Published in final edited form as:

Nat Genet. 2014 May ; 46(5): 430–437. doi:10.1038/ng.2951.

Heritability and Genomics of Gene Expression in Peripheral Blood

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

[1]Bioinformatics Research Center and Department of Statistics, North Carolina State University, Raleigh, NC, USA [2]Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA [3]Department of Genetics, University of North Carolina at Chapel Hill, NC, USA [4]Department of Genetics, Rutgers University, New Brunswick, NJ, USA [5]Department of Biostatistics, University of North Carolina at Chapel Hill, NC, USA [6]Department of Psychiatry, VU Medical Center, Amsterdam, Netherlands [7]Department of Biological Psychology, VU University, Amsterdam, Netherlands [8]Department of Computer Science, University of North Carolina at Chapel Hill, NC [9]Environmental and Occupational Health Sciences Institute, Rutgers University, New Brunswick, NJ, USA [10]Department of Pharmacotherapy & Outcomes Science, Virginia Commonwealth University, Richmond, VA, USA [11]Department of Computer Science, University of California, Los Angeles, USA

Abstract

Correspondence should be addressed to F.A.W. (fred_wright@ncsu.edu) or P.F.S (pfsulliv@email.unc.edu).
[12]These authors contributed equally to this work.

URLs

The fundamental data for this report (Affymetrix 6.0 and U219) are available by application to 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).

Accession numbers.

Expression data and genotypes are available in dbGap, accession: phs000486.v1.p1

Author Contributions

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

Conflicts of Interest

Dr Sullivan was on the SAB of Expression Analysis (Durham, NC, USA). The other authors report no conflicts of interest.

We assessed gene expression profiles in 2,752 twins, using a classic twin design to quantify expression heritability and quantitative trait loci (eQTL) 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/ontology classes, and strongly associated with disease designation. The design enabled a comparison of twin-based heritability to estimates based on dizygotic IBD 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 were further examined in a replication set of 1,895 unrelated subjects. A large number of local eQTLs (6,988) met replication criteria, while a relatively small number of distant eQTLs (165) met quality control and replication standards. Our results provide an important new resource toward understanding the genetic control of transcription.

Keywords

gene expression; peripheral blood; twin study; heritability; expression quantitative trait loci; eQTL

Introduction

Determining the biological significance 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 non-coding. Several lines of evidence suggest that genetic variation implicated in GWAS alters transcription[1–3]. Expression quantitative trait loci (eQTLs)[4,5] overlap markedly with GWAS-identified SNPs, both collectively[6–8] and for specific traits (e.g., 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 $n>1000$,[15,19,20] while $n > 3{,}000$ 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 including tissue,[8,10,23] ancestry,[7] winner's curse effects, and batch effects,[5,7,24–26] and cell heterogeneity.[27,28]

To achieve large sample sizes in humans, tissues must be accessible, and an attractive choice is peripheral venous blood, while 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 EBV copy number and growth rates.[31] A MuTHER LCL study of expression in female twins found a large impact of common "environment" shared by twins: 32% of transcripts showed common environmental effects >30%, compared to 2% in adipose and 8% in skin.[8] The authors attributed the dramatic effect to correlated sample handling rather than environmental exposures shared by co-twins, suggesting possible biases with LCLs.

Despite these challenges, quantifying human transcriptomic heritability is important. Although genes with significant eQTLs are, by definition, "heritable," additional polygenic variation may be widespread and fail to reach statistical significance by standard genotype-expression association. Genes with substantial polygenic variation may also be subject to unique selection pressures not apparent from local eQTLs. The classical twin design, contrasting resemblance in monozygotic (MZ) to dizygotic (DZ) twin pairs, offers distinct advantages in interpretability and efficiency in 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 (3.4X the size of the next largest twin eQTL report[8,15,30]), providing high resolution. Gene expression was assessed in peripheral blood, with careful attention to sample collection, cell type heterogeneity, bias, and control of experimental error. Our goals were to: (1) describe and evaluate the heritability of all transcripts measured in peripheral blood; (2) identify a comprehensive list of local and distant eQTLs and evaluate their characteristics and replicability; and (3) assess their biomedical relevance.

