Supplementary Material

Table of Contents

Supplementary Methods

Complete details of sample recruitment and assessment procedure

Knowledge Networks (KN), Inc., Menlo Park, CA (KN)), sent emails inviting 14,463 individuals to be screened. These individuals were members of an online nationwide U.S. survey and market research panel that is generally representative of the population; note that individuals were excluded who were invited to be screened for the NIMH Molecular Genetics of Schizophrenia control group control group (58) which was also recruited by KN, and who were still members of the panel. Criteria for these invitations were self-reported "Caucasian" ancestry, ages 21-60. The 9,569 respondents were screened online with lifetime self-report versions of the depression and alcohol/substance dependence modules of the Composite International Diagnostic Interview-Short Form (CIDI-SF) (10) in addition to screening items for schizophrenia and bipolar disorder. Of those who did not display current dependence, schizophrenia, or bipolar disorder, 1,959 (prospective cases) reported two or more periods of depressed mood and/or anhedonia (or one period of 52 or more weeks), with five total MDD criteria (or four criteria plus either telling a professional or endorsing that depression interfered with functioning); and 4,700 (prospective controls) reported no two-week period with more than two total criteria. During the same initial online screening session, 1,771 prospective cases and 3,162 prospective controls were invited for interview, of whom 964 and 880 gave online consents to be contacted and answered additional online questions about height; current and highest lifetime weight (when not pregnant); childhood trauma (the scale, which is included at the end of this supplement, included 10 items about physical abuse inside or outside of the home, sexual abuse, physical or emotional neglect, and two screening items for PTSD in response to trauma including avoidance of related thoughts or feelings and physical reactions when reminded of the trauma);

and smoking (the questionnaire is included at the end of this supplement; it permitted deriving a Fagerstrom score for nicotine dependence). A national phlebotomy company then attempted to contact those who have online consent, and 698 candidate cases and 624 candidate controls gave written informed consent and completed the blood draw. Contact information for each of these individuals was then provided to one of the two clinical sites, and 650 and 589 participated in telephone interviews comprised of the Structured Clinical Interview for DSM-IV (SCID) (depression, bipolar, alcohol, substance, and anxiety modules plus the psychosis screening module), the Patient Health Questionnaire (PHQ-9) (11) and Generalized Anxiety Disorder (GAD-7) (13) scales for current depressive and anxiety symptoms, and a screen for family history of MDD, bipolar disorder and suicide. The principal investigators of the clinical recruitment sites (MMW and JBP) reviewed final clinical data, and the overall PI (DFL) performed an additional review. Subjects were included who, based on the SCID interview, had no schizophrenia, bipolar-I or bipolar-II diagnosis, no current substance dependence, and were either cases (MDD with two or more lifetime episodes or one episode lasting two or more years) or controls (same criteria as described above). Note that most clinical interviews were conducted within 1-2 months of the blood draw, but for some individuals there was a delay of several months. After additional exclusions (genotypically non-European ancestry, unusual medical comorbidities, and standard quality control analysis of RNA-sequencing and SNP chip data), 922 individuals (463 cases and 459 control subjects) were included in the analyses (see below).

Phenotypic and drug intake variables

Multiple variables related to depression, anxiety and substance use history are available from the screening and SCID data. These were explored to create factor scores (using Principal

Components Analysis with Varimax rotation) that served as summaries of these features for *post hoc* analyses of their possible relationship with gene expression variables. First, the ten child abuse items were factor analyzed, resulting in three factors, with the explained variance due primarily to two factors interpreted as (i) physical abuse and neglect, and (ii) sexual abuse. Then, a five-factor solution was derived from the following variables: number of depressive criteria during the worst episode; severity of functional impairment during the worst episode; logarithm of the duration of the longest episode in days; logarithm of the number of lifetime episodes (truncated); logarithm of the age at onset; substance abuse or dependence (abuse scored as 1 and dependence as 3); alcohol abuse or dependence (abuse scored as 1 and dependence as 3); Fagerstrom score for nicotine dependence; presence of absence of lifetime panic disorder, social phobia PTSD (separate variables based on SCID); presence of absence of a family history (siblings or parents) or major depression and/or bipolar disorder; the two childhood abuse factor scores; and the current total PHQ score (measuring current depression). Analyzing the most strongly correlated variables, the five factors' appeared to primarily measure depression severity; recurrence and early onset (with physical abuse also loading here); substance use; PTSD and sexual abuse; and comorbid anxiety disorders (see section *Analysis of clinical variables*).

Current medication and substance intake variables were created for use as covariates in the analysis of the association of MDD to gene expression, because of the likely effects of many drugs on gene expression. Our modified SCID included items (for alcohol and for any abused substance reported by the subject) about the number of drinks per day or the number of times the substance was used per day on average during the past 2 weeks (considered present as a covariate if the average was 1 or more per day). Only alcohol and cannabis were being used currently by more than 30 subjects.

Our modified SCID interview required the recording of all current prescribed or over-the-counter medications with notes about their use. D.F.L. reviewed the medication lists and notes for each subject and classified them into a set of relevant pharmacological classes shown in Table S2 for commonly-used classes and in Table S4 for several relevant drugs or classes taken by only a few subjects.

RNA-sequencing quality control

If pooled RNA-sequencing libraries did not produce at least 180M reads, sequencing or library preparation was repeated for all three individuals in the lane (Figure S1a). For each individual with more than one sequencing run, runs with sufficient reads (> 20M), good mappability (> 40%), and good reproducibility of quantified expression data with the other runs (r^2 > 0.9), were merged. The first base of each read was trimmed to account for the stronger sequencing biases at the beginning cycles before mapping (Figure S1b). Genotype calls were made from RNA-seq data based on loci with sufficient read depth, and compared to genotypes from the SNP array; individuals with concordance below 0.85 were removed from the study as potentially mislabeled (Figure S2). Additional quality metrics were evaluated to ensure reasonable RNA Integrity Numbers (RIN) (Figure S1c), low percentage of hemoglobin reads (Figure S1d), and a high proportion of mapped reads in each individual (Figure S3). The results of quality control analysis were also utilized in a data normalization step described later.

SNP array quality control and genotype PCs

Genotype data were filtered for quality as follows. QC was carried out simultaneously for this dataset and a second dataset (Genetics of Recurrent Early-Onset Depression, phase 2, unpublished data) as they were genotyped by the same lab on the same platform a few months

apart. Pairwise estimates of IBD were computed in PLINK (59) , and any sample duplicates were excluded. Samples were excluded for elevated rates of heterozygosity (over 34.5% of SNPs) or if genotypes could not be called for more than 1.4% of SNPs. For SNPs, we evaluated QC metrics in each study, and retained SNPs with a missingness rate below 0.012, a 10% Gencall score above 0.55, and $p>0.001$ for deviation from Hardy-Weinberg equilibrium. To obtain principal component (PC) scores reflecting ancestry differences for use as covariates in association analyses, principal components analysis (PCA) was carried out for all individuals using every fifth autosomal SNP (to reduce the influence of LD among SNPs), and the PC scores examined for relationship to geographical/ethnic origin based on self-report (Figure S4). As shown in previous studies (60), PC1 was interpretable as a North-South gradient (Anglo-Saxon and Scandinavian to the North, Mediterranean and Ashkenazi Jewish to the South) and PC2 as East-West (from Russian/Slavic to Western European), with PC3 further separating Ashkenazi Jewish ancestry from other Mediterranean (Italian, Greek) ancestry. Individuals were excluded who were obvious outliers (by visual inspection of the plot of PC1 vs. PC2, or because multiple smaller components were numerical outliers) and PCA was repeated to ensure that there were no outliers by visual inspection of plots (Figure S4).

Normalizing RNA-seq count data

As the first step of the analysis, we used ridge regression on logarithm of read counts to remove to effect of several technical and biological factors from the quantified expression data (see Table S1). We set the ridge penalty parameter by evaluating the fit of regression model in a cross-validation setting. The technical factors, which include sequencing depth, per-individual GC bias, and percent hemoglobin counts were constructed from the Picard metrics (61) and an in-house QC pipeline. In addition to the technical factors, we also removed the effect of time of blood draw and estimated blood cell type proportions (see next section).

Inference of blood cell-type proportions

We inferred cell type proportions using a compendium of cell-type specific gene signatures (17). We used a non-negative least squares (NNLS) approach (16) for decomposing the observed *mixed* expression profile for each individual into a weighted linear combination of cell-type specific expression profiles. We obtained the profiles of 17 cell types from Supplementary Table S1 of Abbas et al. (16,17). We then estimated per-individual cell type proportions by using non-negative least squares regression, and regressing the observed expression data that reflect a mixture of cells for a given individual onto the cell-type-specific expression signatures. Given the observed logarithm (non-normalized read-counts) expression values A_i of size *s* x 1 for *s* genes in individual i (defined below), we estimated cell type proportions in individual *i* as follows:

$A_i \sim N(X \times C_i, I)$, s.t. $C_i \geq 0$

Where X is a matrix of size *s* x 17, s being the number of genes that are in the cell signatures: each column *k* of X represents the expression levels of the s genes in cell type *k*. The vector C_i then represents the proportions of each of the nine cell types in individual *i*. The non-negative constraint ensures that the estimated cell-type proportions have non-negative values. The non-negativity constraint often results in a sparse solution, especially when there are correlated cell-types. Here, we obtained a non-zero estimate for 11 cell-types (Table S1).

Identifying and accounting for phenotypic, drug intake covariates, and hidden factors

We measured the significance of association of the expression level of each gene (or transcript) with MDD using a likelihood ratio test (LRT), allowing us to account for a large number of possible confounding factors as background covariates. In particular, the LRT compares the log likelihood of two models: $L(y|B)$ and $L(y|B,g)$, the null model and full model, respectively. Here y is a binary n-vector representing MDD status in n individuals, B is an nxc matrix, comprised of c background (or confounding) covariates (each column of B represents a covariate), and g is an n-vector representing the expression level of a gene. In our setting, the likelihood is the likelihood of logistic regression. We computed p-values using permutation analysis (8,000 initial permutations; and 1,000,000 permutations for three genes with p=0 in the initial 8,000).

To identify measured confounding covariates, we carefully curated demographical, phenotypic, and drug intake covariates (see above). Among the set of all covariates (158), we only considered covariates that were relevant to at least 30 individuals. We then examined the correlation between each of the covariates with MDD status and the top ten expression PCs. We identified confounding covariates as those that were correlated either with an expression PC or with MDD status at $p < 0.05$. In agreement with previous studies, we found that BMI, gender, smoking, and age explained the most variability in expression measurements (Table S2, Figure S10). In addition to these standard covariates, we also identified 20 other covariates (mostly medication intake variables) that either correlated with MDD status or expression PCs (see Table S2).

We also accounted for five genotype PCs to represent population structure (62, 63) (see *SNP array quality control and genotype PCs* above), and ten expression PCs to account for *hidden* confounders. We chose to account for the top ten expression PCs, as we observed that they significantly correlated with major confounding covariates in our data (age, BMI, smoking, and gender). Also, we previously observed that removing expression PCs improves identification of

true-positive biological relationships between genes (20). To identify the number of PCs to account for, we evaluated the percentage of explained variance by the top 30 PCs, and identified the point at which the explained variance plateaus (see Figure S6).

eQTL analysis

We identified *cis*-eQTLs by associating expression levels of each gene with SNPs within 1Mb from the gene's transcription start site. Associations were carried out using Spearman Correlation Coefficients. We controlled for multiple testing using 0.05 FDR at the gene level subsequent to a Bonferroni correction per gene to account for the number of tested SNPs (see Battle et al., 2013, in revision Genome Research, for details on the eQTL association study).

