SUPPLEMENTARY INFORMATION

for

Identification of genetic loci shared between major depression and intelligence with mixed directions of effect

by Bahrami *et al*.

The Supplementary Information includes:

Supplementary Note on pages 1-18.

Supplementary References on page 10.

Supplementary Figures 1-10 on pages 11-18

Methods

Participant samples

The GWAS summary statistics on major depression were obtained from the recent publication by Wray and colleagues[1]. The PGC major depression sample consisted of 135,458 cases and 344,901 controls from seven cohorts (PGC29, deCODE, GenScot, GERA, iPSYCH, UK Biobank and 23andMe)[1].

The comparability of the seven cohorts for meta-analysis was evaluated using SNP genotype data and h^2 for each cohort was estimated using LD score (LDSC) regression. The genetic correlations (r_g) attributable to common variants were estimated between the seven cohorts to confirm their comparability and the absence of sample overlaps. The weighted mean r_g was 0.76 (SE=0.03) and the LDSC intercepts were all near zero. The meta-analysis included 9.6 million imputed SNPs. There was no evidence of residual population stratification (LD score regression intercept=1.018, SE=0.009). A leave-one-sample-out strategy for genetic risk scores (GRSs) was used to demonstrate significant differences in case–control GRS distributions of the left-out sample for all. GRS ranked cases higher than controls with probability of 0.57 for any randomly selected case and control and the odds ratio of major depression for those in the tenth versus those in the first GRS decile was 2.4[1].

To do standardized quality control, imputation, and analysis, individual genotype data for three cohorts including PGC29, GERA, and iPSYCH were processed using the PGC ricopili pipeline and other cohorts were processed by the collaborating research teams using comparable procedures. Using the 1000 Genomes Project multi-ancestry reference panel, SNPs and insertion–deletion polymorphisms were imputed[1].

The general intelligence ability GWAS summary statistics were obtained from the publication by Savage and colleagues based on 269,867 individuals drawn from 14 cohorts[2]. The cohorts were assessed using different neuropsychological tests. Fluid intelligence was determined either by Spearman's *g* or by a primary measure of fluid intelligence function that correlated highly with *g*[2,3].

The average genetic correlations (r_g) were calculated across the cohorts to assess their comparability (r_g=0.67) and confirm the absence of sample overlaps. Before the meta-analysis stringent quality control scales were used: imputation quality (INFO/R^2) score<0.6, Hardy-Weinberg equilibrium $p<1\times10^{-6}$, study-specific minor allele frequency (MAF) corresponding to a minor allele count < 100, and mismatch of alleles or allele frequency difference greater than 20% from the Haplotype Reference Consortium genome reference panel[4]. Indels and SNPs that were duplicate, multi-allelic, monomorphic, or ambiguous (A/T or C/G with a MAF>0.4) were also excluded. 9,295,119 SNPs passed quality control and their association p-values in all the cohorts were meta-analyzed using METAL[5].

LD Score Regression

To estimate SNP-based genetic correlation between major depression and INT, we used linkage disequilibrium (LD) score regression[6]. In this case, we used major depression with all seven cohorts (with UKBB) and INT as described in the main text. The analysis was performed using the Python-based package available at (https://github.com/bulik/ldsc). The procedure is described in the documentation of the package (https://github.com/bulik/ldsc/wiki/Heritability-and-Genetic-Correlation).

Since we knew that there is sample overlap between major depression and INT datasets, we didn't constrain the regression intercept. The obtained genetic correlation with all cohorts was significantly below zero (r_g=-0.0282, SE=0.0215).

Fold Enrichment Plot

To visually assess the association enrichment in a primary trait when conditioning on another (conditional) trait, we used fold enrichment plots. Enrichment exists if the degree of upward deflection from the expected null level (horizontal line through 1) depends on the stratum defined by the p-values for association with the condotional trait. Each SNP stratum corresponds to some p-value threshold p_{thresh} for association with trait T (either major depression or INT) (Supplementary Figures 4 A & B).

In order to obtain the fold enrichment plots, we measure the empirical cumulative distribution function (CDF) of the primary trait association p-values for all SNPs. Using the p-values for association with the conditional trait, the CDF of the primary trait association p-values is subsequently estimated for each SNP stratum defined. The fold enrichment is estimated as the ratio $CDF_{stratum}/CDF_{all}$ for each stratum. In the fold enrichment plots, the x-axis shows the nominal $-\log_{10}$(p-value) for the primary trait; the y-axis indicates the fold enrichment. Here we focus on polygenic effects for SNPs not reaching the standard GWAS significance threshold - $\log_{10}(p)<7.30$ (corresponding to $p>5.00\times10^{-8}$).

Conditional/Conjunctional FDR

The following brief description of the conditional/conjunctional false discovery rate (condFDR/conjFDR) approach is based on [7,8].

In empirical Bayesian formulation, the false discovery rate (FDR) can be expressed as follows[9]:

$$FDR(p)= \Pi_0 F_0(p)/F(p),$$

where Π_0 is the a priori fraction of null SNPs, F_0 is the null cumulative distribution function (CDF), and F is the CDF of all SNPs, both null and non-null. Under the null hypothesis, F_0 is the CDF of the uniform distribution on the unit interval [0,1], so $F_0(p) = p$ and the latter formula reduces to

$$FDR(p)= \Pi_0 p/F(p),$$

which was used to determine the FDR for a phenotype conditional on another. The conditional FDR can be expressed as the posterior probability that a given SNP is null for the primary phenotype given that the p-values for both phenotypes are as small as or smaller than the observed p-values. The conditional FDR can be expressed as:

$$condFDR(p_1 p_2)= \Pi_0(p_2) p_1/F(p_1|p_2)$$

where p_1 and p_2 are p-values of association with the first and the second phenotype respectively, $\Pi_0(p_2)$ is the conditional proportion of null SNPs, and $F(p_1|p_2)$ is the conditional CDF for the first phenotype given that p-value for the second phenotype is p_2 or smaller. The conditional FDR for phenotype 1 (pt1) given phenotype 2 (pt2) is defined as $FDR_{pt1|pt2}$. A conservative estimate of $FDR_{pt1|pt2}$ was defined by setting $\Pi_0(p_2)=1$ and replacing $FDR_{pt1|pt2}$ with the empirical conditional CDF.