Results

Twin-based heritability in the peripheral blood transcriptome

We first report the heritability of steady-state transcription in peripheral blood for 43,628 transcripts 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 43K probe sets as a "transcript" (1–18 transcripts per gene, mean 2.4). Careful annotation was performed for the platform, which compares favorably to RNA-Seq (Supplementary Note).[33] Subjects were from 1,444 twin pairs (both members of 1,308 pairs, 95.1% of subjects, and one member from 136 pairs). The 1,308 complete pairs consisted of 690 MZ pairs (52.8%, 209 male and 481 female MZ pairs) and 618 DZ pairs (47.2%, 110 male, 256 female, and 252 opposite-sex DZ pairs). Expression QC included zygosity/sex confirmation, randomization for sex and zygosity balance, sample identity checks, and dropping low-quality samples. Primary analyses are 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 genotypes of blood count-associated SNPs. Supplementary Figure 1 shows 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 environment (e^2), where these values were measured using a covariance (ACE) model that includes additive genetic, common or shared environment, and non-shared environment terms. Variance components were not constrained to be positive), so the model would be unbiased for h^2, and to indicate whether genetic non-additive effects (dominance) may be present (by estimating c^2 as negative). Covariate correction notably increased evidence for highly heritable transcripts, while no transcript was significant for c^2 (Supplementary Figures 1b–e), in contrast to the MuTHER study.[8]

Figure 1a shows a *P*-value Manhattan plot for twin-based h^2 for 18,392 genes (selecting for each gene the transcript with the largest h^2), based on twin zygosity comparisons (inset). The h^2 had mean±SD 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 (Figure 1b, Supplementary Table 1). Mean within-cluster *r* ranged from 0.46 to 0.006. Cluster identity was supported by significantly higher connectivity in protein-protein interaction databases[34] and GO pathways[35] (Supplementary Table 1). Numerous clustered genes displayed expression patterns similar to other tissues, including brain,[35] suggesting broader tissue relevance. Regional clustering indicated enrichment for immune function (Supplementary Table 2, e.g., IgG Fc fragment receptors at chr1:161–162 mb and the MHC region at chr6:31–33 mb), while other regions showed fewer heritable genes (e.g., neuronal protocadherin gene cluster at chr5:140–141 mb and epidermal keratin gene clusters on chr17:39–40 mb and chr21:31–32 mb). Figure 1c shows that heritability is strongly associated with mean expression (r=0.356, $P < 10^{-200}$), with a striking increase above an array-specific detection threshold, with detectable expression for 21,971 transcripts (50.3%).

We next compared h^2 for all genes to multiple external "predictors"[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 mean expression correction. This negative association was surprising, as GC content ± 5kb from the TSS was positively correlated with gene density (r=0.40) and each modestly with mean expression (r=0.11 and r=0.10). Accordingly, after correcting for mean expression, the negative association with gene density was even stronger (Table 2, Figure 2a). Genes with recent evolutionary acceleration in primates and humans[42] showed significant positive association with h^2 after mean correction (Figure 2b and Table 2). HomoloGene conservation was highly significant, although attenuated after correction. Associations between h^2 and numerous KEGG and GO pathways were also highly significant (Supplementary Table 3). Interestingly, all pathway associations with h^2 were positive, except two related to sensory perception and smell (GO: 0050907 and GO:0050911).

To investigate disease relevance, we used the NHGRI GWAS catalog (17 July 2013),[1] identifying the nearest gene ("GWAS-genes") for each of 3,628 significantly disease-associated SNPs ($P \leq 5\times10^{-8}$), for a total of 2,343 GWAS genes. Heritability was highly significantly positively associated with GWAS genes (Figure 2b and Table 2). Enrichment remained elevated for genes nearby, but not necessarily closest, to the GWAS SNP, and genes with numerous nearby GWAS SNPs were especially heritable (Supplementary Figure 2). Enrichment was attenuated by removing chr6 genes (including the MHC region) and for immune-related diseases[43] (Supplementary Table 4). GWAS phenotypes include those relevant to blood/immunity along with central nervous system, bowel, cancers, and morphological traits. Given the GWAS-gene designation based only on proximity to NHGRI SNPs, these results may reflect an even stronger true tendency of disease-causing genes to be highly heritable (see Supplementary Figure 2). These results are complementary

to observations that disease-associated SNPs show eQTL enrichment.[6] Additionally, OMIM shows similar heritability enrichment, even though NHGRI GWAS and OMIM only partly overlapped (of genes in either list, 10% are in both). The OMIM genes with significant heritability are also quite diverse, further supporting potential relevance of peripheral blood to other tissues and developmental processes (Supplementary Table 5). Moreover, the 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 the disease associations should be interpreted with caution. The disease-heritability enrichment may reflect other underlying sources of commonality, but still point to transcription as an important intermediary in disease risk.

