and the Susan G. Komen Breast Cancer Foundation. To use the Swedish

CAHRES data, researchers must obtain approval from the Swedish Ethical Review Board and from the Steering Committee of CAHRES. For further information, contact Per Hall (per.hall@ki.se)

CAPS (Cancer Prostate Sweden) - The CAPS study was supported by grants from the Swedish Research Council, the Swedish Cancer Society, and the National Cancer Institute. Researchers interested in using CAPS data must obtain approval from the Swedish Ethical Review Board and from the Steering Committee of the CAPS. Researchers using the data are required to follow the terms of an Assistance Agreement containing a number of clauses designed to ensure protection of privacy and compliance with relevant laws. For further information, contact Henrik Grönberg (Henrik.Gronberg@ki.se).

CCF (Cleveland Clinic Foundation) - R01 HL090620 and R01 HL111314 from the National Heart, Lung, and Blood Institute (Chung, Barnard, Van Wagoner); NIH/NCRR, CTSA 1UL-RR024989 (Chung, Van Wagoner); Heart and Vascular Institute, Department of Cardiovascular Medicine, Cleveland Clinic (Chung); Leducq Foundation 07-CVD 03 (Van Wagoner, Chung); Atrial Fibrillation Innovation Center, State of Ohio (Van Wagoner, Chung). Researchers interested in using the Cleveland Clinic data must obtain approval from the Cleveland Clinic study group. Researchers using the data are required to follow the terms of a research agreement between them and the Cleveland Clinic investigators. Note that individual level data cannot be released to external investigators, only summary GWAS results. For further information contact Mina Chung (chungm@ccf.org).

CoLaus (Etude Cohorte Lausannoise) - The CoLaus study was supported by research grants from the Swiss National Science Foundation (grant no: 33CSCO-122661) from GlaxoSmithKline and the Faculty of Biology and Medicine of Lausanne, Switzerland. The authors also express their gratitude to the participants in the Lausanne CoLaus study and to the investigators who have contributed to the recruitment, in particular research nurses Yolande Barreau, Anne-Lise Bastian, Binasa Ramic, Martine Moranville, Martine Baumer, Marcy Sagette, Jeanne Ecoffey and Sylvie Mermoud for data collection. Researchers must obtain approval from the Steering Committee of the CoLaus Study and from the Institutional Ethics Committee of the University in Lausanne, Switzerland. Researchers using the data are required to follow the terms of an Assistance Agreement containing a number of clauses designed to ensure protection of privacy and compliance with relevant laws. For further information go to www.colaus.ch or contact Peter Vollenweider (peter.vollenweider@chuv.ch).

Cr_Kor (Croatia Korcula) - The CROATIA-Korcula study was funded by grants from the Medical Research Council (UK), European Commission Framework 6 project EUROSPAN (Contract No. LSHG-CT-2006-018947) and Republic of Croatia Ministry of Science, Education and Sports research grants to I.R. (108-1080315-0302). We would like to acknowledge the invaluable contributions of the recruitment team in Korcula, the administrative teams in Croatia and Edinburgh and the people of Korcula. The SNP genotyping for the CROATIA-Korcula cohort was performed in Helmholtz Zentrum Munchen, Neuherberg, Germany. Researchers interested in using Croatia Korcula data must obtain approval from the QTL Executive Committee at the University of Edinburgh. Researchers requiring access to the data will also be required to complete a data-transfer agreement and agree to conform to the requirements for confidentiality and to strict guidelines for the protection of the data. For further information contact Caroline Hayward (Caroline.Hayward@igmm.ed.ac.uk)

Cr_Spl (Croatia Split) - The CROATIA-Split study iwas funded by grants from the Medical Research Council (UK), European Commission Framework 6 project EUROSPAN (Contract No. LSHG-CT-2006-018947) and

Republic of Croatia Ministry of Science, Education and Sports research grants to I.R. (108-1080315-0302). We would like to acknowledge the staff of several institutions in Croatia that supported the field work, including but not limited to The University of Split and Zagreb Medical Schools and the Croatian Institute for Public Health. The SNP genotyping for the CROATIA-Split cohort was performed by AROS Applied Biotechnology, Aarhus, Denmark. Researchers interested in using Croatia Split data must obtain approval from the QTL Executive Committee at the University of Edinburgh. Researchers requiring access to the data will also be required to complete a data-transfer agreement and agree to conform to the requirements for confidentiality and to strict guidelines for the protection of the data. For further information contact Carline Hayward (Caroline.Hayward@igmm.ed.ac.uk)

Cr_Vis (Croatia Vis) – The CROATIA-Vis study was funded by grants from the Medical Research Council (UK) and Republic of Croatia Ministry of Science, Education and Sports research grants to I.R. (108-1080315-0302). We would like to acknowledge the staff of several institutions in Croatia that supported the field work, including but not limited to The University of Split and Zagreb Medical Schools, the Institute for Anthropological Research in Zagreb and Croatian Institute for Public Health. The SNP genotyping for the CROATIA-Vis cohort was performed in the core genotyping laboratory of the Wellcome Trust Clinical Research Facility at the Western General Hospital, Edinburgh, Scotland. Researchers interested in using Croatia Vis data must obtain approval from the QTL Executive Committee at the University of Edinburgh. Researchers requiring access to the data will also be required to complete a data-transfer agreement and agree to conform to the requirements for confidentiality and to strict guidelines for the protection of the data. For further information contact Carline Hayward (Caroline.Hayward@igmm.ed.ac.uk)

DHS (Dortmund Health Study) - The collection of sociodemographic and clinical data in the Dortmund Health Study was supported by the German Migraine & Headache Society (DMKG) and by unrestricted grants of equal share from Almirall, Astra Zeneca, Berlin Chemie, Boehringer, Boots Health Care, Glaxo-Smith-Kline, Janssen Cilag, McNeil Pharma, MSD Sharp & Dohme and Pfizer to the University of Muenster. Blood collection in the Dortmund Health Study was done through funds from the Institute of Epidemiology and Social Medicine University of Muenster. Genotyping for the Human Omni Chip was supported by the German Ministry of Education and Research (BMBF, grant no. 01ER0816). Researchers interested in using DHS data are required to sign and follow the terms of an Cooperation Agreement that includes a number of clauses designed to ensure protection of privacy and compliance with relevant laws. For further information, contact Klaus Berger (bergerk@uni-muenster.de)

ERF (Erasmus Rucphen Family study) -This study is financially supported by the Netherlands Organization for Scientific Research (NWO), the Internationale Stichting Alzheimer Onderzoek (ISAO), the Hersenstichting Nederland (HSN), and the Centre for Medical Systems Biology (CMSB) in the framework of the Netherlands Genomics Initiative (NGI). We thank the participants from the Genetic Research in Isolated Populations, Erasmus Rucphen Family, who made this work possible. Researchers who wish to use data of the Erasmus Rucphen Family Study must seek approval from the management team of the Erasmus Rucphen Family study. They are advised to contact the study PI, professor Cornelia van Duijn (c.vanduijn@erasmusmc.nl).

EGCUT (Estonian Genome Center, University of Tartu) - EGCUT received financing from FP7 programs (ENGAGE, OPENGENE), targeted financing from Estonian Government SF0180142s08, Estonian Research

Roadmap through Estonian Ministry of Education and Research, Center of Excellence in Genomics (EXCEGEN) and University of Tartu (SP1GVARENG). Also . We acknowledge EGCUT technical personnel, especially Mr V. Soo and S. Smit. Data analyzes were carried out in part in the High Performance Computing Center of University of Tartu. Researchers interested in using the Estonian Biobank (EGCUT) data must obtain approval from the Estonian Genome Center of University of Tartu (EGCUT)) study group. Note that anonymized individual level data can only be released after study has been approved by Research Ethics Committee of University of Tartu and must be carried out in collaboration with EGCUT investigators. Researchers using the data are required to follow the terms of a research agreement between them and the EGCUT investigators. For further information visit www.biobank.ee or contact directly the Director of Estonian Biobank, Prof. Andres Metspalu (Andres.Metspalu@ut.ee).

FINRISK (FINRISK) - We would like to thank all the Finrisk study participants. The Finrisk surveys were mainly funded by the National Institute for Health and Welfare (THL), Finland. Additional support was obtained through funds from the European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE Consortium, grant agreement HEALTH-F4-2007-201413, and BioSHaRE Consortium, grant agreement 261433. K.K. was supported by grant number 250207 from the Academy of Finland and a grant from the Orion-Farmos Research Foundation. M.P. is partly financially supported for this work by the Finnish Academy SALVE program ‘‘Pubgensense’’ 129322 and by grants from Finnish Foundation for Cardiovascular Research. Researchers wishing to use FINRISK data must send a written proposal to the FINRISK Steering Committee. If the proposal is approved, a specific Data Transfer and Collaboration Agreement must be signed before sending the data. For further information, contact Veikko Salomaa (Veikko.Salomaa@thl.fi).