Supplementary Results

Expressed genes and functional annotations

Among the set of 22,339 UCSC protein-coding genes (hg19), 13,857 autosomal genes were considered to be adequately expressed to be included in association analyses (at least 10 reads across the transcript in at least 100 individuals). We investigated the diversity of functional annotations in our set of expressed genes, and compared the over-representation of broad functional categories to those present for all genes with any canonical pathway annotation in MSigDB. In order to determine whether we had detected genes that are representative of all functional classes of genes in this study of whole blood, we manually inspected canonical pathways in MSigDB with more than 200 annotations and identified eight broad sets of pathways that are representative of a large fraction of gene annotations (Table S9). As shown in the table, we found that our set of expressed genes included 37.93-85.04% of genes in each category, and included a majority of the genes in 8 of 9 categories. As expected, the proportion

of genes with an “immune system” annotation is slightly larger in our set than in all annotated CP MSigDB genes (~13% and ~10%, respectively), however, we also detect most genes (~57%) with a “neuronal system” annotation.

Association of genetic variation with MDD

We identified one SNP, *rs11232553*, as significantly associated with MDD in this cohort, passing the genome-wide correction threshold (p<3e-8). To evaluation replication of *rs11232553* in the Psychiatric Genetics Consortium (PGC) cohort, we obtained the meta-analysis results form the PGC web-resource(64). *rs11232553* was not tested itself, but the SNP with the highest LD (*rs7342242*) had p-value>0.5 in PGC. Therefore, the observed association does not appear to replicate in the larger PGC cohort, and could represent a false positive. We performed several additional analyses to explore the possibility that the significant result at *rs11232553* could have resulted from population structure that was not adequately controlled by our PC covariates. First, at a global level, we do not observe inflation of GWAS p-values (Figure S11a), with an inflation factor λ of very close to 1. We do not observed an inflation in associations between gene expression levels and *rs11232553* (Figure S11b), and none of the top 10 genotype PCs were correlated with *rs11232553* (Table S10). Thus we do not see evidence that population structure is likely to account for the finding, and we concluded that it is either a false positive result or a rather extreme example of winner’s curse, where a true but very weak association exists, but evidence for a much stronger association is detected in a very underpowered sample because of the high variance of signals in such samples. In a second analysis, we also investigated the possibility of correlation between *rs11232553* and associations identified from expression analysis (25% FDR genes, and the IFN pathway genes); if *rs11232553* is a false-positive finding, we wanted to ensure that it does not underlie the expression findings. We did not observe a significant association between *rs11232553* and

expression levels of either the top 29 genes (25% FDR) or the genes in the interferon signaling pathway (Figure S12).

We also performed a targeted analysis of the interferon pathway genes, testing whether SNPs affecting interferon genes showed any evidence of association with MDD status. We analyzed association in both our cohort and the PGC cohort using two different sets of SNPs: (1) all SNPs within a 1MB window of each interferon gene; and (2) SNPs significantly associated with expression levels of each interferon gene (Battle et al., 2013, in revision, Genome Research) (*cis*-eQTLs). These analyses did not yield any significant result. We also investigated the possibility of a trans-eQTL impacting the expression levels of interferon genes, however, we did not find a genome-wide significant trans-eQTL association for any of the genes in the interferon signaling pathway, or a trans-eQTL association for the interferon pathway PC1 score (the first PC from a PCA of expression levels of all adequately expressed genes in the interferon α/β signaling pathway with univariate nominal p-values < 0.05).

Additionally, we performed combined analyses of PGC MDD(62) results with eQTLs and RNA-seq gene level associations obtained in this study. (Note that these analyses were exploratory in nature; no genome-wide significant associations between single SNPs and MDD were detected in the PGC MDD GWAS analysis, or in the analyses in the present study of the association between MDD and expression levels of single genes.) First, we performed a meta-analysis that combined p-values reported by PGC (for SNPs) with our results for association of individual genes. For this analysis, we used our eQTL data (see above) to *assign* SNPs to genes (only considering eQTLs that pass genome-wide multiple testing threshold). Next, each gene was assigned a p-value equal to the *minimum* (most significant) p-value in the PGC GWAS for any of its eQTL SNPs. Finally we performed a meta-analysis (using Fisher's method(65)) for each gene, combining the gene's best PGC eQTL SNP p-value with that gene's p-value for

association of its expression levels with MDD. After multiple hypothesis correction (correcting for numbers of eQTLs and of genes), we did not find a genome-wide significant association: the best association has a q-value of 0.22 and implicates the gene CNIH4. We also investigated using only the best eQTL per gene (in the previous analysis, we had considered all eQTLs for each genes), which also did not yield a significant result. Second, we also performed a more targeted analysis of the top 0.1%, 1% and 5% of PGC gene-level p-values (as above), overlapping the corresponding sets with top quantiles of RNA-seq associations. In this analysis, as in the above, we assigned PGC MDD p-values to genes based on associations of eQTLs with genes in our cohort (using the best eQTL for each gene). We then ranked the genes in two ways: using the PGC MDD p-values, or using the LRT expression p-values obtained in this study. Next we compared the set of top 0.1% of genes according to the PGC MDD p-values with the set of top 0.1% of genes according to the expression p-values, and computed the significance of the overlap between the gene sets using the hypergeometric test (this analysis was repeated for the top 1% and 5% of genes). However, this analysis did not yield significant overlaps compared to the expectation.

Additional post-hoc analysis of confounding factors for RNA-sequencing data

As described in the main text, we performed a *post hoc* assessment of possible confounders of interferon pathway results.

First, we annotated steroid intake (including oral or inhaled steroids), which affects the immune system and interferon response, and constructed five covariates (implicating 74 individuals) (Table S6). We also used the five steroid covariates as input to a LRT, and re-evaluated functional enrichment among the top *N* genes. After this correction, we observed that the

interferon α/β signaling pathway was significantly enriched in the top 100, 150, 200, 300, and 500 gene sets (0.05 FDR) (Table S3), but not in the top 60 genes.

In addition, we observed a general enrichment of low p-values for associations of gene expression levels with substance dependence. To ensure that our primary analysis was not confounded by inadequate accounting for substance or alcohol use impacting the immune system, we repeated the analysis using additional covariates -- a PC score reflecting lifetime alcohol and substance dependence, and also the Fagerstrom score for nicotine dependence -- and repeated the pathway enrichment analysis using these adjusted results (Table S3) (note that in the primary analysis we had only included covariates related to current alcohol or cannabis use, and current smoking status, rather than these variables related to lifetime dependence). Although we did observe larger p-values after the adjustment (Table S3), the interferon α/β signaling pathway was still significantly enriched among the top 100, 150, 200, 300, and 500 gene sets, though again enrichment in the top 60 genes did not pass the genome-wide threshold.

We also reassessed the possibility of confounding of results based on differences in cell-type proportions. In the primary analysis, we adjusted for cell-type proportions using NNLS. We decided to use NNLS in the primary analysis because it provides a natural estimate of cell-type proportions (i.e, *positive* values) and was previously tested and experimentally validated in a similar setting (16). Using NNLS results in a sparse solution, often alleviating over-fitting. As part of the post-hoc analysis for assessing the robustness of the results, we also estimated cell-type proportions using ridge regression. In this setting, ridge regression is more aggressive (more prone to over-fitting) and not as interpretive as NNLS (as cell-type estimates can be negative). Here, we observed a correlation between estimates of cell types and expression level of interferon pathway genes (though the re-estimated cell types are again not correlated with

MDD). Using the ridge regression cell-type estimates in the LRT, we performed the functional enrichment test on subsets of top 30, 60, 100, 150, 200, 300, and 500 genes. We observed a genome-wide significant enrichment (0.05 FDR) for three of the subsets: 100, 300, and 500 gene sets (Table S3). The fact that several gene subsets continue to support the results, despite the less desirable properties of ridge regression, is reassuring. However, because we do not have cell-type counts for individuals in this cohort, we can not directly assess the accuracy of our approach in this setting and several issues remain for future research. At the very least, studies are needed in which cell type proportions have been measured directly with proper storage of specimens after blood draw for this purpose. There is also limited biological understanding in general about interferon signaling in different cell types, and which cell types might be more biologically relevant to the role of interferon signaling in disease.

In the primary analysis, we included ten expression PCs in our likelihood ratio test as background covariates. However, we did not directly account for data batches, as we had already removed a large proportion of variability (>75%) (Table S1). We also evaluated the impact of batch in this post-hoc analysis. We observed that although our correction removed much of the correlation between expression levels and batch, we still do observe residual correlations for ~600 genes. In particular >12,000 genes in the raw data were impacted by batch, 1,855 genes were impacted by batch after accounting for technical factors alone (Table S1), and 629 genes were impact by batch after accounting for expression PCs in addition to technical factors (all associations based on Bonferroni threshold of 0.05). To ensure that this residual correlation is not impacting our finding, we performed hypergeometric enrichment tests on batch corrected data (additionally including batch covariates in the LRT), where we again observed a significant enrichment of interferon signaling genes in the top {100, 150, 200, 300, 500} subset of genes ($p<1e-5$ for all cases).

Association between isoform ratio and MDD

Cufflinks(66) was used to quantify isoform expression levels by computing the *isoform ratio*, representing the *fraction* of a gene's expression arising from a particular transcript. In particular, given a gene with *k* transcripts, let e_{ji} represent the expression level of transcript *i* in individual *j.* The isoform ratio for the i^{th} transcript of this gene in individual *j* is given by $\frac{e_{ji}}{\sum_{v=1}^{k} e_{jv}}$. We performed association testing between isoform ratios for each isoform and MDD status, following a procedure that was similar to that for testing association of gene expression levels. We first normalized the isoform ratio by linearly accounting for technical factors listed in Table S1 in addition to 10 PCs derived from isoform ratio data, and then obtained isoform ratio p-values using LRT while accounting for the 29 background covariates listed in TableS2.

In this analysis, we did not identify a statistically significant association after correcting for multiple hypothesis testing (threshold of 6.5283e-06, 7659 tested isoform ratios). Table S8 lists the top 10 isoform ratios associated with MDD and the corresponding q-values (67). We also performed pathway analysis, which did not yield a significant finding.

We also performed a targeted analysis by considering only isoform ratios with significant correlations with (1) expression levels of any of the top 30 genes in the analysis of single-gene associations with MDD; or (2) the IFN pathway PC1 score (see legend to Figure 1, main text). In the first analysis, we identified 342 isoform ratios with significant expression correlation with a top 30 gene. Among these, the best p-value for association with MDD did not pass genome-wide significant threshold (q-value of 0.15 for an isoform of gene WWP2). In the second analysis, we identified 17 isoforms with significant associations with IFN pathway PC1 score, among which the best q-value was 0.2 for association between isoform ratio of IFIT3 and MDD.