Dissecting local genetic contributions to heritability, and bias in h^2 estimation

After genotyping QC and imputation, 8.3 million SNPs were available for eQTL mapping in 2,494 individual twins (90.4% of the expression dataset). We evaluated multiple "predictors" of heritability, including association r^2 based on the most-significant local SNP within ±1Mb with, r^2 for the top distant SNP, local SNP-heritability estimation based on genetic relatedness among unrelated subjects using GCTA,[46] and variance-components results from complete local identity-by-descent inference among the DZ pairs (local IBD). We computed ratios of each component to the overall h^2 estimate (Supplementary Figure 3). Means and medians for $r^2_{\text{local SNP}}/h^2$ (0.04, 0.09) were similar to those reported in the MuTHER study [8], while the ratio $h^2_{\text{local IBD}}/h^2$ 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 in expressed genes of 0.16 (skin), 0.21 (LCLs), and 0.26 (adipose), with >20% of expressed genes displaying $h^2 > 0.3$.[8] Our study, although much larger, produced lower values of 0.14 and 12.3%. Our average h^2 should be unbiased, as we allowed for negative estimates (even if true $h^2 \geq 0$) whereas variance-component methods[8] can produce bias by forcing estimates to be nonnegative, 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 as following a gamma distribution, with sampling variation determined by the ACE model. The result (Figure 3a) is a shrunken distribution with a similar mean h^2 but markedly less variation. The model estimates that the true proportion of expressed genes with heritability > 0.3 is actually only 7.9%. For high heritability thresholds, the differing results across studies can appear to be dramatic – while the MuTHER report estimated >700 expressed genes in both skin and LCLs with heritability >0.5, we estimate the true number in our study as ~100. The studies differ in tissue 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. Removal of age as a covariate (Supplementary Note) suggests 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 fits our estimated

h^2 (Figure 3a). A similar approach quantifies the impact of sample size (Supplementary Note), again using the gamma model obtained from NTR, but inflating the sampling variation to reflect the smaller MuTHER sample size. The resulting estimated h^2 distribution is similar to that reported in MuTHER (Figure 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 conclusion (Supplementary Figure 4a, Supplementary Note). Although we conclude 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. Supplementary Figure 4b shows accuracy prediction as a function of sample size – even with the NTR sample size we predict that the rank correlation between true vs. estimated heritability is only slightly greater than 0.5.

We applied similar modeling approaches to local IBD-based h^2 (Figures 3c, 3d), estimating the proportion of total h^2 attributable to local genetic variation. Our mean local IBD h^2 was 0.03, with mean($h^2_{\text{local IBD}}$)/mean(h^2) = 0.23. This ratio is somewhat lower than those reported for MuTHER (>0.30), perhaps partly attributable to their focus 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. A regression approach (Supplementary Figure 5) shows that ~35% of the variation in estimated h^2 can be explained by the predictors.

eQTL analyses: genome-wide SNPs and the peripheral blood transcriptome

We next analyzed genotypes as predictors of transcription (i.e., a GWAS for each transcript) for 2,494 twins, using a REML (restricted maximum likelihood) model accounting for twin status and covariates. eQTLs within ± 1 Mb of a gene were classified “local” and all others 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 ($P < 10^{-200}$ for both).

Figure 4a shows the effect of sample size on local eQTL identification. The figure includes nearly all published blood-derived eQTL studies[7,8,15,20,31,48–52] (comparisons to the large meta-analysis of Ref. 29 described separately below), the full NTR data (n=2,494), and random subsamples of our data. We reanalyzed the datasets using a common QC 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 eQTL mapping on random subsets of varying sample size, using fewer covariates (i.e., no blood counts or SNPs) and ~600K genotyped SNPs. We also evaluated our robustness approaches (normal quantile transformation, and normal quantile transformation with SNP minor allele frequency, MAF > 0.005 or > 0.01 in each subsample). For local eQTLs, there was little difference among the transformations.

Figure 4a shows there is considerable inter-study variability in the number of significant eQTLs even with consistent QC and analysis.[19,31] With increasing sample size it appears that most expressed genes (>10,000) show evidence of local eQTL influence in peripheral blood. For NTR, the number of significant genes ($q < 0.01$) was 11,834, and after employing final QC steps was 9,640. Replication was checked 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). Supplementary Figure 6 shows reproducibility of eQTLs between NTR and the 1,895 unrelated NESDA samples, and Supplementary Figure 7 shows regulatory feature enrichment/deficits for local eQTL SNPs.