To identify shared genetic variants associated with both traits and assess potential pleiotropic signals, a conjunctional FDR procedure is used[7,8]. The conjunctional FDR is defined as the posterior probability that a given SNP is null for either phenotype or both phenotypes simultaneously when the p-values for both phenotypes are as small as or smaller than the observed p-values.

Conditional quantile-quantile (Q-Q) plot

Q-Q plots compare a nominal probability distribution against an empirical distribution. In the presence of all null relationships, nominal p-values form a straight line on a Q-Q plot when plotted against the empirical distribution.

We used conditional quantile-quantile (Q-Q) plots[7] of -log10 nominal p-values against -log10 empirical p-values to provide a visual pattern of enrichment in associations [7].

The conditional Q-Q plots compare the association with one trait within SNPs strata determined by the significance of their association with a second trait. The empirical cumulative distribution of nominal p-values was computed in one phenotype for all SNPs and for subsets of SNPs with significance levels in another phenotype below determined thresholds ($p \leq 1$, $p \leq 0.1$, $p \leq 0.01$, $p \leq 0.001$, respectively). Enrichment of associations is expected to exist for a given phenotype if the degree of leftward deflection from the expected null line increases with increasing association significance in the second phenotype[7]. In this study we focused on polygenic effects below the standard GWAS significance threshold ($p > 5 \times 10-8$)[7].

The test to assess the statistical significance of the enrichment observed in the conditional Q-Q plots was performed using the partitioned LD-score regression approach[6]. This analysis showed an increase in the enrichment parameter for major depression given INT ranging from 2.20 for $p_{INT}<0.1$ (Enrichment p-value=5.047e-16) to 3.32 for $p_{INT}<0.01$ (Enrichment p-value=3.354e-17), and 4.00 for $p_{INT}<0.001$ (Enrichment p-value=1.574e-07). The partitioned LD-score also showed an increase in the enrichment parameter for INT given major depression ranging from 2.02 for $p_{major\ depression} <0.1$ (Enrichment p-value=1.135e-18) to 3.26 for $p_{major\ depression} <0.01$ (Enrichment p-value=2.235e-12), and 5.58 for $p_{major\ depression} <0.001$ (Enrichment p-value=1.518e-06).

Functional annotation

In this research we functionally annotated all candidate SNPs in the genomic loci with a condFDR or conjFDR value<0.1 having an $r^2 \geq 0.6$ with one of the independent significant SNPs using FUMA[10] (http://fuma.ctglab.nl). In this study we annotated SNPs using Combined Annotation Dependent Depletion (CADD) scores[11], RegulomeDB scores[12], and chromatin states[13,14]. We also used FUMA to evaluate gene ontology (GO)[15] gene-set enrichment for the genes nearest the identified shared loci with concordant effect directions or opposite effect directions in major depression and INT, respectively.

In FUMA, positional mapping is performed based on ANNOVAR[16] annotations by physical distance between SNPs and genes or based on functional consequences of SNPs on genes. ANNOVAR is an online software tool that uses update-to-date information to functionally annotate genetic variants detected in diverse genomes[16].

CADD is a freely available tool for integrating 63 diverse functional genomic annotations and scoring single nucleotide variants and insertions/deletions[11] according to their potential deleteriousness. The CADD score strongly correlates with both molecular functionality and pathogenicity. A CADD score of 12.37 was suggested as a reasonable threshold for considering a variant to be deleterious[11]. The RegulomeDB score is a categorical score measuring the regulatory functionality of SNPs based on eQTLs and chromatin marks[12]. The RegulomeDB annotates and classifies variants from 1a to 7. The scoring system shows with increasing confidence that a variant lies in functional location and likely results in a functional consequences. The scores have the following meanings: 1a=eQTL + Transcription Factor (TF) binding + matched TF motif + matched DNase Footprint + DNase peak; 1b=eQTL + TF binding + any motif + DNase Footprint + DNase peak; 1c=eQTL + TF binding + matched TF motif +

DNase peak; 1d=eQTL + TF binding + any motif + DNase peak; 1e=eQTL + TF binding + matched TF motif; 1f=eQTL + TF binding / DNase peak; 2a=TF binding + matched TF motif + matched DNase Footprint + DNase peak; 2b=TF binding + any motif + DNase Footprint + DNase peak; 2c=TF binding + matched TF motif + DNase peak; 3a=TF binding + any motif + DNase peak; 3b=TF binding + matched TF motif; 4=TF binding + DNase peak; 5=TF binding or DNase peak; 6=other; 7=Not available[12].

The chromatin state shows a genomic region's accessibility for every 200bp with 15 categorical states predicted by ChromHMM based on five histone modification marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3) for all 127 epigenomes[13].

A lower score shows higher accessibility in the chromatin state and refers to a more open state. The 15-core chromatin states as suggested by Roadmap are as follows: 1=Active Transcription Start Site (TSS); 2=Flanking Active TSS; 3=Transcription at gene 5' and 3'; 4=Strong transcription; 5=Weak Transcription; 6=Genic enhancers; 7=Enhancers; 8=Zinc finger genes & repeats; 9=Heterochromatic; 10=Bivalent/Poised TSS; 11=Flanking Bivalent/Poised TSS/Enh; 12=Bivalent Enhancer; 13=Repressed PolyComb; 14=Weak Repressed PolyComb; 15=Quiescent/Low[13].

For the gene-set analyses, we used the genes nearest of the identified shared loci with concordant effect directions or opposite effect directions in major depression and INT respectively (Supplementary Tables 8 and 9).

