A genome-wide association study of educational attainment was conducted in a discovery sample of 101,069 individuals and a replication sample of 25,490. Three independent SNPs are genome-wide significant (rs9320913, rs11584700, rs4851266), and all three replicate. Estimated effects sizes are small ($R^2 \approx 0.02\%$), approximately 1 month of schooling per allele. A linear polygenic score from all measured SNPs accounts for $\approx 2\%$ of the variance in both educational attainment and cognitive function. Genes in the region of the loci have previously been associated with health, cognitive, and central nervous system phenotypes, and bioinformatics analyses suggest the involvement of the anterior caudate nucleus. These findings provide promising candidate SNPs for follow-up work, and our effect size estimates can anchor power analyses in social-science genetics.

This inflation is not driven by specific cohorts and is expected for a highly polygenic phenotype even in the absence of population stratification (16).

From the discovery phase, we identified one genome-wide significant locus (rs9320913, $p = 4.2 \times 10^{-9}$) and three suggestive loci (defined as $p < 10^{-6}$) for EduYears. For College, we identified two genome-wide significant loci (rs11584700, $p = 2.1 \times 10^{-9}$, and rs4851266, $p = 2.2 \times 10^{-9}$) and an additional four suggestive loci (Table 1). We conducted replication analyses in 12 additional, independent cohorts that became available after the completion of the discovery meta-analysis, using the same pre-specified analysis plan. For both EduYears and College, the replication sample comprises 25,490 individuals. For each of the ten loci that reached at least suggestive significance, we brought forward for replication the SNP with the lowest $p$-value. The three genome-wide significant SNPs replicate at the Bonferroni-adjusted 5% level, with point estimates of the same sign and similar magnitude (Fig. 1 and Table 1). The seven loci that did not reach genome-wide significance did not replicate (the effect went in the anticipated direction in 5 out of 7 cases). The meta-analytic findings are not driven by extreme results in a small number of cohorts (see $p_{het}$ in Table 1), by cohorts from a specific geographic region (figs. S7 to S15), or by a single sex (figs. S3 to S6). Given the high correlation between EduYears and College (5), it is unsurprising that the set of SNPs with low $p$-values exhibit considerable overlap in the two analyses (tables S8 and S9).

The observed effect sizes of the three replicated individual SNPs are small (see (5) for discussion). For EduYears, the strongest effect identified (rs9320913) explains 0.022% of phenotypic variance in the replication sample. This $R^2$ corresponds to a difference of ~1 months of schooling per allele. For college completion, the SNP with the strongest estimated effect (rs11584700) has an odds ratio of 0.912 in the replication sample, equivalent to a 1.8 percentage-point difference per allele in the frequency of completing college.

We subsequently conducted a “combined stage” meta-analysis, including both the discovery and replication samples. This analysis revealed additional genome-wide significant SNPs: four for EduYears and three for College. Three of these newly genome-wide significant SNPs (rs1487441, rs11584700, rs4851264) are in linkage disequilibrium with the replicated SNPs. The remaining four are located in different loci and warrant replication attempts in future research: rs7309, a 3′UTR variant.
in TANK; rs11687170, close to Gbx2; rs1056667, a 3’UTR variant in BTN1A1; and rs134001104 in ASB18.

Using the results of the combined meta-analyses of discovery and replication cohorts, we conducted a series of complementary and exploratory supplemental analyses to aid in interpreting and contextualizing the results: gene-based association tests; eQTL analyses of brain and blood tissue data; pathway analysis; functional annotation searches; enrichment analysis for cell-type-specific overlap with H3K4me3 chromatin marks; and predictions of likely gene function using gene-expression data. Table S20 summarizes promising candidate loci identified through follow-up analyses (5). Two regions in particular showed convergent evidence from functional annotation, blood cis-eQTL analyses, and gene-based tests: chromosome 1q32 (including LRRN2, MDM4, and PIK3C2B) and chromosome 6 near the Major Histocompatibility Complex (MHC). We also find evidence that in anterior caudate cells, there is enrichment of H3K4me3 chromatin marks (believed to be more common in active regulatory regions) in the genomic regions implicated by our analyses (fig. S20). Many of the implicated genes have previously been associated with health, central nervous system, or cognitive-process phenotypes in either human-GWAS or model-animal studies (table S22). Gene coexpression analysis revealed that several implicated genes (including BSN, Gbx2, LRRN2, and PIK3C2B) are likely involved in pathways related to cognitive processes (such as learning and long-term memory) and neuronal development or function (table S21).

