Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2017 November 22.

Published in final edited form as:

Nat Genet. 2017 July ; 49(7): 1107–1112. doi:10.1038/ng.3869.

Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence

Suzanne Sniekers[1], **Sven Stringer**[1], **Kyoko Watanabe**[1], **Philip R Jansen**[1,2], **Jonathan RI Coleman**[3,4], **Eva Krapohl**[3], **Erdogan Taskesen**[1,5], **Anke R Hammerschlag**[1], **Aysu Okbay**[1,6], **Delilah Zabaneh**[3], **Najaf Amin**[7], **Gerome Breen**[3,4], **David Cesarini**[8], **Christopher F Chabris**[9], **William G Iacono**[10], **M. Arfan Ikram**[11], **Magnus Johannesson**[12], **Philipp Koellinger**[1,6], **James J Lee**[13], **Patrik K. E. Magnusson**[14], **Matt McGue**[10], **Mike B. Miller**[10], **William ER Ollier**[15], **Antony Payton**[15], **Neil Pendleton**[16], **Robert Plomin**[3], **Cornelius A Rietveld**[6,17], **Henning Tiemeier**[2,11,18], **Cornelia M van Duijn**[7,19], and **Danielle Posthuma**[1,20,*]

[1]Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University Amsterdam, The Netherlands [2]Department of Child and Adolescent Psychiatry, Erasmus Medical Center, 3000 CB, Rotterdam, The Netherlands [3]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, UK [4]NIHR Biomedical Research Centre for Mental Health, South London and Maudsley NHS Trust, London, UK [5]Alzheimer Centrum, Amsterdam Neuroscience, VU University Medical Center, Amsterdam, The Netherlands [6]Erasmus University Rotterdam Institute for Behavior and Biology, 3062 PA Rotterdam, The Netherlands [7]Genetic Epidemiology Unit, Department of Epidemiology, Erasmus Medical Center, 3000 CA Rotterdam,

*Correspondence to: Danielle Posthuma: Department of Complex Trait Genetics, VU University, De Boelelaan 1085, 1081 HV, Amsterdam, The Netherlands. Phone: +31 20 598 2823, Fax: +31 20 5986926, d.posthuma@vu.nl.

URLs:

http://ukbiobank.ac.uk
http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=155580
http://ssgac.org/documents/CHIC_Summary_Benyamin2014.txt.gz
https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html
http://ctg.cncr.nl/software/magma
http://software.broadinstitute.org/gsea/msigdb/collections.jsp
http://genome.sph.umich.edu/wiki/METAL_Program
https://github.com/bulik/ldsc

Summary statistics have been made available for download from http://ctg.cncr.nl/software/summary_statistics.

Author Contributions: S.Sn. performed the analyses. D.P. conceived the study. S.St. QC-ed the UKB data. K.W. and E.T. conducted in silico follow-up analyses. P.R.J., E.K. and J.R.I.C. conducted PRS analyses. P.K., C.A.R., D.Z., H.T, C.v.D, N.A., P.M., D.C., M.J., M.McG, M.B.M., W.G.I., J.J.L., G.B., R.P., N.P., A.P.,W.O., M.A.I. and C.F.C contributed data. A.R.H. provided scripts for the pathway analyses. A.O. performed the educational attainment meta-analysis. S.Sn. and D.P. wrote the paper. All authors discussed the results and commented on the paper.

The other authors declare no competing financial interests.

Data Availability Statement:

Summary statistics have been made available for download from http://ctg.cncr.nl/software/summary_statistics. Genotype data that underlie the findings of this study are available from UK Biobank but restrictions apply to the availability of these data, which were used under license for the current study (application number 16406), and so are not publicly available. Summary statistics from the CHIC consortium are available from http://ssgac.org/documents/CHIC_Summary_Benyamin2014.txt.gz. Source data for Figs. 2b–d, 3a, d and Supplementary Fig. 5 have been provided in Supplementary Tables 4, 8, 14 and 7 respectively.

The Netherlands [8]Center for Experimental Social Science, Department of Economics, New York University, New York, NY 10012 [9]Department of Psychology, Union College, Schenectady, NY 12308 (currently at: Geisinger Health System, Danville, PA 17822) [10]Department of Psychology, University of Minnesota, Minneapolis, MN 55455-0344 [11]Department of Epidemiology, Erasmus Medical Center, 3000 CB, Rotterdam, The Netherlands [12]Department of Economics, Stockholm School of Economics, 113 83 Stockholm, Sweden [13]Department of Psychology, Harvard University, Cambridge, MA 02138; Department of Psychology, University of Minnesota, Minneapolis, MN 55455-0344 [14]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 171 77 Stockholm, Sweden [15]Centre for Epidemiology, Division of Population Health, Health Services Research & Primary Care, The University of Manchester [16]Division of Neuroscience and Experimental Psychology, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester M13 9PL [17]Department of Applied Economics, Erasmus School of Economics, Erasmus University Rotterdam, 3062 PA Rotterdam, The Netherlands [18]Department of Psychiatry, Erasmus Medical Center, 3000 CB, Rotterdam, The Netherlands [19]Translational Epidemiology, Faculty Science, Leiden University, The Netherlands [20]Department of Clinical Genetics, Amsterdam Neuroscience, VU University Medical Center, Amsterdam, The Netherlands

Intelligence is associated with important economic and health-related life outcomes[1]. Despite substantial heritability[2] (0.54) and confirmed polygenic nature, initial genetic studies were mostly underpowered[3–5]. Here we report a meta-analysis for intelligence of 78,308 individuals. We identify 336 single nucleotide polymorphisms (SNPs) (METAL $P<5\times10^{-8}$) in 18 genomic loci, of which 15 are novel. Roughly half are located inside a gene, implicating 22 genes, of which 11 are novel findings. Gene-based analyses identified an additional 30 genes (MAGMA $P<2.73\times10^{-6}$), of which all but one have not been implicated previously. We show that identified genes are predominantly expressed in brain tissue, and pathway analysis indicates the involvement of genes regulating cell development (MAGMA competitive $P=3.5\times10^{-6}$). Despite the well-known difference in twin-based heritability for intelligence in childhood (0.45) and adulthood[2] (0.80), we show substantial genetic correlation ($r_g=0.89$, LD Score regression $P=5.4\times10^{-29}$). These findings provide novel insight into the genetic architecture of intelligence.

