nature genetics

Heritability and genomics of gene expression in peripheral blood

Fred A Wright[1–3,13], Patrick F Sullivan[4,13], Andrew I Brooks[5], Fei Zou[6], Wei Sun[6], Kai Xia[6], Vered Madar[6], Rick Jansen[7], Wonil Chung[6], Yi-Hui Zhou[1,2], Abdel Abdellaoui[8], Sandra Batista[9], Casey Butler[9], Guanhua Chen[6], Ting-Huei Chen[6], David D'Ambrosio[10], Paul Gallins[4], Min Jin Ha[6], Jouke Jan Hottenga[8], Shunping Huang[9], Mathijs Kattenberg[8], Jaspreet Kochar[10], Christel M Middeldorp[8], Ani Qu[10], Andrey Shabalin[11], Jay Tischfield[5], Laura Todd[4], Jung-Ying Tzeng[1,2], Gerard van Grootheest[7], Jacqueline M Vink[8], Qi Wang[10], Wei Wang[12], Weibo Wang[9], Gonneke Willemsen[8], Johannes H Smit[7], Eco J de Geus[8], Zhaoyu Yin[6], Brenda W J H Penninx[7] & Dorret I Boomsma[8]

We assessed gene expression profiles in 2,752 twins, using a classic twin design to quantify expression heritability and quantitative trait loci (eQTLs) in peripheral blood. The most highly heritable genes (~777) were grouped into distinct expression clusters, enriched in gene-poor regions, associated with specific gene function or ontology classes, and strongly associated with disease designation. The design enabled a comparison of twin-based heritability to estimates based on dizygotic identity-by-descent sharing and distant genetic relatedness. Consideration of sampling variation suggests that previous heritability estimates have been upwardly biased. Genotyping of 2,494 twins enabled powerful identification of eQTLs, which we further examined in a replication set of 1,895 unrelated subjects. A large number of non-redundant local eQTLs (6,756) met replication criteria, whereas a relatively small number of distant eQTLs (165) met quality control and replication standards. Our results provide a new resource toward understanding the genetic control of transcription.

Determining the biological relevance of findings from genome-wide association studies (GWAS) has emerged as a major challenge for complex trait analysis, as over 90% of significant associations are noncoding. Several lines of evidence suggest that genetic variation implicated in GWAS alters transcription[1–3]. eQTLs[4,5] overlap markedly with GWAS-identified SNPs, both collectively[6–8] and for specific traits (for example, height, adiposity, cardiovascular risk factors, chemotherapy-induced cytotoxicity, autism, schizophrenia and Crohn's disease)[9–16]. An estimated 55% of eQTL SNPs lie in DNase I hypersensitivity sites, and 77% of significant GWAS SNPs are in or correlated with these sites[2,17,18]. Although understanding of eQTLs has progressed rapidly, important questions remain. Most eQTL catalogs are incomplete, and few studies have had sample sizes of $n > 1{,}000$ (refs. 15,19,20), although $n > 3{,}000$ may be necessary for more complete eQTL identification[21]. Many eQTLs do not replicate, even using the same HapMap lymphoblastoid cell lines (LCLs) under standardized procedures[19]. Replication of distant (*trans*) eQTLs has been particularly elusive[22]. Potential sources of variation include tissue type[8,10,23], ancestry[7], winner's curse and batch effects[5,7,24–26], and cell heterogeneity[27,28].

To achieve large sample sizes in humans, tissues must be accessible. An attractive choice is peripheral venous blood, although most but not all[20,29] human blood-derived eQTL studies have used LCLs. However, gene expression differs between LCLs and peripheral blood[30], and LCLs can be influenced by factors such as Epstein-Barr virus (EBV) copy number and growth rates[31]. The Multiple Tissue Human Expression Resource (MuTHER) LCL study of expression in female twins found a large impact of the common 'environment' shared by twins: 32% of transcripts showed common environmental effects of >30%, compared to 2% in adipose and 8% in skin[8]. The authors attributed this dramatic effect to correlated sample handling rather than to environmental exposures shared by twins, suggesting possible biases with LCLs.

Despite these challenges, quantifying human transcriptomic heritability is important. Although genes with genome-wide significant eQTLs are by definition 'heritable', additional polygenic variation may be widespread and fail to reach statistical significance by standard

[1]Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA. [2]Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA. [3]Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina, USA. [4]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [5]Department of Genetics, Rutgers University, New Brunswick, New Jersey, USA. [6]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [7]Department of Psychiatry, VU Medical Center, Amsterdam, The Netherlands. [8]Department of Biological Psychology, VU University, Amsterdam, The Netherlands. [9]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [10]Environmental and Occupational Health Sciences Institute, Rutgers University, New Brunswick, New Jersey, USA. [11]Department of Pharmacotherapy and Outcomes Science, Virginia Commonwealth University, Richmond, Virginia, USA. [12]Department of Computer Science, University of California, Los Angeles, Los Angeles, California, USA. [13]These authors contributed equally to this work. Correspondence should be addressed to F.A.W. (fred_wright@ncsu.edu) or P.F.S. (pfsulliv@email.unc.edu).

Received 28 October 2012; accepted 14 March 2014; published online 13 April 2014; doi:10.1038/ng.2951

genotype-expression association. Genes with substantial polygenic variation may also be subject to unique selection pressures not apparent from the analysis of local eQTLs. The classical twin design, contrasting resemblance in monozygotic twin pairs to that in dizygotic twin pairs, offers distinct advantages in the interpretability and efficiency of heritability estimation[32].

To address these questions, we conducted a combined study of twin heritability of expression and eQTLs that is the largest yet reported (3.4 times the size of the next largest twin eQTL report[8,15,30]), providing high resolution. We assessed gene expression in peripheral blood, with careful attention to sample collection, cell type heterogeneity, bias and control of experimental error. Our goals were to (i) describe and evaluate the heritability of all transcripts measured in peripheral blood; (ii) identify a comprehensive list of local and distant eQTLs and evaluate their characteristics and replicability; and (iii) assess the biomedical relevance of the identified eQTLs.

RESULTS

Twin-based heritability in the peripheral blood transcriptome

We first report the heritability of steady-state transcription in peripheral blood for 43,628 probe sets from 18,392 genes from 2,752 individual twins in the Netherlands Twin Registry (NTR; **Table 1**). The U219 platform includes alternate 3′ sequences of well-annotated genes, and we refer to each of the probe sets as a 'transcript' (1–18 transcripts per gene, mean of 2.4). We performed careful annotation for the platform, which compares favorably to RNA sequencing (RNA-seq) (**Supplementary Note**)[33]. Subjects were from 1,444 twin pairs (both members of 1,308 pairs, 95.1% of subjects and 1 member from 136 pairs). The 1,308 complete pairs consisted of 690 monozygotic pairs (52.8%; 209 male and 481 female monozygotic pairs) and 618 dizygotic pairs (47.2%; 110 male, 256 female and 252 opposite-sex dizygotic pairs). Expression quality control included zygosity and sex confirmation, randomization for sex and zygosity balance, sample identity checks and removal of low-quality samples. Primary analyses were based on robust multi-array average (RMA) expression estimates, filtered to exclude probes containing SNPs or mapping non-uniquely, with each transcript transformed to an exact normal distribution for robust analysis.

The **Supplementary Note** lists ~140 covariates used, including blood cell counts and the genotypes of blood count–associated SNPs. We computed the proportion of variance explained (R^2) attributable to covariates and the effect of covariate control on heritability (h^2) and variance explained by common (c^2) and unique (e^2) environment, where these values were measured using a covariance (ACE) model that includes additive genetic, common or shared environment, and non-shared environment terms (**Supplementary Fig. 1**). Variance components were not constrained to be positive, so the model would be unbiased for h^2 estimation, and to indicate whether genetic non-additive effects (dominance) might be present (by estimating c^2 as negative). Covariate correction notably increased evidence for highly heritable transcripts, whereas no transcript was significant for c^2 values (**Supplementary Fig. 1b–e**), in contrast to the MuTHER study[8].

Table 1 Demography of 2,752 subjects from 1,444 twin pairs for twin-based heritability analyses

Variable	Median or proportion	Quartiles
Age (years)	32	28–39
Body mass index (kg/m^2)	23.3	21.3–25.8
White blood cell count (10^9/l)	6.3	5.3–7.4
Hematocrit (fraction)	0.42	0.40–0.45
Female sex	0.658	
Blood draw between 7:00 and 11:00 a.m.	0.940	
Fasting at time of blood draw	0.947	
Current smoker	0.216	
Alcohol user (12 drinks/year)	0.771	

Quartiles, interquartile range.