With the help of FUMA's GENE2FUNC procedure, we also assessed the GTEx and Braineac eQTLs resources[17,18] to investigated the effect of the two genetic variants (rs301807 and rs9620039) with lowest RegulomeDB scores. The GTEx database is a publicly aaccessible database built upon whole-genome sequencing[17]. The Braineac is an online web service

providing access to data from the UK Brain Expression Consortium to aid the investigation of genes and SNPs associated with neurological disorders. Braineac provides information on cis-eQTLs for 10 brain regions: cerebellar cortex, frontal cortex, hippocampus, medulla (specifically inferior olivary nucleus), occipital cortex (specifically primary visual cortex), putamen, substantia nigra, thalamus, temporal cortex, and intralobular white matter[18].

We determined human tissue expression of the genes nearest to the identified shared loci with concordant effect directions and opposite effect direction in major depression and INT, respectively (Supplementary Figures 6A and B). The generated gene heatmap figures were based on the Genotype Tissue Expression (GTEx v7) resource[17], and the figure was generated using the GENE2FUNC procedure of FUMA[10]. We also tested the tissue specificity using the differentially expressed genes (DEG) sets defined for the genes nearest of the identified shared loci with corcordant effect directions (Supplementary Figures 7 and 8) and opposite effect directions (Supplementary Figures 9 and 10) in major depression and INT, respectively. DEG sets were pre-determined by performing a two-sided t-test for each tissue against all remaining tissues for each of the expression data sets. Genes with a Bonferroni corrected p-value < 0.05 and absolute log fold change ≥ 0.58 were deemed differentially expressed. The up-regulated DEG and down-regulated DEG were also pre-determined by taking the signs of the t-statistics into account.

We also used the ConsensusPathDB resource[19] to perform pathway over-representation analyses for the genes nearest of the identified shared loci for the loci with corcordant effect directions and opposite effect directions in major depression and INT, respectively (Supplementary Table 10). ConsensusPathDB integrates interaction networks including binary and complex protein-

protein, genetic, metabolic, signaling, gene regulatory and drug-target interactions, as well as biochemical pathways[19].

References:

1 Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* **50**, 668-681, doi:10.1038/s41588-018-0090-3 (2018).

2 Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet* **50**, 912-919, doi:10.1038/s41588-018-0152-6 (2018).

3 Deary, I. J., Penke, L. & Johnson, W. The neuroscience of human intelligence differences. *Nat Rev Neurosci* **11**, 201-211, doi:10.1038/nrn2793 (2010).

4 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-1283, doi:10.1038/ng.3643 (2016).

5 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191, doi:10.1093/bioinformatics/btq340 (2010).

6 Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-1241, doi:10.1038/ng.3406 (2015).

7 Andreassen, O. A. *et al.* Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am J Hum Genet* **92**, 197-209, doi:10.1016/j.ajhg.2013.01.001 (2013).

8 Andreassen, O. A. *et al.* Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genet* **9**, e1003455, doi:10.1371/journal.pgen.1003455 (2013).

9 Efron, B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. (Cambridge University Press, 2010).

10 Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826, doi:10.1038/s41467-017-01261-5 (2017).

11 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).

12 Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-1797, doi:10.1101/gr.137323.112 (2012).

13 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

14 Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-487, doi:10.1038/ng.3538 (2016).

15 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).

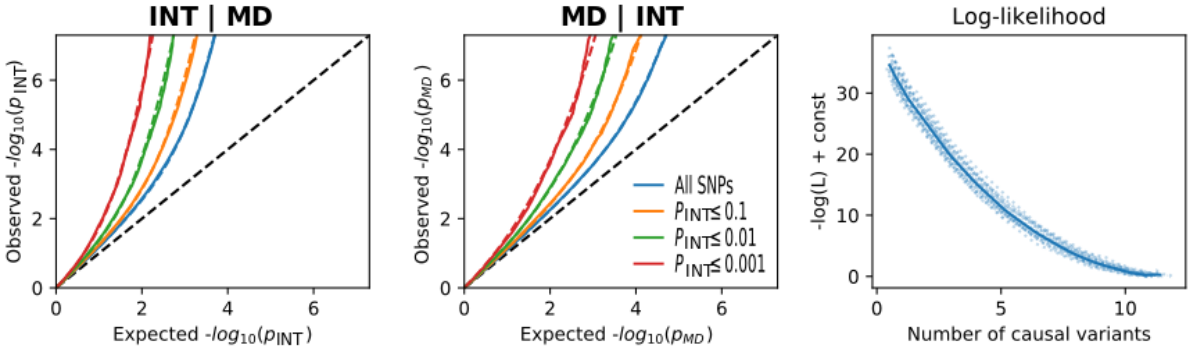
16 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164, doi:10.1093/nar/gkq603 (2010).

17 Consortium, G. T. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213, doi:10.1038/nature24277 (2017).

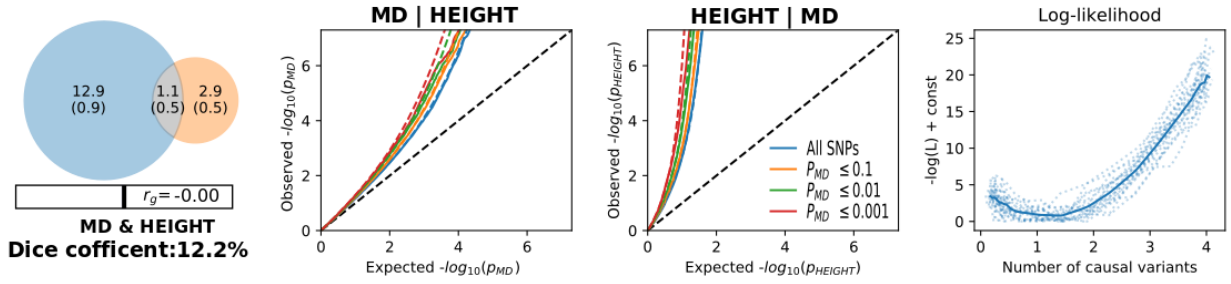
18 Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat Neurosci* **17**, 1418-1428, doi:10.1038/nn.3801 (2014).