Analysis of clinical variables

Factor analyses were carried out to reduce the number of variables, using Principal Components Analysis with Varimax rotation (SYSTAT software). The most parsimonious solution contained 5 rotated factors with the following variable loadings as shown in Table S11, with the following interpretation of the factors based on examination of the variables that loaded most highly on each factor score: PC1 is interpreted as Recurrent/Early-Onset/FH/Persistent(PHQ), PC2 as Substance use, PC3 as PTSD/sexual abuse/chronicity, PC4 as Anxiety disorders, and PC5 as Severity/impairment.

The two child abuse scores that were entered into the above factor analysis were themselves generated by a separate factor analysis of the responses to the Childhood Trauma Questionnaire (reproduced at the end of this supplement). Including factor scores in a PCA is not considered ideal, but in this case it permitted us to summarize the main clinical data points in a single set of scores, showing, for example, the closer relationship between sexual abuse and lifetime PTSD rather than with other anxiety disorders. The most parsimonious rotated solution contained the 3 factors and variable loadings shown in Table S12. We interpreted these factors as follows (again by examining the mostly highly-loaded variables): AbuseF1 is interpreted as Physical/Emotional Abuse, AbuseF2 as Sexual Abuse, and AbuseF3 as External threat (this was not entered into the overall clinical PCA above because it explained very little of the variance and did not correlate well with other variables).

Pearson correlations between IFN-I pathway score (Figure 1 legend in main text) and clinical factor scores were not significant (p-values > 0.05). Additional exploration of individual variables (not shown) failed to identify significant predictors of IFN-I pathway PC1 scores.

Tables S1-S7

Table S1. Technical and biological covariates (see legend next page)

	Cases: mean (std)	Controls: mean (std)	Association with MDD (p-value)
Sum of log2 reads (SD)	1.33E+05 (9.13E+03)	1.32E+05 (9.90E+03)	4.21E-01
Percent Hemoglobin reads	1.28E-02 (1.84E-02)	1.18E-02 (1.90E-02)	6.30E-01
Individual-specific GC bias	1.10E-01 (5.42E-02)	1.11E-01 (5.54E-02)	9.20E-01
Individual-specific length bias	-8.66E-02 (1.77E-02)	-8.86E-02 (1.77E-02)	6.68E-02
RNA yield	7.17E+03 (3.80E+03)	7.12E+03 (3.65E+03)	8.14E-01
Globin flag (technician)	1.96E-02 (1.39E-01)	3.02E-02 (1.71E-01)	3.01E-01
Percent duplicated reads	7.93E+01 (4.17E+00)	8.00E+01 (4.26E+00)	2.39E-02
Number of coding bases	1.36E+09 (5.51E+08)	1.32E+09 (5.37E+08)	2.76E-01
Number of intergenic bases	4.05E+08 (1.89E+08)	4.20E+08 (2.48E+08)	8.74E-01
Number of intronic bases	3.57E+08 (1.69E+08)	3.72E+08 (2.10E+08)	6.25E-01
Median 3' bias	4.33E-01 (6.02E-02)	4.38E-01 (6.03E-02)	1.33E-01
Median 5' bias	2.05E-01 (3.60E-02)	2.02E-01 (3.41E-02)	2.19E-01
Median 5' to 3' bias	5.53E-01 (1.35E-01)	5.39E-01 (1.30E-01)	1.52E-01
Median CV coverage	5.23E-01 (5.59E-02)	5.28E-01 (5.85E-02)	5.57E-02
Percent coding bases	4.34E-01 (6.33E-02)	4.26E-01 (7.89E-02)	3.63E-01
Percent intergenic bases	1.34E-01 (5.83E-02)	1.40E-01 (7.34E-02)	7.71E-01
Percent intronic bases	1.17E-01 (4.72E-02)	1.23E-01 (5.81E-02)	1.91E-01
Percent mRNA bases	7.49E-01 (1.04E-01)	7.37E-01 (1.31E-01)	4.16E-01
Percent usable bases	7.49E-01 (1.04E-01)	7.37E-01 (1.31E-01)	4.16E-01
Percent UTR bases	3.15E-01 (4.57E-02)	3.10E-01 (5.57E-02)	7.67E-01
Percent aligned bases	3.10E+09 (1.12E+09)	3.07E+09 (1.06E+09)	6.90E-01
Number of BF bases	3.10E+09 (1.12E+09)	3.07E+09 (1.06E+09)	6.89E-01
Number of UTR bases	9.81E+08 (3.90E+08)	9.58E+08 (3.80E+08)	4.14E-01
Cell-type proportion: Th	3.55E-01 (6.43E-02)	3.56E-01 (6.68E-02)	8.36E-01
Cell-type proportion: Tc	2.56E-03 (1.14E-02)	3.32E-03 (1.78E-02)	3.65E-01
Cell-type proportion: Tc_act	1.39E-04 (2.98E-03)	0.00E+00 (0.00E+00)	3.16E-01
Cell-type proportion: B	8.61E-02 (6.52E-02)	9.09E-02 (6.97E-02)	4.72E-01
Cell-type proportion: PC	0.00E+00 (0.00E+00)	4.73E-04 (8.21E-03)	1.59E-01
Cell-type proportion: NK	7.78E-02 (5.72E-02)	6.79E-02 (5.31E-02)	9.02E-03
Cell-type proportion: NK_act	1.40E-03 (1.18E-02)	1.02E-03 (1.01E-02)	7.88E-01
Cell-type proportion: mono	1.25E-01 (4.58E-02)	1.23E-01 (4.55E-02)	9.94E-01
Cell-type proportion: DC	5.47E-03 (1.35E-02)	5.57E-03 (1.39E-02)	6.92E-01
Cell-type proportion: DC_act	1.26E-03 (9.71E-03)	1.54E-03 (1.23E-02)	7.10E-01
Cell-type proportion: neutron	5.55E-01 (6.46E-02)	5.54E-01 (6.39E-02)	8.22E-01
Time of day for blood draw	1.25E+03 (3.51E+02)	1.23E+03 (3.32E+02)	5.11E-01

Table S1 lists all technical and biological covariates that were used in the normalization of the raw logarithm read counts. Seventeen of these are technical factors (yellow) obtained from Picard metrics, two technical factors obtained from the technicians (orange), we estimated four other technical factors from the quantified reads (purple), eleven factors are the inferred cell type proportions, and one factor representing the time of the day that the individual's blood was drawn. Individual-specific exon length, and individual-specific GC are estimated as the proportion of read variance in each individual (mapped to exons) that can be explained by GC compositions of the exons or the length of the exons, respectively---these factors are estimated per individual, by correlating mapped reads to exons with a vector of exon GC composition or exon lengths. Note that cell type frequencies (estimated using NNLS) are not normalized to represent fractions between [0-1], the estimates are positive scalars that depend on sequencing depth and other variables that impact read counts.

Table S2. **Correlation of demographical and medication intake variables with MDD status.** All variables shown in this table were used as background in the LRT, in addition to ten expression PCs (Figure S6). For binary variables, p-values are obtained using fisher's exact test. For non-binary variables, p-values represent significance of Spearman's rank correlation. Sorted by p-value for association with expression PCs.

	Number of individuals or mean (sd)		Association with	
Covariate name	**Controls**	**Cases**	**MDD**	**Expression PC**
Age at interview	44.64 (11.0)	44.77 (10.69)	9.58E-01	1.0E-18 (PC 8)
Female gender	288	360	4.08E-07	6.66E-16 (PC 6)
BMI	27.87 (6.45)	30.25 (7.74)	2.63E-07	5.36E-10 (PC 7)
Number of cigarettes per day	1.12 (4.55)	2.49 (6.76)	8.78E-06	1.87E-08 (PC 7)
Number of blood pressure meds	0.22 (0.53)	0.31 (0.66)	6.02E-02	4.17E-08 (PC 8)
Smoked before blood draw	34	67	4.01E-04	2.85E-07 (PC 7)
Cholesterol lowering meds	50	64	1.05E-01	2.04E-05 (PC 8)
Oral hypoglycemic meds	16	30	2.58E-02	2.44E-05 (PC 8)
Ate before blood draw	339	323	1.68E-01	1.45E-04 (PC 10)
Current alcohol use	0.69 (0.76)	0.61 (0.78)	3.96E-02	4.22E-04 (PC 6)
Diuretic	25	29	3.49E-01	4.58E-04 (PC 8)
OBCP meds	39	45	2.98E-01	2.09E-03 (PC 3)
Protein Pump Inhibitor (PPI)	32	52	1.62E-02	2.83E-03 (PC 3)
ACE Inhibitor	30	36	2.74E-01	3.36E-03 (PC 8)
Opiate use	9	46	1.18E-07	3.50E-03 (PC 8)
Thyroid medication	34	54	1.82E-02	4.81E-03 (PC 3)
Cannabis use (past 2 weeks)	9	23	9.61E-03	9.16E-03 (PC 4)
Beta blocker meds	14	44	3.45E-05	1.26E-02 (PC 8)
Decongestants or Stimulants	10	52	1.27E-08	2.95E-02 (PC 5)
NSAID meds	19	46	4.05E-04	3.67E-02 (PC 6)
Decongestant meds	10	41	4.94E-06	3.99E-02 (PC 6)
Exercise before blood draw	92	74	6.43E-02	4.34E-02 (PC 2)
Anticholinergic meds	12	23	4.41E-02	5.97E-02 (PC 1)
Antihistaminic meds	39	60	1.84E-02	8.70E-02 (PC 8)
Genotype PC5	0.0014 (0.020)	0.00025 (0.021)	3.78E-01	1.25E-01 (PC 5)
Genotype PC3	0.00038 (0.019)	-0.00041 (0.019)	4.00E-01	1.55E-01 (PC 10)
Genotype PC1	0.0017 (0.019)	0.0032 (0.016)	8.47E-01	1.56E-01 (PC 1)
Genotype PC2	0.0018 (0.022)	-0.0014 (0.021)	2.72E-02	1.58E-01 (PC 9)
Genotype PC4	-0.00042 (0.021)	0.0011 (0.022)	4.27E-01	2.06e-01 (PC 7)

Table S3. Enrichment (p-values) of interferon α/β signaling in sets of N genes with the strongest evidence for association with MDD. Enrichment (p-values) of interferon α/β signaling pathway among the top N genes with lowest association p-value to MDD. To ensure that medication intake or substance dependence is not primarily driving the enrichment, we performed the analysis multiple times, each time including additional covariates (beyond those presented in Table S1) or excluding certain individuals based on their medication history. M1 refers to a model that includes all covariates in Table S1. Orange indicates significance at 0.05 FDR.

	Number of most strongly associated genes in the tested set						
Covariates in addition to M1:	**30**	**60**	**100**	**150**	**200**	**300**	**500**
(M1 alone)	4.00E-03	1.00E-06	7.00E-07	3.00E-11	2.00E-11	3.00E-13	3.00E-16
Fagerstrom score	9.00E-02	9.00E-04	1.00E-05	4.00E-07	3.00E-06	2.00E-09	3.00E-10
Substance abuse/depend score	9.00E-02	9.00E-04	1.00E-05	4.00E-07	3.00E-06	2.00E-09	3.00E-09
Hepatitis C (14 cases, 11 controls)	NA	1.00E-06	7.00E-07	3.00E-11	2.00E-11	3.00E-13	5.00E-15
5 steroid use variables	9.00E-02	9.00E-04	2.00E-05	8.00E-06	1.00E-08	2.00E-09	1.00E-12
Re-estimated cell type proportions	9.00E-02	2.00E-02	2.00E-05	2.00E-04	4.4E-04	7.00E-09	5.00E-08
(92 Ss excluded - Table S5)	9.00E-02	1.00E-06	2.00E-05	3.00E-08	2.00E-08	7.00E-12	2.00E-13

Table S4. Medication and diagnostic criteria for exclusion of 92 subjects for *post hoc* analysis of association of MDD to Interferon α/β signaling pathway. Manually examining the diagnostic questionnaires, including medication lists and narrative notes, we identified unusual medication intake or medical diagnosis as shown in this table. We repeated the analysis of enrichment of sets of top genes for Interferon α/β Signaling Pathway genes after removing the 92 individuals shown in the table.