Figure 1a shows a *P*-value Manhattan plot for twin-based h^2 estimation for 18,392 genes (selecting for each gene the transcript with the largest h^2 value), based on twin zygosity comparisons. The h^2 value had mean ± s.d. of 0.101 ± 0.142 (0.138 ± 0.153 for expressed genes), with maximum estimated h^2 = 0.905. We conservatively highlight 777 genes with significant heritability (q<0.05; 4.2% of the genes on the microarray), applying *k*-means clustering and analysis of genomic location. The 777 genes yielded 9 expression clusters (**Fig. 1b** and **Supplementary Table 1**). Mean within-cluster expression correlation r ranged from 0.46 to 0.006. Cluster identity was supported by significantly ($P < 0.05$) higher connectivity in protein-protein interaction databases[34] and Gene Ontology (GO) pathways[35] (**Supplementary Table 1**). Numerous clustered genes showed expression patterns similar to those observed in other tissues, including brain[35], suggesting broader tissue relevance. Regional clustering indicated enrichment for immune function (**Supplementary Table 2**; for example, IgG Fc fragment receptors encoded at chr. 1: 161–162 Mb and the major histocompatibility complex (MHC) region at chr. 6: 31–33 Mb), whereas other regions showed fewer heritable genes (for example, the neuronal protocadherin gene cluster at chr. 5: 140–141 Mb and epidermal keratin gene clusters on chr. 17: 39–40 Mb and chr. 21: 31–32 Mb). Heritability was strongly associated with mean expression ($r = 0.356$, $P < 1 \times 10^{-200}$; **Fig. 1c**), with a striking increase above an array-specific detection threshold, showing detectable expression for 21,971 transcripts (50.3%).

We next compared h^2 values for all genes to multiple external 'predictors' (refs. 1,36–44) using an enrichment statistic rigorously evaluated under permutations of twin zygosity (**Table 2**). Heritability was strongly associated with expression mean and variance. Regional GC content was negatively associated with h^2 after correction for mean expression. This negative association was surprising, as GC content ±5 kb from the transcription start site (TSS) was positively correlated with gene density ($r = 0.40$), and each correlated modestly with mean expression ($r = 0.11$ and 0.10, respectively). Accordingly, after correction for mean expression, the negative association with gene density was even stronger (**Fig. 2a** and **Table 2**). Genes with recent evolutionary acceleration in primates and humans[42] showed significant ($P = 7.11 \times 10^{-5}$ and 3.73×10^{-5}) positive association with h^2 after correction for mean expression (**Fig. 2b** and **Table 2**). HomoloGene conservation was highly significant ($P = 2.00 \times 10^{-17}$), although it was attenuated after correction. Associations between h^2 and numerous Kyoto Encyclopedia of Genes and Genomes (KEGG) and GO pathways were also highly significant (**Supplementary Table 3**). Interestingly, all pathway associations with h^2 were positive, except for two related to sensory perception and smell (GO:0050907 and GO:0050911, respectively).

To investigate disease relevance, we used the National Human Genome Research Institute (NHGRI) GWAS catalog (17 July 2013)[1], identifying the nearest gene (GWAS genes) for each of 3,628 significantly disease-associated SNPs ($P \leq 5 \times 10^{-8}$), for a total of 2,343 GWAS genes. There was a highly significantly positive association between heritability and GWAS genes (**Fig. 2b** and **Table 2**). Enrichment remained elevated for genes that were nearby but not necessarily the closest to the GWAS-identified SNP, and genes with numerous nearby GWAS SNPs were especially heritable (**Supplementary Fig. 2**). Enrichment was attenuated by removing chromosome 6 genes (including the MHC region) and for immune-related diseases[43]

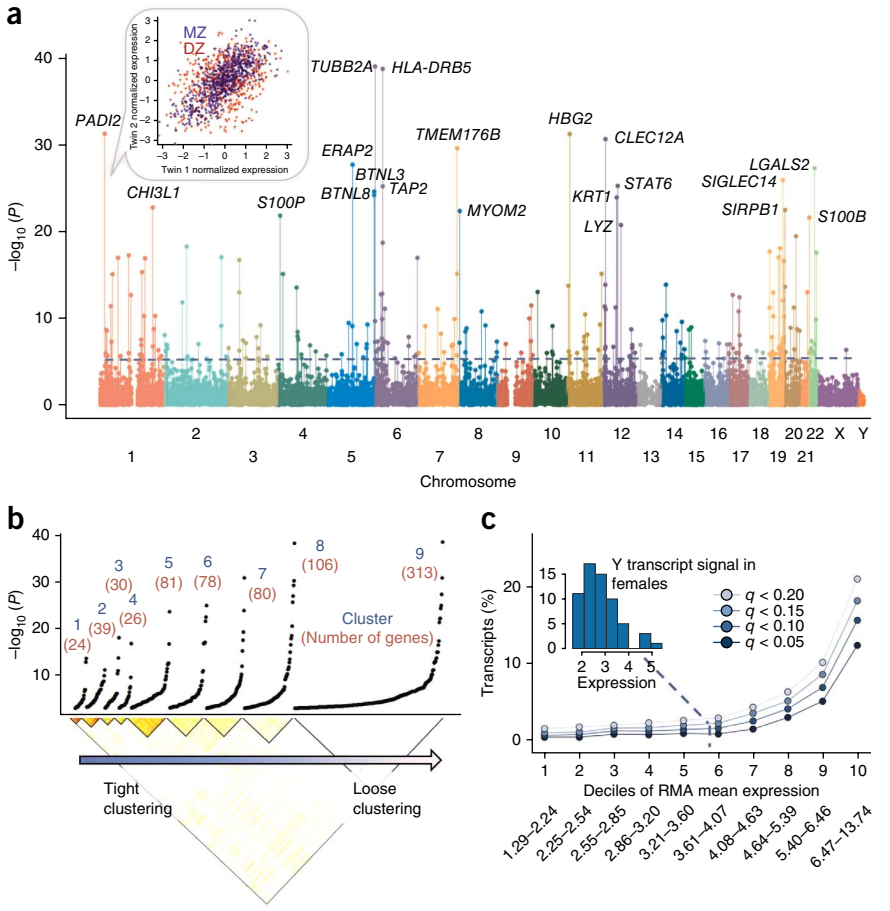

Figure 1 Transcriptome-wide estimates of heritability based on 2,752 twins. (**a**) Manhattan plot of heritability *P* values for the transcript with the highest h^2 estimate for each of 18,392 genes. The inset (*PADI2*) shows that the evidence for heritability is based on higher correlation between monozygotic pairs (MZ) than between dizygotic pairs (DZ). The dashed line marks the threshold for genes with $q < 0.05$. (**b**) Clustering of 777 genes with $q < 0.05$ for h^2 estimates. The most heritable genes belong to the cluster with the lowest intergene correlation, but many significant genes belong to clusters with high intergene correlation. (**c**) Among 43,628 transcripts, the significant proportion (in terms of FDR *q* value) is dependent on mean transcript expression, increasing rapidly for transcripts above an approximate detection threshold (RMA expression ≥3.584, determined as the 90th percentile of chromosome Y RMA 'expression' in females).

(**Supplementary Table 4**). GWAS phenotypes included ones relevant to blood and immunity along with the central nervous system, the bowel, cancers and morphological traits. Given that GWAS genes were designated only on the basis of proximity to NHGRI-listed SNPs, these results may reflect an even stronger true tendency of disease-causing genes to be highly heritable (**Supplementary Fig. 2**). These results are complementary to observations that disease-associated SNPs show eQTL enrichment[6]. Additionally, the Online Mendelian Inheritance in Man (OMIM) database shows similar heritability enrichment, even though NHGRI GWAS and OMIM only partly overlap (of genes in either list, 10% are in both). The OMIM genes with significant heritability ($q < 0.05$) are also quite diverse, further supporting the potential relevance of peripheral blood to other tissues and developmental processes (**Supplementary Table 5**). Moreover, evolutionary associations are consistent with the observation that heritability is necessary for responsiveness to selection[45].

We emphasize that these results do not imply causality, and, in particular, disease associations should be interpreted with caution. Enrichment of disease-associated heritability may reflect other underlying sources of commonality but still point to transcription as an important intermediary in disease risk.

Local genetic contributions and bias in h^2 estimation

After genotyping quality control and imputation, 8.3 million SNPs were available for eQTL mapping in 2,494 individual twins (90.4% of the expression data set). We evaluated multiple predictors of heritability, including association r^2 values based on the most significant local SNP within 1 Mb, r^2 values for the top distant SNP, local SNP heritability estimation based on genetic relatedness among unrelated subjects using Genome-wide Complex Trait Analysis (GCTA)[46] and variance-component results from complete local identity-by-descent inference among the dizygotic pairs (local IBD). We computed ratios of each component to the overall h^2 estimate (**Supplementary Fig. 3**). Mean and median values for $r^2_{\text{local SNP}} / h^2$ (0.04 and 0.09, respectively) were similar to those reported in the MuTHER study[8], whereas the $h^2_{\text{local IBD}} / h^2$ ratio was higher (median = 0.11, mean = 0.30), consistent with higher explained variation when the total local contribution was considered. However, in published studies, estimates have been complicated by bias and variability in h^2 estimation. MuTHER reported mean h^2 values in expressed genes of 0.16 (skin), 0.21 (LCLs) and 0.26 (adipose), with >20% of expressed genes displaying $h^2 > 0.3$ (ref. 8). Our study, although much larger, produced lower values of 0.14 and 12.3%. Each of our h^2 estimates should be unbiased, as we allowed for negative estimates (even if $h^2 \geq 0$), whereas variance-component methods[8] can produce bias by forcing estimates to be non-negative, and sampling variability further complicates the view.