19 Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res* **41**, D793-800, doi:10.1093/nar/gks1055 (2013).

Figures

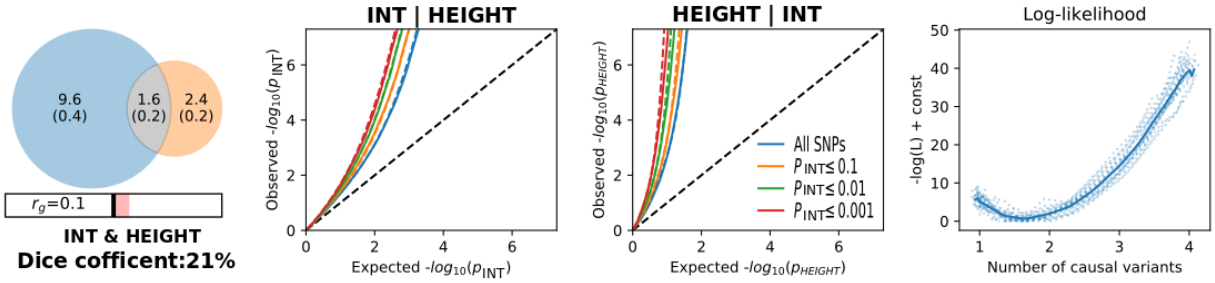


Supplementary Figure 1. Conditional Q-Q plots and Likelihood cost. **Conditional QQ plots** of observed versus expected -log10 p-values in the primary trait as a function of significance of association with the secondary trait at the level of $p \leq 0.1$ (orange lines), $p \leq 0.01$ (green lines) and $p \leq 0.001$ (red lines). Blue lines indicate all SNPs. Dotted lines indicate model predictions for each stratum. Black dotted line is the expected Q-Q plot under the null hypothesis (no SNPs associated with the phenotype). **Likelihood cost**: Log-likelihood of the bivariate fit as a function of π parameter. The remaining parameters of the model were constrained to their fitted values.

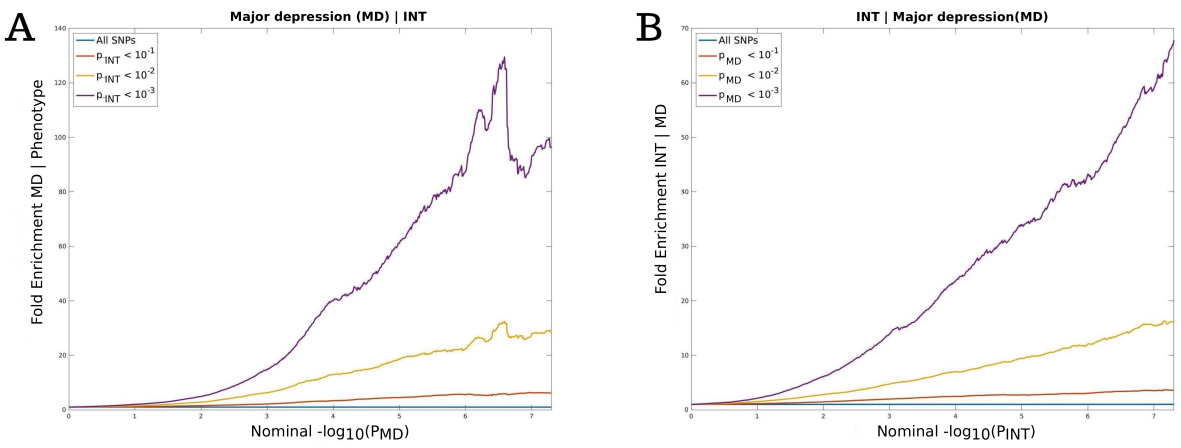


Supplementary Figure 2. Venn Diagram, conditional Q-Q plots and Likelihood cost. **Venn diagram** of unique and shared polygenic components at the causal level, showing polygenic overlap (gray) between major depression (MD) (blue) and height (orange). The numbers in the Venn diagram indicate the estimated quantity of causal variants (in thousands) per component, explaining 90% of SNP heritability in each phenotype, followed by standard error. The

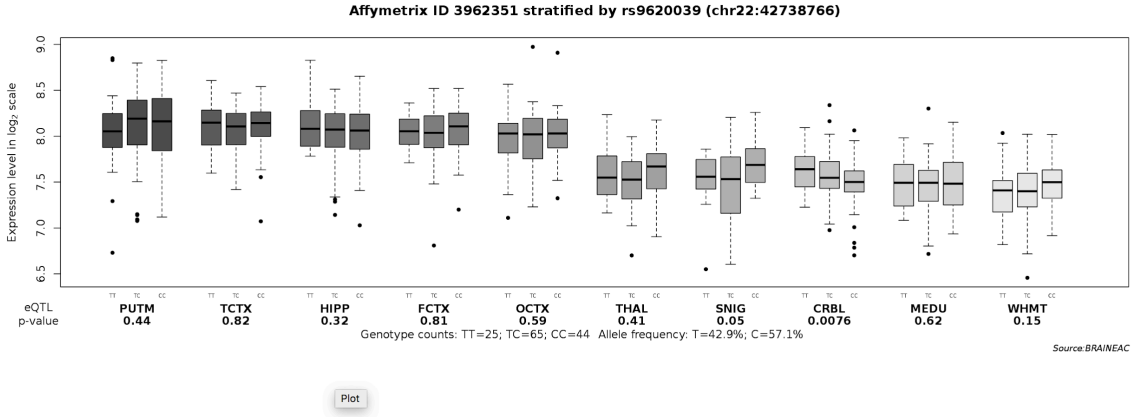
size of the circles reflects the degree of polygenicity. **Conditional QQ plots** of observed versus expected -log10 p-values in the primary trait as a function of significance of association with the secondary trait at the level of $p \leq 0.1$ (orange lines), $p \leq 0.01$ (green lines) and $p \leq 0.001$ (red lines). Blue lines indicate all SNPs. Dotted lines indicate model predictions for each stratum. Black dotted line is the expected Q-Q plot under the null hypothesis (no SNPs associated with the phenotype. **Likelihood cost**: Log-likelihood of the bivariate fit as a function of π parameter. The remaining parameters of the model were constrained to their fitted values.