Of 9,640 genes with local eQTLs in NTR (q < 0.01), 9,148 (94.9) replicated ($q < 0.1$) in NESDA (with the less-stringent replication q threshold to allow for winner's curse attenuation). Of genes with the strongest local eQTL evidence in NTR ($q < 0.001$), 6,756 of 6,941 genes (97.3%) replicated in NESDA. There was strong overlap ($P = 1\times10^{-180}$) 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 ($P=1\times10^{-18}$), along with higher expression ($P=2\times10^{-119}$), and higher heritability ($P=8\times10^{-131}$) in NTR. The lack of overlap among smaller HapMap samples is likely an example of the "winner's curse": in the larger Zeller et al.[15] and Fehrmann et al.[20] studies, among genes annotated in all three studies, replication in NTR was 66.8% (2,799/4,189) and 77.2% (3,404/4,412) (Figure 4b). Similarly, for local gene-SNP pairs with $q<0.05$ from the peripheral blood eQTL meta-analysis of Westra et al.[29] (n=5,311), estimated true-discovery rates in NTR and NESDA were 59.6% and 59.7%, respectively (Supplementary Note and Supplementary Figure 8).

Characteristics of distant eQTLs: many are false, hotspots are few

Robust distant eQTL results (Figure 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 QC (see below). Extrapolating to larger sample sizes, we anticipate identifying <1,000 replicating eQTLs even for samples sizes exceeding 5,000. Figure 4d shows overlap of genes with significant distant eQTLs ($q<0.001$) among the large studies, with much lower overlap for distant than local eQTLs. For significant distant gene-SNP pairs from Westra et al.[29] (n=5,311), estimated true-discovery rates in NTR and NESDA were 23.1% and 23.0% (Supplementary Figure 8).

Our 601 distant eQTLs with $q<0.001$ (Figure 5) involved 581 genes and 538 non-redundant SNPs (for each gene, only the most significant SNP per chromosome was retained). We applied additional QC 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 ($q<0.01$) (Figure 5, Supplementary Figure 6 and Supplementary Table 6). Genes in the 348 eQTLs were analyzed using 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 based on the most severe predicted consequence. Most of the SNPs were intronic, followed by intergenic, up- or downstream of protein coding sequence, and exonic (Figure 5b). The 53 intergenic SNPs had the lowest rate of overlapping regulatory features, or replication in NESDA (Supplementary Figure 8). SNPs in up/downstream sequences were more likely to overlap with regulatory elements, and SNPs in intronic/exonic regions were more likely to replicate in NESDA. Only 6 of the 348 distant eQTLs were exonic, suggesting they influence expression rather than modify proteins, consistent with our finding that these distant eQTL SNPs are more likely to be local eQTLs (Supplementary Figure 9).

We next sought to identify eQTL hotspots (SNPs influencing numerous transcripts). We grouped the 304 distant eQTL SNPs into 203 regional clusters (Supplementary Figure 10). 160 clusters included only one SNP and the other 43 clusters spanned 2 kb to 2 Mb (median size 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 (Figure 5c), many times lower than reported for 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 chr19 is also a local eQTL of *MYOF1*. Given the expression of *SOX13*, *MYOF1* expression is independent of other distant eQTL genes (Figure 5d), suggesting the eQTL signals are mediated by *SOX13*. *MYOF1* encodes unconventional myosins which bind to membranous compartments and serve in intracellular movements. *SOX13* is a transcription factor that modulates the Wnt/TCF signaling pathway,[56] and several other distant eQTL genes are involved in cellular signaling (e.g., *TMEM134*, *RGS12*, and *SYT13*). Although significant networks were found (Supplementary Figure 11), the relatively few genes influenced by hotspots or distant eQTLs suggests such networks do not play 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 ($P < 1\times10^{-8}$), there were significant results for 3,415 SNPs, 498 traits, and 4,167 SNP-trait pairs from 927 papers. The greatest numbers of SNP-trait/disease associations were for height (248), HDL cholesterol (92), Crohn's disease (155), Type 2 diabetes (98), and ulcerative colitis (81). The extended MHC region (chr6:25–34 mb, 0.3% of the genome) is the second most gene-dense region of the genome and contained the greatest number of SNPs implicated by GWAS (6.8%). Of 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 3,118 genes in OMIM, 74.4% were part of a SNP-gene local eQTL pair ($q < 0.05$). These include 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), amyotropic lateral sclerosis (15 genes), Charcot-Marie-Tooth disease (25 genes), epilepsies (21 genes), and candidate genes for schizophrenia (*DISC1*, *DAOA*, and *RGS4*). Of 517 genes implicated in Mendelian autism spectrum disorders[57] or mental retardation,[44,58,59] 69.6% are part of a local eQTL SNP-gene pair. Of 3,294 genes with a copy number variant implicated in autism spectrum disorders,[57] developmental delay,[60] or a psychiatric disorder,[61] 72.4% are part of a local eQTL SNP-gene pair.