To more definitively assess the true extent of transcriptomic heritability for our study, we modeled true h^2 values as following a gamma distribution, with sampling variation determined by the ACE model. The result (**Fig. 3a**) was a shrunken distribution with a similar mean h^2 value but markedly less variation. The model estimated that the true proportion of expressed genes with heritability of >0.3 was actually only 7.9%. With high heritability thresholds, differing results across studies can appear dramatic—whereas the MuTHER report estimated >700 expressed genes in both skin and LCLs with heritability of >0.5, we estimate the true number in our study as ~100. The studies differed in tissue type and platform (the MuTHER study used the Illumina HT-12 BeadChip platform), NTR mean age was ~20 years younger and the NTR samples included both sexes. Results when age was removed as a covariate (**Supplementary Note**) suggested that it was not an important heritability determinant in NTR. However, the important effect of sampling variation has not been fully explored. First, we assessed the gamma fit by artificially adding sampling error to the true distribution, showing that it fit our estimated h^2 distribution (**Fig. 3a**). A similar approach quantified the impact of sample size (**Supplementary Note**), again using the gamma model

Table 2 Predictors of high heritability expression levels

Predictor	Mean h^2 change	Enrichment z	P	Expression-corrected enrichment z	P
Mean expression		**11.25**	**2.43×10^{-29}**	–	–
Variance in expression		**14.14**	**2.23×10^{-45}**	**14.89**	**4.02×10^{-50}**
GC content, +5 kb of TSS		–1.42	0.155	**–5.33**	**9.60×10^{-8}**
GC content, –5 kb of TSS		–0.72	0.471	**–5.00**	**5.73×10^{-7}**
DHS near TSS[a]		**9.45**	**3.55×10^{-21}**	**4.01**	**6.00×10^{-5}**
DHS near TSS, blood		**8.87**	**7.02×10^{-16}**	1.30	0.195
Gene density[b]		**–6.98**	**2.98×10^{-12}**	**–10.85**	**2.09×10^{-27}**
Gene size[c]		**8.07**	**7.02×10^{-16}**	**11.30**	**1.27×10^{-29}**
Local recombination rate[d]		0.73	0.464	3.01	0.0026
Size of LD block[e]		–0.05	0.959	–0.49	0.622
Gene conservation score[f]		**8.49**	**2.00×10^{-17}**	1.14	0.255
Genes under selection (185)[g]	0.013	1.60	0.109	1.82	0.068
Genes under positive selection (549)[h]	0.007	1.32	0.186	1.78	0.074
Genes under balancing selection (47)[i]	0.042	2.65	0.0081	2.83	0.0046
Genes under adaptive selection (174)[j]	0.019	2.26	0.024	1.13	0.260
Human accelerated genes (161)[k]	0.024	3.05	0.0023	**4.12**	**3.73×10^{-5}**
Primate accelerated genes (137)[k]	0.024	2.86	0.0042	**3.97**	**7.11×10^{-5}**
NHGRI GWAS catalog (2,343)[l]	**0.018**	**7.42**	**1.14×10^{-13}**	**7.52**	**5.53×10^{-14}**
NHGRI, chr. 6 genes removed (2,142)	**0.016**	**6.06**	**1.37×10^{-9}**	**6.42**	**1.36×10^{-10}**
NHGRI, immune diseases (720)[m]	**0.032**	**7.22**	**5.02×10^{-13}**	**5.77**	**7.99×10^{-9}**
NHGRI, non-immune diseases (1,623)	**0.011**	**3.71**	**0.0002**	**4.88**	**1.03×10^{-6}**
OMIM disease entries (3,089)[n]	**0.018**	**8.87**	**7.63×10^{-19}**	**7.54**	**4.81×10^{-14}**
NHGRI + OMIM (4,809)	**0.019**	**10.81**	**2.96×10^{-27}**	**9.84**	**7.81×10^{-23}**

TSS, transcription start site; DHS, DNase I–hypersensitive site; NHGRI, National Human Genome Research Institute; GWAS, genome-wide association study; OMIM, Online Mendelian Inheritance in Man. Values in bold correspond to $P < 0.0022$, for Bonferroni significance at $\alpha = 0.05$ for the 23 tests in each of uncorrected and corrected analyses.

[a]From Encyclopedia of DNA Elements (ENCODE) Duke UCSC tracks. [b]Defined as the reversed rank of the variance of the base-pair positions of the gene and two flanking genes. [c]The end transcription base-pair position minus the start transcription base-pair position. [d]deCODE sex-averaged standardized recombination maps in 10-kb bins. [e]Linkage disequilibrium block boundaries as described in the **Supplementary Note**. [f]NCBI HomoloGene (Build 66) score, defined as the ratio of the number of appearances in other organisms to the total of 21. [g]Ref. 36, genes with the property shown in parentheses. [h]Refs. 36–38. [i]Ref. 39. [j]Ref. 35. [k]Ref. 41. [l]Ref. 1, for SNPs with $P < 5 \times 10^{-8}$. [m]Following classification in ref. 42. [n]Ref. 43.

obtained from NTR but inflating the sampling variation to reflect the smaller MuTHER sample size. The resulting estimated h^2 distribution was similar to that reported by MuTHER (**Fig. 3b**). We suggest that, despite other differences between the studies, much of the apparent differences may be attributable to sample size effects. Analysis of the recent Brisbane Systems Genetics twin study[47] suggested a similar effect of sampling variation (**Supplementary Fig. 4a** and **Supplementary Note**). Although we conclude that the underlying heritability in all of these studies may be comparable, this is a distributional statement, and larger sample sizes are desirable in terms of accuracy. Accuracy prediction as a function of sample size is shown in **Supplementary Figure 4b**—even with the NTR sample size, we predict that the rank correlation between true and estimated heritability is only slightly greater than 0.5.

We applied similar modeling approaches to local IBD–based h^2 values (**Fig. 3c,d**), estimating the proportion of total h^2 attributable to local genetic variation. Our mean local IBD–derived h^2 estimate was 0.03, with $\text{mean}(h^2_{\text{local IBD}})/\text{mean}(h^2) = 0.23$. The value for this ratio was somewhat lower than those reported by MuTHER (>0.30), which is perhaps partly attributable to the focus of their study on genes with higher total heritability[8]. A definitive statement of average per-gene ratios ($h^2_{\text{local IBD}}/h^2$) will require more complex modeling to handle correlation structures in the measurements and underlying true structure. However, the results from our large sample support the view that local genetic variation explains only a minority of transcriptomic heritability and much of the unexplained variation is among genes with modest h^2 estimates. A regression approach (**Supplementary Fig. 5**) showed that ~35% of the variation in estimated h^2 values could be explained by the predictors.

eQTL analyses of peripheral blood

We next analyzed genotypes as predictors of transcription (a GWAS for each transcript) for 2,494 twins, using an REML (restricted maximum-likelihood) model accounting for twin status and covariates. eQTLs within 1 Mb upstream of the TSS and 1 Mb downstream of the transcription end site of a gene were classified as 'local', and all others were classified as 'distant', with separate false discovery rate (FDR) control. Genes with at least one local eQTL ($q < 0.01$) had significantly higher expression levels and heritability ($P < 1 \times 10^{-200}$ for both).

The effect of sample size on local eQTL identification is shown in **Figure 4a**, which includes nearly all published blood-derived eQTL studies[7,8,15,20,31,48–52] (comparisons to the large meta-analysis in ref. 29 are described separately), the full NTR data (n = 2,494) and random subsamples of our data. We reanalyzed the data sets using a common quality control pipeline on inverse quantile-normalized data[19] (except where unavailable[8,15]). For comparison, we selected a set of unrelated twins (1,263 individuals) and performed local

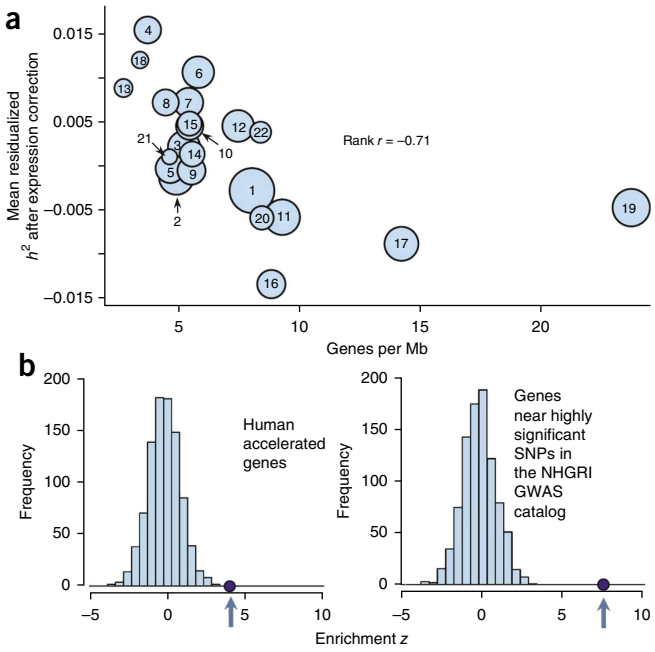

Figure 2 Gene density and other predictors of heritability, using 2,616 paired twins and 18,392 genes. (**a**) Mean h^2 estimates (corrected for gene expression levels) versus density of protein-coding genes per autosome, showing that heritability is considerably higher for gene-poor chromosomes. Plot symbol area is proportional to the number of genes present on the array per chromosome. (**b**) Histograms of the permuted enrichment z statistics for two predictors listed and defined in **Table 2**. Observed values (blue circles) are extreme compared to the permutations.