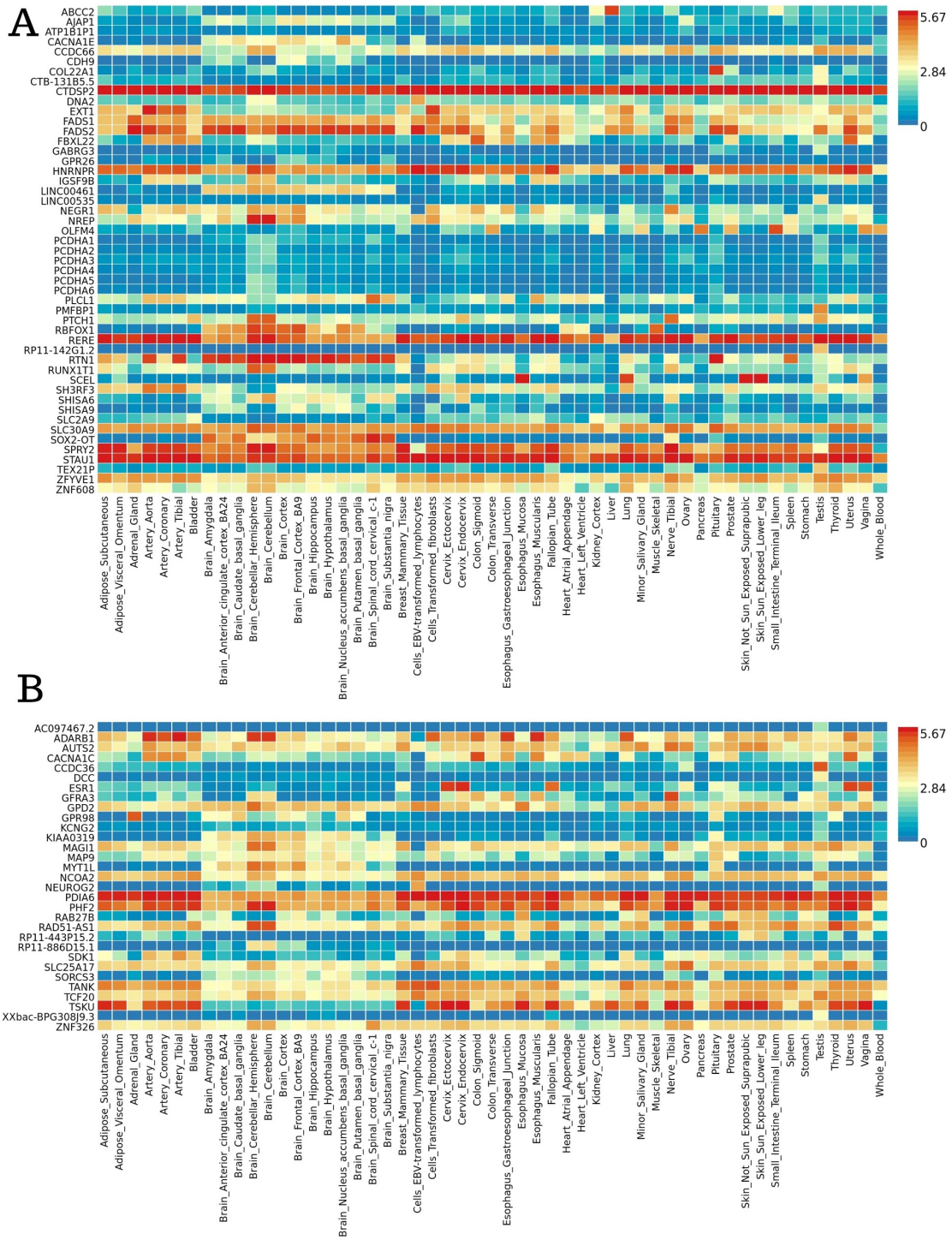
Supplementary Figure 3. Venn Diagram, conditional Q-Q plots and Likelihood cost. **Venn diagram** of unique and shared polygenic components at the causal level, showing polygenic overlap (gray) intelligence (INT) (blue) and height (orange). The numbers in the Venn diagram indicate the estimated quantity of causal variants (in thousands) per component, explaining 90% of SNP heritability in each phenotype, followed by standard error. The size of the circles reflects the degree of polygenicity. **Conditional QQ plots** of observed versus expected -log10 p-values in the primary trait as a function of significance of association with the secondary trait at the level of $p \leq 0.1$ (orange lines), $p \leq 0.01$ (green lines) and $p \leq 0.001$ (red lines). Blue lines indicate all SNPs. Dotted lines indicate model predictions for each stratum. Black dotted line is the expected Q-Q plot under the null hypothesis (no SNPs associated with the phenotype. **Likelihood cost**: Log-likelihood of the bivariate fit as a function of π parameter. The remaining parameters of the model were constrained to their fitted values.