Although the effects of individual SNPs on educational attainment are small, many of their potential uses in social science depend on their combined explanatory power. To evaluate the combined explanatory power, we constructed a linear polygenic score (5) for each of our two education measures using the meta-analysis results (combining discovery and replication), excluding one cohort. We tested these scores for association with educational attainment in the excluded cohort. We constructed the scores using SNPs whose nominal p-values fall below a certain threshold, ranging from 5 × 10^−4 (only the genome-wide significant SNPs were included) to 1 (all SNPs were included).

We replicated this procedure with two of the largest cohorts in the study, both of which are family-based samples (QIMR and STR). The results suggest that educational attainment is a highly polygenic trait (Fig. 2 and table S23): the amount of variance accounted for increases as the p-value threshold becomes less conservative (i.e., includes more SNPs). The linear polygenic score from all measured SNPs accounts for ≈ 2% (p = 1.0 × 10^−25) of the variance in EduYears in the STR sample and ≈ 3% (p = 7.1 × 10^−23) in the QIMR sample.

To explore one of the many potential mediating endophenotypes, we examined how much the same polygenic scores (constructed to explain EduYears or College) could explain individual differences in cognitive function. While it would have been preferable to explore a richer set of mediators, this variable was available in STR, a dataset where we had access to the individual-level genotypic data. Cognitive function had been measured in a subset of males using the Swedish Enlistment Battery (used for conscription) (5, 17). The estimated R^2 ≈ 2.5% (p < 1.0 × 10^−5) for cognitive function is actually slightly larger than the fraction 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 (5). A mediation analysis (table S24) provides tentative evidence consistent with this interpretation.

The polygenic score remains associated with educational attainment and cognitive function in within-family analyses (table S25). Thus, these results appear robust to possible population stratification.

If the size of the training sample used to estimate the linear polygenic score increased, the explanatory power of the score in the prediction sample would be larger because the coefficients used for constructing the score would be estimated with less error. In (5), we report projections of this increase. We also assess, at various levels of explanatory power, the benefits from using the score as a control variable in a randomized educational intervention (5). An asymptotic upper bound for the explanatory power of a linear polygenic score is the additive genetic variance across individuals captured by current SNP microarrays. Using combined data from STR and QIMR, we estimate that this upper bound is 22.4% (S.E. = 4.2%) in these samples (table S12).

Placed in the context of the GWAS literature (10), our largest estimated SNP effect size of 0.02% is over an order of magnitude smaller than those observed for height and BMI: 0.4% (15) and 0.3% (18) respectively. While our linear polygenic score for education achieves an R^2 of 2% estimated from a sample of 120,000, a score for height reached 10% estimated from a sample of 180,000 (15), and a score for BMI using only the top 32 SNPs reached 1.4% (18). Taken together, our findings suggest that the genetic architecture of complex behavioral traits is far more diffuse than that of complex physical traits.

Existing claims of “candidate gene” associations with complex social–cognitive processes are widely varying effect sizes—many with R^2 values more than one hundred times larger than those we find (4, 6). For complex social–science phenotypes that are likely to have a genetic architecture similar to educational attainment, our estimate of 0.02% can serve as a benchmark for conducting power analyses and evaluating the plausibility of existing findings in the literature.

The few GWAS studies conducted to date in social–science genetics have not found genome-wide significant SNPs that replicate consistently (19, 20). One commonly proposed solution is to gather better measures of the phenotypes in more environmentally homogenous samples. Our findings demonstrate the feasibility of a complementary approach: identify a phenotype that, although more distal from genetic influences, is available in a much larger sample [see (5) for a simple theoretical framework and power analysis]. The genetic variants uncovered by this “proxy-phenotype” methodology can then serve as a set of empirically-based candidate genes in follow-up work, such as tests for associations with well-measured endophenotypes (e.g., personality, cognitive function), research on gene–environment interactions, or explorations of biological pathways.

In social–science genetics, researchers must be especially vigilant to avoid misinterpretations. One of the many concerns is that a genetic association will be mischaracterized as “the gene for X,” encouraging misperceptions that genetically influenced phenotypes are immune to environmental intervention [for rebuttals, see (21, 22)] and misperceptions that individual SNPs have large effects (which our evidence contradicts). If properly interpreted, identifying SNPs and constructing polygenic scores are steps toward usefully incorporating genetic data into social–science research.