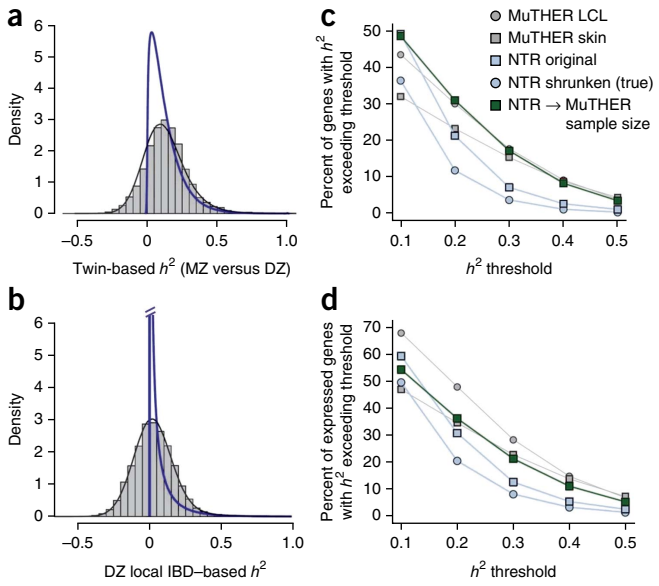

Figure 3 Apparent heritability and local IBD effects versus true underlying distributions. (**a**) For twin-based h^2 estimates (n = 2,752; 8,818 expressed genes shown), subtracting the effects of sampling variation produces an estimated true distribution (blue curve). Resimulating from the fitted true assumed distribution closely approximates the observed h^2 estimates (black curve). (**b**) Analogous expressed gene results for local IBD effect estimation. (**c**) Proportions of all 18,392 genes exceeding h^2 thresholds for observed data and for the estimated true h^2 distribution. The MuTHER study (n = 856) reported many more extreme h^2 values, but the observation is consistent with greater sampling variation due to smaller sample size. (**d**) Analogous plot using only expressed genes from both studies.

eQTL mapping on random subsets of varying sample size, using fewer covariates (no blood counts or SNPs) and ~600,000 genotyped SNPs. We also evaluated our robustness approaches (normal quantile transformation and normal quantile transformation with SNP minor allele frequency (MAF) of >0.005 or >0.01 in each subsample). For local eQTLs, there was little difference among the transformations.

There is considerable interstudy variability in the number of significant eQTLs (**Fig. 4a**), even with consistent quality control and analysis[19,31]. With increasing sample size, it seems that most expressed genes (>10,000) show evidence of local eQTL influence in peripheral blood. For NTR, the number of genes with significant eQTLs (q < 0.01) was 11,834, and, after employing final quality control steps, there were 9,640 significant genes. Replication was examined in 1,895 unrelated samples from the Netherlands Study of Depression and Anxiety (NESDA), which had a similar sex distribution (68% female) and age range (from 18 through 65 years). Reproducibility of eQTLs between NTR and the 1,895 unrelated NESDA samples is shown in **Supplementary Figure 6**, and enrichment and deficits of regulatory features for local eQTL SNPs are shown in **Supplementary Figure 7**.

Of the 9,640 genes with local eQTLs in NTR (at least 1 SNP with q < 0.01), 9,148 (94.9%) replicated (q < 0.1) in NESDA (using a less stringent replication q-value threshold to allow for winner's curse attenuation). This approach was not intended to control the per-gene FDR but to focus on genes with the greatest evidence of replication. Of the genes with the strongest evidence of local eQTLs in NTR (q < 0.001), 6,756 of 6,941 (97.3%) replicated in NESDA. There was strong overlap ($P = 1 \times 10^{-180}$) of genes with local eQTLs in the full NTR sample, with the same gene having a local eQTL in a meta-analysis of HapMap LCL studies[19]. For genes with local eQTLs (q < 0.1) in the LCL meta-analysis, 56.1% (2,417/4,306) also had significant local eQTLs in NTR. Genes that replicated had smaller meta-analysis q values ($P = 1 \times 10^{-18}$), along with higher expression ($P = 2 \times 10^{-119}$) and higher heritability ($P = 8 \times 10^{-131}$) in NTR. The lack of overlap among smaller HapMap samples is likely an example of winner's curse: considering two larger studies[15,20], among the genes annotated in all three studies, replication in NTR was 66.8% (2,799/4,189 genes) and 77.2% (3,404/4,412 genes), respectively (**Fig. 4b**). Similarly, for local gene-SNP pairs with q < 0.05 from the peripheral blood eQTL

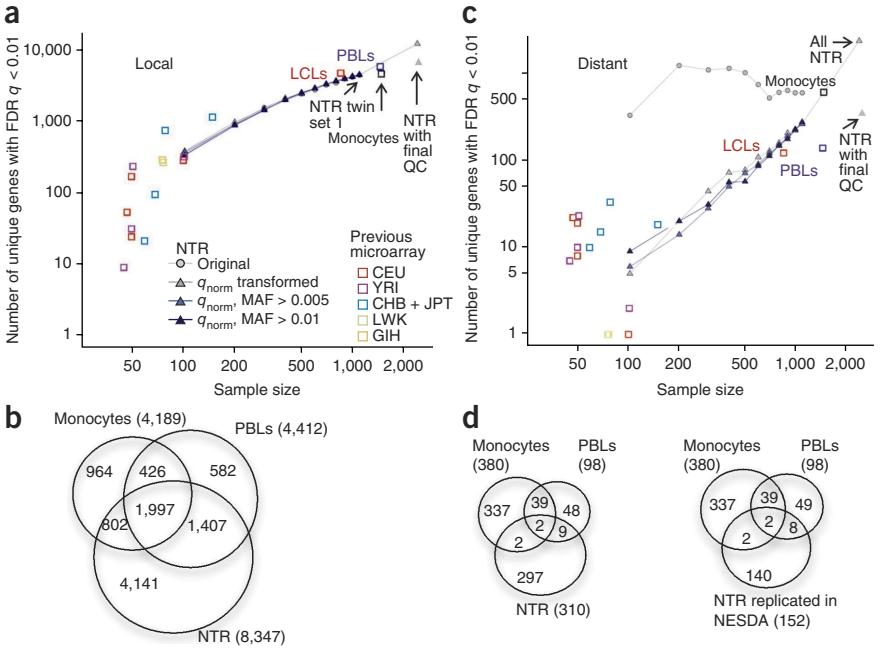

Figure 4 Comparison and replication of eQTL results. (**a**) Number of unique genes with evidence of local association (q < 0.01; within 1 Mb of gene), depicted for published leukocyte eQTL studies (LCLs[8], monocytes[15] and peripheral blood leukocytes (PBLs)[20]), as well as subsampling of NTR data (PBLs) using only genotyped markers and moderate quality control (n = 2,494; 43,628 transcripts examined). Sample sizes are corrected for the number of covariates used. The "NTR with final QC" value applies q < 0.001. q_{norm} refers to the rank inverse quantile-normal transformed expression data. Ancestries are CEU (Northern and Western European), YRI (Yoruba in Ibadan, Nigeria), CHB (Han Chinese in Beijing, China), JPT (Japanese in Tokyo, Japan), LWK (Luhya in Webuye, Kenya) and GIH (Gujarati Indians in Houston, Texas). (**b**) Overlap of local eQTL findings with two other large blood studies, at q < 0.01. (**c**) Number of unique genes with evidence (q < 0.01) for distant association (>1 Mb from gene). The implausible non-monotone pattern for NTR on original expression values shows the importance of robust association methods. Using final quality control on NTR data and q < 0.001 drops the number of distant eQTLs from over 800 to ~300. The results suggest that many distant associations remain to be discovered, but careful quality control is essential. (**d**) Overlap of distant eQTL findings (q < 0.001) with previous studies (within 1 Mb of gene).