	Number of individuals	Number of cases
Allopurinol	5	0
Anti-HIV	2	1
Triptan	5	3
uc_salicylates	4	2
Provigil	1	0
Decongestant	1	0
Asthma_inhaler_unknown	1	1
Arava	1	1
Antibiotic_unspec	2	1
Interferon	2	1
Enbrel	1	0
Cimzia	1	1
TNF_inhib	4	3
Cipro	1	1
Flagyl	1	1
cephalosporin	1	1
Asthma_unknown	1	1
Byetta	1	1
Hydrochloroquine	1	1
Insulin	17	10
Anti-platelet	8	4
H2_antagonist	14	9
Antiherpes	9	4
Leukotriene_antag	20	11
Tetracycline	7	6
Immunosuppressant	7	6
MS or Lupus	8	8

Table S5. Major categories of anti-depressant medications taken by cases and controls.

Drug class	Number of cases	Number of controls
SSRI	124	6
SNRI	43	3
Bupropion	52	1

Note: 10 control subjects reported taking these antidepressant drugs for reasons unrelated to depression (pain; smoking cessation; anxiety symptoms without depression), and were judged to meet clinical criteria for controls.

Table S6. Variables related to steroid intake (*post hoc* analysis). Five covariates were manually curated from data on use of steroid medications, for *post hoc* analyses of the relationship between MDD and interferon α/β signaling pathway expression and MDD. The table shows case-control Ns (total cases+controls and cases alone) and p-values for case-control comparisons (Fisher's exact tests) for .

	N (total)	N (cases)	Case-control p-value
Oral or inhaled steroids	27	34	0.2
Oral steroid	9	6	0.85
Current oral steroids	7	3	0.94
Anabolic steroids	3	1	0.94
Inhaled nasal topical steroids	20	31	0.08

Table S7. **Interferon α/β signaling pathway (REACTOME curated).** For the 45 genes with adequate expression levels, p-values are shown for the primary analysis of association of individual gene expression levels with MDD, with genome-wide rank and direction (+ means increased expression in cases). Yellow background indicates genes with nominal p-values <0.05.

Gene name	p-value	Rank	Direction
MX1	1.26E-04	7	+
OAS1	2.52E-04	15	+
IFIT3	3.78E-04	22	+
PTPN6	1.26E-03	43	+
ADAR	1.26E-03	45	+
IRF7	1.64E-03	53	+
IFIT1	3.15E-03	100	+
USP18	3.78E-03	121	+
ISG15	3.78E-03	122	+
OAS2	4.16E-03	127	+
IRF8	6.93E-03	185	+
IFIT2	7.30E-03	201	+
OAS3	7.56E-03	208	+
MX2	8.44E-03	226	+
IRF9	1.07E-02	274	+
IFI6	1.13E-02	285	+
OASL	1.44E-02	352	+
XAF1	1.49E-02	362	+
IFI35	1.65E-02	404	+
IFNAR2	4.99E-02	1030	-
SOCS1	6.80E-02	1347	+
PSMB8	9.41E-02	1798	+
STAT2	1.36E-01	2432	+
IRF5	1.37E-01	2442	+
IRF4	1.59E-01	2774	-
IFI27	1.78E-01	3056	+
HLA-A	2.08E-01	3479	-
SOCS3	2.42E-01	3958	-
HLA-B	2.89E-01	4667	-
IFNAR1	2.90E-01	4692	-
IFITM2	3.37E-01	5347	-
STAT1	3.38E-01	5367	+
ISG20	3.51E-01	5545	+
HLA-F	3.58E-01	5649	+
IP6K2	3.71E-01	5839	-
IRF3	4.35E-01	6697	-
IRF6	6.34E-01	9250	-
JAK1	6.53E-01	9499	-
IRF2	6.77E-01	9806	+
GBP2	6.79E-01	9833	+
PTPN1	7.22E-01	10416	-
EGR1	7.62E-01	10905	-
IFITM1	7.65E-01	10957	+
IRF1	7.76E-01	11094	+
RNASEL	8.41E-01	11914	+
HLA-C	9.03E-01	12708	+
HLA-G	9.70E-01	13505	+
IFITM3	9.94E-01	13787	-
TYK2	9.97E-01	13826	-

Table S8. Top 10 isoform ratio associations with MDD. We did not identify genome-wide significant associations. The table lists the top 10 associations and the corresponding q-values.

Gene name	Isoform name	Qvalue
DLST	NM_001933_DLST_DLST_TSS25245_chr14:75348593-75370450	0.6631
CR2	NM_001877_CR2_CR2_TSS18205_chr1:207627644-207663240	0.6631
PRCP	NM_005040_PRCP_PRCP_TSS12107_chr11:82535408-82611557	0.6631
THAP5	NM_182529_THAP5_THAP5_TSS4186_chr7:108202670-108209897	0.6631
RPS24	NM_001026_RPS24_RPS24_TSS1929_chr10:79793517-79800473	0.6631
SIGIRR	NM_021805_SIGIRR_SIGIRR_TSS14971_chr11:405715-417397	0.6631
MDH1	NM_005917_MDH1_MDH1_TSS18119_chr2:63815742-63834330	0.6631
CCDC77	NM_001130148_CCDC77_CCDC77_TSS21223_chr12:498515-551806	0.6794
C6orf106	NM_024294_C6orf106_C6orf106_TSS8045_chr6:34555065-34664625	0.6794
HNRNPK	NM_031262_HNRNPK_HNRNPK_TSS7115_chr9:86582997-86595569	0.6794

Table S9. Expressed genes and functional annotations. The table summarizes the representation of 10 major MSigDB canonical pathway (CP) category in the set of expressed genes in this study (see Supplementary Results).

	genes with MSigDB_CP annotation	% of all annotated genes	expressed genes (% of genes in category)	% of expressed annotated genes	% of expressed genes
REACTOME IMMUNE SYSTEM	933	10.63%	765 (81.99%)	13.42%	5.52%
REACTOME SIGNALING BY GPCR	920	10.48%	349 (37.93%)	6.12%	2.52%
REACTOME METABOLISM OF PROTEINS	518	5.90%	365 (70.46%)	6.40%	2.63%
REACTOME METABOLISM OF LIPIDS AND LIPOPROTEINS	478	5.44%	370 (77.41%)	6.49%	2.67%
REACTOME CELL_CYCLE	421	4.79%	358 (85.04%)	6.28%	2.58%
REACTOME METABOLISM OF RNA	330	3.76%	240 (72.73%)	4.21%	1.73%
REACTOME NEURONAL SYSTEM	279	3.18%	160 (57.35%)	2.81%	1.15%
REACTOME DEVELOPMENTAL BIOLOGY	396	4.51%	290 (73.23%)	5.09%	2.09%

Table S10. The table shows the association p-values between *rs11232553* and the top genotype PCs.

PC	p-value
geno PC1	0.9676
geno PC2	0.2055
geno PC3	0.8168
geno PC4	0.2795
geno PC5	0.6871
geno PC6	0.4728
geno PC7	0.6202
geno PC8	0.0568
geno PC9	0.3387
geno PC10	0.337

Table 11. Factor analysis of clinical variables. The table shows the rotated loading matrix. The highly weighted clinical covariates for each PC (column) are highlighted in yellow.

Clinical variable	PC1	PC2	PC3	PC4	PC5
Impairment_Severity_Worst_MDE	-0.011	0.006	-0.013	0.030	0.612
Num_Criteria_Worst_MDE	-0.029	-0.057	0.004	-0.054	0.615
SubstAbuse1/SubstDepend3	0.011	0.597	0.022	-0.009	-0.059
AlcAbuse1/AlcDepend3	-0.089	0.574	-0.168	0.106	-0.006
Fagerstrom_score(NicDepend)	0.041	0.517	0.114	-0.055	0.035
PTSD(lifetime)	-0.117	-0.033	0.529	0.018	0.037
ChildSexualAbuse	0.005	0.088	0.543	-0.234	-0.144
Log(LongestMDE)	0.095	-0.025	0.564	0.242	0.076
Panic_Disorder(lifetime)	0.048	0.066	0.028	-0.607	0.235
Social_Phobia(lifetime)	-0.302	-0.049	-0.002	-0.514	-0.007
Log(EstimatedNumberMDEs)	-0.561	0.045	-0.155	0.024	0.109
FamHist(MDDorBipolar)	-0.494	-0.124	-0.013	0.021	-0.307
Log(AgeAtOnset)	0.369	0.015	-0.123	-0.234	-0.142
PHQ-9(total)	-0.370	0.069	0.149	-0.148	-0.016
ChildAbuse_Physical-Emotional	-0.201	0.069	0.049	0.398	0.181
Proportion of variance explained (PVE)	0.17	0.11	0.08	0.07	0.07

Table S12. Factor analysis of childhood trauma questionnaire responses. The table shows the rotated loading matrix. For each PC (column), the most highly weighted covariates are highlighted in yellow.