References and Notes

5. Please see the supplementary materials on Science Online.

Downloaded from www.sciencemag.org on May 31, 2013
32. P. I. W. de Bakker


20. B. Benyamin

19. M. H. de Moor

17. B. Carlstedt ,

16. J. Yang


15. A. C. J. Lager , J. Torssander , Causal effect of education on mortality in a


3. L. E. Duncan, M. C. Keller, A critical review of the first 10 years of candidate

2. L. E. Duncan, M. C. Keller, A critical review of the first 10 years of candidate

1. L. E. Duncan, M. C. Keller, A critical review of the first 10 years of candidate
M. Imielinski 

D. P. McGovern 

L. Jostins 

J. C. Barrett 

J. C. Barrett 

M. Parkes 


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 474, 661 (2007). doi:10.1038/nature09511 


J. C. Barrett et al.; UK IBID 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 

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 


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 

M. Inuielinski et al.; Western Regional Alliance for Pediatric IBID; International IBID Genetics Consortium; NIDDK IBID Genetics Consortium; Belgian-French IBID 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.480 


Italy. 66Department of Sciences Biomedical, Università di Sassari, 07100 SS, Italy. 67Department of Health Sciences Research, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN 55905, USA. 68Program in Translational Neuropsychiatric Genomics, Department of Neurology, Brigham and Women’s Hospital, Boston, MA 02115, USA. 69School of Public Policy, University College London, London WC1H 9QU, UK. 70Centre for Economic Performance, London School of Economics, London WC2A 2AE, UK. 71Department of Psychology, The University of Edinburgh, Edinburgh EH8 9JZ, Scotland, UK. 72Department of Nutrition and Dietetics, Harokopio University of Athens, Athens 17671, Greece. 73LifeLines Cohort Study, University of Groningen, University Medical Center Groningen, 9700 RB Groningen, The Netherlands. 74Department of General Practice and Primary Health Care, University of Helsinki, Helsinki 00014, Finland. 75Unit of General Practice, Helsinki University Central Hospital, Helsinki 00280, Finland. 76Folkhälsoanstalt Research Center, Helsinki 00250, Finland. 77Vaasa Central Hospital, Vaasa 65130, Finland. 78MRC Centre for Causal Analyses in Translational Epidemiology, School of Social and Community Medicine, University of Bristol, Bristol BS8 2PR, UK. 79Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, USA. 80Division of Biology and Biomedical Sciences, Washington University, St. Louis, MO 63110-1093, USA. 81Faculty of Behavioral and Social Sciences, University of Groningen, 9747 AD Groningen, The Netherlands. 82Department of Psychology, University of Minnesota, Minneapolis, MN 55455-0344, USA. 83Institute of Health Economics and Health Care Management, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany. 84Hannover Unified Biobank, Hannover Medical School, 30625 Hannover, Germany. 85Department of Epidemiology and Biostatistics, MRC-HPA Centre for Environment and Health, Imperial College London, London W2 1PG, UK. 86Unit of Primary Care, Oulu University Hospital, Oulu 90220, Finland. 87Department of Children and Young People and Families, National Institute for Health and Welfare, Oulu 90101, Finland. 88Department of Clinical Physiology, Tampere University Hospital and University of Tampere School of Medicine, Tampere 33520, Finland. 89Department of Public Health, University of Helsinki, 00014 Helsinki, Finland. 90Department of Mental Health and Substance Abuse Services, National Institute for Health and Welfare, 00300 Helsinki, Finland. 