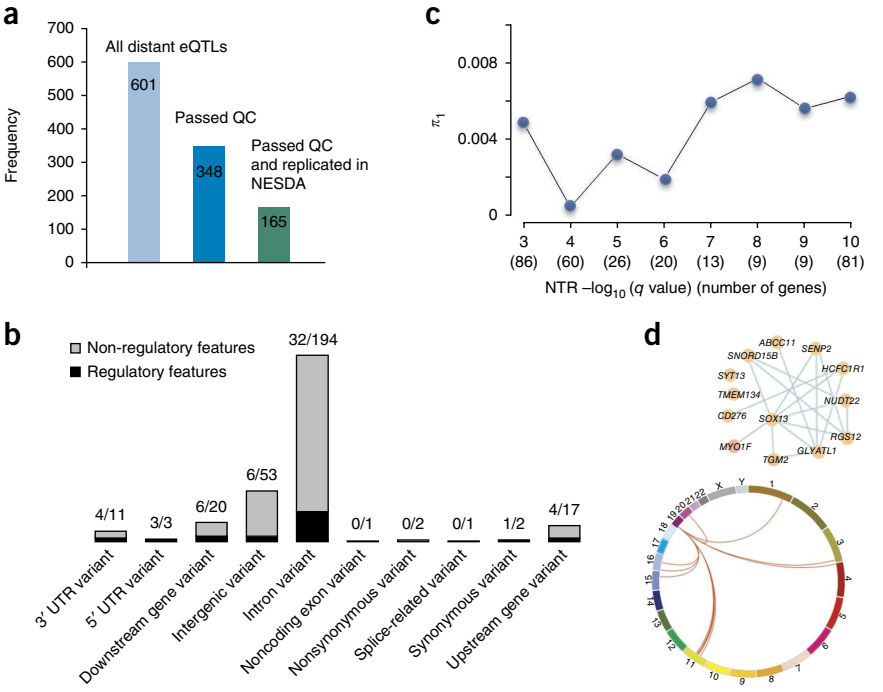

Figure 5 Properties of distant eQTLs. (**a**) In total, 348 eQTLs (gene-SNP pairs) were significant ($q < 0.001$) and passed the quality control procedures; of these, 165 replicated ($q < 0.1$) in 1,895 NESDA individuals. (**b**) We examined 304 SNPs in significant eQTLs for overlap with regulatory features, including DNase I/FAIRE (formaldehyde-assisted isolation of regulatory elements) and transfactor-binding sites, using the Ensembl Variant Effect Predictor (version 2.8)[54]. Most features were not enriched, although the three SNPs annotated as 5′ UTR variants all overlap regulatory features, representing a significant enrichment ($P < 0.01$) compared to the total 18.4% overlap of distant eQTL SNPs with regulatory features. The overall proportion of regulatory features is 56/304 = 0.184. (**c**) The π_1 value represents the estimated proportion of the transcriptome influenced by the 304 SNPs that passed quality control in significant eQTLs. Across all significant bins, the cumulative proportion is only ~3%. (**d**) A distant eQTL hotspot on chromosome 19 was associated with the expression of 12 distant genes and 1 local gene (*MYO1F*). The partial correlation graph suggests that *MYO1F* expression is independent of the expression of the other distant genes, given the expression of the transcription factor *SOX13*.

meta-analysis of Westra *et al.*[29] (n = 5,311), the estimated true discovery rates in NTR and NESDA were 59.6% and 59.7%, respectively (**Supplementary Fig. 8** and **Supplementary Note**).

Characteristics of distant eQTLs

Robust distant eQTL results (**Fig. 4c**, expression transformed to an exact normal) were again consistent with published studies, roughly linear (log-log scale) with sample size[15]. For NTR, we obtained a robust set of 348 distant eQTLs by applying stricter significance criteria ($q < 0.001$) followed by additional careful quality control. Extrapolating to larger sample sizes, we anticipate the identification of <1,000 replicating eQTLs, even for sample sizes exceeding 5,000. Overlap of genes with significant distant eQTLs ($q < 0.001$) among the large studies is shown in **Figure 4d**, with much lower overlap for distant eQTLs than for local eQTLs. For significant distant gene-SNP pairs from Westra *et al.*[29] (n = 5,311), the estimated true discovery rates in NTR and NESDA were 23.1% and 23.0%, respectively (**Supplementary Fig. 8**).

Our 601 distant eQTLs with $q < 0.001$ (**Fig. 5**) involved 581 genes and 538 non-redundant SNPs (for each gene, only the most significant SNP per chromosome was retained). We applied additional quality control to these highly significant distant eQTLs (**Supplementary Note**), reducing the number of eQTLs to 348 (57.9%), of which 165 (47.4%) replicated in NESDA ($q < 0.01$) (**Fig. 5**, **Supplementary Fig. 6** and **Supplementary Table 6**). Genes in the 348 eQTLs were analyzed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) P values for KEGG and GO enrichment, which have been shown to be liberal[53], but only GO:0003779 (actin binding) was declared significant ($P = 0.0001$, FDR $q = 0.046$).

The 304 SNPs among the 348 eQTLs were examined using the Ensembl Variant Effect Predictor[54] (**Supplementary Table 6**), with each SNP assigned on the basis of the most severe predicted consequence. Most of the SNPs were intronic, and the next most frequent SNPs included intergenic SNPs, SNPs upstream or downstream of protein-coding sequence, and exonic SNPs (**Fig. 5b**). The 53 intergenic SNPs had the lowest rate of overlap with regulatory features or replication in NESDA (**Supplementary Fig. 9**). SNPs in upstream or downstream sequences were more likely to overlap with regulatory elements, and SNPs in intronic or exonic regions were more likely to replicate in NESDA. Only 6 of the 348 distant eQTLs were exonic, suggesting that they influence expression rather than modify proteins, consistent with our finding that these distant eQTL SNPs are more likely to be local eQTLs than randomly selected comparable SNPs (**Supplementary Fig. 10**).

We next sought to identify eQTL hotspots (SNPs influencing numerous transcripts). We grouped the 304 distant eQTL SNPs into 203 regional clusters (**Supplementary Fig. 11**), of which 160 included only 1 SNP and the other 43 spanned 2 kb to 2 Mb of DNA (median size of 89 kb). Eleven clusters associated with ≥6 genes were considered potential hotspots, showing agreement with analogous results from NESDA. For each of the 304 SNPs, we estimated the proportion of associated transcripts, using NESDA data to avoid selection bias. These values were <0.008 for a wide range of NTR eQTL strengths (**Fig. 5c**), many times lower than reported for the three tissues in the MuTHER study[8]. We conclude that eQTL hotspots and significant distant eQTLs influence relatively few genes in peripheral blood.

We analyzed each putative eQTL hotspot using a penalized partial correlation graph[55]. We highlight a network where a distant eQTL located on chromosome 19 is also a local eQTL of *MYO1F*. Given the expression of *SOX13*, *MYO1F* expression is independent of that in other distant eQTL genes (**Fig. 5d**), suggesting that eQTL signals are mediated by *SOX13*. *MYO1F* encodes unconventional myosins, which bind to membranous compartments and serve in intracellular movements. *SOX13* encodes a transcription factor that modulates the WNT-TCF signaling pathway[56], and several other distant eQTL genes are involved in cellular signaling (for example, *TMEM134*, *RGS12* and *SYT13*). Additional network estimation was performed for the replicating hotspots (**Supplementary Fig. 12**), but the relatively few genes influenced by hotspots or distant eQTLs suggest that such networks do not have a predominant role in steady-state transcription in peripheral blood.

Biomedical relevance

This catalog of eQTLs can be used to generate *in silico* hypotheses for biomedical follow-up using peripheral blood as a proxy tissue. Using the NHGRI GWAS catalog[1], after stringent filtering ($P < 1 \times 10^{-8}$), there were significant results for 3,415 SNPs, 498 traits and 4,167 SNP-trait pairs from 927 reports. The greatest numbers of SNPs associated with a trait or disease were found for height ($n = 248$), high-density lipoprotein (HDL) cholesterol ($n = 92$), Crohn's disease ($n = 155$), type 2 diabetes ($n = 98$) and ulcerative colitis ($n = 81$). The extended MHC region (chr. 6: 25–34 Mb, 0.3% of the genome) was the second most gene-dense region of the genome and contained the greatest number of SNPs implicated by GWAS (6.8%). Of the 4,167 SNP-trait pairs implicated by GWAS, 534 (12.8%) were part of a local eQTL (either directly or via a proxy SNP with $r^2 > 0.5$).

To complement the analyses, we evaluated genes cataloged in OMIM (downloaded 17 July 2013)[44]. Of the 3,118 genes in OMIM, 74.4% were part of a SNP-gene local eQTL pair ($q < 0.05$). These included many genes related to immune and hematological abnormalities, muscular dystrophy (21 genes) and genes implicated in nervous system diseases. Examples include Alzheimer's disease (*APP* and *PSEN2*), deafness (42 genes), amyotrophic lateral sclerosis (15 genes), Charcot-Marie-Tooth disease (25 genes), epilepsies (21 genes) and candidate genes for schizophrenia (*DISC1*, *DAOA* and *RGS4*). Of the 517 genes implicated in mendelian autism spectrum disorders[57] or mental retardation[44,58,59], 69.6% were part of a local eQTL SNP-gene pair. Of the 3,294 genes with a copy number variant implicated in autism spectrum disorders[57], developmental delay[60] or a psychiatric disorder[61], 72.4% were part of a local eQTL SNP-gene pair.

Finally, we combined heritability predictors and gene-disease designations into several multiple regressions (**Supplementary Table 7**). Predictors were as shown in **Table 2**, with the addition of eQTL evidence (best local and distant r^2 values), chromosome 6 (human leukocyte antigen (HLA) genes), chromosome 19 (an outlier in gene density analysis), the X chromosome (under-represented in GWAS) and a blood DNase I hypersensitivity–gene conservation interaction (identified in exploratory analyses). eQTL evidence alone (top local and distant SNPs) explained 23.9% of the variation in h^2 estimates, and the full model explained 32.9%. h^2 estimates remained significantly predictive of OMIM and NHGRI GWAS disease status except for the smaller sets of NHGRI-cataloged genes subdivided by immune designation, even when the best local and distant eQTLs were no longer significant. Gene conservation was highly predictive of OMIM status. Gene density showed strong negative association with disease status, but this effect was attenuated for OMIM. NHGRI disease status was significantly enriched ($P = 1.70 \times 10^{-10}$; **Supplementary Table 7**) for chromosome 6 loci and showed a deficit on the X chromosome ($P = 1.87 \times 10^{-13}$), which we attribute to the neglect of the X chromosome in GWAS[62]. OMIM showed enrichment of the X chromosome, consistent with the importance of X-linked disorders in medical genetics.