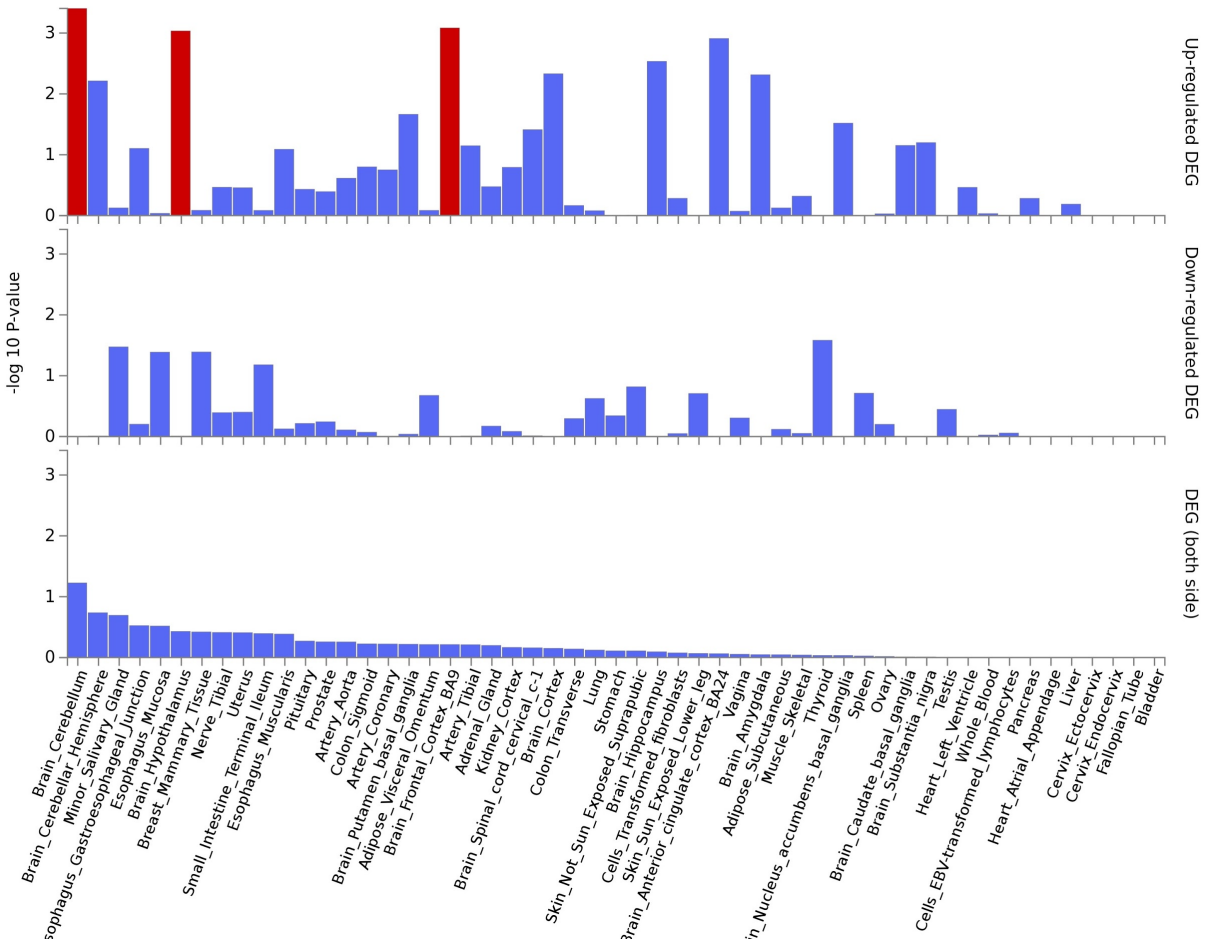
Supplementary Figure 4. Fold enrichment of association for major depression and INT. Note: Fold enrichment plots of the observed $-\log_{10}(p)$ below the standard genome-wide association study (GWAS) threshold (corresponding to $p>5\times10^{-8}$) in the primary trait stratified based on the association with the conditional trait. A sequence of four nested strata is presented: all single nucleotide polymorphisms (SNPs; i.e. p-values of the conditional trait≤1), $p_{conditional\ trait}<10^{-1}$, $p_{conditional\ trait}<10^{-2}$ and $p_{conditional\ trait}<10^{-3}$. Successive upward elevation compared to all SNPs demonstrates polygenic enrichment both cases. **A**) major depression conditioned on INT. **B**) INT conditioned on major depression.



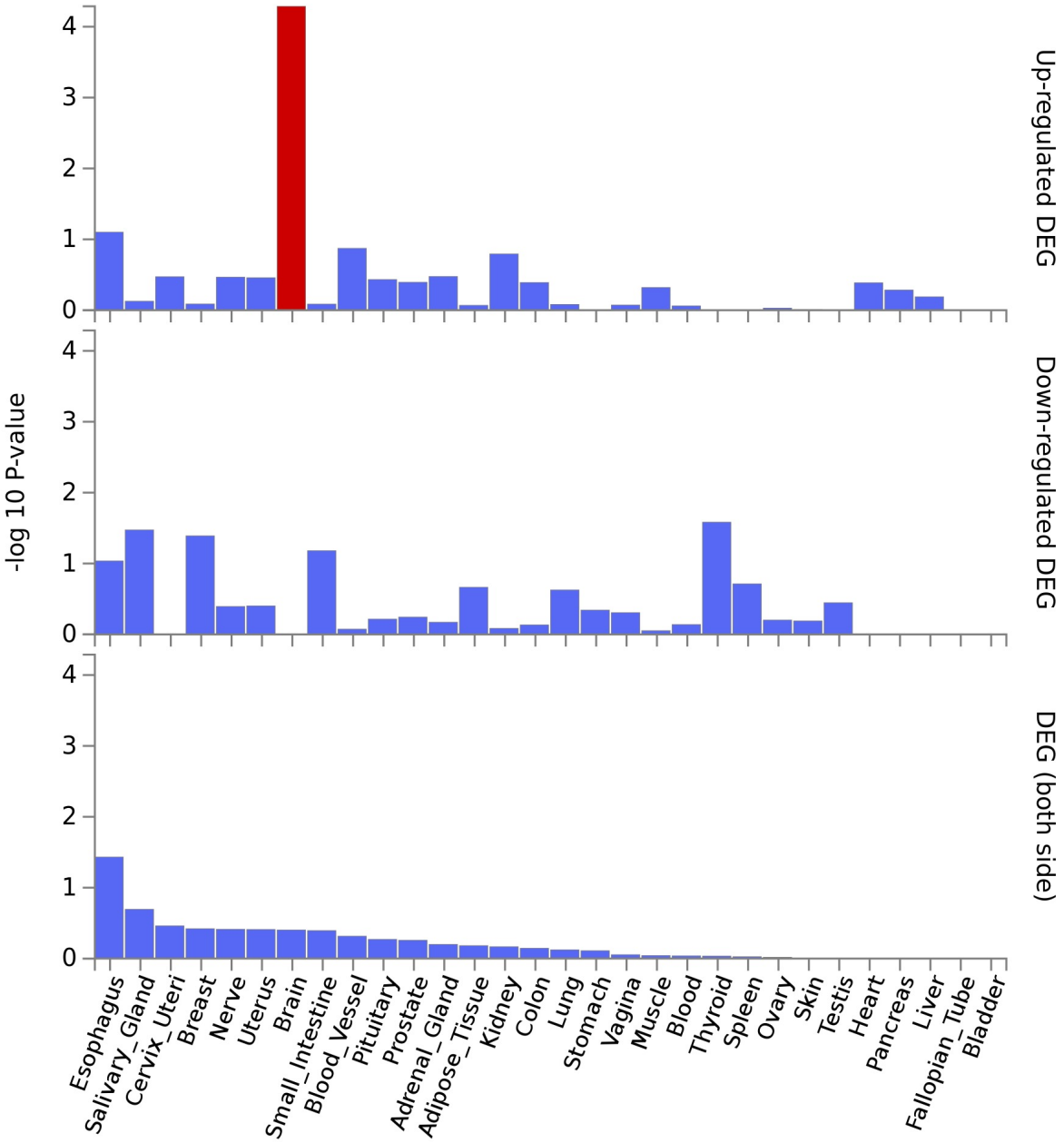
Supplementary Figure 5. The rs9620039 variant shared between major depressive disorder and intelligence is significantly associated with brain grene expression of *TCF20* in different parts of brain[18].