We combined GWAS data for intelligence in 78,308 unrelated individuals from 13 cohorts (Online Methods). Of these, full GWAS results for intelligence on N=48,698 have been published in two different studies[5,6] (N=12,441 and N=36,257 respectively), while GWAS results on the remaining 29,610 individuals have not been published previously. Across the different cohorts, various tests to measure intelligence were used. Therefore – following previous publications on combining intelligence phenotypes across different cohorts[5,7] – the cohorts either calculated Spearman's g or used a primary measure of fluid intelligence (Supplementary Table 1), which is known to correlate highly with g[8]. Previous research has shown that many different aspects of intelligence are highly correlated to each other, and that Spearman's g captures the latent general intelligence trait, irrespective of the specific tests used to construct it[9,10].

All association studies were performed on individuals of European descent; standard quality-control procedures included correcting for population stratification and filtering on minor allele frequency and imputation quality (Online Methods). As eight out of the 13 cohorts consisted of children (aged < 18; total N=19,509) and five of adults (N=58,799, aged 18–78), we first meta-analyzed the children- and adult-based cohorts separately using METAL software[11], and subsequently calculated the r_g using LD Score regression[12]. The estimated r_g was 0.89 (SE=0.08, P=5.4×10^{-29}), indicating substantial overlap between the genetic variants influencing intelligence in childhood and adulthood, and warranting a combined meta-analysis. The genetic correlations between all individual cohorts were generally larger than 0.80 except for those involving some of the smaller sized cohorts (N<4,000), which, given the large standard errors of the r_g's, is likely due to the relatively low sample sizes in some of the individual cohorts (Supplementary Table 2). The full meta-analysis of all 13 cohorts (maximum N=78,308) included 12,104,294 SNPs. The quantile-quantile (Q-Q) plot of all SNPs exhibited some inflation (λ_{ALL}=1.21; Supplementary Fig. 1; Supplementary Table 3), which is within the expected range for a polygenic trait at the current sample size and heritability[13]. We performed LD Score regression to quantify the proportion of inflation in the mean χ^2 that was due to confounding biases. An intercept of 1.01 and mean χ^2 of 1.30 were obtained, suggesting that more than 95% of the inflation was caused by true polygenic signal. SNP-based heritability was estimated at 0.20 (SE=0.01) in the total sample, and this was comparable in adults (0.21, SE=0.01) and children (0.20; SE=0.03). These estimates were obtained using LD Score regression and are likely to be biased downwards.

The meta-analysis identified 18 independent genome-wide significant loci (Fig. 1; Fig. 2a; Table 1), including 336 top SNPs (i.e. below the genome-wide threshold of significance; Supplementary Table 4). Of the 18 identified loci, three have been implicated in intelligence previously: 6q16.1[14], 7p14.3 and 22q13.2[6] (Supplementary Table 5). The top SNPs implicated 22 genes of which 11 were novel. Functional annotation of the 336 genome-wide significant SNPs showed that a large proportion was intronic (162/336) (Fig. 2b). Of the 18 lead SNPs, 10 were intronic (Fig 2b), all were in an active chromatin state (Fig. 2c; Supplementary Fig. 2–4) and 8 SNPs were expression quantitative trait loci (eQTLs; Fig. 2d; Supplementary Table 4; Supplementary Table 6). Lead SNPs *rs12928404* (located in the intronic region of *ATXN2L*) had the highest probability of being a regulatory SNP based on the Regulome database score[15] and of the eight lead SNPs that were eQTLs, this SNP was associated with differential expression of the largest number of genes (i.e.14). Focusing on brain tissue, the T allele of this SNP, which was associated with higher intelligence scores, was associated with lower expression of *TUFM* (Supplementary Table 6).

We calculated the variance explained (R^2) in intelligence by the GWAS results in four independent samples, using LDpred[16] (Online Methods and Supplementary Table 7 and Supplementary Fig. 5). Our results show that the current results explain up to 4.8% of the variance in intelligence and that on average across the four samples there is a 1.9-fold increase in explained variance compared to the most recent GWAS on intelligence[6].

Apart from a SNP-by-SNP GWAS we conducted a genome-wide gene association analysis (GWGAS) as implemented in MAGMA[17] (Online Methods). GWGAS relies on converging

evidence from multiple genetic variants in the same gene and can yield novel genome-wide significant signals on a gene-based level that are not necessarily picked up by a standard GWAS. The GWGAS identified 47 genes (Fig. 3a, Supplementary Table 8). The GWGAS and GWAS identified 17 overlapping genes, thus the total number of implicated genes either by a SNP hit or by GWGAS was 22+47−17=52. Twelve out of 52 genes have been associated with intelligence previously (Supplementary Table 9). Tissue expression analyses (Online Methods) of the 52 genes using the GTEx data resource showed that 14 out of 44 genes for which GTEx data was available were more strongly expressed in the brain than in other tissues (Fig. 3b). Epigenetic states were calculated for 51 out of 52 implicated genes (Online Methods) and showed that 57% of genes were at least weakly transcribed in at least 50% of tissues (Fig. 3c; Supplementary Fig. 6). Pathway analysis for 6,166 gene ontology (GO[18]) and 674 Reactome[19] gene-sets (obtained from MSigDB[20]) resulted in one associated gene-set (GO: regulation of cell development, which is defined as any process that modulates the rate, frequency or extent of the progression of the cell over time, from its formation to the mature structure.) (MAGMA competitive $P=3.5\times10^{-6}$; corrected $P=0.03$, Supplementary Tables 10, 11). This gene-set contains four genes that were genome-wide significant: *BMPR2, SHANK3, DCC* and *ZFHX3*, and many other genes that showed weaker association (Supplementary Table 12). Three of the genome-wide significant genes are involved in neuronal function: *SHANK3* is involved in synapse formation, *DCC* encodes a netrin receptor involved in axon guidance and is associated with putamen volume, and *ZFHX3* is known to regulate myogenic and neuronal differentiation. The fourth gene, *BMPR2*, plays a role in embryogenesis and endochondral bone formation and has been linked to pulmonary arterial hypertension. The four GO pathways with the subsequent smallest *P*-values are not independent from the top associated gene-set and provide insight in more specific functions of the genes driving the observed gene-set association. These four gene-sets are: regulation of nervous system development ($P=3.0\times10^{-5}$; 87% of genes overlapping with the regulation of cell development pathway, including the four genome-wide significant genes), negative regulation of dendrite development ($P=7.9\times10^{-5}$; 100% overlapping, thus a complete subset), myelin sheath ($P=8.5\times10^{-5}$; 14% overlapping) and neuron spine ($P=1.5\times10^{-4}$; 34% overlapping).