DISCUSSION

We have established clear patterns underlying the heritability of steady-state gene transcription in peripheral blood and have demonstrated strong connections to disease annotation. The use of peripheral blood enables further investigation of immune-related diseases[63] but may also be useful for other tissues. Our results supply mechanistic hypotheses that can be evaluated in subsequent experiments. In comparisons across four mouse tissues, we found that genes expressed in multiple tissues tended to have *cis* regulatory elements (J.J. Crowley, V. Zhabotynsky, W. Sun, S. Huang, I.K. Pakatci *et al.* unpublished data).

Examination of h^2 estimates relative to gene density builds upon a literature demonstrating that essential genes expressed in many tissues can occur in dense clusters of high expression, including instances of transcriptional colocalization[64–66]. Essential genes (those that when mutated cause lethality at or before birth) identified in mouse mutagenesis screens show high linkage conservation[67], and intergenic regions in humans have higher SNP densities than in introns, along with higher rates of neutral polymorphisms[68,69]. Our observations seem concordant with these reports, whether selection directly inhibits heritability in gene-dense regions or the inhibition is due to the relative paucity of genotype variation in such regions.

The ability of h^2 estimates to predict OMIM and NHGRI designations may suggest new approaches to augment association mapping, as current approaches generally focus on the sequence context of associated SNPs rather than the genes themselves. The ability to detect heritability only in expressed genes somewhat complicates interpretation, given the higher average expression in high-density clusters and the lack of information for genes not expressed in this tissue. Critically, full elucidation of these relationships may be possible only with careful cross-tissue eQTL analysis of a large number of individuals[15].

URLs. The fundamental data for this report (Affymetrix 6.0 and U219) are available by application to the database of Genotypes and Phenotypes (dbGaP; http://www.ncbi.nlm.nih.gov/gap/). Summary results are available in the seeQTL browser (http://gbrowse.csbio.unc.edu/cgi-bin/gb2/gbrowse/seeqtl/) or in downloadable GFF3 files (https://pgc.unc.edu/). deCODE sex-averaged standardized recombination maps, http://www.decode.com/addendum/; phased genotype calls on 379 European samples from the 1000 Genomes Project, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521; GENCODE, http://www.gencodegenes.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Expression data and genotypes are available in dbGaP under accession phs000486.v1.p1.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The work described in this paper was funded by the US National Institute of Mental Health (RC2 MH089951, principal investigator P.F.S.) as part of the American Recovery and Reinvestment Act of 2009. Transport, extraction and preparation of the NTR samples were carried out under a supplement to the NIMH Center for Collaborative Genomics Research on Mental Disorders (U24 MH068457, principal investigator J.T.). We thank T. Lehner (National Institute of Mental Health) for his support. Additional analytic support was provided by grants R01 MH090936, R01 GM074175 and P42 ES005948 and by a Gillings Innovations Award. The Netherlands Study of Depression and Anxiety (NESDA) and the Netherlands Twin Register (NTR) were funded by the Netherlands Organization for Scientific Research (MagW/ZonMW; grants 904-61-090, 985-10-002, 904-61-193, 480-04-004, 400-05-717 and 912-100-20; Spinozapremie 56-464-14192; and Geestkracht program grant 10-000-1002), the Center for Medical Systems Biology (CMSB2; NWO Genomics), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL), the VU University EMGO+ Institute for Health and Care Research and the Neuroscience Campus Amsterdam, NBIC/BioAssist/RK (2008.024), the European Science Foundation (EU/QLRT-2001-01254), the European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE (HEALTH-F4-2007-201413) and the European Research Council (ERC; 230374).

AUTHOR CONTRIBUTIONS

Study design and writing: F.A.W., P.F.S., A.I.B., F.Z., W.S., B.W.J.H.P. and D.I.B. Analysis: F.A.W., P.F.S., F.Z., W.S., K.X., V.M., R.J., W.C., Y.-H.Z., A.A., G.C., T.-H.C., P.G., M.J.H., J.J.H., S.H., M.K., J.K., C.M.M., A.Q., A.S., J.-Y.T., Q.W., Wei Wang, Weibo Wang, G.W., J.H.S., E.J.d.G. and Z.Y. Genomic assays: A.I.B., D.D., J.T., A.Q. and Q.W. Phenotype collection: G.v.G. and J.M.V. Project management: L.T. Database design and management: S.B. and C.B.

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details are available in the online version of the paper.

1. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
2. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
3. Hardy, J. Psychiatric genetics: are we there yet? *JAMA Psychiatry* **70**, 569–570 (2013).
4. Majewski, J. & Pastinen, T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.* **27**, 72–79 (2011).
5. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10**, 184–194 (2009).
6. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
7. Stranger, B.E. *et al.* Patterns of *cis* regulatory variation in diverse human populations. *PLoS Genet.* **8**, e1002639 (2012).
8. Grundberg, E. *et al.* Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).
9. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
10. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
11. de Jong, S. *et al.* Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes. *Eur. J. Hum. Genet.* **20**, 1004–1008 (2012).
12. Fransen, K. *et al.* Analysis of SNPs with an effect on gene expression identifies *UBE2L3* and *BCL3* as potential new risk genes for Crohn's disease. *Hum. Mol. Genet.* **19**, 3482–3488 (2010).
13. Luo, R. *et al.* Genome-wide transcriptome profiling reveals the functional impact of rare *de novo* and recurrent CNVs in autism spectrum disorders. *Am. J. Hum. Genet.* **91**, 38–55 (2012).
14. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
15. Zeller, T. *et al.* Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* **5**, e10693 (2010).
16. Gamazon, E.R., Huang, R.S., Cox, N.J. & Dolan, M.E. Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. *Proc. Natl. Acad. Sci. USA* **107**, 9287–9292 (2010).
17. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
18. Degner, J.F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
19. Xia, K. *et al.* seeQTL: a searchable database for human eQTLs. *Bioinformatics* **28**, 451–452 (2012).
20. Fehrmann, R.S. *et al.* *Trans*-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* **7**, e1002197 (2011).
21. Min, J.L. *et al.* The use of genome-wide eQTL associations in lymphoblastoid cell lines to identify novel genetic pathways involved in complex traits. *PLoS ONE* **6**, e22070 (2011).
22. Grundberg, E. *et al.* Population genomics in a disease targeted primary cell model. *Genome Res.* **19**, 1942–1952 (2009).
23. Gibbs, J.R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6**, e1000952 (2010).
24. Leek, J.T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
25. Akey, J.M., Biswas, S., Leek, J.T. & Storey, J.D. On the design and analysis of gene expression studies in human populations. *Nat. Genet.* **39**, 807–808 (2007).
26. Innocenti, F. *et al.* Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet.* **7**, e1002078 (2011).
27. Fairfax, B.P. *et al.* Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).
28. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* **9**, e1003486 (2013).
29. Westra, H.J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
30. Powell, J.E. *et al.* Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome Res.* **22**, 456–466 (2012).
31. Choy, E. *et al.* Genetic analysis of human traits *in vitro*: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.* **4**, e1000287 (2008).
32. van Dongen, J., Slagboom, P.E., Draisma, H.H., Martin, N.G. & Boomsma, D.I. The continuing value of twin studies in the omics era. *Nat. Rev. Genet.* **13**, 640–653 (2012).
33. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41**, D48–D55 (2013).
34. Rossin, E.J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7**, e1001273 (2011).
35. Huang, W. *et al.* DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* **35**, W169–W175 (2007).
36. Grossman, S.R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883–886 (2010).
37. Nickel, G.C., Tefft, D. & Adams, M.D. Human PAML browser: a database of positive selection on human genes using phylogenetic methods. *Nucleic Acids Res.* **36**, D800–D808 (2008).
38. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170 (2005).
39. Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
40. Andrés, A.M. *et al.* Targets of balancing selection in the human genome. *Mol. Biol. Evol.* **26**, 2755–2764 (2009).
41. Grossman, S.R. *et al.* Identifying recent adaptations in large-scale genomic data. *Cell* **152**, 703–713 (2013).
42. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
43. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89**, 607–618 (2011).
44. McKusick, V.A. Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007).
45. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
46. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
47. Powell, J.E. *et al.* Congruence of additive and non-additive effects on gene expression estimated from pedigree and SNP data. *PLoS Genet.* **9**, e1003502 (2013).
48. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
49. Montgomery, S.B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
50. Pickrell, J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
51. Price, A.L. *et al.* Effects of *cis* and *trans* genetic ancestry on gene expression in African Americans. *PLoS Genet.* **4**, e1000294 (2008).
52. Spielman, R.S. *et al.* Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.* **39**, 226–231 (2007).
53. Gatti, D.M., Barry, W.T., Nobel, A.B., Rusyn, I. & Wright, F.A. Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics* **11**, 574 (2010).
54. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
55. Sun, W., Ibrahim, J.G. & Zou, F. Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. *Genetics* **185**, 349–359 (2010).
56. Marfil, V. *et al.* Interaction between Hhex and SOX13 modulates Wnt/TCF activity. *J. Biol. Chem.* **285**, 5726–5737 (2010).
57. Betancur, C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380**, 42–77 (2011).
58. Chiurazzi, P., Schwartz, C.E., Gecz, J. & Neri, G. XLMR genes: update 2007. *Eur. J. Hum. Genet.* **16**, 422–434 (2008).
59. Inlow, J.K. & Restifo, L.L. Molecular and comparative genetics of mental retardation. *Genetics* **166**, 835–881 (2004).
60. Cooper, G.M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).
61. Sullivan, P.F., Daly, M.J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet.* **13**, 537–551 (2012).
62. Wise, A.L., Gyi, L. & Manolio, T.A. eXclusion: toward integrating the X chromosome in genome-wide association analyses. *Am. J. Hum. Genet.* **92**, 643–647 (2013).
63. Xavier, R.J. & Rioux, J.D. Genome-wide association studies: a new window into immune-mediated diseases. *Nat. Rev. Immunol.* **8**, 631–643 (2008).
64. Hurst, L.D., Pal, C. & Lercher, M.J. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5**, 299–310 (2004).
65. Osborne, C.S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.* **36**, 1065–1071 (2004).
66. Sproul, D., Gilbert, N. & Bickmore, W.A. The role of chromatin structure in regulating the expression of clustered genes. *Nat. Rev. Genet.* **6**, 775–781 (2005).
67. Hentges, K.E., Pollock, D.D., Liu, B. & Justice, M.J. Regional variation in the density of essential genes in mice. *PLoS Genet.* **3**, e72 (2007).
68. Cai, J.J., Macpherson, J.M., Sella, G. & Petrov, D.A. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genet.* **5**, e1000336 (2009).
69. Davidson, S., Starkey, A. & MacKenzie, A. Evidence of uneven selective pressure on different subsets of the conserved human genome; implications for the significance of intronic and intergenic DNA. *BMC Genomics* **10**, 614 (2009).