Finally, we combined heritability predictors and gene disease designations into several multiple regressions (Supplementary Table 7). The predictors were as shown in Table 2, with the addition of eQTL evidence (best local and distant r^2), chromosomes 6 (*HLA* genes), 19 (which was an outlier in gene density analysis), and X (underrepresented in GWAS), and a blood DNase 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, the full model 32.9%. h^2 remained significantly predictive of OMIM/NHGRI disease status except for the smaller sets of NHGRI genes subdivided by immune designation, even while the best local and distant eQTLs were no longer significant. Gene conservation was highly predictive for OMIM status. Gene density was strongly negatively associated with disease status, but this effect was attenuated for OMIM. NHGRI disease status was significantly enriched for chromosome 6, and showed a deficit on chrX, which we attribute to the neglect of chrX in GWAS.[62] OMIM showed enrichment of chromosome X, consistent with the importance of X-linked disorders in medical genetics.

Discussion

We have established clear patterns underlying heritability of steady-state gene peripheral blood transcription, and demonstrated strong connections to disease annotation. The use of peripheral blood enables further investigation to immune-related diseases,[63] but may also be useful for other tissues. For example, there were 78 genome-wide significant loci for inflammatory bowel disease (IBD) in cohorts genotyped with the custom "immunochip". In 10 of 78 instances, there was near perfect overlap of the local eQTL results from this study with the IBD association (excludes numerous other regions which overlapped but not as precisely, B Bulik-Sullivan and M Daly, personal communication). These 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.[64]

Examination of h^2 vs. 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 co-localization.[65–67] Essential genes identified in mouse mutagenesis screens show high linkage conservation;[68] and intergenic regions in humans have higher SNP densities than in introns, along with higher rates of neutral

polymorphisms.[69,70] Our observations appear concordant with these reports, whether selection directly inhibits heritability in gene-dense regions or due to the relative paucity of genotype variation in such regions.

The ability of h^2 to predict OMIM/NHGRI designation 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 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]

Online Methods

Subjects and biological sampling

Subjects were ascertained and sampled using harmonized protocols from two longitudinal cohort studies, the Netherlands Twin Registry (NTR) [71] and the Netherlands Study of Depression and Anxiety (NESDA). [72] NTR is an observational, 25-year longitudinal study of twins and their families [73–75] The study protocol was approved by Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Center, Amsterdam[71]. 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[72,74,75]. The study protocol was approved by the Ethical Review Board of the VU University Medical Centre and subsequently by local review boards of each participating center[72]. Informed consent was obtained from all participants in both studies.

Peripheral venous blood samples were drawn in the morning (NTR 0700–1100, NESDA 0830–0930) after an overnight fast. For fertile women in NTR, samples were obtained on day 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 minutes (60 minutes 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 for NTR and NESDA were conducted at the Rutgers University Cell and DNA Repository. Total RNA was extracted at Rutgers (for NESDA, at VU Medical Center) using the PAXgene Blood RNA MDx Kit protocol in 96 well format using the BioRobot Universal System (Qiagen). RNA quality and quantity was assessed by Caliper AMS90 with HT DNA5K/RNA LabChips. Samples were randomized to plates, with checks to ensure sex/zygosity balance. Co-twins were randomized without respect to relationship to avoid bias in family correlation estimates. For cDNA synthesis, 50ng 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, labeled with biotin (Encore Biotin Module, NuGEN), and size distributions verified (Caliper AMS90, HT DNA 5K/RNA LabChips). Samples were hybridized to Affymetrix U219 (Supplementary Note) array plates

to enable expression profiling in 96-sample sets. Array hybridization, washing, staining, and scanning were carried out in an Affymetrix GeneTitan System per the manufacturer's protocol.