Intelligence has been associated with many socio-economic and health-related outcomes. We used whole-genome LD Score Regression[12] to calculate the genetic correlation with 32 traits from these domains for which GWAS summary statistics were available for download. Significant genetic correlations were observed with 14 traits. The strongest, positive genetic correlation was with Educational attainment ($r_g=0.70$, SE=0.02, $P=2.5\times10^{-287}$). Moderate, positive genetic correlations were observed with smoking cessation, intracranial volume, head circumference in infancy, Autism spectrum disorder and height. Moderate negative genetic correlations were observed with Alzheimer's disease, depressive symptoms, having ever smoked, schizophrenia, neuroticism, waist-to-hip ratio, body mass index, and waist circumference (Fig. 3d; Supplementary Table 13).

To examine the robustness of the 336 SNPs and 47 genes that reached genome-wide significance in the primary analyses, we sought replication. Since there are no reasonably large GWAS for intelligence available and given the high genetic correlation with educational attainment, which has been used previously as a proxy for intelligence[7], we used

the summary statistics from the latest GWAS for educational attainment (EA[21]) for proxy-replication (Online Methods). We first deleted overlapping samples, resulting in a sample of 196,931 individuals for EA. Out of the 336 top SNPs for intelligence, 306 were available for look-up in EA, and 16 out of 18 independent lead SNPs. We found that the effects of 305 out of 306 available SNPs in EA were sign concordant between EA and intelligence, and the effects of all 16 independent lead SNPs (exact binomial $P<10^{-16}$; Supplementary Table 14). This approach resulted in nine proxy-replicated loci ($P<0.05/16$): seven for which the lead SNP was significant (16p11.2, 1p34.2, 2q11.2, 2q22.3, 3p24.3, 6q16.1 and 7q33) and two for which another correlated top SNP in the same locus was significant (3p24.2 and 7p14.3). Of the 47 genes that were significantly associated with intelligence in the GWGAS, 15 were also significantly associated with EA ($P<0.05/47$, Supplementary Table 15). Given the high (0.70) but not perfect genetic correlation between EA and intelligence, these results strongly support the involvement of the proxy-replicated SNPs and genes in intelligence.

The strongest emerging association with intelligence is with *rs2490272* (6q21) in an intronic region of *FOXO3* and neighboring SNPs in the promotor of the same gene. This gene is part of the insulin/insulin-like growth factor 1 signaling pathway and is believed to trigger apoptosis, including neuronal cell death as a result of oxidative stress[22]. Moreover, it has been shown to be associated with longevity[23,24]. The gene with the strongest association in the GWGAS is *CSE1L*, which also plays a role in apoptosis and cell proliferation[25]. Of all 52 genes that were implicated, 35 were reported in the GWAS catalog for a previous association with at least one of 67 distinct traits. Nine genes (*ATP2A1, NEGR1, SKAP1, FOXO3, COL16A1, YIPF7, DCC, SH2B1* and *TUFM*) were previously implicated with body mass index[26–29], seven (*CYP2D6, NAGA, NDUFA6, TCF20* and *SEPT3, FAM109B* and *MEF2C*) with schizophrenia[30] and four (*NEGR1, SH2B1, DCC* and *WNT4*) with obesity[31–33]. *EXOC4* and *MEF2C* have been associated previously with Alzheimer's disease (Supplementary Tables 16, 17). Many of the implicated genes are involved in neuronal function: *DCC, APBA1, PRR7, ZFHX3, HCRTR1, NEGR1, MEF2C, SHANK3* and *ATXN2*L (see Supplementary Note for the GeneCards summaries).

In conclusion, we conducted a meta-analysis GWAS and GWGAS for intelligence, including 13 cohorts and 78,308 individuals. We confirmed three loci and 12 genes, and identified 15 novel genomic loci and 40 novel genes for intelligence. Pathway analysis demonstrated the involvement of genes regulating cell development. We showed genetic overlap with several neuropsychiatric and metabolic disorders. These findings provide starting points for understanding the molecular neurobiological mechanisms underlying intelligence, one of the most investigated traits in humans.

Online Methods

Discovery sample

The current study was based on 78,308 individuals. The origin of the samples is as follows:

1. UK Biobank web-based measure (UKB-wb; N=17,862), GWAS results have not yet been published previously, raw genotypic data is available for the present study.

2. UK Biobank touchscreen measure (UKB-ts; N=36,257, non-overlapping with UKB-wb) has been published before[6], raw genotypic data is available for the present study.

3. CHIC consortium[5] (N=12,441) has been published before, meta-analysis summary statistics are available for the present study.

4. Five additional cohorts (N=11,748), of which 69 SNP associations with IQ have previously been published as part of a lookup effort[7], but full GWAS results have not been published previously. Per cohort full GWAS summary statistics are available for the present study.

We describe these datasets in more detail below.

UK Biobank samples (UKB-wb, UKB-ts)

We used the data provided by the UK Biobank Study[35] resource (see URLs), which is a major national health resource including >500,000 participants. All participants provided written informed consent; the UK Biobank received ethical approval from the National Research Ethics Service Committee North West–Haydock (reference 11/NW/0382), and all study procedures were performed in accordance with the World Medical Association Declaration of Helsinki ethical principles for medical research. The current study was conducted under the UK Biobank application number 16406.

The study design of the UK Biobank has been described in detail elsewhere[35,36]. Briefly, invitation letters were sent out in 2006–2010 to ~9.2 million individuals including all people aged 40–69 years who were registered with the National Health Service and living up to ~25 miles from one of the 22 study assessment centers. A total of 503,325 participants were subsequently recruited into the study[35]. Apart from registry based phenotypic information, extensive self-reported baseline data have been collected by questionnaire, in addition to anthropometric assessments and DNA collection. For the present study we used imputed data obtained from UK Biobank (May 2015 release) including ~73 million genetic variants in 152,249 individuals. Details on the data are provided elsewhere (see URLs). In summary, the first ~50,000 samples were genotyped on the UK BiLEVE Axiom array, and the remaining ~100,000 samples were genotyped on the UK Biobank Axiom array. After standard quality control of the SNPs and samples, which was centrally performed by UK Biobank, the dataset comprised 641,018 autosomal SNPs in 152,256 samples for phasing and imputation. Imputation was performed with a reference panel that included the UK10K haplotype panel and the 1000 Genomes Project Phase 3 reference panel.