ONLINE METHODS

Subjects and biological sampling. Subjects were ascertained and sampled using harmonized protocols from two longitudinal cohort studies, the Netherlands Twin Registry (NTR)[70] and the Netherlands Study of Depression and Anxiety (NESDA)[71]. NTR is an observational, 25-year longitudinal study of twins and their families[72–74]. The study protocol was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Center[70]. NESDA is a cohort study to investigate the long-term course and consequences of depressive and anxiety disorders and includes persons both with and without emotional disorders[71,73,74]. The study protocol was approved by the Ethical Review Board of the VU University Medical Center and, subsequently, by the local review board of each participating center[71]. Informed consent was obtained from all participants in both studies.

Peripheral venous blood samples were drawn in the morning (NTR, 7:00—11:00 a.m.; NESDA, 8:30–9:30 a.m.) after an overnight fast. For fertile women in NTR, samples were obtained on days 3–5 of their menstrual cycle or in the pill-free week if on oral contraception. Heparinized whole blood was transferred into PAXgene Blood RNA tubes (Qiagen) within 20 min (60 min for NESDA), incubated and stored at −20 °C or −30 °C (NTR). High-molecular-weight genomic DNA was isolated using PureGene DNA isolation kits (Qiagen).

Gene expression assays. Gene expression assays for NTR and NESDA were conducted at the Rutgers University Cell and DNA Repository. Total RNA was extracted at Rutgers (for NESDA, at the VU Medical Center) using the PAXgene Blood RNA MDx kit protocol in 96-well format with the BioRobot Universal System (Qiagen). RNA quality and quantity were assessed by Caliper AMS90 with HT DNA 5K/HT RNA LabChips. Samples were randomized to plates, with checks to ensure sex and zygosity balance. Co-twins were randomized without respect to relationship to avoid bias in family correlation estimates. For cDNA synthesis, 50 ng of RNA was reverse transcribed and amplified in a plate format on a Biomek FX liquid-handling robot (Beckman Coulter) using Ovation Pico WTA reagents (NuGEN). Products purified from single-primer isothermal amplification (SPIA) were fragmented and labeled with biotin (Encore Biotin Module, NuGEN), and size distributions were verified (Caliper AMS90, HT DNA 5K/RNA LabChips). Samples were hybridized to Affymetrix U219 array plates (**Supplementary Note**) to enable expression profiling in 96-sample sets. Array hybridization, washing, staining and scanning were carried out in an Affymetrix GeneTitan System according to the manufacturer's protocol.

Quality control was conducted on NTR and NESDA data in parallel. Expression data were required to pass standard Affymetrix Expression Console quality metrics before further undergoing quality control. The array superset consisted of 6,526 U219 arrays (3,516 NTR samples, 2,783 NESDA samples, divided into baseline samples and a smaller portion after 2-year follow-up, and 227 controls) on 69 plates, including 417 samples that were identified as having reduced quality ($D < -5.0$) and were rehybridized. Expression values were obtained using RMA normalization (Affymetrix Power Tools, v1.12.0). Probe sequences were mapped to the human genome (hg19) using Bowtie[75], and probes with sequences not mapping, mapping to multiple locations or intersecting a polymorphic SNP (HapMap 3 and 1000 Genomes Project data) were removed[76,77]. We mapped and annotated all Affymetrix U219 probe sets with reference to GENCODE (v14) gene models.

The large sample size enabled additional quality control metrics involving intersample comparisons. First, samples showing sex inconsistency were removed (on the basis of X-chromosome and Y-chromosome probe sets). Second, we examined the pairwise correlation matrix of expression profiles. Using r_{ij} as the correlation between arrays i and j, we computed

$$\bar{r}_i = \sum_j r_{ij} / n$$

the average correlation of array i with all others of the total n arrays. Lower $\bar{r}_i$ values correspond to lower quality and were expressed in terms of median absolute deviations $D_i = (\bar{r}_i - \bar{\bar{r}}) / \text{median}(|\bar{r}_i - \bar{\bar{r}}|)$ to provide a sense of distance from the grand correlation mean $\bar{\bar{r}}$. Third, we verified sample identity on the basis of U219 gene expression data and Affymetrix 6.0 genotypes, having previously discovered genotype-expression mismatch rates of up to 5% in published eQTL studies[19]. Briefly, 500 of the most significant SNP-transcript local eQTL pairs[19] were used to estimate a posterior probability for a match between gene expression and genotype profile (similar to in ref. 78). This approach identified sex-mismatched samples and additional samples of poor quality.

Fourth, initial analysis using unrelated participants showed the potential for spurious eQTL identification owing to expression outliers. Thus, conservatively, we transformed expression values using inverse quantile normal transformation, which results in values that precisely fit a normal distribution. These values were used for all primary analyses. Fifth, we evaluated the effects of covariates on gene expression and found significant associations for plate, hybridization well position, age at blood sampling, sex, time interval between extraction and hybridization steps, total white and red blood cell counts, hematocrit and the top five expression principal components (similar to that of surrogate variables)[79]. Imputation was performed to estimate a small proportion of missing covariates (2.1%). All heritability and eQTL analyses corrected for these covariates (93 degrees of freedom), and eQTL analyses additionally corrected for the first 3 genotype principal components.

Sixth, we observed that D values and the posterior probability of mismatch were highly correlated, and we reasoned that D values might be useful in removing additional low-quality samples. To determine the optimal threshold for D, we successively removed individual samples according to D value, and recomputed the intraclass correlation coefficient (ICC)-based estimate of heritability $2(\hat{\rho}_{MZ} - \hat{\rho}_{DZ})$ and accompanying P values[80] for all transcripts using covariate-residualized expression data. Benjamini-Hochberg FDR q values for transcripts were computed using p.adjust in R (v.2.14). Removal of 19 samples with the lowest D values resulted in the largest number of significant transcripts ($q < 0.10$; **Supplementary Note**). The optimal choice of samples to remove was largely robust to the q-value threshold in the range $q = 0.05$–0.20 and to the use of non-normalized expression data.

After expression quality control, the U219 gene expression set consisted of 2,752 NTR subjects. An additional 1,895 NESDA subjects (representing the NESDA baseline set) were used for replication in this report. Expression quality control for NESDA followed the same steps as for NTR (except that zygosity did not apply). Expression distributions for monozygotic and dizygotic twins were compared for differing mean expression (t test) and differing variances (F test for normally distributed data), performed separately within twin sets 1 and 2. No transcript showed significantly different mean expression between monozygotic and dizygotic twins, but four transcripts showed significantly different (FDR $q < 0.05$) variances. However, of these four transcripts, none showed $q < 0.05$ for h^2 estimates.

Genome-wide SNP assays. Genomic DNA was tested using 96 TaqMan SNP Genotyping assays (RUID panel) with Fluidigm 96.96 GT Dynamic Array chips, a BioMark Genetic Analysis instrument and SNP Genotyping Analysis Software (v3.0.2). After the quality, sex and identity of genomic DNA samples were verified, all samples were randomized to plates. Genotyping was conducted using Affymetrix Genome-Wide Human SNP Array 6.0 (**Supplementary Note**) according to the manufacturer's protocol. Resulting data were required to pass standard Affymetrix quality control metrics (contrast QC > 0.4) before further analysis.