Supplementary Figures 6 A & B. Gene expression heatmap across 53 human tissues for genes nearest the most strongly associated SNP in a distinct genomic locus shared between major depression and INT. **A**) the genes with consistent effect directions in major depression and general INT and **B**) the genes with opposite effect directions in major depression and general INT at conjunctional false discovery rate <0.05 (Supplementary Table 3).

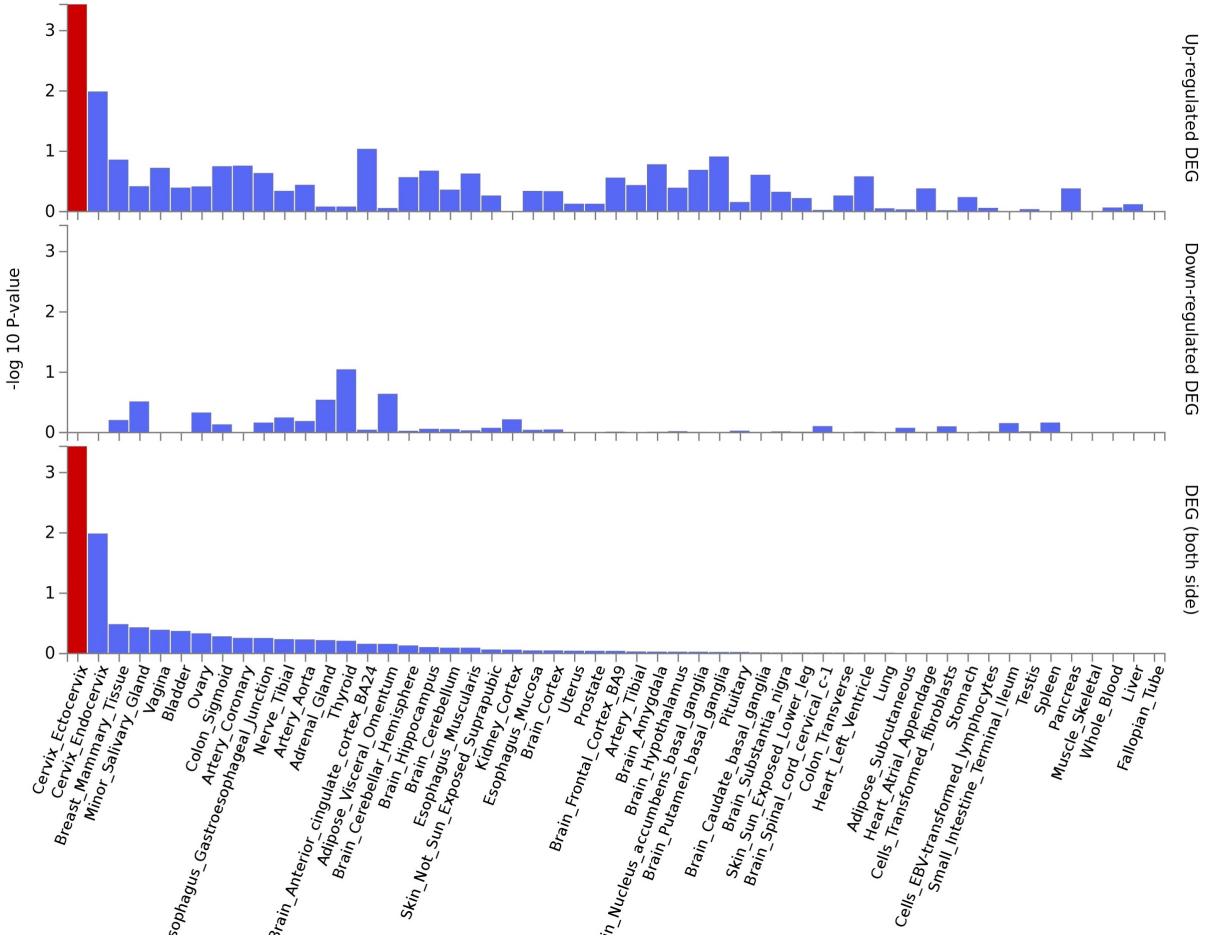


Supplementary Figure 7. Tissue enrichment for differential gene expression (DEG) in 53 GTEx tissue types of genes nearest lead SNPs in distinct genomic loci significantly associated with both major depression and INT at conjFDR<0.05 (Supplementary Table 8). Enrichment of differential gene expression is shown for higher (up-regulated DEG), lower (down-regulated DEG), or two-sided differences in gene expression [DEG (both side)]. Significantly enriched DEG sets with p<0.05 after Bonferroni correction are highlighted in red. The analysis was performed using FUMA[10].