We used two fluid intelligence phenotypes from the Biobank data set. These are based on questionnaires that were taken either in the assessment center at the initial intake (‘touchscreen’, field 20016) or at a later moment at home (‘web-based’, field 20191). The measures indicate the number of correct answers out of 13 fluid intelligence questions. The data distribution roughly approximates a normal distribution.

For the analyses in our study, we only included individuals of Caucasian descent. After removal of related individuals, discordant sex, withdrawn consent, and missing phenotype

data, 36,257 individuals remained for analysis for the fluid intelligence touchscreen measure and 28,846 for the web-based version. As 10,984 individuals had taken both the touchscreen and the web-based test, we only included the data from the touchscreen test for these individuals. This resulted in 54,119 individuals with a score on either the fluid intelligence web-based (UKB-wb) or touchscreen (UKB-ts) version (Supplementary Table 1). At the time of taking the test, participants' ages ranged between 40 and 78. Half of the participants were between 40 and 60 years old, 44% between 60 and 70 and 6% were older than 70. The mean age was 58.98 with a standard deviation of 8.19.

Summary statistics from CHIC consortium

We downloaded the publicly available combined GWAS results from the meta-analyses as reported by CHIC[5] (see URLs). Details on the included cohorts and performed analyses are reported in the original publication[5]. Briefly, CHIC includes 6 cohorts totaling 12,441 individuals: the Avon Longitudinal Study of Parents and Children (ALSPAC, N = 5,517), the Lothian Birth Cohorts of 1921 and 1936 (LBC1921, N = 464; LBC1936, N = 947), the Brisbane Adolescent Twin Study subsample of Queensland Institute of Medical Research (QIMR, N = 1,752), the Western Australian Pregnancy Cohort Study (Raine, N = 936), and the Twins Early Development Study (TEDS, N = 2,825). All individuals are children aged between 6–18 years. Within each cohort the cognitive performance measure was adjusted for sex and age and principal components were included to adjust for population stratification. See also Supplementary Table 1.

Full GWAS data from additional cohorts

We used the same additional (non-CHIC) cohorts as described in detail in ref.[7], which included 11,748 individuals from 5 cohorts. In ref.[7], results were only reported for 69 SNPs, as these served as a secondary analysis for a look-up effort. In the current study we use the full genome-wide results from these cohorts. GWAS were conducted in 2013 and summary statistics were obtained from the PIs of the 5 cohorts. The quality control protocol entailed excluding SNPs with MAF < 0.01, imputation quality score < 0.4, Hardy-Weinberg *P*-value < 10^{-6} and call rate < 0.95[7]. The five cohorts included the Erasmus Rucphen Family Study (ERF, N = 1,076), the Generation R Study (GenR, N = 3,701), the Harvard/Union Study (HU, N = 389), the Minnesota Center for Twin and Family Research Study (MCTFR, N = 3,367) and the Swedish Twin Registry Study (STR, N = 3,215). Detailed descriptions of these cohorts are provided in ref.[7], and summarized in Supplementary Table 1. Within each cohort the cognitive performance measure was adjusted for sex and age and principal components were included to adjust for population stratification.

SNP analysis in UK Biobank sample

Association tests were performed in SNPTEST[37] (see URLs), using linear regression. Both phenotypes were corrected for a number of covariates, including age, sex and a minimum of five genetically determined principal components, depending on how many were associated with the phenotype (i.e. 5 for the web-based test and 15 for the touchscreen version, tested by linear regression). Additionally we included the Townsend deprivation index as a covariate, which is based on postal code and measures material deprivation. The touchscreen version of the phenotype was also corrected for assessment center and genotyping array.

SNPs with imputation quality < 0.8 and MAF < 0.001 (based on all Caucasians present in the total sample) were excluded after the association analysis, resulting in 12,573,858 and 12,595,966 SNPs for the touchscreen and web-based test respectively.

Gene analysis

The SNP based *P*-values from the meta-analysis were used as input for the gene-based analysis. We used all 19,427 protein-coding genes from the NCBI 37.3 gene definitions as basis for a genome-wide gene association analysis (GWGAS) in MAGMA (see URLs). After SNP annotation there were 18,338 genes that were covered by at least one SNP. Gene-association tests were performed taking LD between SNPs into account. We applied a stringent Bonferroni correction to account for multiple testing, setting the genome-wide threshold for significance at 2.73×10^{-6}.

Pathway analysis

We used MAGMA to test for association of predefined gene-sets with intelligence. A total of 6166 Gene Ontology and 674 Reactome gene-sets were obtained (see URLs). We computed competitive *P*-values, which are less likely to be below the threshold of significance compared to self-contained *P*-values. Competitive *P*-values are the outcomes of the test that the combined effect of genes in a gene-set is significantly larger than the combined effect of all other genes, whereas self-contained *P*-values are informative when testing against the null hypothesis of no association. Self-contained *P*-values are not interpreted and not reported by us. Competitive *P*-values were corrected for multiple testing using MAGMA's built in empirical multiple testing correction with 10,000 permutations.

Meta-analysis

Meta-analysis of the results of the 13 cohorts was performed in METAL[11] (see URLs). We did not include SNPs that were not present in the UK Biobank sample. The analysis was based on *P*-values, taking sample size and direction of effect into account using the samplesize scheme.

Genetic correlations

Genetic correlations (r_g) were calculated between intelligence and 32 other traits for which summary statistics from GWAS were publicly available, using LD Score regression (see URLs). This method corrects for sample overlap, by estimating the intercept of the bivariate regression. A conservative Bonferroni-corrected threshold of 1.56×10^{-3} was used to determine significant correlations.

Functional annotation

We identified all SNPs that had an r^2 of 0.1 or higher with the 18 independent lead SNPs and were included in the METAL output. We used the 1000G phase 3 reference panel to calculate r^2. We further filtered on SNPs with a *P*-value < 0.05. In addition, we only annotated SNPs with MAF > 0.01.

Positional annotations for all lead SNPs and SNPs in LD with the lead SNPs were obtained by performing ANNOVAR gene-based annotation using refSeq genes. In addition, CADD

scores[38], and RegulomeDB[15] scores were annotated to SNPs by matching chromosome, position, reference and alternative alleles. For each SNP eQTLs were extracted from GTEx (44 tissue types)[39], Blood eQTL browser[40] and BIOS gene-level eQTLs[41]. The eQTLs obtained from GTEx were filtered on gene *P*-value < 0.05 and eQTLs obtained from the other two databases were filtered on FDR < 0.05. The FDR values were provided by GTEx, BIOS and Blood eQTL browser. For GTEx eQTLs, there is one FDR value available per gene-tissue pair. As such, the FDR is identical for all eQTLs belonging to the same gene-tissue pair. For BIOS and Blood eQTL browser, an FDR value was computed per SNP.