FTC (Finnish Twin Cohort) -The FTC was supported by Academy of Finland Center of Excellence in Complex Disease Genetics (grant numbers: 213506, 129680), US P.H.S. NIDA 12854, Global Research Awards for Nicotine Dependence (GRAND), and ENGAGE – European Network for Genetic and Genomic Epidemiology,FP7-HEALTH-F4-2007, grant agreement number 201413. Researchers interested in using FTC data must obtain approval from an Ethical Review Board if not covered by existing ethical approvals, and from the principal investigators of the Finnish Twin Study. To ensure protection of privacy and compliance with national data protection legislation, a data use/transfer agreement is needed, the content and specific clauses of which will depend on the nature of the requested data. It is also possible that requested analyses are run in-house. For further information please contact Jaakko Kaprio (jaakko.kaprio@helsinki.fi).

GAIN (Genetic Association Information Network Schizophrenia-Controls) / nonGAIN (Non-Genetic Association Information Network Schizophrenia-Controls) / NIA (National Institute of Aging) - Funding support for the companion studies, Genome-Wide Association Study of Schizophrenia (GAIN) and Molecular Genetics of Schizophrenia - nonGAIN Sample (MGS_nonGAIN), was provided by Genomics Research Branch at NIMH see below) and the genotyping and analysis of samples was provided through the Genetic Association Information Network (GAIN) and under the MGS U01s: MH79469 and MH79470. Assistance with data cleaning was provided by the National Center for Biotechnology Information. The MGS dataset(s) used for the analyses described in this manuscript were obtained from the database of Genotype and Phenotype (dbGaP) found at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers phs000021.v2.p1 (GAIN) and phs000167.v1.p1 (nonGAIN). Samples and associated phenotype data for the MGS GWAS study were collected

under the following grants: NIMH Schizophrenia Genetics Initiative U01s: MH46276 (CR Cloninger), MH46289 (C Kaufmann), and MH46318 (MTTsuang); and MGS Part 1 (MGS1) and Part 2 (MGS2) R01s: MH67257 (NG Buccola), MH59588 (BJ Mowry), MH59571 (PV Gejman), MH59565 (Robert Freedman), MH59587 (F Amin), MH60870 (WF Byerley), MH59566 (DW Black), MH59586 (JM Silverman), MH61675 (DF Levinson), and MH60879 (CR Cloninger). Further details of collection sites, individuals, and institutions may be found in data supplement Table 1 of Sanders et al.(2008; PMID: 18198266) and at the study dbGaP pages. Funding support for the "Genetic Consortium for Late Onset Alzheimer's Disease" was provided through the Division of Neuroscience, NIA. The Genetic Consortium for Late Onset Alzheimers Disease includes a genome-wide association study funded as part of the Division of Neuroscience, NIA. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by Genetic Consortium for Late OnsetAlzheimers Disease. Statistical analyses of GAIN, nonGAIN and NIA data were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org) which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003). We further wish to acknowledge the financial support of NWO-VI-016-065-318, NWO 645-000-003 and the Center for Neurogenomics and Cognitive Research (CNCR) at the VU University. None of the authors have a conflict of interest to declare. Genotype data are deposited in the NIH GWAS repository (dbGaP). Researchers wishing to use the GAIN/nonGAIN/NIA genetic data must apply to dbGaP for access. The process to request access to any dbGaP study is done via the dbGaP authorized access system.

GENOA (Genetic Epidemiology Network of Arteriopathy) – Genetic Epidemiology Network of Arteriopathy (GENOA) is supported by the National Institutes of Health, grant numbers HL087660 and HL100245 from the National Heart, Lung and Blood Institute and grant number P60MD002249 from the National Institute on Minority Health and Health Disparities. We thank Eric Boerwinkle, PhD from the Human Genetics Center and Institute of Molecular Medicine and Division of Epidemiology, University of Texas Health Science Center, Houston, Texas, USA and Julie Cunningham, PhD from the Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN, USA for their help with genotyping. We thank Min A. Jhun, M.S. from the Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan, USA, for her help with conducting genome-wide association analyses for GENOA. Regretfully, she was omitted as a GENOA co-author and should have been included. Researchers interested in using GENOA data should submit a written proposal to the GENOA Steering Committee which includes details on the data being requested and plans for data security. For further information, contact Sharon Kardia (skardia@umich.edu).

H2000 (Health 2000) - We would like to thank all the Health 2000 Survey participants. The Health 2000 Study was funded by the National Institute for Health and Welfare (THL), the Finnish Centre for Pensions (ETK), the Social Insurance Institution of Finland (KELA), the Local Government Pensions Institution (KEVA) and other organizations listed on the website of the survey (http://www.terveys2000.fi). V.S. was supported by grants number 129494 and 139635 from the Academy of Finland and a grant from the Finnish Foundation for Cardiovascular Research. Researchers wishing to use H2000 data must send a written proposal to the H2000 Steering Committee. If the proposal is approved, a specific Data Transfer and Collaboration Agreement must be signed before sending the data. For further information, contact Veikko Salomaa (Veikko.Salomaa@thl.fi).

HABC (Health ABC) - The Health ABC Study was supported by NIA contracts N01AG62101, N01AG62103, and N01AG62106 and, in part, by the NIA Intramural Research Program. The genome-wide association study was funded by NIA grant 1R01AG032098-01A1 to Wake Forest University Health Sciences and genotyping

services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096C. Genotype data have been deposited in the NIH GWAS repository (dbGaP) and are available through the dbGaP application process. For phenotype data that may not be part of dbGaP, researchers may request the information from the Health ABC Executive Committee by submitting an analysis plan per the guidelines outlined on this website, http://www.grc.nia.nih.gov/branches/ledb/healthabc/index.htm. For further information, contact Dr. Tamara B. Harris (harris99@mail.nih.gov).

HBCS (Helsinki Birth Cohort Study) – We thank all study participants as well as everybody involved in the Helsinki Birth Cohort Study. Helsinki Birth Cohort Study has been supported by grants from the Academy of Finland, the Finnish Diabetes Research Society, Folkhälsan Research Foundation, Novo Nordisk Foundation, Finska Läkaresällskapet, Signe and Ane Gyllenberg Foundation, University of Helsinki, Ministry of Education, Ahokas Foundation, Emil Aaltonen Foundation, Juho Vainio Foundation, and Wellcome Trust (grant number WT089062). Researchers interested in using HBCS data must obtain approval from the Steering Committee of the Helsinki Birth Cohort Study. Researchers using the data are required to follow the terms in a number of clauses designed to ensure protection of privacy and compliance with relevant Finnish laws. For further information, contact Johan Eriksson (johan.eriksson@helsinki.fi).

HCS (Hunter Community Study) - EGH is supported by the Australian NHMRC Fellowship scheme. Researchers interested in using the HCS data must obtain approval from Principal Investigators involved with the HCS and from the Institutional Ethics Committee of the University of Newcastle, Australia. For further information, contact John Attia (john.attia@newcastle.edu.au).

HRS (Health and Retirement Study) - HRS is supported by the National Institute on Aging (NIA U01AG009740). The genotyping was funded as a separate award from the National Institute on Aging (RC2 AG036495). Our genotyping was conducted by the NIH Center for Inherited Disease Research (CIDR) at Johns Hopkins University. Genotyping quality control and final preparation of the data were performed by the Genetics Coordinating Center at the University of Washington. HRS genotype data have been deposited in the NIH GWAS repository (dbGaP). Researchers wishing to use the HRS genetic data must first apply to dbGaP for access. The process to request access to any dbGaP study is done via the dbGaP authorized access system. Researchers who wish to obtain HRS phenotype measures that are not in dbGaP must submit a data access use agreement to HRS. For further information, contact hrsquestions@umich.edu. Relevant websites describing HRS genotype and phenotype data are: www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000428.v1.p1 and http://hrsonline.isr.umich.edu.

InCHIANTI (Invecchiare in Chinati) - The InCHIANTI study baseline (1998-2000) was supported as a "targeted project" (ICS110.1/RF97.71) by the Italian Ministry of Health and in part by the U.S. National Institute on Aging (Contracts: 263 MD 9164 and 263 MD 821336). Researchers interested in using InCHIANTI data should know that individual level data cannot be released to external investigators, only summary GWAS results; and that they are required to follow the terms of a research agreement between them and InCHIANTI investigators, submitting an IRB-approved protocol and specific plan to the Steering Committee for consideration (as specified at the website http://inchiantistudy.net).

KORA (Kooperative Gesundheitsforschung in der Region Augsburg) - The KORA Augsburg studies were financed by the Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany and supported by grants from the German Federal Ministry of Education and Research (BMBF). Part of this work was financed by the German National Genome Research Network (NGFN). Our research was supported within the Munich Center of Health Sciences (MC Health) as part of LMUinnovativ. Researchers interested in using the KORA data must obtain approval from the KORA study group. Researchers using the data are required to follow the terms of a research agreement between them and the KORA investigators. Note that individual level data cannot be released to external investigators, only summary GWAS results. For further information contact the KORA speaker, Prof. Annette Peters (peters@helmholtz-muenchen.de). More information can be found at http://www.helmholtz-muenchen.de/en/kora-en/information-for-scientists/participating-in-kora/index.html.