	AbuseF1	AbuseF2	AbuseF3
Emot_abuse	0.846	0.142	0.143
No_comfort	0.768	0.166	0.093
Parent_parent_abuse	0.724	0.064	0.114
Physical_abuse	0.652	0.061	0.328
Avoidance	0.635	0.515	-0.249
Physical_reactions	0.609	0.458	-0.237
Neglect	0.539	0.160	0.404
Touched	0.173	0.880	0.084
Sexual_abuse	0.076	0.842	0.262
Attacked_threatened (outside home)	0.144	0.131	0.796
Percent variance explained	3.38	2.06	1.14

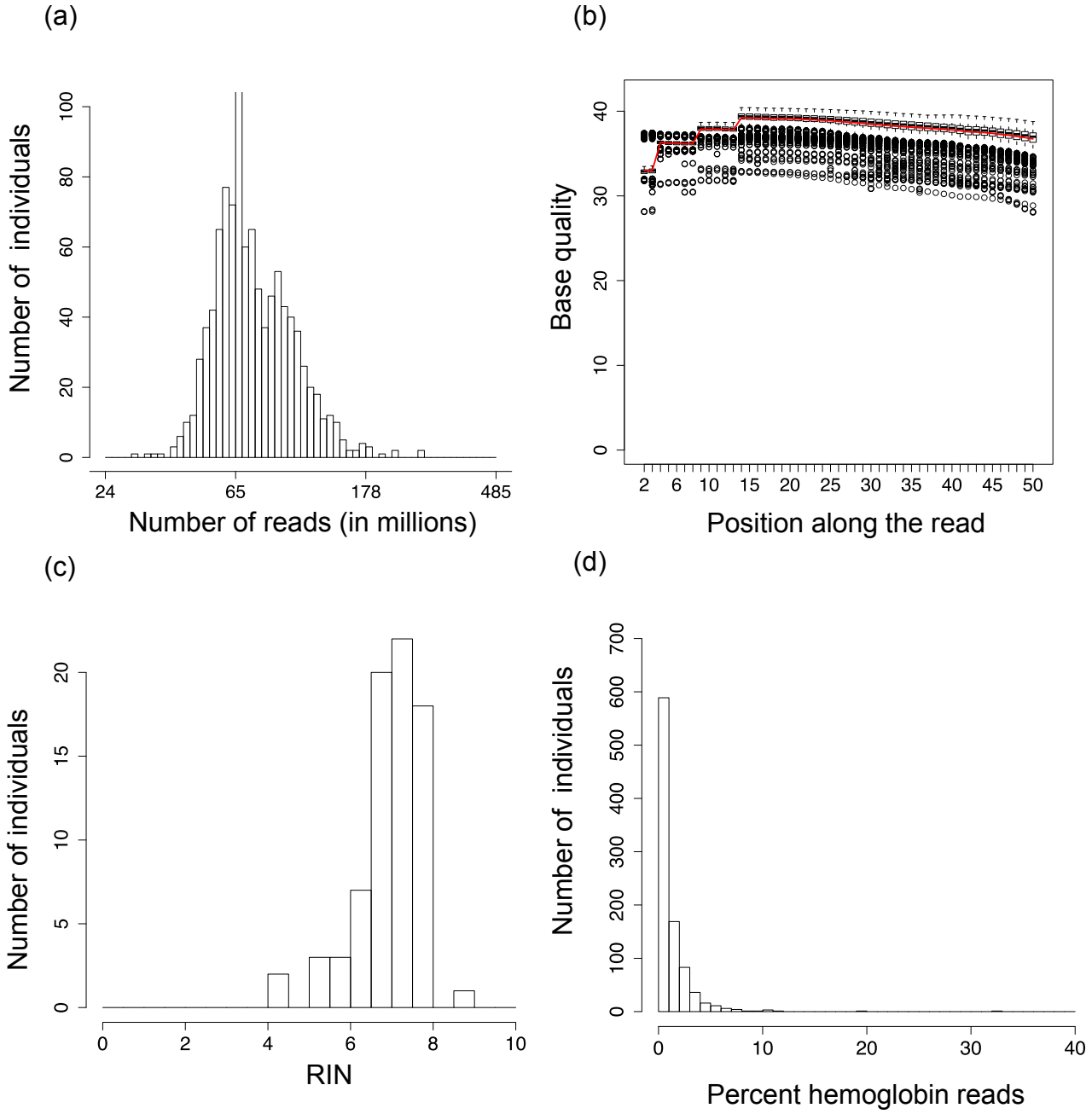


Figure S1. RNA-sequencing quality control. (a) The distribution of the number of sequenced reads is plotted in log scale. The distribution is skewed to the right because of the extra sequencing runs for poorly sequenced individuals. (b) Boxplot of base quality scores along the sequencing reads (from base 2 to base 50). Average score at each position is marked in red. The base quality reaches its maximum at base 14 and begins to decrease slowly after base 25. (c) RNA Integrity Numbers (RIN) for post-GlobinClear RNA. We recorded the RINs for 12 samples from each 96 well plate containing RNAs. (d) Using the GlobinClear protocol, hemoglobin RNA was removed from each sample before sequencing. A histogram of the percent reads coming from hemoglobin transcripts demonstrates the effectiveness of the GlobinClear procedure amongst our individuals (median percent hemoglobin read is 0.7%).

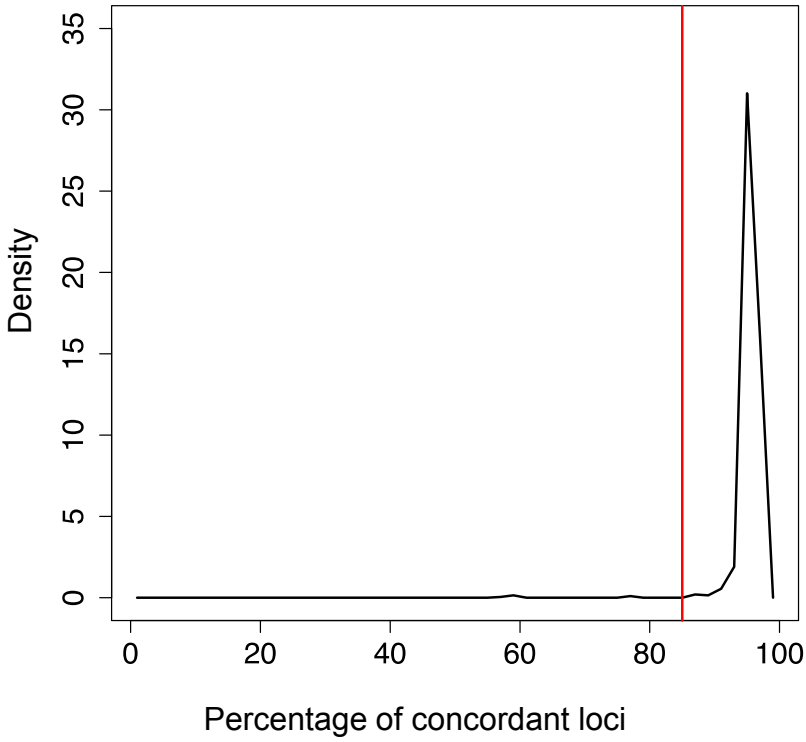


Figure S2. Concordance between SNP array and RNA-seq called genotypes. SNP genotypes were called using RNA-seq reads in deep covered regions and compared with the SNP array data. Low concordance (<85%, shown as a red line) suggests a potential labeling error, and such individuals were removed from this study. Most individuals show high estimates of concordance. We removed 6 subjects at the cutoff of 85% concordance.

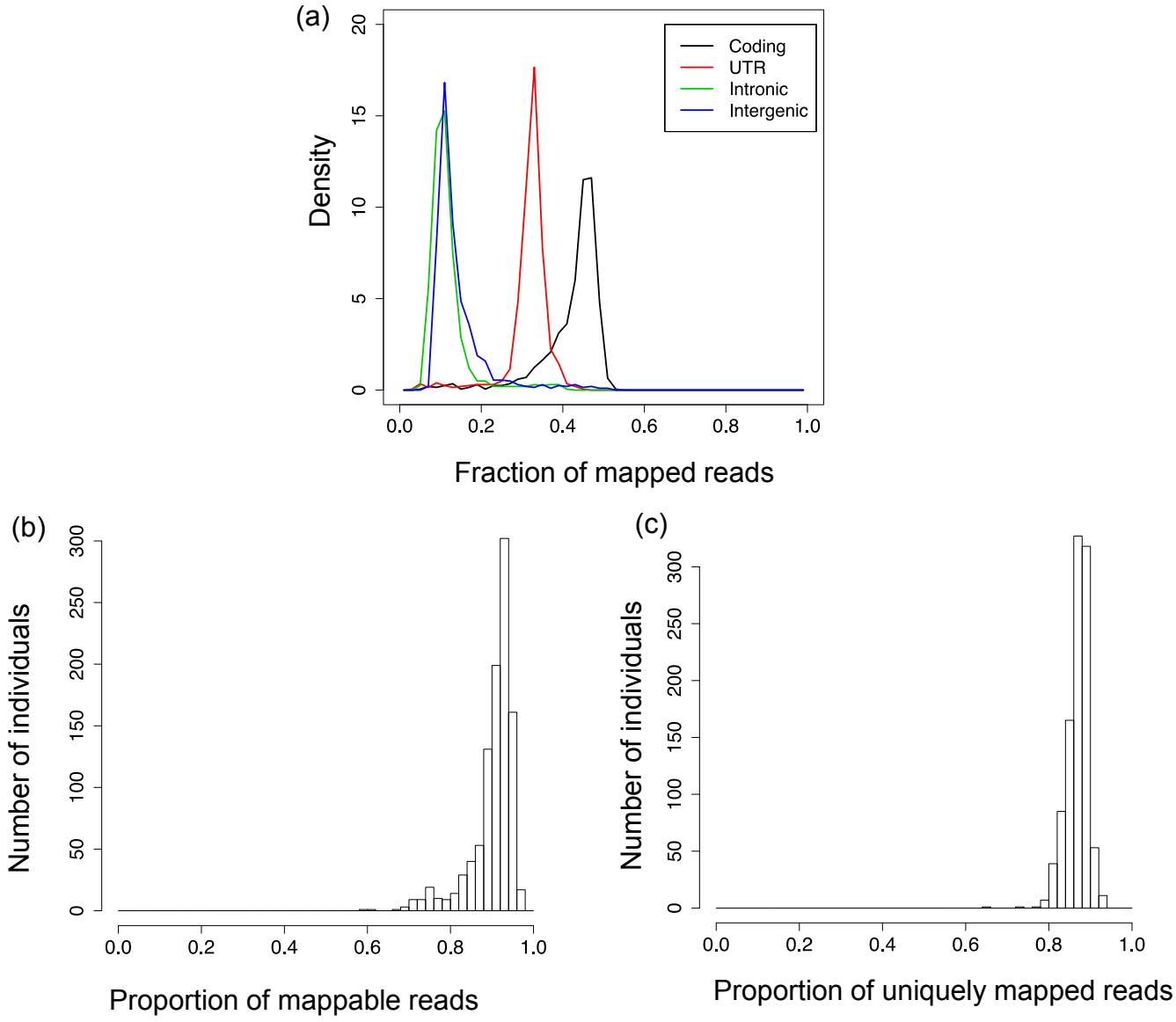
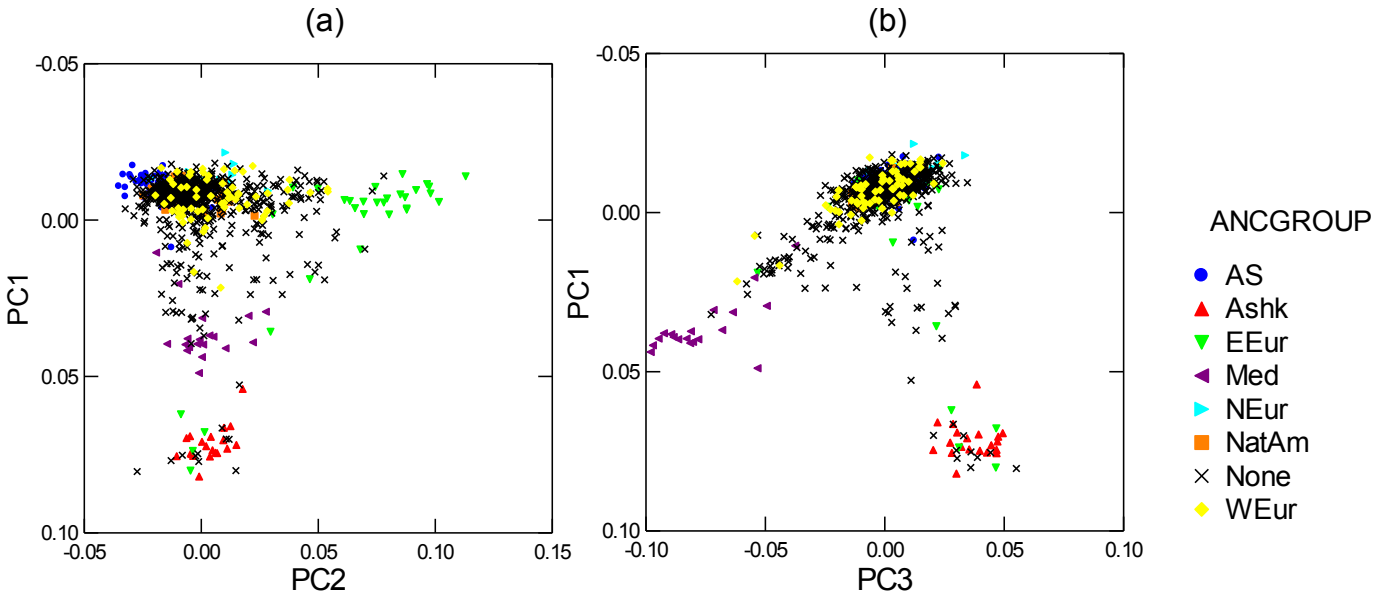