To test whether the SNPs were functionally active by means of histone modifications, we obtained epigenetic data from the NIH Roadmap Epigenomics Mapping Consortium[42] and ENCODE[43]. For every 200bp of the genome a 15-core chromatin state was predicted by a Hidden Markov Model based on 5 histone marks (i.e. H3K4me3, H3K4me1, H3K27me3, H3K9me3, and H3K36me3) for 127 tissue/cell types[44]. We annotated chromatin states (15 states in total) to SNPs by matching chromosome and position for every tissue/cell type. We computed the minimum state (1: the most active state) and the consensus state (majority of states) across 127 tissue/cell types for each SNP.

Chromatin states were also determined for the 52 genes (47 from the gene-based test + 5 additional genes implicated by single SNP GWAS). For each gene and tissue, the chromatin state was obtained per 200 bp interval in the gene. We then annotated the genes by means of a consensus decision when multiple states were present for a single gene; i.e. the state of the gene was defined as the modus of all states present in the gene.

Tissue expression of genes

RNA sequencing data of 1,641 tissue samples with 45 unique tissue labels was derived from the GTEx consortium[39]. This set includes 313 brain samples over 13 unique brain regions (see Supplementary Table 18 for sample size per tissue). Of the 52 genes implicated by either the GWAS or the GWGWAS, 44 were included in the GTEx data. Normalization of the data was performed as described previously[45]. Briefly, genes with RPKM (Reads Per Kilobase Million) value smaller than 0.1 in at least 80% of the samples were removed. The remaining genes were $\log_2$ transformed (after using a pseudocount of 1), and finally a zero-mean normalization was applied.

Proxy-replication in educational attainment

For the replication analysis we used a subset of the data from ref. 21. In particular, we excluded the Erasmus Rucphen Family, the Minnesota Center for Twin and Family Research Study, the Swedish Twin Registry Study, the 23andMe data and all individuals from UK Biobank, to make sure there was no sample overlap with our IQ dataset. Genetic correlation between intelligence and EA in this non-overlapping subsample was r_g=0.73, SE=0.03, P=1.4×10^{-163}. The replication analysis was based on the phenotype *EduYears*, which measures the number of years of schooling completed. A total of 306 out of our 336 top SNPs (and 16 out of 18 independent lead SNPs) was available in the educational attainment sample. We performed a sign concordance analysis for the 16 independent lead SNPs, using the exact binomial test. For each independent signal we determined whether either the lead

SNP had a *P*-value smaller than 0.05/16 in the educational attainment analysis, or another (correlated) top SNP in the same locus if this was not the case. All 47 genes implicated in the GWGAS for intelligence were available for look-up in the EA sample. For each gene we determined whether it had a *P*-value smaller than 0.05/47 in the EA analysis.

Polygenic Risk Score analysis

We used LDpred[16] to calculate the variance explained in intelligence in independent samples by a polygenic risk score based on our discovery analysis, as well as based on two previous GWAS studies for intelligence[5,6]. LDpred adjusts GWAS summary statistics for the effects of linkage disequilibrium (LD) by using an approximate Gibbs sampler that calculates posterior means of effects, conditional on LD information, when calculating polygenic risk scores. We used varying priors for the fraction of SNPs with non-zero effects (prior: 0.01, 0.05, 0.1, 0.5, 1, and an infinitesimal prior). Independent datasets available for PRS analyses are described in the Supplementary Note.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by The Netherlands Organization for Scientific Research (NWO VICI 453-14-005). The analyses were carried out on the Genetic Cluster Computer, which is financed by the Netherlands Scientific Organization (NWO: 480-05-003), by the VU University, Amsterdam, The Netherlands, and by the Dutch Brain Foundation, and is hosted by the Dutch National Computing and Networking Services SurfSARA. This research has been conducted using the UK Biobank Resource under application number 16406. We thank the participants and researchers who collected and contributed to the data.

References

1. Deary IJ. Intelligence. Annu. Rev. Psychol. 2012; 63:453–82. [PubMed: 21943169]
2. Polderman TJC, et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nat. Genet. 2015; 47:702–709. [PubMed: 25985137]
3. Chabris CF, et al. Most reported genetic associations with general intelligence are probably false positives. Psychol. Sci. 2012; 23:1314–23. [PubMed: 23012269]
4. Davies G, et al. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. Mol. Psychiatry. 2011; 16:996–1005. [PubMed: 21826061]
5. Benyamin B, et al. Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. Mol. Psychiatry. 2014; 19:253–8. [PubMed: 23358156]
6. Davies G, et al. Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112 151). Mol. Psychiatry. 2016; 21:758–67. [PubMed: 27046643]
7. Rietveld CA, et al. Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. Proc. Natl. Acad. Sci. 2014; 111:13790–13794. [PubMed: 25201988]
8. Deary IJ, Penke L, Johnson W. The neuroscience of human intelligence differences. Nat Rev Neurosci. 2010; 11:201–211. [PubMed: 20145623]
9. Johnson W, Bouchard TJ, Krueger RF, McGue M, Gottesman II. Just one g: Consistent results from three test batteries. Intelligence. 2004; 32:95–107.
10. Ree MJ, Earles JA. The stability of g across different methods of estimation. Intelligence. 1991; 15:271–278.
11. Willer CJ, Li Y, Abecasis GR. METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010; 26:2190–2191. [PubMed: 20616382]