QC was conducted on NTR and NESDA data in parallel. Expression data were required to pass standard Affymetrix Expression Console quality metrics before further QC. The array superset consisted of 6,526 U219 arrays (3,516 NTR, 2,783 NESDA samples, divided into baseline samples and a smaller portion after 2-year followup, and 227 controls) on 69 plates, including 417 samples which were identified as having reduced quality ($D < -5.0$, described below) and re-hybridized. Expression values were obtained using robust multichip averaging (RMA) normalization (Affymetrix Power Tools, v1.12.0). Probe sequences were mapped to the human genome (hg19) using BOWTIE, [76] and probes with sequences not mapping, mapping to multiple locations, or intersecting a polymorphic SNP (HapMap3 and 1000 Genomes Project data) were removed. [77,78] We mapped and annotated all Affymetrix U219 probesets with reference to GENCODE (v14) gene models as we were dissatisfied with the standard Affymetrix annotations.

The large sample size enabled additional QC metrics involving inter-sample comparisons. First, samples showing sex inconsistency were removed (based on chrX and chrY probesets). Second, we examined the pairwise correlation matrix of expression profiles. Using r_{ij} as the correlation between arrays i and j, we computed $\overline{r_i}=\sum_j r_{ij}/n$, the average correlation of array i with all others of the total n arrays. Lower $\bar{r}_i$ corresponds to lower quality, and were expressed in terms of median absolute deviations $D_i=(\overline{r}_i-\overline{\overline{r}})/median(|\overline{r}_i-\overline{\overline{r}}|)$ to provide a sense of distance from the grand correlation mean $\overline{\overline{r}}$. Third, we verified sample identity between U219 gene expression and Affymetrix 6.0 genotypes (see below), having previously discovered up to 5% genotype-expression mismatch rates 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 profiles (similar to reference [79]). This approach identified sex-mismatched samples and additional samples of poor quality.

Fourth, initial analysis using unrelated participants illustrated the potential for spurious eQTL identification due to expression outliers. Thus, conservatively, we transformed the expression values using the 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 intervals between extraction and hybridization steps, total white and red cell counts, hematocrit, and the top five expression principal components (PCs) (similar to that of surrogate variables). [80] 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 three genotype PCs.

Sixth, we observed that D and the posterior probability of "mismatch" were highly correlated, and reasoned that D might be useful for dropping additional low-quality samples.

To determine the optimal threshold for *D*, we successively dropped individual samples according to *D*, and recomputed the intraclass correlation coefficient (ICC)-based estimate of heritability $2(\hat{\rho}_{MZ} - \hat{\rho}_{DZ})$ and accompanying *p*-values [81] for all transcripts using covariate-residualized expression data. A Benjamini-Hochberg false discovery-rate *q*-value for transcripts was computed using p.adjust in R (v.2.14). Dropping 19 samples with the lowest *D* values resulted in the largest number of significant transcripts ($q < 0.10$) (Supplementary Note). This choice was largely robust to the *q* threshold in the range q=0.05 – 0.20, and to the use of unnormalized expression data.

After expression QC, 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 QC for NESDA followed the same steps as for NTR (except zygosity did not apply). The expression distributions of monozygotic and dizygotic twins were compared for differing mean expression (t-test) and differing variances (F-test for normally distribute 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 (FDR q<0.05) different variances. However, of these four transcripts, none showed h^2 q<0.05.

Genome-wide SNP assays

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

SNP QC is detailed in the Supplementary Note. Briefly, QC included removal of SNPs for non-unique probe mapping to NCBI Build 37/UCSC hg19, low minor allele frequency (< 0.005, determined empirically), substantial deviation from HapMap3 CEU founder allele frequencies, deviation from Hardy-Weinberg equilibrium ($p_{HWE} < 1\times10^{-8}$), 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. The resulting genotypes were of high quality, with relatively low SNP and subject missingness (97.5th percentiles of 0.035 and 0.020. Among 714 monozygotic twin pairs, the intrapair agreement for 686,895 autosomal SNPs was 0.9985. Prior genome-wide genotyping using a Perlegen four-chip platform was available for 2,219 subjects and 110,588 SNPs, [75] and had 0.9996 agreement with Affymetrix 6.0 genotyping.

Phased genotype calls on 379 European samples from 1000 Genomes were used as the reference set (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521) for imputation. The NTR samples were split into two unrelated sets. SNPs with call rate <95%, or HWE *P*<1E-09 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 defined such that the 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 used here.