91Ontario Institute for Cancer Research, Toronto, Ontario MSG 0A3, Canada. 92Department of Clinical Chemistry, Fimlab Laboratories, Tampere University Hospital, Tampere 33520, Finland. 93Human Genetics, Genome Institute of Singapore, Singapore 138672, Singapore. 94Institute of Human Genetics, Helmholtz Centre Munich, German Research Center for Environmental Health, 85764 Neuherberg, Germany. 95Department of Health, Functional Capacity and Welfare, National Institute for Health and Welfare, Helsinki 00271, Finland. 96Western Australia Sleep Disorders Research Institute, Sir Charles Gairdner Hospital, Perth, Western Australia 6009, Australia. 97Department of Medicine, University of Toronto, Toronto, Ontario M5S 1A8, Canada. 98Women’s College Research Institute, University of Toronto, Toronto, Ontario MSG 1N8, Canada. 99Department of Medical Genetics, University of Helsinki, 00014 Helsinki, Finland. 100Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku 20520, Finland. 101Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku 20520, Finland. 102Department of Internal Medicine, Erasmus Medical Center, Rotterdam 3000 CA, The Netherlands. 103Department of Economics, University of Minnesota, Minneapolis, MN 55455-0462, USA. 104Chronic Disease Epidemiology Unit, Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki 00271, Finland. 105School of Oral and Dental Sciences, University of Bristol, Bristol BS1 2LY, UK. 106Alzheimer Scotland Dementia Research Centre, The University of Edinburgh, Edinburgh EH8 9JZ, Scotland, UK. 107Department of Computer Science, University of California, Los Angeles, CA 90095, USA. 108Department of Economics, Oulu Business School, University of Oulu, Oulu 90014, Finland. 109Interfaculty Institute for Genetics and Functional Genomics, Department of Functional Genomics, University Medicine Greifswald, Greifswald 17487, Germany. 110Department of Child and Adolescent Psychiatry, Erasmus Medical Center, 3000 CB Rotterdam, The Netherlands. 111Division of Welfare and Health Promotion, National Institute for Health and Welfare, Helsinki 00271, Finland. 112Department of Medicine, Turku University Hospital, Turku 20520, Finland. 113Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology, Ludwig-Maximilians-Universität, 81377 Munich, Germany. 114Klinikum Grosshadern, 81377 Munich, Germany. 115Institute of Epidemiology I, Helmholtz Zentrum München, German Research Centre for Environmental Health, 85764 Neuherberg, Germany. 116Department of Sociology, New York University, New York, NY 10012, USA. 117Department of Economics, Harvard University, Cambridge, MA 02138, USA. 118Petrie-Flom Center for Health Law Policy, Biotechnology, & Bioethics, Harvard Law School, Cambridge, MA 02138, USA. 119Nelson A. Rockefeller Institute of Government, State University of New York, Albany, NY 12203–1003, USA. 120Department of Clinical Genetics, VU University Medical Center, 1081 BT Amsterdam, The Netherlands. 121Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam 3000 DR, The Netherlands. 122Department of Economics, Stockholm School of Economics, Stockholm 113 86, Sweden. 123Panteia, Zoetermeer 2701 AA, Netherlands. 124GSCM-Montpellier Business School, Montpellier 34185, France. 125Centre for Medical Systems Biology, Leiden University Medical Center, 2300 RC Leiden, The Netherlands. 126Department of Economics, Cornell University, Ithaca, NY 14853, USA. 127Center for Experimental Social Science, Department of Economics, New York University, New York, NY 10012, USA. 128Division of Social Science, New York University Abu Dhabi, PO Box 129188, Abu Dhabi, UAE. 129Research Institute of Industrial Economics, Stockholm 102 15, Sweden.