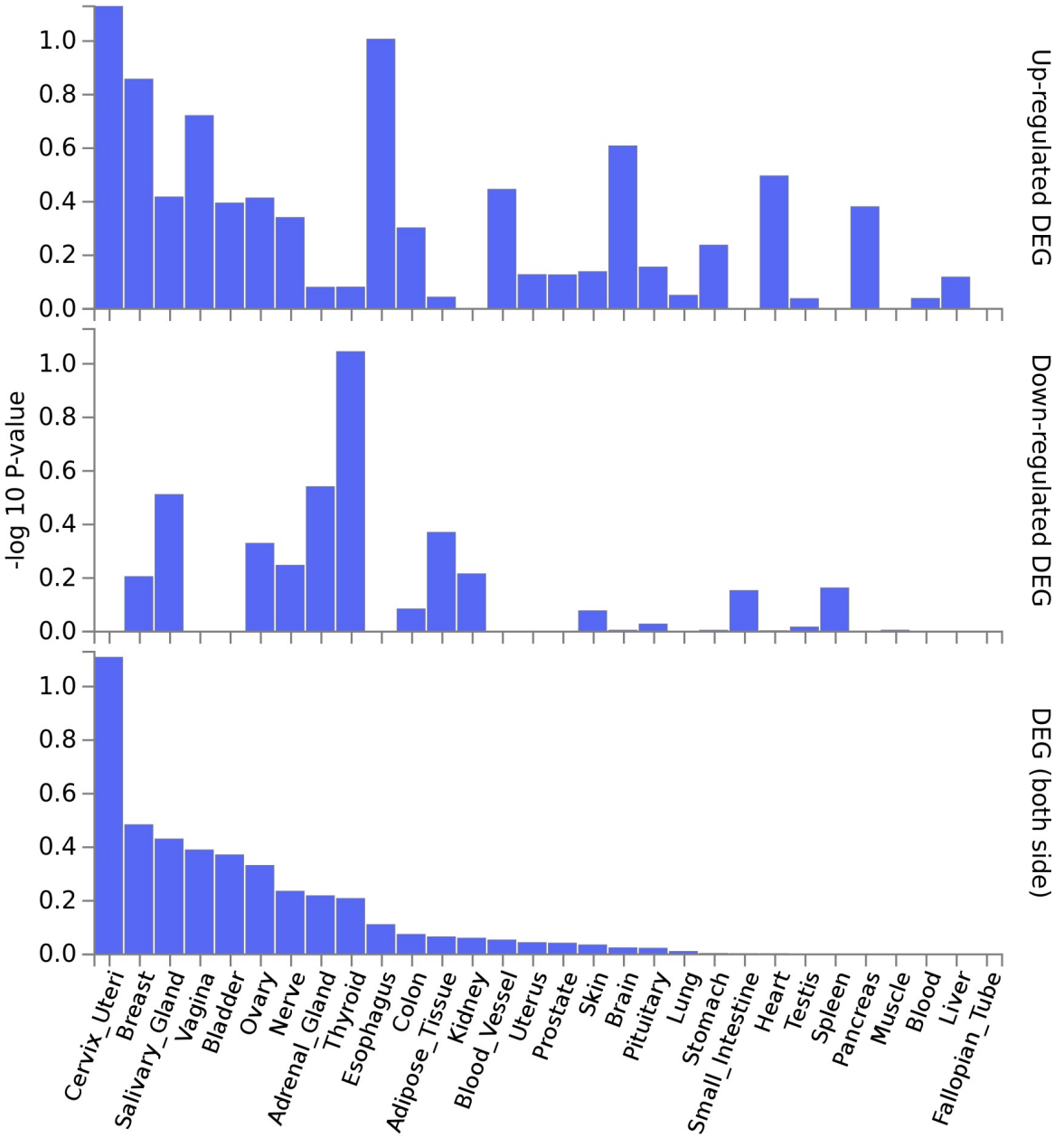


Supplementary Figure 8. Tissue enrichment for differential gene expression (DEG) in 53 GTEx tissue types of genes nearest lead SNPs in distinct genomic loci significantly associated with both major depression and INT at conjFDR<0.05 (Supplementary Table 8). Enrichment of differential gene expression is shown for higher (up-regulated DEG), lower (down-regulated DEG), or two-sided differences in gene expression [DEG (both side)].

Significantly enriched DEG sets with p<0.05 after Bonferroni correction are highlighted in red. The analysis was performed using FUMA[10].



Supplementary Figure 9. Tissue enrichment for differential gene expression (DEG) in 53 GTEx tissue types of genes nearest lead SNPs in distinct genomic loci significantly associated with both major depression and INT at conjFDR<0.05 (Supplementary Table 9). Enrichment of differential gene expression is shown for higher (up-regulated DEG), lower (down-regulated DEG), or two-sided differences in gene expression [DEG (both side)]. Significantly enriched DEG sets with p<0.05 after Bonferroni correction are highlighted in red. Analysis performed using FUMA[10].



Supplementary Figure 10. Tissue enrichment for differential gene expression (DEG) in 53 GTEx tissue types of genes nearest lead SNPs in distinct genomic loci significantly associated with both major depression and INT at conjFDR<0.01 (Supplementary Table 9). Enrichment of differential gene expression is shown for higher (up-regulated DEG), lower (down-regulated DEG), or two-sided differences in gene expression [DEG (both side)]. Significantly enriched DEG sets with p<0.05 after Bonferroni correction are highlighted in red. Analysis performed using FUMA[10].