LifeLines (LifeLines) - The LifeLines Cohort Study, and generation and management of GWAS genotype data for the LifeLines Cohort Study is supported by the Netherlands Organization of Scientific Research NWO (grant 175.010.2007.006), the Economic Structure Enhancing Fund (FES) of the Dutch government, the Ministry of Economic Affairs, the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the Northern Netherlands Collaboration of Provinces (SNN), the Province of Groningen, University Medical Center Groningen, the University of Groningen, Dutch Kidney Foundation and Dutch Diabetes Research Foundation. We thank Behrooz Z. Alizadeh, Annemieke Boesjes, Marcel Bruinenberg, Noortje Festen, Pim van der Harst, Ilja Nolte, Lude Franke, Mitra Valimohammadi for their help in creating the GWAS database, and Rob Bieringa, Joost Keers, René Oostergo, Rosalie Visser, Judith Vonk for their work related to data-collection and validation. The authors are grateful to the study participants, the staff from the LifeLines Cohort Study and Medical Biobank Northern Netherlands, and the participating general practitioners and pharmacists. Researchers interested in using the LifeLines data must obtain approval for a specific analysis plan from the scientific board of LifeLines to obtain access to the data. Researchers using the data are required to follow the terms of a signed agreement containing a number of clauses designed to ensure protection of privacy and compliance with relevant laws. For further information, contact Harold Snieder (h.snieder@umcg.nl).

LBC1921 and LBC1936 (Lothian Birth Cohorts 1921 and Lothian Birth Cohorts 1936) - We thank the cohort participants who contributed to the Lothian Birth Cohorts. The genotyping was supported by the UK's Biotechnology and Biological Sciences Research Council (BBSRC). Phenotype collection in the LBC1921 was supported by the BBSRC, The Royal Society and The Chief Scientist Office of the Scottish Government. Phenotype collection in the LBC1936 was supported by Research Into Ageing (continues as part of Age UK's The Disconnected Mind project). The work was undertaken by The University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology, part of the cross council Lifelong Health and Wellbeing Initiative (G0700704/84698). Funding from the Biotechnology and Biological Sciences Research Council (BBSRC), Engineering and Physical Sciences Research Council (EPSRC), Economic and Social Research Council (ESRC) and Medical Research Council (MRC) is gratefully acknowledged. Researchers interested in using LBC data should approach the Directors of the Lothian Birth Cohorts to obtain a Data Request Form. Approval is required from the relevant medical research ethics committee in Scotland. A data transfer agreement is required. For further information, contact Ian Deary (i.deary@ed.ac.uk).

MoBa (Mother and Child Cohort of NIPH) – This work was supported by grants from the Norwegian Research Council (FUGE 183220/S10, FRIMEDKLI-05 ES236011), Swedish Medical Society (SLS 2008-21198), Jane and Dan Olsson Foundations and Swedish government grants to researchers in the public health service (ALFGBG-2863, ALFGBG-11522), and the European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE Consortium, grant agreement HEALTH-F4-2007-201413. The Norwegian Mother and Child Cohort Study was also supported by the Norwegian Ministry of Health and the Ministry of Education and Research, NIH/NIEHS (contract no N01-ES-75558), NIH/NINDS (grant no.1 UO1 NS 047537-01 and grant no.2 UO1 NS 047537-06A1), and the Norwegian Research Council/FUGE (grant no. 151918/S10). We are grateful to all the participating families in Norway who take part in this ongoing cohort study. Researchers interested in using MoBa data must obtain approval from the Scientific Management Committee of MoBa and from the Regional Committee for Medical and Health Research Ethics for access to data and biological material. Researchers are required to follow the terms of an Assistance Agreement containing a number of clauses designed to ensure protection of privacy and compliance with relevant laws. For further information, contact the principal investigator of MoBga, Per Magnus (per.magnus@fhi.no).

MCTFR (Minnesota Center for Twin and Family Research) - The MCTFR is supported by USPHS Grants from the National Institute on Alcohol Abuse and Alcoholism (AA09367 and AA11886), the National Institute on Drug Abuse (DA05147, DA13240, and DA024417), and the National Institute on Mental Health (MH066140). Jaime Derringer was supported by NIH grants DA029377 and MH016880. The genotype data from MCTFR is in the process of being deposited in the Database of Genotypes and Phenotypes (dbGaP) maintained by the US National Center on Bioinformatics (NCBI). Once the process for submitting the MCTFR genotypes is complete and dbGaP is setup to accept additional phenotype data from the MCTFR, the education phenotypes used in the analyses reported here will be deposited there. Requests to use the data could then be made through dbGaP.

NESDA (Netherlands Study of Depression and Anxiety) - We acknowledge financial support from the Geestkracht program of ZonMW (10-000-1002); matching funds from universities and mental health care institutes involved in NESDA; Center for Medical Systems Biology (NWO Genomics), Neuroscience Campus Amsterdam. Genotyping was funded by the Genetic Association Information Network (GAIN) of the Foundation for the US National Institutes of Health. Genotype data were obtained from dbGaP (http://www.ncbi.nlm.nih.gov/dbgap, accession number phs000020.v1.p1). Researchers interested in using the NESDA data must obtain approval from the NESDA study group. Researchers using the data are required to follow the signed terms of a research agreement between them and the NESDA investigators. Note that individual level data cannot be released to external investigators, only summary GWAS results. For further information contact B.W.J.H. Penninx (b.penninx@vumc.nl).

NFBC1966 (Northern Finland Birth Cohorts (1966 Cohort)) – We thank Professor Paula Rantakallio (launch of NFBC1966 and initial data collection), Ms Sarianna Vaara (data collection), Ms Tuula Ylitalo (administration), Mr Markku Koiranen (data management), Ms Outi Tornwall and Ms Minttu Jussila (DNA biobanking). This work was supported by the Academy of Finland [project grants 104781, 120315, 129418, Center of Excellence in Complex Disease Genetics and Public Health Challenges Research Program (SALVE)], University Hospital Oulu, Biocenter, University of Oulu, Finland (75617), the European Commission [EURO-

BLCS, Framework 5 award QLG1-CT-2000-01643], The National Heart, Lung and Blood Institute [5R01HL087679-02] through the SNP Typing for Association with Multiple Phenotypes from Existing Epidemiologic Data (STAMPEED) program [1RL1MH083268-01], The National Institute of Health/The National Institute of Mental Health [5R01MH63706:02], European Network of Genomic and Genetic Epidemiology (ENGAGE) project and grant agreement [HEALTH-F4-2007-201413], and the Medical Research Council, UK [G0500539, G0600705, PrevMetSyn/ Public Health Challenges Research Program (SALVE)]. Researchers interested in using NFBC1966 data must obtain approval from the Ethical Committee of Northern Ostrobothnia Hospital District and from the Data and Publication Committee of the Northern Finland Birth Cohorts. Researchers using the data are required to follow The Declaration of Helsinki and rules of practice containing a number of clauses designed to ensure protection of privacy and compliance with relevant laws. For further information, contact Marjo-Riitta Jarvelin (m.jarvelin@imperial.ac.uk).

NTR (Netherlands Twin Register) – Funding was obtained from the Netherlands Organization for Scientific Research (NWO: MagW/ZonMW grants 904-61-090, 985-10-002,904-61-193,480-04-004, 400-05-717, Addiction-31160008 Middelgroot-911-09-032, Spinozapremie 56-464-14192), Center for Medical Systems Biology (CSMB, NWO Genomics), NBIC/BioAssist/RK(2008.024), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI –NL, 184.021.007), the VU University's Institute for Health and Care Research (EMGO+) and Neuroscience Campus Amsterdam (NCA), the European Science Foundation (ESF, EU/QLRT-2001-01254), the European Community's Seventh Framework Program (FP7/2007-2013), ENGAGE (HEALTH-F4-2007-201413); the European Science Council (ERC Advanced, 230374), Rutgers University Cell and DNA Repository (NIMH U24 MH068457-06) and the National Institutes of Health (NIH, R01D0042157-01A). Part of the genotyping and analyses were funded by the Genetic Association Information Network (GAIN) of the Foundation for the US National Institutes of Health, the (NIMH, MH081802) and by the Grand Opportunity grants 1RC2MH089951-01 from the NIMH. AA was supported by CSMB/NCA. Statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org), which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003), the Dutch Brain Foundation and the department of psychology and education of the VU University Amsterdam. Genotype data have been deposited in the NIH GWAS repository (dbGaP). Researchers wishing to use the NTR genetic data must first apply to dbGaP for access. The process to request access to any dbGaP study is done via the dbGaP authorized access system. For further information, contact Dorret Boomsma (d.i.boomsma@vu.nl).