Figure S3. Mappability and distribution of mapped bases. (a) For each individual, we computed the fraction of mapped reads in coding regions (black), UTRs (red), introns (green) or intergenic regions (blue). This figure shows the distribution of fraction of mapped reads in each of these regions. As expected, majority of the mapped bases are within the coding regions or UTRs, while ~10% of the bases are within introns or intergenic regions. (b) Histogram of proportion of mappable reads in each individual. (c) Histogram of proportion of uniquely mapped reads (among the reads that were mapped) in each individual. As shown, in the majority of the individuals, at least 80% of the mapped reads were mapped uniquely.



(c)

ANCGROUP	Frequency	Cumulative Frequency	Percent	Cumulative Percent
Anglo-Saxon	104	104	11.05	11.05
Ashkenazi	20	124	2.13	13.18
Eastern Eur	31	155	3.29	16.47
Mediterranean	19	174	2.02	18.49
Northern Eur	14	188	1.49	19.98
Native Amer	28	216	2.98	22.95
None	662	878	70.35	93.3
Western Eur	63	941	6.7	100

Figure S4. **Ancestry and Principal Components of genotype data.** The plot shows Principal Component (PC) 1 and 2 scores for 941 individuals with genotype data, of which 279 reported that 3 or 4 of their grandparents were of the same ethnic background, as shown in the table above; the predominant ancestry of these individuals is indicated in the legend, while the other 662 are labeled "None" (no known predominant ancestry). (a) PC1 reflects a North (here, more negative) to South gradient with Anglo-Saxons and Northern Europeans (Scandinavians) at the North end and Ashkenazi Jews at the South end, with Mediterranean (Italians, Greeks) in between. PC2 reflects West to East (non-Jewish Slavic/Russian). Note that, consistent with our previous observations in similar samples, individuals with self-reported predominantly Native American ancestry had PC scores in the main cluster of Western European ancestries, probably reflecting a reporting bias (i.e., over-estimation of the proportion of Native American ancestry in the family). (b) The plot shows that PC3 separated Ashkenazi from Mediterranean ancestry. PCs 1-5 were used to correct expression data for population structure.

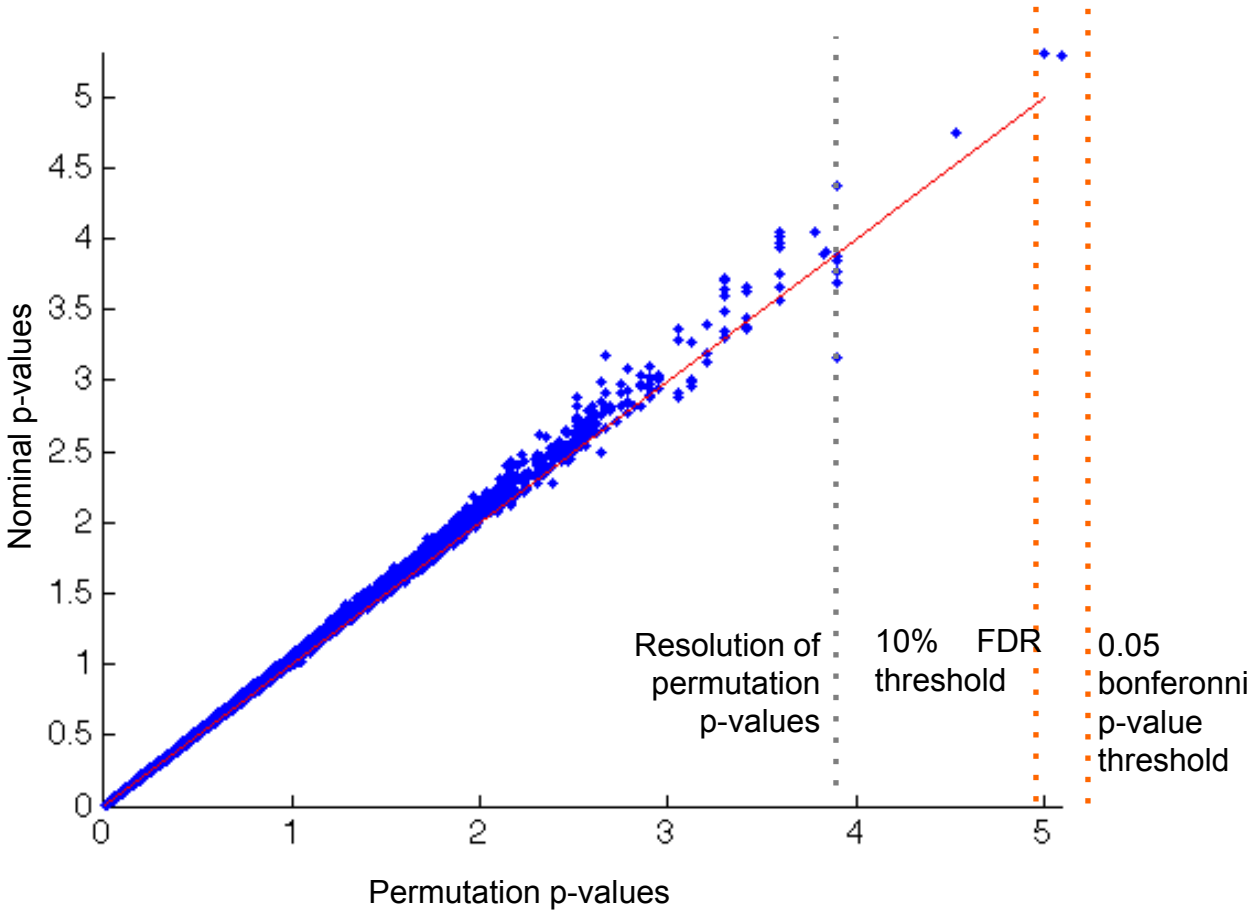


Figure S5. Permutation p-values for LRT. Figure shows the estimate of p-values for associations between each gene and MDD status when accounting for 10 expression PCs in addition to all the known covariates. We performed 8,000 permutations of phenotype initially for each gene, and 1,000,000 permutations for 3 genes for which we did not have enough resolution to assign p-values using 8000 permutations.

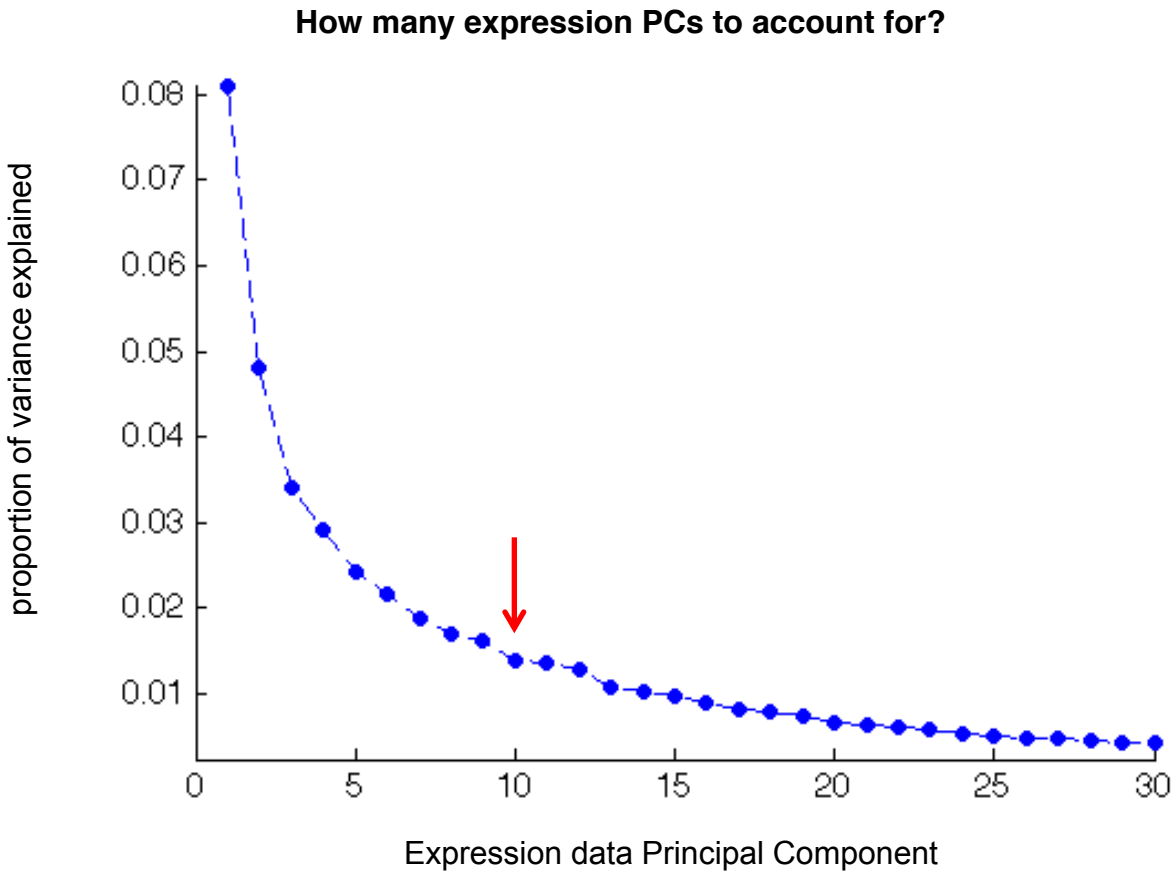


Figure S6. Proportion of variance explained and expression PCs. Figure shows the proportion of variance (out of 1) (PVE) of expression data that is explained by each of the top 30 expression principal components (PCs).The PCs were computed on normalized data, where we removed the effects of technical factors (i.e., not phenotypic variables) by using ridge regression (Table S1). Based on visual inspection of this plot, we decided to account for the first 10 PCs, as after this point PVE starts to slowly plateau. The top PCs likely represent broad trends not representative of specific variability associated with MDD status (Figure S9).