SNP quality control is detailed in the **Supplementary Note**. Briefly, quality control included the removal of SNPs for non-unique probes mapping to NCBI Build 37/UCSC hg19, low MAF (<0.005, determined empirically), substantial deviation from HapMap 3 CEU (Utah residents of Northern and Western European ancestry) founder allele frequencies, deviation from Hardy-Weinberg equilibrium ($P < 1 \times 10^{-8}$) or high missingness (>0.05). Subjects were eliminated from analysis for high missingness (>0.05), outlying genome-wide homozygosity or ancestry, discrepant genetic and phenotypic sex, or twin relatedness inconsistent with monozygosity or dizygosity. Resulting genotypes were of high quality, with relatively low SNP and subject missingness (97.5 percentiles of 0.035 and 0.020, respectively). Among 714 monozygotic twin pairs, the intrapair agreement for 686,895 autosomal SNPs was 0.9985. Previous genome-wide genotyping using a Perlegen 4-chip platform was available for 2,219 subjects and 110,588 SNPs[74] and showed 0.9996 agreement with Affymetrix 6.0 genotyping.

doi:10.1038/ng.2951

Phased genotype calls on 379 European samples from the 1000 Genomes Project were used as the reference set for imputation. NTR samples were split into two unrelated sets. SNPs with call rate of <95% or Hardy-Weinberg equilibrium $P < 1 \times 10^{-9}$ were excluded. Imputation was performed using MACH. For each NTR set, MAF bins of (0.005, 0.1), (0.01, 0.03), (0.03, 0.05) and (0.05, 0.5) were defined, and within each bin an r^2 threshold was defined such that average $r^2 = 0.8$. The r^2 thresholds were 0.55, 0.4, 0.3 and 0.3, respectively. The final SNP numbers were 8.4 million for each of the twin sets, with the intersection of 8.3 million SNPs used here.

Heritability. Three methods for estimating heritability are detailed in the **Supplementary Note**. The primary approach was twin-based heritability via an REML mixed model, with random additive genetic components of variation, along with shared and individual-specific environmental effects plus selected covariates as fixed effects. Random terms were assumed to be mutually independent and normally distributed with mean of 0 and variances σ_a^2, σ_c^2 and σ_e^2. This corresponds to a standard ACE model and assumes that dizygotic twins have an average IBD proportion of 0.5 (refs. 81,82). For each transcript, the twin-based heritability and shared environmental effects were estimated, respectively, as $\hat{a}^2 = \hat{\sigma}_a^2/(\hat{\sigma}_a^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2)$ and $\hat{c}^2 = \hat{\sigma}_c^2/(\hat{\sigma}_a^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2)$. The ACE model can be fit using either variance-component maximization, constraining $\hat{a}^2$ and $\hat{c}^2$ to be non-negative or using an unconstrained general covariance structure. After establishing that results from the two approaches were highly concordant, we used the unconstrained approach to best match the intraclass correlation approach[80] used for pathway analysis. Under additive assumptions, $\hat{a}^2$ is the heritability estimate h^2, and P values are reported for the right tail (positive $\hat{a}^2$) except where noted. P values for the X chromosome were obtained using separate heritability analysis for males and females (using identical methods as for autosomes) and then combined using Fisher's method. For the analyses in **Supplementary Table 7**, h^2 values for the X chromosome were obtained by ignoring twin sex, producing an approximate average across the sexes. After calculating results for all 47,628 transcripts, a unique 'best h^2' set used the most significant transcript for each of the 18,293 genes, with FDR control applied to the best h^2 set in a manner accounting for all transcripts.

The second heritability estimation approach was dizygotic-only heritability following a constrained ACE mixed-model approach for full siblings[83]. For this approach, an REML mixed model was used to relate observed variation in true IBD proportions among dizygotic pairs to expression phenotypes. P values were obtained using likelihood ratio tests. The third approach was heritability estimated from the genetic relatedness matrix, as implemented in GCTA[46]. For this approach, we divided the NTR subjects into unrelated sets (twin set 1, n = 1,370; twin set 2, n = 1,372) and averaged the h^2 estimates from the 2 twin sets. Results showed almost no correlation with twin-based heritability (data not shown), and we reasoned that genome-wide IBD might have reduced power for those genes influenced largely locally. Thus, we ran GCTA again, using IBD estimation performed in the local region within 1 Mb of each transcript.

Local IBD analysis. Residualized expression data showed a nearly perfect normal distribution, and the bivariate normal model of Wright[84] for sibling pair IBD mapping was therefore applied to the dizygotic pairs, offering a potential improvement over the Haseman-Elston approach[8]. MERLIN[85] was run on the thinned set of markers used for stratification analysis, and probabilistic IBD estimates were produced at each marker closest to or within each gene. A full maximum-likelihood approach was applied for an additive model for the effect of each increment of IBD on dizygotic twin correlation as a function of IBD status, thus extracting maximum information. The approach in ref. 84 provides a regression coefficient, which was then converted to local h^2 equivalents as the proportion of variation in the trait explained by local IBD status.

Heritability enrichment and pathway analysis. A primary question is whether heritability associates with gene sets, pathways or quantitative gene features, which we generically refer to as heritability enrichment. We employed DAVID/EASE as a descriptive tool to investigate heritable gene clusters[35]. However, simple methods that ignore transcriptomic correlation produce very high false positive rates[53]. Furthermore, a large number of genes are heritable, necessitating 'competitive' enrichment testing[86], contrasting the heritability of each set of genes with that of a complementary set. Accordingly, we devised a rigorous testing approach for each gene set. We used a covariate-residualized version of the expression data, computing the ICC-based estimate for complete twin pairs as $h^2 = 2(\hat{\rho}_{MZ} - \hat{\rho}_{DZ})$ for all genes using the transcripts with the best h^2 values. For the observed data, this approach was highly consistent with REML estimates (r = 0.992; **Supplementary Note**). Twin zygosity status was permuted 1,000 times, and h^2 was computed for all genes for each permutation, along with the difference in mean h^2 for the gene set versus the complementary set. As this difference showed a nearly normal distribution, an enrichment z statistic was calculated as the observed difference divided by its permutation s.d., and a two-sided P value was computed assuming normality. A similar approach was used for continuous predictors in which the correlation between h^2 and the predictor was computed (with z as the correlation divided by its s.d.). By permuting only zygosity status, the enrichment z-score approach preserves mean twin pair correlations, as well as gene-gene correlations. To control for the complicating effects of mean expression, some analyses (including all KEGG and GO pathway analyses) were performed in which h^2 values were corrected for the effect of mean expression in the original and permuted data sets.

eQTL analysis. We refer to eQTLs as local (SNP-transcript associations within 1 Mb of the transcription start and end sites) or distant (the remaining findings). We prefer these terms to *cis* and *trans* designations, which connote a greater understanding of underlying mechanisms.

The REML twin-based model can be used for eQTL analysis by including SNP genotype (additive coding as copies of the minor allele) and computing the corresponding Wald statistic, in this manner properly handling covariates and twin correlation structure. This approach is computationally prohibitive for full eQTL analysis, so we used Matrix eQTL[87] to rapidly screen for local or distant eQTL relationships. To account for dependence, the full REML model was then applied to all transcript-SNP associations with nominal $P < 1 \times 10^{-5}$ (a liberal threshold for the ~3×10^{10} tests performed). Separate FDR q-value error control was performed for local and distant eQTLs. After FDR correction, it was apparent that all significant results with true REML $q < 0.10$ had indeed been captured. Some of the eQTL findings are reported in terms of unique genes, i.e., the most significant transcript-SNP combination for each gene, and in such instances the full testing multiplicity was considered.

70. Willemsen, G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res. Hum. Genet.* **13**, 231–245 (2010).
71. Penninx, B.W. *et al.* The Netherlands Study of Depression and Anxiety (NESDA): rationales, objectives and methods. *Int. J. Methods Psychiatr. Res.* **17**, 121–140 (2008).
72. Boomsma, D.I. *et al.* Netherlands Twin Register: from twins to twin families. *Twin Res. Hum. Genet.* **9**, 849–857 (2006).
73. Boomsma, D.I. *et al.* Genome-wide association of major depression: description of samples for the GAIN major depressive disorder study: NTR and NESDA Biobank Projects. *Eur. J. Hum. Genet.* **16**, 335–342 (2008).
74. Sullivan, P.F. *et al.* Genomewide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Mol. Psychiatry* **14**, 359–375 (2009).
75. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
76. Altshuler, D.M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
77. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
78. Schadt, E.E., Woo, S. & Hao, K. Bayesian method to predict individual SNP genotypes from gene expression data. *Nat. Genet.* **44**, 603–608 (2012).
79. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).
80. Falconer, D.S. & Mackay, T.F.C. *Introduction to Quantitative Genetics* (Longman Group, Ltd., London, 1996).
81. Neale, M.C. & Cardon, L.R. *Methodology for the Study of Twins and Families* (Kluwer Academic Publisher Group, Dordrecht, The Netherlands, 1992).
82. Wang, X., Guo, X., He, M. & Zhang, H. Statistical inference in mixed models and analysis of twin and family data. *Biometrics* **67**, 987–995 (2011).
83. Visscher, P.M. *et al.* Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.* **2**, e41 (2006).
84. Wright, F.A. The phenotypic difference discards sib-pair QTL linkage information. *Am. J. Hum. Genet.* **60**, 740–742 (1997).
85. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97–101 (2002).
86. Barry, W.T., Nobel, A.B. & Wright, F.A. A statistical framework for testing functional categories in microarray data. *Ann. Appl. Stat.* **2**, 286–315 (2008).
87. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).