ORCADES (The Orkney Complex Disease Study) - ORCADES was supported by the Chief Scientist Office of the Scottish Government, the Royal Society, the MRC Human Genetics Unit, Arthritis Research UK and the European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947). DNA extractions were performed at the Wellcome Trust Clinical Research Facility in Edinburgh. We would like to acknowledge the invaluable contributions of Lorraine Anderson and the research nurses in Orkney, the administrative team in Edinburgh and the people of Orkney. Researchers interested in using the individual level data should contact the ORCADES investigators, who will consider the application. Approval from an appropriate ethics committee must be in place and researchers must then follow the guidelines for ORCADES data. For further information, contact Jim Wilson (jim.wilson@ed.ac.uk).

QIMR (Queensland Institute of Medical Research) - Funding was provided by the Australian National Health and Medical Research Council (241944, 339462, 389927, 389875, 389891, 389892, 389938, 442915, 442981, 496739, 552485, 552498), the Australian Research Council (A7960034, A79906588, A79801419, DP0770096, DP0212016, DP0343921), the FP-5 GenomEUtwin Project (QLG2-CT-2002-01254), and the U.S. National Institutes of Health (NIH grants AA07535, AA10248, AA13320, AA13321, AA13326, AA14041, DA12854, MH66206). A portion of the genotyping on which the QIMR study was based (Illumina 370K scans) was carried out at the Center for Inherited Disease Research, Baltimore (CIDR), through an access award to the authors' late colleague Dr. Richard Todd (Psychiatry, Washington University School of Medicine, St Louis). Imputation was carried out on the Genetic Cluster Computer, which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003). N.W.H.M was supported by a PhD scholarship from the ANZ trust. S.E.M., is supported by the National Health and Medical Research Council (NHMRC) Fellowship Scheme. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Researchers interested in using QIMR data can contact Nick Martin (Nick.Martin@qimr.edu.au) and Sarah Medland (medlandse@gmail.com).

RS (The Rotterdam Study) - The GWA study of the Rotterdam Study was funded by the Netherlands Organisation for Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012), the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Consortium for Healthy Aging (NCHA) project nr. 050-060-810. We thank Pascal Arp, Mila Jhamai, Michael Moorhouse, Marijn Verkerk, and Sander Bervoets for their assistance in creating the GWAS database. The Rotterdam Study is funded by the Erasmus Medical Center; Erasmus University, Rotterdam; the Netherlands Organization for the Health Research and Development (ZonMw); the Research Institute for Diseases in the Elderly (RIDE); the Ministry of Education, Culture, and Science; the Ministry for Health, Welfare, and Sports; the European Commission (DG XII); and the Municipality of Rotterdam. The authors are very grateful to the participants and staff from the Rotterdam Study, participating general practioners and the pharmacists. We would like to thank Dr. Tobias A. Knoch, Anis Abuseiris, Karol Estrada, Luc V. de Zeeuw, and Rob de Graaf, as well as their institutions, the Erasmus Grid Office, Erasmus MC Rotterdam, The Netherlands, and especially the national German MediGRID and Services@MediGRID part of the German D-Grid, both funded by the German Bundesministerium fuer Forschung und Technology under grants # 01 AK 803 A-H and # 01 IG 07015 G for access to their grid resources. Some of the statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org) which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003 PI: Posthuma) along with a supplement from the Dutch Brain Foundation and the VU University Amsterdam. Researchers who wish to use data of the Rotterdam Study must obtain approval from the Rotterdam Study Management Team. They are advised to contact the PI of the Rotterdam Study, Dr Albert Hofman (a.hofman@erasmusmc.nl).

RUSH-MAP (Rush University Medical Center - Memory and Aging Project) / RUSH-ROS (Rush University Medical Center - Religious Orders Study) - The MAP and ROS data used in this analysis was supported by National Institute on Aging grants P30AG10161, R01AG17917, R01AG15819, R01AG30146, the Illinois Department of Public Health, and the Translational Genomics Research Institute. Researchers interested in accessing the clinical and genomic data, in addition to other data and biospecimens, must obtain approved from the Rush Alzheimer's Disease Center resource distribution committee following scientific review of a

submitted request. Resource requests can be made via the portal at http://www.rush.edu/radc. Here you find additional information regarding access policies and instructions for submitting requests.

SAGE (Study of Addiction: Genetics and Environment) -Funding support for the Study of Addiction: Genetics and Environment (SAGE) was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01 HG004422). SAGE is one of the genome-wide association studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Support for collection of datasets and samples was provided by the Collaborative Study on the Genetics of Alcoholism (COGA; U10 AA008401), the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392), and the Family Study of Cocaine Dependence (FSCD; R01 DA013423, R01 DA019963). Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438), the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Drug Abuse, and the NIH contract "High throughput genotyping for studying the genetic contributions to human disease" (HHSN268200782096C). LJB also receives support from K02DA021237. The phenotypic variables and genotype data may be obtained via request to and approval of dbGAP (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1). Access to imputed genotypic data used for the present analysis or other variables may be requested from the PI, Laura Bierut (bierutl@psychiatry.wustl.edu) and is contingent on approval by the Steering Committees of contributing studies COGEND and FSCD.

SardiNIA (SardiNIA Study of Aging) - This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Institute on Aging. Researchers interested in using SardiNIA data should know that individual level data cannot be released to external investigators, only summary GWAS results; and that they are required to follow the terms of a research agreement between them and SardiNIA investigators, submitting an IRB-approved protocol and specific plan to the Steering Committee for consideration (as specified at the website http://sardinia.nia.nih.gov).

SHIP (Study of Health in Pomerania) - SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine (GANI_MED)' funded by the Federal Ministry of Education and Research (grant 03IS2061A). Genome-wide data have been supported by the Federal Ministry of Education and Research (grant no. 03ZIK012) and a joint grant from Siemens Healthcare, Erlangen, Germany and the Federal State of Mecklenburg- West Pomerania. The University of Greifswald is a member of the 'Center of Knowledge Interchange' program of the Siemens AG and the Caché Campus program of the InterSystems GmbH. SHIP data are owned by the Research Network of Community Medicine of the University Medicine Greifswald, Germany. Researches who wish to use SHIP data can apply via http://www.community-medicine.de. Data usage applications have to be approved by the Steering Committee of the research network.

STR (Swedish Twin Registry) – The Jan Wallander and Tom Hedelius Foundation, the Ragnar Söderberg Foundation, the Swedish Council for Working Life and Social Research, the Ministry for Higher Education, the Swedish Research Council (M-2005-1112), GenomEUtwin (EU/QLRT-2001-01254; QLG2-CT-2002-01254), NIH DK U01-066134, The Swedish Foundation for Strategic Research (SSF), the Heart and Lung foundation no. 20070481. Researchers interested in using STR data must obtain approval from the Swedish Ethical Review Board and from the Steering Committee of the Swedish Twin Registry. Researchers using the data are required to follow the terms of an Assistance Agreement containing a number of clauses designed to ensure protection of privacy and compliance with relevant laws. For further information, contact Patrik Magnusson (patrik.magnusson@ki.se).

THISEAS (The Hellenic study of Interactions between SNPs & Eating in Atherosclerosis Susceptibility) - Recruitment for THISEAS was partially funded by a research grant (PENED 2003) from the Greek General Secretary of Research and Technology; we thank all the dieticians and clinicians for their contribution to the project. The genotyping was funded by the Wellcome Trust. We like to thank the members of the WTSI GenotypingFacility in particular Sarah Edkins and Cordelia Langford. Researchers interested in using the THISEAS data must obtain approval from the THISEAS study group. Researchers using the data are required to follow the terms of a research agreement between them and the THISEAS investigators. Note that individual level data cannot be released to external investigators, only summary GWAS results. For further information contact George Dedoussis (dedousi@hua.gr).

TwinsUK (St Thomas' UK Adult Twin Registry) - The study was funded by the Wellcome Trust; European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE project grant agreement (HEALTH-F4-2007-201413). The study also receives support from the Dept of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's & St Thomas' NHS Foundation Trust in partnership with King's College London. Genotyping was performed by The Wellcome Trust Sanger Institute, support of the National Eye Institute via an NIH/CIDR genotyping project. Researchers interested in using TwinsUK data must obtain approval for data access from the TwinsUK Resource Executive Committee (TREC). All research is approved by the London - Westminster Research Ethics Committee. The Department of Twin Research facilitates and encourages the access and sharing of data with the scientific community to promote and contribute to further scientific research. For further information, please contact Victoria Vazquez (victoria.vazquez@kcl.ac.uk) or go to our website www.twinsuk.ac.uk/data-access/open-access/.

UQ (The University of Queensland) The Australian Research Council (DP130102666), the Australian National Health and Medical Research Council (APP1048853 and APP1052684) and the National Institutes of Health (GM099568, GM075091, MH100141).