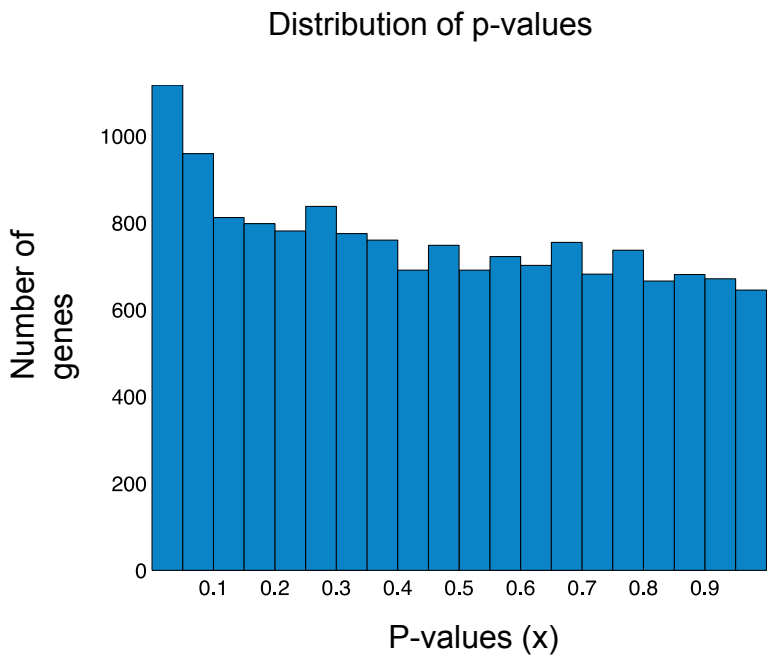


Figure S7. P-value distribution for univariate associations between MDD and each gene. p-value distribution for association between expression levels of genes and MDD status (accounting for all covariates shown in Table S2 and 10 expression PCs). The figure shows an excess of low p-values, with π_1=0.13 estimating the proportion of true positive associations assuming a uniform null distribution.

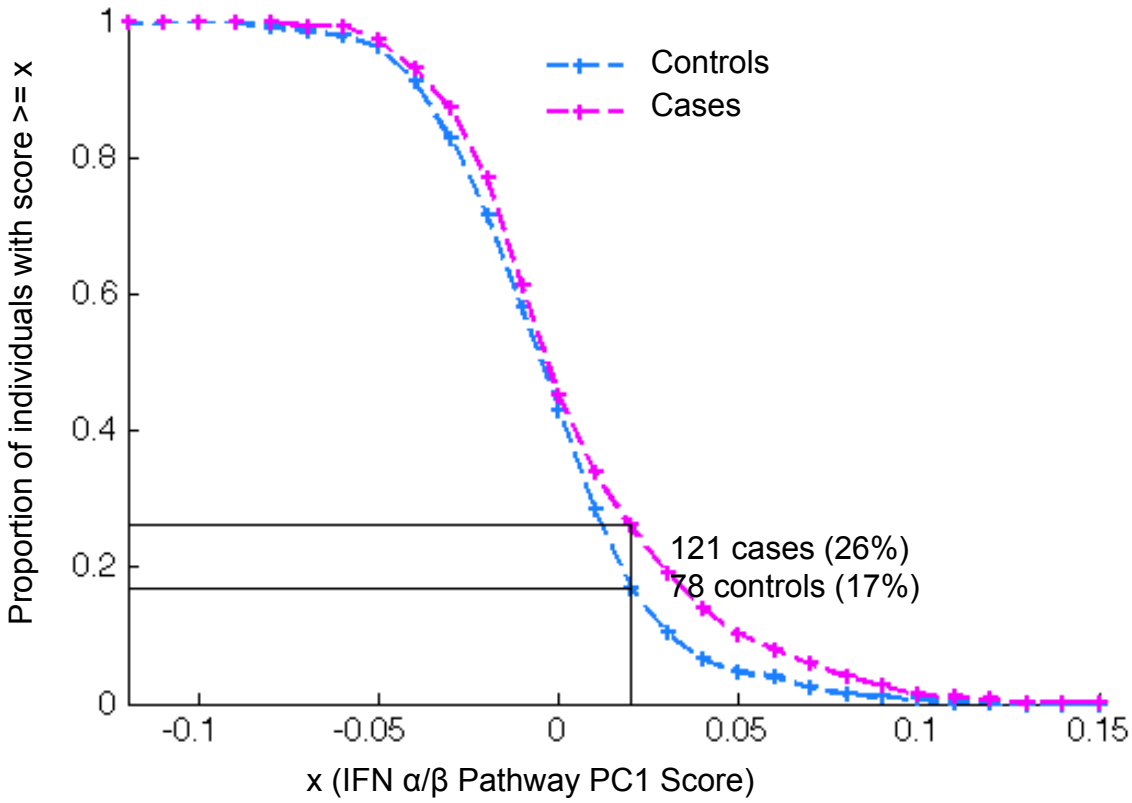


Figure S8. Cumulative distribution of IFN α/β pathway PC1 score in cases and controls. Figure shows the cumulative probability of observing cases (magenta) and controls (blue) at increasing levels of IFN pathway PC1 score. The pathway PC1 score is obtained as the first principal component of the top 20 genes (with $p<0.05$) in the IFN α/β signaling pathway (Table 3) using the normalized read counts. Y-axis depicts the proportion of individuals with scores greater than or equal to values on the x-axis.

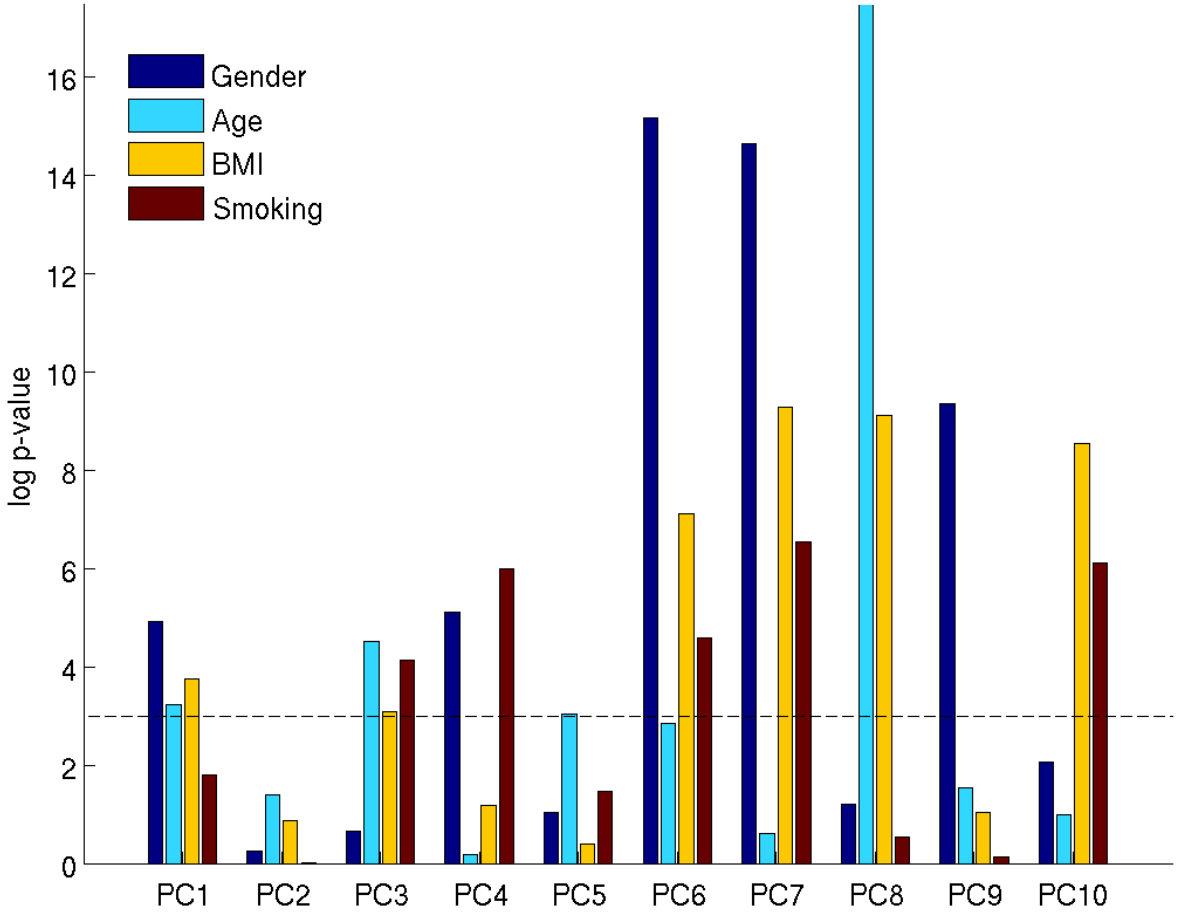
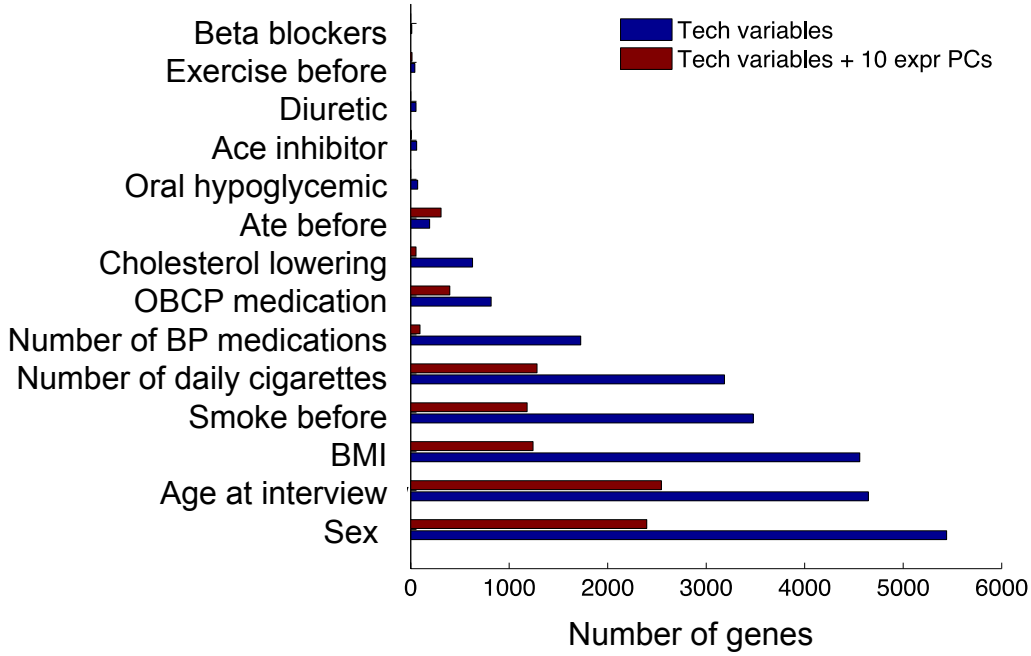


Figure S9. Association of expression PCs with confounding factors. Figure shows the association strength (log p-value) between top ten expression PCs (obtained on normalized data, Table S1) and the four major confounders of expression data (gender, age, BMI, and smoking indicator) .