*These authors contributed equally to this work.
†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.)
Table 1. The results of the GWAS meta-analysis for the independent signals reaching $p < 10^{-6}$ in the discovery stage. The rows in bold are the independent signals reaching $p < 5 \times 10^{-8}$ in the discovery stage. "Frequency" refers to allele-frequency in the combined-stage meta-analysis. "Beta/OR" refers to the effect size in the EduYears analysis and to the Odds Ratio in the College analysis. All $p$-values are from the sample-size-weighted meta-analysis (fixed effects). The $p$-value in the replication stage meta-analysis was calculated from a one-sided test. $I^2$ represents the % heterogeneity of effect size between the discovery stage studies. $p_{\text{het}}$ is the heterogeneity $p$-value.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr</th>
<th>Position (bp)</th>
<th>Nearest gene</th>
<th>Effective allele</th>
<th>Frequency</th>
<th>Beta/OR</th>
<th>$P$-value</th>
<th>$I^2$</th>
<th>$p_{\text{het}}$</th>
<th>Beta/OR</th>
<th>$P$-value</th>
<th>$I^2$</th>
<th>$p_{\text{het}}$</th>
<th>Beta/OR</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>EduYears</td>
<td>College</td>
<td></td>
<td></td>
<td>College</td>
<td></td>
<td></td>
<td></td>
<td>College</td>
<td></td>
</tr>
<tr>
<td>rs9320913</td>
<td>6</td>
<td>98691454</td>
<td>LOC100129158</td>
<td>A</td>
<td>0.483</td>
<td>0.016</td>
<td>4.19×10^{-8}</td>
<td>18.3</td>
<td>0.097</td>
<td>0.077</td>
<td>0.012</td>
<td></td>
<td></td>
<td></td>
<td>0.350</td>
</tr>
<tr>
<td>rs3783006</td>
<td>13</td>
<td>97909210</td>
<td>STK24</td>
<td>C</td>
<td>0.454</td>
<td>0.096</td>
<td>2.29×10^{-7}</td>
<td>0</td>
<td>0.982</td>
<td>0.056</td>
<td>0.055</td>
<td></td>
<td></td>
<td></td>
<td>0.959</td>
</tr>
<tr>
<td>rs8049439</td>
<td>16</td>
<td>28745016</td>
<td>ATXN2L</td>
<td>T</td>
<td>0.581</td>
<td>0.090</td>
<td>7.12×10^{-7}</td>
<td>10.7</td>
<td>0.229</td>
<td>0.065</td>
<td>0.026</td>
<td></td>
<td></td>
<td></td>
<td>0.205</td>
</tr>
<tr>
<td>rs1318378</td>
<td>5</td>
<td>101958587</td>
<td>SLCO6A1</td>
<td>A</td>
<td>0.878</td>
<td>-0.136</td>
<td>7.49×10^{-7}</td>
<td>0</td>
<td>0.791</td>
<td>0.091</td>
<td>0.141</td>
<td></td>
<td></td>
<td></td>
<td>0.646</td>
</tr>
<tr>
<td>rs11584700</td>
<td>1</td>
<td>202843606</td>
<td>LRRN2</td>
<td>A</td>
<td>0.780</td>
<td>0.921</td>
<td>2.07×10^{-9}</td>
<td>13.8</td>
<td>0.179</td>
<td>0.912</td>
<td>4.86×10^{-4}</td>
<td></td>
<td></td>
<td></td>
<td>0.221</td>
</tr>
<tr>
<td>rs4851266</td>
<td>2</td>
<td>100184911</td>
<td>LOC150577</td>
<td>T</td>
<td>0.396</td>
<td>1.050</td>
<td>2.20×10^{-9}</td>
<td>23.7</td>
<td>0.049</td>
<td>1.049</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
<td>1.050</td>
</tr>
<tr>
<td>rs2054125</td>
<td>2</td>
<td>199093966</td>
<td>PLC1L1</td>
<td>T</td>
<td>0.064</td>
<td>1.468</td>
<td>5.55×10^{-8}</td>
<td>7</td>
<td>0.325</td>
<td>1.098</td>
<td>0.225</td>
<td></td>
<td></td>
<td></td>
<td>1.054</td>
</tr>
<tr>
<td>rs3227</td>
<td>6</td>
<td>33770273</td>
<td>ITPR3</td>
<td>C</td>
<td>0.498</td>
<td>1.043</td>
<td>6.02×10^{-8}</td>
<td>5</td>
<td>0.363</td>
<td>1.010</td>
<td>0.280</td>
<td></td>
<td></td>
<td></td>
<td>1.046</td>
</tr>
<tr>
<td>rs4073894</td>
<td>7</td>
<td>104254200</td>
<td>LHFP1L3</td>
<td>A</td>
<td>0.207</td>
<td>1.076</td>
<td>4.41×10^{-7}</td>
<td>0</td>
<td>0.765</td>
<td>1.003</td>
<td>0.467</td>
<td></td>
<td></td>
<td></td>
<td>1.050</td>
</tr>
<tr>
<td>rs12640626</td>
<td>4</td>
<td>17683266</td>
<td>GPM6A</td>
<td>A</td>
<td>0.580</td>
<td>1.041</td>
<td>4.94×10^{-7}</td>
<td>10.9</td>
<td>0.234</td>
<td>1.000</td>
<td>0.495</td>
<td></td>
<td></td>
<td></td>
<td>1.038</td>
</tr>
</tbody>
</table>
Fig. 1. Regional association plots of replicated loci associated with educational attainment [(A): rs9320913, (B): rs11584700, (C): rs4851266]. The plots are centered on the SNPs with the lowest p-values in the discovery stage (purple diamond). The $R^2$ values are from the CEU HapMap 2 samples. The CEU HapMap 2 recombination rates are indicated with a blue line on the right-hand y-axis. The figures were created with LocusZoom (http://csg.sph.umich.edu/locuszoom/).
Fig. 2. Solid lines show results from regressions of EduYears on linear polygenic scores in a set of unrelated individuals from the QIMR (N = 3526) and STR (N = 6770) cohorts. Dashed lines show results from regressions of Cognitive function on linear polygenic scores in a sample from STR (N = 1419). The scores are constructed from the meta-analysis for either EduYears or College, excluding the QIMR and STR cohorts.