YFS (The Cardiovascular Risk in Young Finns Study) – The Young Finns Study has been financially supported by the Academy of Finland: grants 126925, 121584, 124282, 129378 (Salve), 117787 (Gendi), and 41071 (Skidi), the Social Insurance Institution of Finland, Kuopio, Tampere and Turku University Hospital Medical Funds (grant 9M048 and 9N035 for TeLeht), Juho Vainio Foundation, Paavo Nurmi Foundation, Finnish Foundation of Cardiovascular Research and Finnish Cultural Foundation, Tampere Tuberculosis Foundation and Eemil Aaltonen Foundation (T.L). The expert technical assistance in the statistical analyses by

Irina Lisinen and Ville Aalto are gratefully acknowledged. Researchers interested in using the YFS data must obtain approval from the YFS study group. Researchers using the data are required to follow the terms of a research agreement between them and the YFS investigators. For further information contact Olli Raitakari (olli.raitakari@utu.fi).

WASHS (Western Australia Sleep Health Study) – Sir Charles Gairdner and Hollywood Private Hospital Research Foundations and the Western Australian Sleep Disorders Research Institute, Western Australia. Researchers interested in using the WASHS data must request and obtain approval from the Western Australian Sleep Health Study Management Committee through completion of a Data Access Application. Researchers using the data are required to follow the terms of the Data Access Policy containing a number of clauses designed to ensure protection of privacy and compliance with relevant laws. For further information, contact the Chair of the Western Australian Sleep Health Study Management Committee, Dr. Sutapa Mukherjee (sutapameister@gmail.com).

References and Notes

1. R. Plomin, J. DeFries, V. Knopik, J. Neiderhiser, *Behavioral Genetics* (Worth Publishers, ed. 6, 2013).

2. D. Cesarini, C. T. Dawes, M. Johannesson, P. Lichtenstein, B. Wallace, Genetic variation in preferences for giving and risk taking. *Q. J. Econ.* **124**, 809 (2009). doi:10.1162/qjec.2009.124.2.809

3. D. J. Benjamin *et al*., The genetic architecture of economic and political preferences. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8026 (2012). doi:10.1073/pnas.1120666109 Medline

4. J. P. Beauchamp *et al*., Molecular genetics and economics. *J. Econ. Perspect.* **25**, 57 (2011). doi:10.1257/jep.25.4.57 Medline

5. See the supplementary materials on *Science* Online.

6. D. J. Benjamin *et al*., The promises and pitfalls of genoeconomics. *Annu. Rev. Econ.* **4**, 627 (2012). doi:10.1146/annurev-economics-080511-110939 Medline

7. R. P. Ebstein, S. Israel, S. H. Chew, S. Zhong, A. Knafo, Genetics of human social behavior. *Neuron* **65**, 831 (2010). doi:10.1016/j.neuron.2010.02.020 Medline

8. L. E. Duncan, M. C. Keller, A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *Am. J. Psychiatry* **168**, 1041 (2011). doi:10.1176/appi.ajp.2011.11020191 Medline

9. J. P. Ioannidis, Why most published research findings are false. *PLoS Med.* **2**, e124 (2005). doi:10.1371/journal.pmed.0020124 Medline

10. P. M. Visscher, M. A. Brown, M. I. McCarthy, J. Yang, Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7 (2012). doi:10.1016/j.ajhg.2011.11.029 Medline

11. J. P. Mackenbach *et al*.; European Union Working Group on Socioeconomic Inequalities in Health, Socioeconomic inequalities in health in 22 European countries. *N. Engl. J. Med.* **358**, 2468 (2008). doi:10.1056/NEJMsa0707519 Medline

12. I. J. Deary, S. Strand, P. Smith, C. Fernandes, Intelligence and educational achievement. *Intelligence* **35**, 13 (2007). doi:10.1016/j.intell.2006.02.001

13. J. J. Heckman, Y. Rubinstein, The importance of noncognitive skills: Lessons from the GED testing program. *Am. Econ. Rev.* **91**, 145 (2001). doi:10.1257/aer.91.2.145

14. UNESCO Institute for Statistics, *International Standard Classification of Education* (UNESCO Institute for Statistics, Montreal, 2006).

15. H. Lango Allen *et al*., Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832 (2010). doi:10.1038/nature09410 Medline

16. J. Yang *et al*.; GIANT Consortium, Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807 (2011). doi:10.1038/ejhg.2011.39 Medline

17. B. Carlstedt, *Cognitive Abilities: Aspects of Structure, Process and Measurement* (Acta Universitatis Gothoburgensis, Göteborg, Sweden, 2000).

18. E. K. Speliotes *et al*.; MAGIC; Procardis Consortium, Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937 (2010). doi:10.1038/ng.686 Medline

19. M. H. de Moor *et al*., Meta-analysis of genome-wide association studies for personality. *Mol. Psychiatry* **17**, 337 (2012). doi:10.1038/mp.2010.128 Medline

20. B. Benyamin *et al*.; Wellcome Trust Case Control Consortium 2 (WTCCC2), Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. *Mol. Psychiatry* 10.1038/mp.2012.184 (2013). doi:10.1038/mp.2012.184 Medline

21. C. Jencks, Heredity, environment, and public policy reconsidered. *Am. Sociol. Rev.* **45**, 723 (1980). doi:10.2307/2094892 Medline

22. A. S. Goldberger, Heritability. *Economica* **46**, 327 (1979). doi:10.2307/2553675

23. B. L. Browning, S. R. Browning, A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210 (2009). doi:10.1016/j.ajhg.2009.01.005 Medline

24. B. Servin, M. Stephens, Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genet.* **3**, e114 (2007). doi:10.1371/journal.pgen.0030114 Medline

25. J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly, A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906 (2007). doi:10.1038/ng2088 Medline

26. Y. Li, C. J. Willer, J. Ding, P. Scheet, G. R. Abecasis, MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816 (2010). doi:10.1002/gepi.20533 Medline

27. S. Purcell *et al*., PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559 (2007). doi:10.1086/519795 Medline

28. B. Devlin, K. Roeder, Genomic control for association studies. *Biometrics* **55**, 997 (1999). doi:10.1111/j.0006-341X.1999.00997.x Medline

29. C. J. Willer, Y. Li, G. R. Abecasis, METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190 (2010). doi:10.1093/bioinformatics/btq340 Medline

30. SCAN: SNP and CNV Annotation Database (2012); www.scandb.org/

31. M. L. Freedman *et al*., Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* **36**, 388 (2004). doi:10.1038/ng1333 Medline

32. P. I. W. de Bakker *et al*., Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122 (2008). doi:10.1093/hmg/ddn288 Medline

33. P. Taubman, Earnings, education, genetics, and environment. *J. Hum. Resour.* **11**, 447 (1976). doi:10.2307/145426 Medline

34. A. R. Branigan, K. J. McCallum, J. Freese, "Variation in the heritability of educational attainment: An international meta-analysis," Working paper 13-09, Northwestern University Institute for Policy Research, Evanston, IL, 2013.

35. D. Cesarini, *Essays on Genetic Variation and Economic Behavior* (Massachusetts Institute of Technology, Cambridge, MA, 2010).

36. P. Lichtenstein *et al*., Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78 (2000). doi:10.1056/NEJM200007133430201 Medline

37. E. Turkheimer, Three laws of behavior genetics and what they mean. *Curr. Dir. Psychol. Sci.* **9**, 160 (2000). doi:10.1111/1467-8721.00084

38. J. Yang *et al*., Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565 (2010). doi:10.1038/ng.608 Medline

39. C. E. Ross, C. Wu, The links between education and health. *Am. Sociol. Rev.* **60**, 719 (1995). doi:10.2307/2096319

40. D. M. Cutler, A. Lleras-Muney, in *Making Americans Healthier: Social and Economic Policy as Health Policy*, J. House, R. Schoeni, G. Kaplan, H. Pollack, Eds. (Russell Sage Foundation, New York, 2008).