12. Bulik-Sullivan BK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015:291–295. **advance on**. [PubMed: 25642630]
13. Yang J, et al. Genomic inflation factors under polygenic inheritance. Eur. J. Hum. Genet. 2011; 19:807–812. [PubMed: 21407268]
14. Davies G, et al. Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53949). Mol. Psychiatry. 2015; 20:183–92. [PubMed: 25644384]
15. Boyle AP, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012; 22:1790–1797. [PubMed: 22955989]
16. Vilhjalmsson B, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am. J. Hum. Genet. 2015; 97:576–592. [PubMed: 26430803]
17. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Comput. Biol. 2015; 11
18. Ashburner M, et al. Gene Ontology: tool for the unification of biology. Nat. Genet. 2000; 25:25–29. [PubMed: 10802651]
19. Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2016; 44:D481–D487. [PubMed: 26656494]
20. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 2005; 102:15545–50. [PubMed: 16199517]
21. Okbay A, et al. Genome-wide association study identifies 74 loci associated with educational attainment. Nature. 2016; 533:539–542. [PubMed: 27225129]
22. Newman AB, Murabito JM. The epidemiology of longevity and exceptional survival. Epidemiol. Rev. 2013; 35:181–197. [PubMed: 23372024]
23. Willcox BJ, et al. FOXO3A genotype is strongly associated with human longevity. Proc. Natl. Acad. Sci. U. S. A. 2008; 105:13987–13992. [PubMed: 18765803]
24. Flachsbart F, et al. Association of FOXO3A variation with human longevity confirmed in German centenarians. Proc. Natl. Acad. Sci. U. S. A. 2009; 106:2700–2705. [PubMed: 19196970]
25. Behrens P, Brinkmann U, Wellmann A. CSE1L/CAS: Its role in proliferation and apoptosis. Apoptosis. 2003; 8:39–44. [PubMed: 12510150]
26. Velez Edwards DR, et al. Gene-environment interactions and obesity traits among postmenopausal African-American and Hispanic women in the Women's Health Initiative SHARe Study. Hum. Genet. 2013; 132:323–336. [PubMed: 23192594]
27. Speliotes EK, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010; 42:937–948. [PubMed: 20935630]
28. Willer CJ, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet. 2009; 41:25–34. [PubMed: 19079261]
29. Locke AE, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015; 518:197–206. [PubMed: 25673413]
30. Ripke S, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511:421–427. [PubMed: 25056061]
31. Comuzzie AG, et al. Novel Genetic Loci Identified for the Pathophysiology of Childhood Obesity in the Hispanic Population. PLoS One. 2012; 7
32. Berndt SI, et al. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat. Genet. 2013; 45:501–12. [PubMed: 23563607]
33. Wheeler E, et al. Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nat. Genet. 2013; 45:513–7. [PubMed: 23563609]
34. Pruim RJ, et al. LocusZoom: Regional visualization of genome-wide association scan results. Bioinformatics. 2011; 27:2336–2337.
35. Sudlow C, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 2015; 12

36. Allen NE, Sudlow C, Peakman T, Collins R. UK Biobank Data: Come and Get It. Sci. Transl. Med. 2014; 6:224ed4–224ed4.

37. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 2007; 39:906–13. [PubMed: 17572673]

38. Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 2014; 46:310–315. [PubMed: 24487276]

39. GTEx Consortium, Gte. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348:648–60. [PubMed: 25954001]

40. Westra H-J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. 2013; 45:1238–43. [PubMed: 24013639]

41. Bonder MJ, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2016 advance on.

42. Consortium RE, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]

43. Bernstein BE, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

44. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature Methods. 2012; 9:215–6. [PubMed: 22373907]

45. Taskesen E, Reinders MJT. 2D representation of transcriptomes by t-SNE exposes relatedness between human tissues. PLoS One. 2016; 11

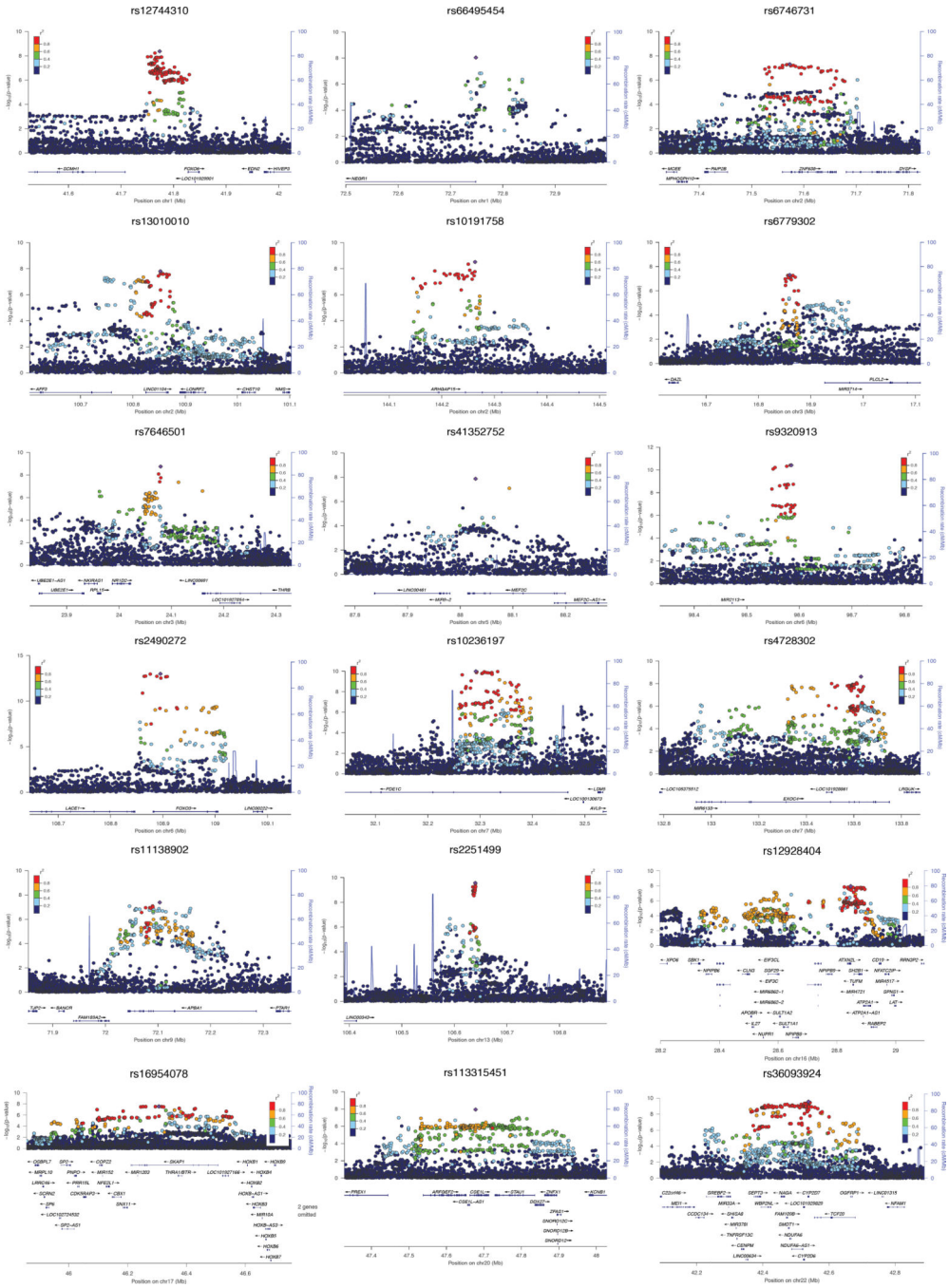

Fig. 1. Regional association and linkage disequilibrium plots for 18 genome-wide significant loci

The *y*-axis represents the negative logarithm (base 10) of the SNP *P*-value and the *x*-axis the position on the chromosome, with the name and location of genes in the UCSC Genome Browser in the bottom panel. The SNP with the lowest *P*-value in the region is marked by a purple diamond. The colors of the other SNPs indicate the r^2 of these SNPs with the lead SNP. Plots are generated with LocusZoom[34].