Heritability

Three methods for estimating heritability are detailed in the Supplementary Note. The primary approach was *twin-based heritability* via a 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 mutually independent and normally distributed with mean 0 and variances σ_a^2, σ_c^2,and σ_e^2. This corresponds to a standard ACE model, and assumes DZ twins have an average identity-by-descent proportion of 0.5. [82,83] For each transcript, the twin-based heritability and shared environmental effects were estimated as $\hat{a}^2=\hat{\sigma}_a^2/(\hat{\sigma}_a^2+\hat{\sigma}_c^2+\hat{\sigma}_e^2)$ and $\hat{c}^2=\sigma_c^2/(\hat{\sigma}_a^2+\hat{\sigma}_c^2+\hat{\sigma}_e^2)$. 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 in order to best match the intra-class correlation approach[81] 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), then combining 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 in a manner accounting for all transcripts.

The second heritability estimation approach was *DZ-only heritability* following a constrained ACE mixed model approach for full siblings. [84] For this approach, a REML mixed model was used to relate the observed variation in true IBD proportions among DZ pairs to the 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*=1370, twin set 2, *n*=1372), and averaged the h^2 estimates from the two twin sets. The results showed almost no correlation with twin-based heritability (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 ±1Mb of each transcript.

Local IBD analysis

The residualized expression data were nearly perfectly normally distributed, and so the bivariate normal model of Wright [85] for sib-pair IBD mapping was applied to the DZ pairs, offering a potential improvement over the Haseman-Elston approach.[8] MERLIN [86] was run on the thinned set of markers used for stratification analysis, and probabilistic IBD estimates

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 DZ twin correlation as a function of IBD status, thus extracting maximum information, and 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 genes 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, [87] contrasting heritability of each set of genes with the 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 best-h^2 transcripts. For the observed data, this approach was highly consistent with the REML estimates (r=0.992, Supplementary Note). Twin zygosity status was permuted 1,000 times, and for each permutation h^2 was computed for all genes, along with the difference in mean h^2 for the gene set versus the complementary set. As this difference is nearly normally distributed, an *enrichment z-statistic* was calculated as the observed difference divided by its permutation standard deviation, and a two-sided P-value 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 standard deviation). By permuting only zygosity status, the enrichment-z approach preserves the mean twin pair correlations, as well as gene-gene correlation. 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 datasets.

eQTL analysis

We refer to eQTLs as *local* (SNP-transcript associations ± 1Mb of the transcription start/end sites) or *distant* (the remaining findings). We prefer these terms to "cis/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 [88] 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 $P < 10^{-5}$ (a liberal threshold for the ~3×10^{10} tests performed). Separate false discovery rate (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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The work described in this paper was funded by the US National Institute of Mental Health (RC2 MH089951, PI Sullivan) as part of the American Recovery and Reinvestment Act of 2009. We thank Dr Thomas Lehner (NIMH) for his support. Additional analytic support provided by R01 MH090936, R01 GM074175, P42 ES005948, and 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, 912-100-20; Spinozapremie 56-464-14192; Geestkracht program grant 10-000-1002); the Center for Medical Systems Biology (CMSB2; NWO Genomics), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL), 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 Program (FP7/2007-2013); ENGAGE (HEALTH-F4-2007-201413); and the European Research Council (ERC, 230374).

References

1. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–7. [PubMed: 19474294]
2. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 237:1190–1195. [PubMed: 22955828]
3. Hardy J. Psychiatric genetics: are we there yet? JAMA psychiatry. 2013
4. Majewski J, Pastinen T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends in genetics: TIG. 2011; 27:72–9. [PubMed: 21122937]
5. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nat Rev Genet. 2009; 10:184–94. [PubMed: 19223927]
6. Nicolae DL, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS genetics. 2010; 6:e1000888. [PubMed: 20369019]
7. Stranger BE, et al. Patterns of cis regulatory variation in diverse human populations. PLoS genetics. 2012; 8:e1002639. [PubMed: 22532805]
8. Grundberg E, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nature genetics. 2012
9. Lango Allen H, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–8. [PubMed: 20881960]
10. Emilsson V, et al. Genetics of gene expression and its effect on disease. Nature. 2008; 452:423–8. [PubMed: 18344981]
11. de Jong S, et al. Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes. European journal of human genetics: EJHG. 2012; 20:1004–1008. [PubMed: 22433715]
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. Human molecular genetics. 2010; 19:3482–8. [PubMed: 20601676]
13. Luo R, et al. Genome-wide Transcriptome Profiling Reveals the Functional Impact of Rare De Novo and Recurrent CNVs in Autism Spectrum Disorders. American journal of human genetics. 2012; 91:38–55. [PubMed: 22726847]
14. Speliotes EK, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010; 42:937–48. [PubMed: 20935630]