41. W. Johnson *et al*., Does education confer a culture of healthy behavior? Smoking and drinking patterns in Danish twins. *Am. J. Epidemiol.* **173**, 55 (2011). doi:10.1093/aje/kwq333 Medline

42. W. Johnson *et al*., Education reduces the effects of genetic susceptibilities to poor physical health. *Int. J. Epidemiol.* **39**, 406 (2010). doi:10.1093/ije/dyp314 Medline

43. A. P. Vermeiren *et al*., Do genetic factors contribute to the relation between education and metabolic risk factors in young adults? A twin study. *Eur. J. Public Health* 10.1093/eurpub/cks167 (2012). doi:10.1093/eurpub/cks167 Medline

44. A. Lleras-Muney, The relationship between education and adult mortality in the United States. *Rev. Econ. Stat.* **72**, 189 (2005). doi:10.1111/0034-6527.00329

45. A. C. J. Lager, J. Torssander, Causal effect of education on mortality in a quasi-experiment on 1.2 million Swedes. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8461 (2012). doi:10.1073/pnas.1105839109 Medline

46. J. N. Arendt, Does education cause better health? A panel data analysis using school reforms for identification. *Econ. Educ. Rev.* **24**, 149 (2005). doi:10.1016/j.econedurev.2004.04.008

47. T. Illig *et al*., A genome-wide perspective of genetic variation in human metabolism. *Nat. Genet.* **42**, 137 (2010). doi:10.1038/ng.507 Medline

48. J. Z. Liu *et al*.; AMFS Investigators, A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87**, 139 (2010). doi:10.1016/j.ajhg.2010.06.009 Medline

49. S. H. Lee, J. Yang, M. E. Goddard, P. M. Visscher, N. R. Wray, Estimation of pleiotropy between complex diseases using using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540 (2012). doi:10.1093/bioinformatics/bts474

50. G. Trynka *et al*., Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124 (2013). doi:10.1038/ng.2504 Medline

51. A. Cvejic *et al*., SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat. Genet.* **45**, 542 (2013). doi:10.1038/ng.2603 Medline

52. L. C. Andreae, A. Lumsden, J. D. Gilthorpe, Chick Lrrn2, a novel downstream effector of Hoxb1 and Shh, functions in the selective targeting of rhombomere 4 motor neurons. *Neural Dev.* **4**, 27 (2009). doi:10.1186/1749-8104-4-27 Medline

53. E. L. Heinzen *et al*., Tissue-specific genetic control of splicing: Implications for the study of complex traits. *PLoS Biol.* **6**, e1 (2008). doi:10.1371/journal.pbio.1000001 Medline

54. J. A. Webster *et al.*; NACC-Neuropathology Group, Genetic control of human brain transcript expression in Alzheimer disease. *Am. J. Hum. Genet.* **84**, 445 (2009). doi:10.1016/j.ajhg.2009.03.011 Medline

55. R. S. N. Fehrmann *et al.*, Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* **7**, e1002197 (2011). doi:10.1371/journal.pgen.1002197 Medline

56. M. Nelis *et al.*, Genetic structure of Europeans: A view from the North-East. *PLoS ONE* **4**, e5472 (2009). doi:10.1371/journal.pone.0005472 Medline

57. H. J. Westra *et al.*, MixupMapper: Correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27**, 2104 (2011). doi:10.1093/bioinformatics/btr323 Medline

58. P. H. Lee, C. O'Dushlaine, B. Thomas, S. M. Purcell, INRICH: Interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* **28**, 1797 (2012). doi:10.1093/bioinformatics/bts191 Medline

59. M. Ashburner *et al.*; The Gene Ontology Consortium, Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25 (2000). doi:10.1038/75556 Medline

60. C. M. Koch *et al.*, The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res.* **17**, 691 (2007). doi:10.1101/gr.5704207 Medline

61. A. C. Need *et al.*, A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB. *Hum. Mol. Genet.* **18**, 4650 (2009). doi:10.1093/hmg/ddp413 Medline

62. M. W. Logue *et al.*; Multi-Institutional Research on Alzheimer Genetic Epidemiology (MIRAGE) Study Group, A comprehensive genetic association study of Alzheimer disease in African Americans. *Arch. Neurol.* **68**, 1569 (2011). doi:10.1001/archneurol.2011.646 Medline

63. J. Burroughs-Garcia, V. Sittaramane, A. Chandrasekhar, S. T. Waters, Evolutionarily conserved function of Gbx2 in anterior hindbrain development. *Dev. Dyn.* **240**, 828 (2011). doi:10.1002/dvdy.22589 Medline

64. K. M. Wassarman *et al.*, Specification of the anterior hindbrain and establishment of a normal mid/hindbrain organizer is dependent on Gbx2 gene function. *Development* **124**, 2923 (1997). Medline

65. L. Chen, M. Chatterjee, J. Y. Li, The mouse homeobox gene Gbx2 is required for the development of cholinergic interneurons in the striatum. *J. Neurosci.* **30**, 14824 (2010). doi:10.1523/JNEUROSCI.3742-10.2010 Medline

66. M. Muers, Complex disease: Ups and downs at the MHC. *Nat. Rev. Genet.* **12**, 456 (2011). doi:10.1038/nrg3021 Medline

67. D. Migliorini *et al.*, Mdm4 (Mdmx) regulates p53-induced growth arrest and neuronal cell death during early embryonic mouse development. *Mol. Cell. Biol.* **22**, 5527 (2002). doi:10.1128/MCB.22.15.5527-5538.2002 Medline

68. J. A. Grahn, J. A. Parkinson, A. M. Owen, The cognitive functions of the caudate nucleus. *Prog. Neurobiol.* **86**, 141 (2008). doi:10.1016/j.pneurobio.2008.09.004 Medline

69. W. D. Altrock *et al*., Functional inactivation of a fraction of excitatory synapses in mice deficient for the active zone protein bassoon. *Neuron* **37**, 787 (2003). doi:10.1016/S0896-6273(03)00088-6 Medline

70. P. R. Burton *et al*.; Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661 (2007). doi:10.1038/nature05911 Medline

71. M. Parkes *et al*.; Wellcome Trust Case Control Consortium, Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat. Genet.* **39**, 830 (2007). doi:10.1038/ng2061 Medline

72. J. C. Barrett *et al*.; UK IBD Genetics Consortium; Wellcome Trust Case Control Consortium 2, Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat. Genet.* **41**, 1330 (2009). doi:10.1038/ng.483 Medline

73. A. Franke *et al*., Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118 (2010). doi:10.1038/ng.717 Medline

74. J. C. Barrett *et al*.; NIDDK IBD Genetics Consortium; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium, Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955 (2008). doi:10.1038/ng.175 Medline

75. L. Jostins *et al*.; International IBD Genetics Consortium (IIBDGC), Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119 (2012). doi:10.1038/nature11582 Medline

76. D. P. McGovern *et al*.; NIDDK IBD Genetics Consortium, Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat. Genet.* **42**, 332 (2010). doi:10.1038/ng.549 Medline

77. C. A. Anderson *et al*., Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246 (2011). doi:10.1038/ng.764 Medline

78. M. Imielinski *et al*.; Western Regional Alliance for Pediatric IBD; International IBD Genetics Consortium; NIDDK IBD Genetics Consortium; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium, Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat. Genet.* **41**, 1335 (2009). doi:10.1038/ng.489 Medline

79. E. A. Stahl *et al*.; BIRAC Consortium; YEAR Consortium, Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508 (2010). doi:10.1038/ng.582 Medline

80. A. Ferguson, D. M. Sedgwick, J. Drummond, Morbidity of juvenile onset inflammatory bowel disease: Effects on education and employment in early adult life. *Gut* **35**, 665 (1994). doi:10.1136/gut.35.5.665 Medline

81. L. M. Mackner, D. P. Sisson, W. V. Crandall, Review: Psychosocial issues in pediatric inflammatory bowel disease. *J. Pediatr. Psychol.* **29**, 243 (2004). doi:10.1093/jpepsy/jsh027 Medline

82. K. A. Frazer *et al*.; International HapMap Consortium, A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851 (2007). doi:10.1038/nature06258 Medline

83. J. Yang *et al*.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369, S1 (2012). doi:10.1038/ng.2213 Medline

84. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76 (2011). doi:10.1016/j.ajhg.2010.11.011 Medline

85. H. D. Daetwyler, B. Villanueva, J. A. Woolliams, Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* **3**, e3395 (2008). doi:10.1371/journal.pone.0003395 Medline

86. B. J. Hayes, P. M. Visscher, M. E. Goddard, Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* **91**, 47 (2009). doi:10.1017/S0016672308009981 Medline

87. M. E. Goddard, N. R. Wray, K. Verbyla, P. M. Visscher, Estimating effects and making predictions from genome-wide marker data. *Stat. Sci.* **24**, 517 (2009). doi:10.1214/09-STS306

88. P. M. Visscher, J. Yang, M. E. Goddard, A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin Res. Hum. Genet.* **13**, 517 (2010). doi:10.1375/twin.13.6.517 Medline

89. R. G. Fryer, Financial incentives and student achievement: Evidence from randomized trials. *Q. J. Econ.* **126**, 1755 (2011). doi:10.1093/qje/qjr045

90. J. Heckman, S. H. Moon, R. Pinto, P. Savelyev, A. Yavitz, Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quant. Econ.* **1**, 1 (2010). doi:10.3982/QE8 Medline

91. J. Eckenrode *et al*., Long-term effects of prenatal and infancy nurse home visitation on the life course of youths: 19-year follow-up of a randomized trial. *Arch. Pediatr. Adolesc. Med.* **164**, 9 (2010). doi:10.1001/archpediatrics.2009.240 Medline

92. L. N. Masse, W. S. Barnett, "A benefit-cost analysis of the Abecedarian early childhood intervention," in *Cost-Effectiveness and Educational Policy* (Eye on Education, Larchmont, NY, 2002), pp. 157–173.