(a)



(b)

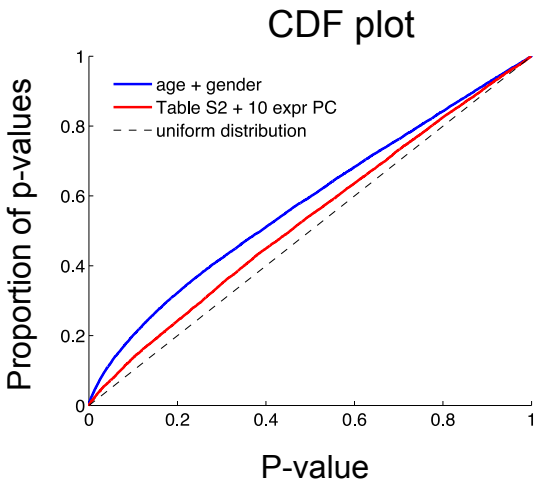


Figure S10. (a) Figure shows the number of significantly associated genes for each covariate in Table S2 (0.05 FDR) (only covariates with at least one significant association are shown). (b) Figure shows the CDF plots of the p-value distributions for (i) associations between MDD and gene expression levels while accounting for covariates in Table S2, (ii) associations between MDD and gene expression levels while only accounting for age and gender. As shown, the large deviation from the expected uniform distribution (dotted black line) for (ii) compared to (i) suggests inflation of p-values caused by confounding factors.

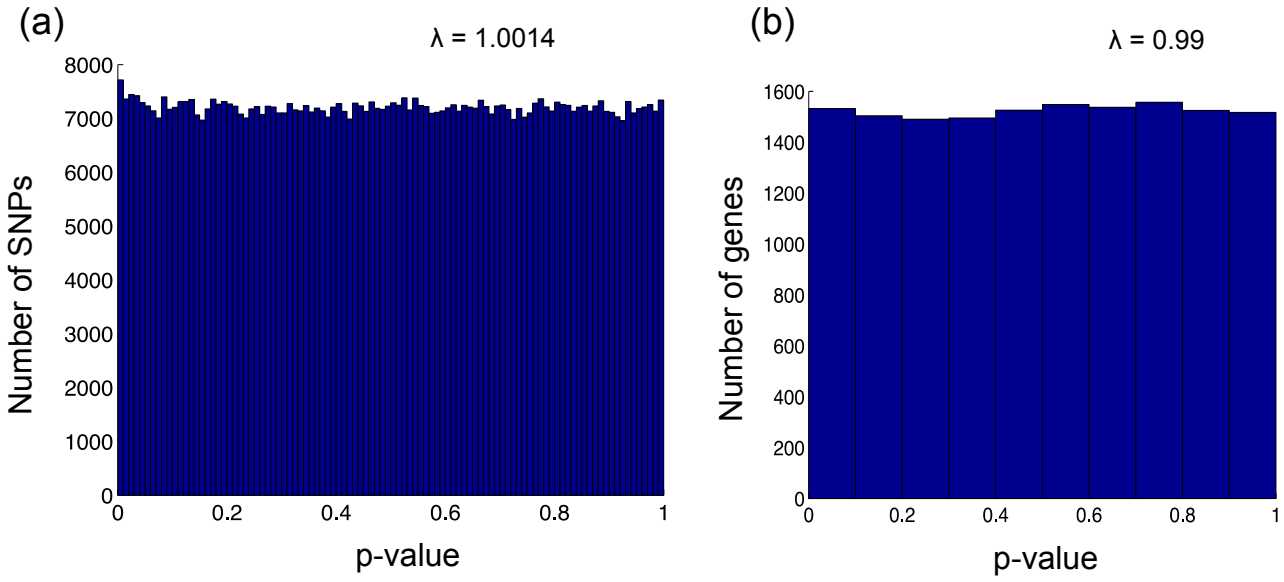


Figure S11. (a) Figure shows the distribution of GWAS p-values in this cohort. The p-values are computed using a likelihood ratio test to account for five genotype PCs (reflecting population structure). As shown, we do not observe an inflation of p-values (inflation factor λ ~= 1). (b) Figure shows the distribution of p-values for association between the SNP rs11232553 and gene expression values. As shown, we do not observe an inflation of p-values, suggesting that there this SNP does not have a broad impact on gene expression levels.

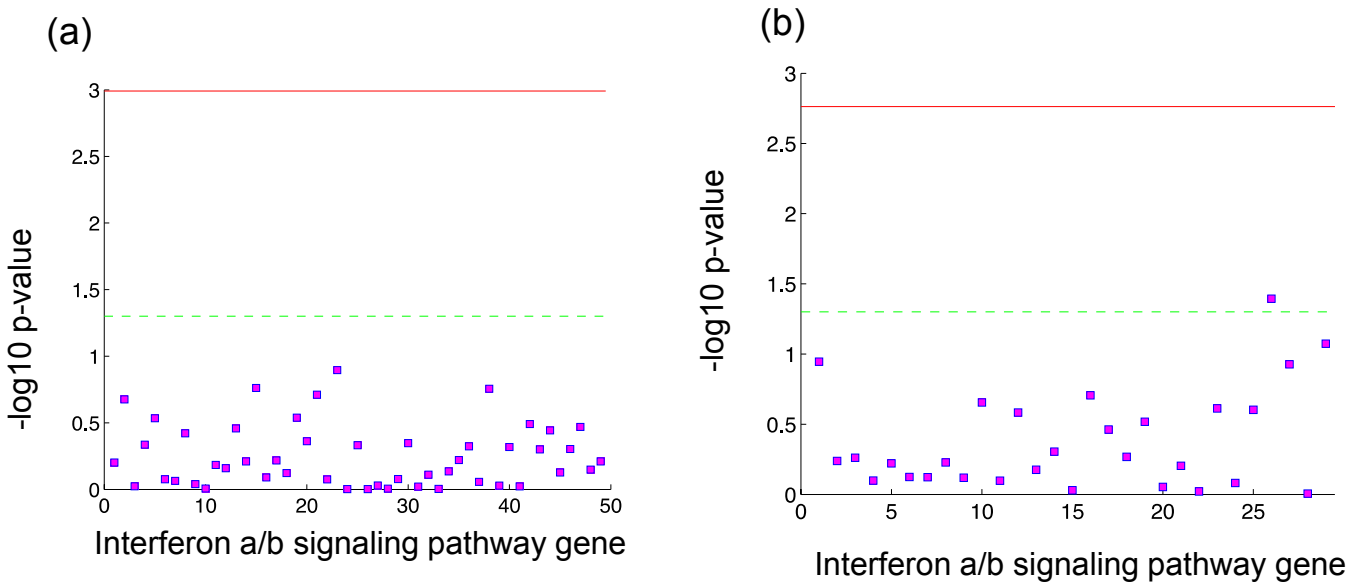


Figure S12. (a) Figure shows the log p-values for associations of rs11232553 with the 49 interferon alpha/beta signaling genes. (b) Figure shows the log p-values for associations between rs11232553 and the top 29 genes associated with MDD based on expression data. Red lines depict the corrected p-value threshold for each analysis. Green lines depict the nominal p-value threshold (-log10 of 0.05).

Childhood trauma scale

This scale was developed for the GenRED study by D. Levinson in collaboration with Dr. Elliot Nelson (Washington University at St. Louis). It is based on screening items from the National Comorbidity Survey, plus screening items for PTSD in reaction to traumatic experiences. Subjects in the present study completed this survey as part of the online screen, after being determined to be eligible for, and giving consent to be contacted about, having blood drawn and being interviewed for the study.

Before the age of 18, how often did any of the following things happen to you when you did not want it to happen:

(Items 1-3 were scored as 1-5 based on the following response choices: Never, Once, 2-5 times, 6-10 times, More than 10 times):

1. Someone outside your household physically attacked or assaulted you, threatened you with a weapon or held you captive

2. Someone touched parts of your body in a sexual way, or had you touch parts of the person in a sexual way

3. Someone had or attempted to have oral sex, anal sex, or sexual intercourse with you

(Items 4-8 were scored as 1-4 based on the following response choices: Never, Rarely, Sometimes, Frequently):

4. Your mother, father or another adult household member hurt you on purpose (for example, beat, choked, kicked, cut or burned you)

5. You observed your parents (or other caretakers) screaming at each other in anger or being physically aggressive with each other or with others

6.Parents (or other caretakers) screamed or yelled at you when you did not deserve it, or called you stupid, lazy or other names that upset you

7. Your parents failed to make sure that you were going to school, or to know what you were doing when they were not around, or to care who your friends were.

8.Your parents failed to comfort you when you were upset.

As a result of any of those childhood experiences (NO or YES):

9. Did you ever have to avoid thoughts or feelings that reminded you of this kind of experience?

10. Did you ever have physical reactions when reminded of this kind of experience?

Smoking questionnaire (to derive Fagerstrom equivalent score):

1. Have you smoked more than 100 cigarettes in your lifetime? (Yes, No) (If yes, skip to 3a).

2. Have you ever smoked a whole cigarette? (Yes, No)

2a. How many cigarettes have you smoked in your lifetime? (1-100)

For the next set of questions, think back to the time in your life when you smoked the most.

3a. Did you smoke cigarettes on a daily basis? (Yes, No)

3b. How many cigarettes did you smoke on a typical day? (0-500)

3c. During this time when you smoked the most, how soon after waking did you smoke your first cigarette?

(1) 5 minutes or less
(2) 6 to 30 minutes
(3) 31 to 60 minutes
(4) More than 60 minutes

3d. Did you find it difficult to refrain from smoking in places where it is forbidden, e.g. in church, at the library, in the cinema etc.? (Yes, No)

3e. During this period, which one cigarette would you have hated most to give up?

(1) 1st one in the A.M.
(2) All others

3f. During this period, did you smoke more frequently during the first hours after waking than during the rest of the day? (Yes, No)

3g. During this period, did you smoke if you are so ill that you are in bed most of the day? (Yes, No)

4. Do you currently smoke cigarettes? (Yes, No)

4a. How many cigarettes do you smoke in a typical day? (1-500)

4b. How long has it been since you quit smoking? (__year and __months)

References

58. Sanders AR, Levinson DF, Duan J, Dennis JM, Li R, Kendler KS, et al. The Internet-based MGS2 control sample: self report of mental illness. The American journal of psychiatry. 2010;167(7):854-65. Epub 2010/06/03.
59. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81(3):559-75. Epub 2007/08/19.
60. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature. 2009;460(7256):753-7. Epub 2009/07/03.
61. Raison CL, Borisov AS, Majer M, Drake DF, Pagnoni G, Woolwine BJ, et al. Activation of central nervous system inflammatory pathways by interferon-alpha: relationship to monoamines and depression. Biological psychiatry. 2009;65(4):296-303. Epub 2008/09/20.
62. Ripke S, Wray NR, Lewis CM, Hamilton SP, Weissman MM, Breen G, et al. A mega-analysis of genome-wide association studies for major depressive disorder. Molecular psychiatry. 2013;18(4):497-511. Epub 2012/04/05.
63. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006;38(8):904-9. Epub 2006/07/25.
64. Mazin P, Xiong J, Liu X, Yan Z, Zhang X, Li M, et al. Widespread splicing changes in human brain development and aging. Molecular systems biology. 2013;9:633. Epub 2013/01/24.
65. Fisher R. Statistical Methods for Research Worker1925.
66. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome biology. 2011;12(3):R22. Epub 2011/03/18.
67. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(16):9440-5. Epub 2003/07/29.