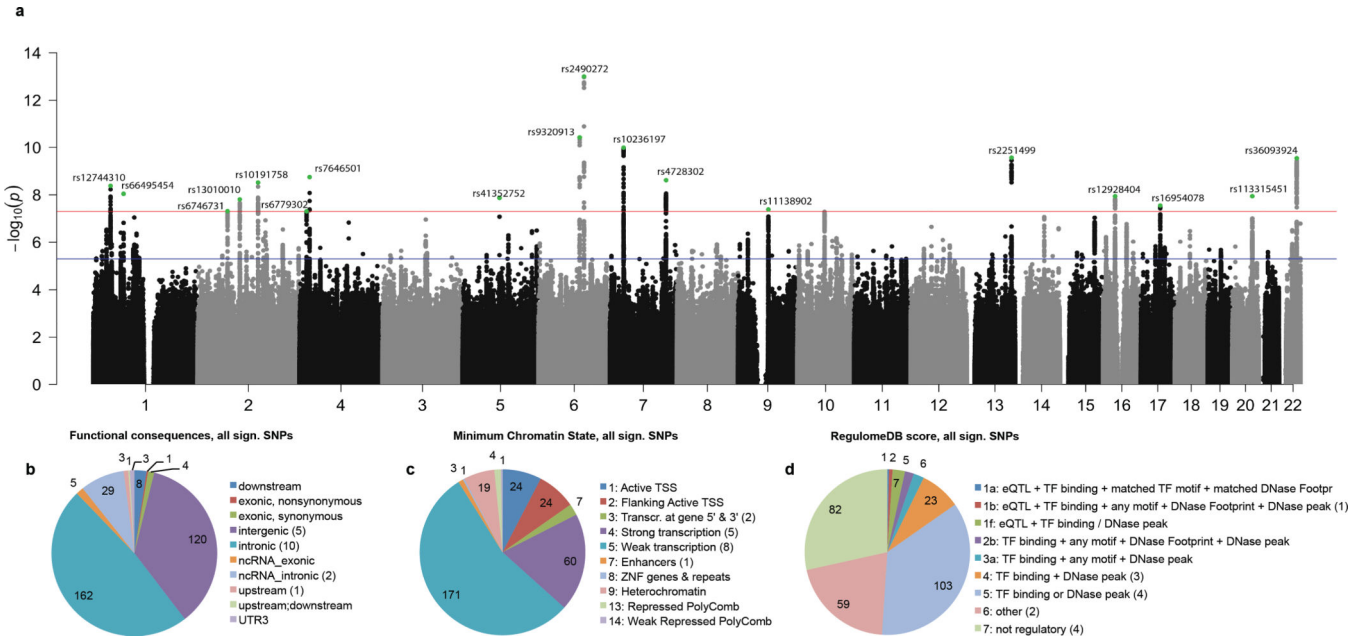

Fig. 2. Results of SNP-based meta-analysis for intelligence based on 78,308 individuals

Association results from the GWAS meta-analysis pertaining to individuals of European descent. (**a**) Negative $\log_{10}$-transformed *P*-values for each SNP (*y*-axis) are plotted by chromosomal position (*x*-axis). The red and blue lines represent the thresholds for genome-wide statistical significant associations ($P=5\times10^{-8}$) and suggestive associations ($P=1\times10^{-5}$) respectively. Green dots represent the independent hits. (**b**) Functional categories for 336 genome-wide significant SNPs. (**c**) The minimum (most active) chromatine state across 127 tissues for 336 genome-wide significant SNPs. (**d**) The Regulome database score for 336 genome-wide significant SNPs. The lower the score the more likely it is that a SNP has a regulatory function. For **b–d** the numbers in brackets in the legends refer to the number of lead SNPs for that category.