93. J. J. Heckman, S. H. Moon, R. Pinto, P. A. Savelyev, A. Yavitz, The rate of return to the HighScope Perry Preschool Program. *J. Public Econ.* **94**, 114 (2010). doi:10.1016/j.jpubeco.2009.11.001 Medline

94. T. B. Harris *et al*., Age, Gene/Environment Susceptibility–Reykjavik Study: Multidisciplinary applied phenomics. *Am. J. Epidemiol.* **165**, 1076 (2007). doi:10.1093/aje/kwk115 Medline

95. A. Fraser *et al*., Cohort profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int. J. Epidemiol.* **42**, 97 (2012). 10.1093/ije/dys066 Medline

96. R. Schmidt *et al*., Assessment of cerebrovascular risk profiles in healthy persons: Definition of research goals and the Austrian Stroke Prevention Study (ASPS). *Neuroepidemiology* **13**, 308 (1994). doi:10.1159/000110396 Medline

97. R. Schmidt, F. Fazekas, P. Kapeller, H. Schmidt, H. P. Hartung, MRI white matter hyperintensities: Three-year follow-up of the Austrian Stroke Prevention Study. *Neurology* **53**, 132 (1999). doi:10.1212/WNL.53.1.132 Medline

98. N. W. Shock *et al*., "Normal human aging: The Baltimore Longitudinal Study of Aging," NIH Publication 84-2450, National Institutes of Health, Bethesda, MD 1984.

99. K. Einarsdóttir *et al*., Linkage disequilibrium mapping of *CHEK2*: Common variation and breast cancer risk. *PLoS Med.* **3**, e168 (2006). doi:10.1371/journal.pmed.0030168 Medline

100. E. T. Chang, M. Hedelin, H. O. Adami, H. Grönberg, K. A. Bälter, Alcohol drinking and risk of localized versus advanced and sporadic versus familial prostate cancer in Sweden. *Cancer Causes Control* **16**, 275 (2005). doi:10.1007/s10552-004-3364-2 Medline

101. M. Hedelin *et al*., Dietary phytoestrogen, serum enterolactone and risk of prostate cancer: The cancer prostate Sweden study (Sweden). *Cancer Causes Control* **17**, 169 (2006). doi:10.1007/s10552-005-0342-2 Medline

102. F. Lindmark *et al*., H6D polymorphism in macrophage-inhibitory cytokine-1 gene associated with prostate cancer. *J. Natl. Cancer Inst.* **96**, 1248 (2004). doi:10.1093/jnci/djh227 Medline

103. M. Firmann *et al*., The CoLaus study: A population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovasc. Disord.* **8**, 6 (2008). doi:10.1186/1471-2261-8-6 Medline

104. I. Rudan *et al*., "10001 Dalmatians:" Croatia launches its national biobank. *Croat. Med. J.* **50**, 4 (2009). doi:10.3325/cmj.2009.50.4 Medline

105. G. B. Ehret *et al*.; International Consortium for Blood Pressure Genome-Wide Association Studies; CARDIoGRAM consortium; CKDGen Consortium; KidneyGen Consortium; EchoGen consortium; CHARGE-HF consortium, Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103 (2011). doi:10.1038/nature10405 Medline

106. K. Sleegers *et al*., Cerebrovascular risk factors do not contribute to genetic variance of cognitive function: The ERF study. *Neurobiol. Aging* **28**, 735 (2007). doi:10.1016/j.neurobiolaging.2006.03.012 Medline

107. F. A. Sayed-Tabatabaei *et al*., Heritability of the function and structure of the arterial wall: Findings of the Erasmus Rucphen Family (ERF) study. *Stroke* **36**, 2351 (2005). doi:10.1161/01.STR.0000185719.66735.dd Medline

108. E. Vartiainen *et al*., Thirty-five-year trends in cardiovascular risk factors in Finland. *Int. J. Epidemiol.* **39**, 504 (2010). doi:10.1093/ije/dyp330 Medline

109. J. Kaprio, L. Pulkkinen, R. J. Rose, Genetic and environmental factors in health-related behaviors: Studies on Finnish twins and twin families. *Twin Res.* **5**, 366 (2002). Medline

110. S. M. Purcell *et al.*; International Schizophrenia Consortium, Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748 (2009). Medline

111. FBPP Investigators, Multi-center genetic study of hypertension: The Family Blood Pressure Program (FBPP). *Hypertension* **39**, 3 (2002). doi:10.1161/hy1201.100415 Medline

112. T. B. Harris *et al.*, Waist circumference and sagittal diameter reflect total body fat better than visceral fat in older men and women. The Health, Aging and Body Composition Study. *Ann. N. Y. Acad. Sci.* **904**, 462 (2000). doi:10.1111/j.1749-6632.2000.tb06501.x Medline

113. D. J. P. Barker, C. Osmond, T. J. Forsén, E. Kajantie, J. G. Eriksson, Trajectories of growth among children who have coronary events as adults. *N. Engl. J. Med.* **353**, 1802 (2005). doi:10.1056/NEJMoa044160 Medline

114. L. Ferrucci *et al.*, Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. *J. Am. Geriatr. Soc.* **48**, 1618 (2000). Medline

115. H.-E. Wichmann, C. Gieger, R. Illig; MONICA/KORA Study Group, KORA-gen - Resource for population genetics, controls and a broad specturm of disease phenotypes. *Gesundheitswesen* **67**, 26 (2005). doi:10.1055/s-2005-858226 Medline

116. R. P. Stolk *et al.*, Universal risk factors for multifactorial diseases. LifeLines: A three-generation population-based study. *Eur. J. Epidemiol.* **23**, 67 (2008). doi:10.1007/s10654-007-9204-4 Medline

117. I. J. Deary, M. C. Whiteman, J. M. Starr, L. J. Whalley, H. C. Fox, The impact of childhood intelligence on later life: Following up the Scottish mental surveys of 1932 and 1947. *J. Pers. Soc. Psychol.* **86**, 130 (2004). doi:10.1037/0022-3514.86.1.130 Medline

118. I. J. Deary *et al.*, The Lothian Birth Cohort 1936: A study to examine influences on cognitive ageing from age 11 to age 70 and beyond. *BMC Geriatr.* **7**, 28 (2007). doi:10.1186/1471-2318-7-28 Medline

119. P. Magnus *et al.*; MoBa Study Group, Cohort profile: The Norwegian Mother and Child Cohort Study (MoBa). *Int. J. Epidemiol.* **35**, 1146 (2006). doi:10.1093/ije/dyl170 Medline

120. L. M. Irgens, The Medical Birth Registry of Norway. Epidemiological research and surveillance throughout 30 years. *Acta Obstet. Gynecol. Scand.* **79**, 435 (2000). doi:10.1080/j.1600-0412.2000.079006435.x Medline

121. B. W. J. H. Penninx *et al.*; NESDA Research Consortium, The Netherlands Study of Depression and Anxiety (NESDA): Rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* **17**, 121 (2008). doi:10.1002/mpr.256 Medline

122. P. Rantakallio, Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatr. Scand.* **193** (suppl.), 193, 1 (1969). Medline

123. C. Sabatti *et al.*, Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41**, 35 (2009). doi:10.1038/ng.271 Medline

124. N. W. Martin *et al.*, Educational attainment: A genome wide association study in 9538 Australians. *PLoS ONE* **6**, e20128 (2011). doi:10.1371/journal.pone.0020128 Medline

125. K. Estrada *et al*., GRIMP: A web- and grid-based tool for high-speed analysis of large-scale genome-wide association using imputed data. *Bioinformatics* **25**, 2750 (2009). doi:10.1093/bioinformatics/btp497 Medline

126. A. Hofman *et al*., The Rotterdam Study: 2012 objectives and design update. *Eur. J. Epidemiol.* **26**, 657 (2011). doi:10.1007/s10654-011-9610-5 Medline

127. D. A. Bennett *et al*., Overview and findings from the rush Memory and Aging Project. *Curr. Alzheimer Res.* **9**, 646 (2012). Medline

128. L. J. Bierut *et al*.; Gene, Environment Association Studies Consortium, A genome-wide association study of alcohol dependence. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 5082 (2010). doi:10.1073/pnas.0911109107 Medline

129. G. Pilia *et al*., Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet.* **2**, e132 (2006). doi:10.1371/journal.pgen.0020132 Medline

130. H. Völzke *et al*., Cohort profile: The study of health in Pomerania. *Int. J. Epidemiol.* **40**, 294 (2011). doi:10.1093/ije/dyp394 Medline

131. P. K. E. Magnusson *et al*., The Swedish Twin Registry: Establishment of a biobank and other recent developments. *Twin Res. Hum. Genet.* **16**, 317 (2013). doi:10.1017/thg.2012.104 Medline

132. A. Moayyeri, C. J. Hammond, A. M. Valdes, T. D. Spector, Cohort profile: TwinsUK and Healthy Ageing Twin Study. *Int. J. Epidemiol.* **42**, 76 (2013). Medline

133. O. T. Raitakari *et al*., Cohort profile: The cardiovascular risk in Young Finns Study. *Int. J. Epidemiol.* **37**, 1220 (2008). doi:10.1093/ije/dym225 Medline

134. V. Pfaffenrath *et al*., Regional variations in the prevalence of migraine and tension-type headache applying the new IHS criteria: The German DMKG Headache Study. *Cephalalgia* **29**, 48 (2009). doi:10.1111/j.1468-2982.2008.01699.x Medline

135. M. M. Vennemann, T. Hummel, K. Berger, The association between smoking and smell and taste impairment in the general population. *J. Neurol.* **255**, 1121 (2008). doi:10.1007/s00415-008-0807-9 Medline

136. A. Aromaa, *Health and Functional Capacity in Finland: Baseline Results of the Health 2000 Health Examination Survey* (National Public Health Institute, Helsinki, 2004).