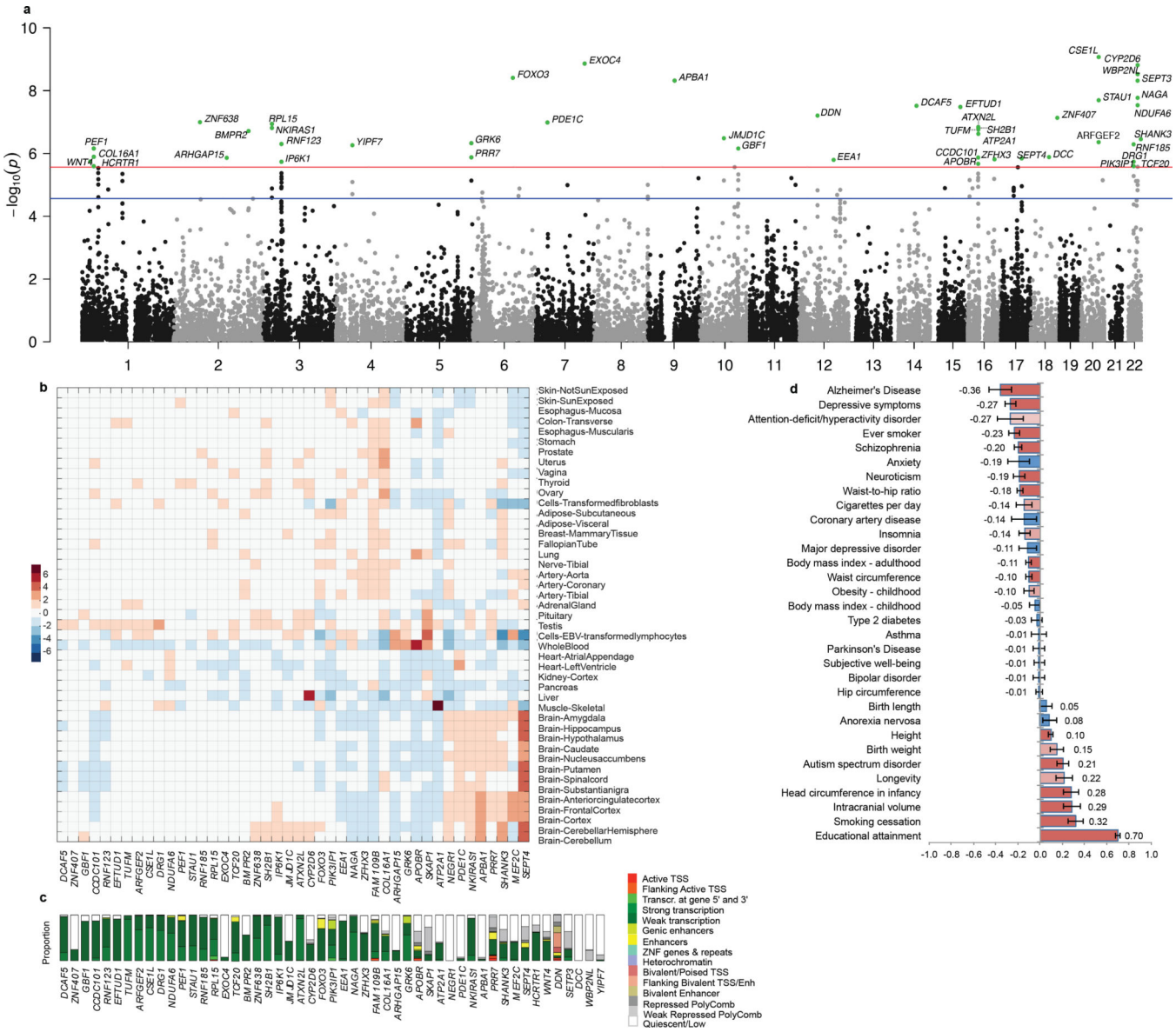

Fig. 3. Gene-based genome wide analysis for intelligence and genetic overlap with other traits

(**a**). Negative $\log_{10}$-transformed *P*-values for each gene are plotted. Green dots represent significantly associated genes from GWGAS. The threshold for gene-wide statistical significant associations was set at the Bonferroni threshold of $P=2.73\times10^{-6}$, the suggestive threshold was set at $P=2.73\times10^{-5}$. (**b**) Heatmap of gene-expression levels of genes for intelligence in 45 tissue types (see Supplementary Table 18 for N per tissue). A value above zero (red) depicts a relatively high expression level with respect to the mean expression level of the gene over all tissues, whereas a value below zero (blue) depicts a relatively low expression level. (**c**) Epigenetic states of genes. The bars denote the proportions of epigenetic states across 127 tissue types. (**d**) Genetic correlations between intelligence and 32 health-related outcomes. Error bars show 95% confidence intervals for estimates of r_g. Red bars represent the traits that showed a significant genetic correlation after correction for multiple testing ($P<1.56\times10^{-3}$), pink bars the traits that showed a nominal significant

correlation (P<0.05), and blue bars the traits that did not show a genetic correlation significantly different from zero. Note: as Alzheimer's disease is an age-related disorder we calculated the r_g with this phenotype across three age groups and found no difference in r_g's (Supplementary Note).

Table 1

Genomic loci and lead SNPs associated with intelligence in the meta-analysis based on N=78,308.

rsID	Annotation	Locus[a]	Ref	Alt	RefF	*Z*	*P*-value	Direction[b]	N	N_{GWS}
rs2490272	FOXO3 intronic	6q21	t	c	0.63	7.44	9.96E-14	++++-+++	78307	28
rs9320913	intergenic	6q16.1	a	c	0.48	6.61	3.79E-11	++++-+++	78307	13
rs10236197	PDE1C intronic	7p14.3	t	c	0.63	6.46	1.03E-10	+++++-++	78286	35
rs2251499	intergenic	13q33.2	t	c	0.26	6.31	2.74E-10	++++++++	78307	22
rs36093924	CYP2D7 ncRNA_intr	22q13.2	t	c	0.46	−6.31	2.87E-10	?--?????	54119	100
rs7646501	intergenic	3p24.2	a	g	0.74	6.02	1.79E-09	?++-++++	65866	5
rs4728302	EXOC4 intronic	7q33	t	c	0.60	−5.97	2.42E-09	---+--+-	78307	45
rs10191758	ARHGAP15 intronic	2q22.3	a	g	0.61	−5.93	3.06E-09	?--?????	54119	17
rs12744310	intergenic	1p34.2	t	c	0.22	−5.88	4.20E-09	?-------	65866	28
rs66495454	NEGR1 upstream	1p31.1	g	gtcct	0.62	−5.75	9.08E-09	?--?????	54119	1
rs113315451	CSE1L intronic	20q13.13	a	attat	0.43	5.71	1.15E-08	?++?????	54119	1
rs12928404	ATXN2L intronic	16p11.2	t	c	0.59	5.71	1.15E-08	++++++++	78307	19
rs41352752	MEF2C intronic	5q14.3	t	c	0.97	−5.68	1.35E-08	?--?????	54119	1
rs13010010	LINC01104 ncRNA_intr	2q11.2	t	c	0.38	5.65	1.56E-08	++++++++	78308	11
rs16954078	SKAP1 intronic	17q21.32	a	t	0.21	−5.55	2.84E-08	?----+--	65866	7
rs11138902	APBA1 intronic	9q21.11	a	g	0.54	5.49	4.12E-08	+++++-++	78307	1
rs6746731	ZNF638 intronic	2p13.2	t	g	0.43	−5.46	4.88E-08	-----+--	78307	1
rs6779302	intergenic	3p24.3	t	g	0.37	−5.45	4.99E-08	?--?????	54119	1

SNP *P*-values and *Z*-scores were computed in METAL by a weighted *Z*-score method. A total of 336 SNPs reached genome-wide significance ($P<5\times10^{-8}$); 18 independent signals were obtained by LD-based clumping, using an r^2 threshold of 0.1 and a window of 300 kb.

Ref, effect or reference allele; Alt, non-effect or alternative allele; RefF, effect allele frequency in UK Biobank, based on individuals of Caucasian ancestry; *Z, Z*-score from METAL; Direction, Direction of the effect in each of the cohorts; N, sample size; N GWS; number of genome-wide significant SNPs in the locus.

[a]Cytogenetic band, build hg19.

[b]Order: CHIC, UKB-wb, UKB-ts, ERF, GENR, HU, MCTFR, STR.