137. M. McEvoy *et al*., Cohort profile: The Hunter Community Study. *Int. J. Epidemiol.* **39**, 1452 (2010). doi:10.1093/ije/dyp343 Medline

138. D. Weir, in *Biosocial Surveys*, M. Weinstein, J. W. Vaupel, K. W. Wachter, Eds. (Committee on Advances in Collecting and Utilizing Biological Indicators and Genetic Information in Social Science Surveys, Washington, DC, 2007), pp. 78, chap. 4.

139. M. B. Miller *et al*., The Minnesota Center for Twin and Family Research genome-wide association study. *Twin Res. Hum. Genet.* **15**, 767 (2012). doi:10.1017/thg.2012.62 Medline

140. J. H. Lee, R. Cheng, N. Graff-Radford, T. Foroud, R. Mayeux; National Institute on Aging Late-Onset Alzheimer's Disease Family Study Group, Analyses of the National Institute on Aging late-onset Alzheimer's disease family study: Implication of additional loci. *Arch. Neurol.* **65**, 1518 (2008). doi:10.1001/archneur.65.11.1518 Medline

141. D. I. Boomsma *et al*., Netherlands Twin Register: From twins to twin families. *Twin Res. Hum. Genet.* **9**, 849 (2006). doi:10.1375/twin.9.6.849 Medline

142. R. McQuillan *et al*., Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83**, 359 (2008). doi:10.1016/j.ajhg.2008.08.007 Medline

143. E. V. Theodoraki *et al*., Fibrinogen beta variants confer protection against coronary artery disease in a Greek case-control study. *BMC Med. Genet.* **11**, 28 (2010). doi:10.1186/1471-2350-11-28 Medline

144. S. Mukherjee *et al*., Cohort profile: The Western Australian Sleep Health Study. *Sleep Breath.* **16**, 205 (2012). doi:10.1007/s11325-011-0491-3 Medline

145. L. A. Baker, S. A. Treloar, C. A. Reynolds, A. C. Heath, N. G. Martin, Genetics of educational attainment in Australian twins: Sex differences and secular changes. *Behav. Genet.* **26**, 89 (1996). doi:10.1007/BF02359887 Medline

146. P. Miller, C. Mulvey, N. Martin, The return to schooling: Estimates from a sample of young Australian twins. *Labour Econ.* **13**, 571 (2006). doi:10.1016/j.labeco.2004.10.008

147. K. Silventoinen, R. F. Krueger, T. J. Bouchard Jr., J. Kaprio, M. McGue, Heritability of body height and educational attainment in an international context: Comparison of adult twins in Minnesota and Finland. *Am. J. Hum. Biol.* **16**, 544 (2004). doi:10.1002/ajhb.20060 Medline

148. A. C. Heath *et al*., Education policy and the heritability of educational attainment. *Nature* **314**, 734 (1985). doi:10.1038/314734a0 Medline

149. G. Isacsson, Estimating the economic return to educational levels using data on twins. *J. Appl. Econ.* **19**, 99 (2004). doi:10.1002/jae.724

150. P. Taubman, The determinants of earnings: Genetics, family, and other environments: A study of white male twins. *Am. Econ. Rev.* **66**, 858 (1976).

151. D. T. Lykken, T. J. Bouchard Jr., M. McGue, A. Tellegen, The Minnesota Twin Family Registry: Some initial findings. *Acta Genet. Med. Gemellol. (Roma)* **39**, 35 (1990). Medline

152. J. R. Behrman, P. Taubman, T. Wales, in *Kinometrics: Determinants of Socioeconomic Success Within and Between Families* (North-Holland Publishing Company, New York, 1977), pp. 35.

153. M. Soler Artigas *et al*.; GIANT consortium, Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat. Genet.* **43**, 1082 (2011). doi:10.1038/ng.941 Medline

154. T. Thye *et al*.; African TB Genetics Consortium; Wellcome Trust Case Control Consortium, Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat. Genet.* **42**, 739 (2010). doi:10.1038/ng.639 Medline

155. J. R. Shaffer *et al*., GWAS of dental caries patterns in the permanent dentition. *J. Dent. Res.* **92**, 38 (2013). doi:10.1177/0022034512463579 Medline

156. R. A. Eeles *et al*.; UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology; UK ProtecT Study Collaborators; PRACTICAL Consortium, Identification of seven new prostate cancer susceptibility

loci through a genome-wide association study. *Nat. Genet.* **41**, 1116 (2009). doi:10.1038/ng.450 Medline

157. N. M. Pajewski *et al*., A genome-wide association study of host genetic determinants of the antibody response to Anthrax Vaccine Adsorbed. *Vaccine* **30**, 4778 (2012). doi:10.1016/j.vaccine.2012.05.032 Medline

158. N. Sandholm *et al*.; DCCT/EDIC Research Group, New susceptibility loci associated with kidney disease in type 1 diabetes. *PLoS Genet.* **8**, e1002921 (2012). doi:10.1371/journal.pgen.1002921 Medline

159. B. Benyamin *et al*., Variants in *TF* and *HFE* explain ~40% of genetic variation in serum-transferrin levels. *Am. J. Hum. Genet.* **84**, 60 (2009). doi:10.1016/j.ajhg.2008.11.011 Medline

160. R. Qayyum *et al*., A meta-analysis and genome-wide association study of platelet count and mean platelet volume in african americans. *PLoS Genet.* **8**, e1002491 (2012). doi:10.1371/journal.pgen.1002491 Medline

161. C. Gieger *et al*., New gene functions in megakaryopoiesis and platelet formation. *Nature* **480**, 201 (2011). doi:10.1038/nature10659 Medline

162. C. S. Fox *et al*.; GIANT Consortium; MAGIC Consortium; GLGC Consortium, Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS Genet.* **8**, e1002695 (2012). doi:10.1371/journal.pgen.1002695 Medline

163. M. Kolz *et al*.; EUROSPAN Consortium; ENGAGE Consortium; PROCARDIS Consortium; KORA Study; WTCCC, Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS Genet.* **5**, e1000504 (2009). doi:10.1371/journal.pgen.1000504 Medline

164. M. Man *et al*., Beyond single-marker analyses: Mining whole genome scans for insights into treatment responses in severe sepsis. *Pharmacogenomics J.* **13**, 218 (2012). Medline

165. J. E. Landers *et al*., Reduced expression of the Kinesin-Associated Protein 3 (KIFAP3) gene increases survival in sporadic amyotrophic lateral sclerosis. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9004 (2009). doi:10.1073/pnas.0812937106 Medline

166. C. Cotsapas *et al*.; GIANT Consortium, Common body mass index-associated variants confer risk of extreme obesity. *Hum. Mol. Genet.* **18**, 3502 (2009). doi:10.1093/hmg/ddp292 Medline

167. K. Nakabayashi *et al*., Identification of independent risk loci for Graves' disease within the MHC in the Japanese population. *J. Hum. Genet.* **56**, 772 (2011). doi:10.1038/jhg.2011.99 Medline

168. B. Kestenbaum *et al*., Common genetic variants associate with serum phosphorus concentration. *J. Am. Soc. Nephrol.* **21**, 1223 (2010). doi:10.1681/ASN.2009111104 Medline

169. M. J. Barber *et al*., Genome-wide association of lipid-lowering response to statins in combined study populations. *PLoS ONE* **5**, e9763 (2010). doi:10.1371/journal.pone.0009763 Medline

170. A. I. Yashin, D. Wu, K. G. Arbeev, S. V. Ukraintseva, Joint influence of small-effect genetic variants on human longevity. *Aging* **2**, 612 (2010). Medline

171. G. Thorleifsson *et al*., Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat. Genet.* **41**, 18 (2009). doi:10.1038/ng.274 Medline

172. C. J. Willer *et al*.; Wellcome Trust Case Control Consortium; Genetic Investigation of ANthropometric Traits Consortium, Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41**, 25 (2009). doi:10.1038/ng.287 Medline

173. E. Melum *et al*., Genome-wide association analysis in primary sclerosing cholangitis identifies two non-HLA susceptibility loci. *Nat. Genet.* **43**, 17 (2011). doi:10.1038/ng.728 Medline

174. J. M. Robins, S. Greenland, Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143 (1992). doi:10.1097/00001648-199203000-00013 Medline

175. M. J. H. M. van der Loos *et al*., The molecular genetic architecture of self-employment. *PLoS ONE* **8**, e60542 (2013). doi:10.1371/journal.pone.0060542 Medline