15. Zeller T, et al. Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. PloS one. 2010; 5:e10693. [PubMed: 20502693]
16. Gamazon ER, Huang RS, Cox NJ, Dolan ME. Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107:9287–92. [PubMed: 20442332]
17. Thurman RE, et al. The accessible chromatin landscape of the human genome. Nature. 2012; 489:75–82. [PubMed: 22955617]
18. Degner JF, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012; 482:390–4. [PubMed: 22307276]
19. Xia K, et al. seeQTL: A searchable database for human eQTLs. Bioinformatics. 2011; 28:451–2. [PubMed: 22171328]
20. Fehrmann RS, 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 genetics. 2011; 7:e1002197. [PubMed: 21829388]
21. Min JL, et al. The use of genome-wide eQTL associations in lymphoblastoid cell lines to identify novel genetic pathways involved in complex traits. PloS one. 2011; 6:e22070. [PubMed: 21789213]
22. Grundberg E, et al. Population genomics in a disease targeted primary cell model. Genome research. 2009; 19:1942–52. [PubMed: 19654370]
23. Gibbs JR, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS genetics. 2010; 6:e1000952. [PubMed: 20485568]
24. Leek JT, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature reviews. Genetics. 2010; 11:733–9.
25. Akey JM, Biswas S, Leek JT, Storey JD. On the design and analysis of gene expression studies in human populations. Nature genetics. 2007; 39:807–8. author reply 808–9. [PubMed: 17597765]
26. Innocenti F, et al. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS genetics. 2011; 7:e1002078. [PubMed: 21637794]
27. Fairfax BP, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nature genetics. 2012; 44:502–10. [PubMed: 22446964]
28. Flutre T, Wen X, Pritchard J, Stephens M. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. PLoS genetics. 2013; 9:e1003486. [PubMed: 23671422]
29. Westra HJ, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013; 45:1238–43. [PubMed: 24013639]
30. Powell JE, et al. Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. Genome research. 2012; 22:456–66. [PubMed: 22183966]
31. Choy E, et al. Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. PLoS Genet. 2008; 4:e1000287. [PubMed: 19043577]
32. van Dongen J, Slagboom PE, Draisma HH, Martin NG, Boomsma DI. The continuing value of twin studies in the omics era. Nature reviews. Genetics. 2012; 13:640–53.
33. Flicek P, et al. Ensembl 2013. Nucleic acids research. 2013; 41:D48–55. [PubMed: 23203987]
34. Rossin EJ, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS genetics. 2011; 7:e1001273. [PubMed: 21249183]
35. Huang DW, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007; 35:W169–75. [PubMed: 17576678]
36. Grossman SR, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science. 2010; 327:883–6. [PubMed: 20056855]
37. Nickel GC, Tefft D, Adams MD. Human PAML browser: a database of positive selection on human genes using phylogenetic methods. Nucleic acids research. 2008; 36:D800–8. [PubMed: 17962310]

38. Nielsen R, et al. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS biology. 2005; 3:e170. [PubMed: 15869325]
39. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS biology. 2006; 4:e72. [PubMed: 16494531]
40. Andres AM, et al. Targets of balancing selection in the human genome. Molecular biology and evolution. 2009; 26:2755–64. [PubMed: 19713326]
41. Grossman SR, et al. Identifying recent adaptations in large-scale genomic data. Cell. 2013; 152:703–13. [PubMed: 23415221]
42. Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011; 478:476–82. [PubMed: 21993624]
43. Sivakumaran S, et al. Abundant pleiotropy in human complex diseases and traits. American journal of human genetics. 2011; 89:607–18. [PubMed: 22077970]
44. McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet. 2007; 80:588–604. [PubMed: 17357067]
45. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. Nature reviews. Genetics. 2008; 9:255–66.
46. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. American journal of human genetics. 2011; 88:76–82. [PubMed: 21167468]
47. Powell JE, et al. Congruence of additive and non-additive effects on gene expression estimated from pedigree and SNP data. PLoS genetics. 2013; 9:e1003502. [PubMed: 23696747]
48. Stranger BE, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007; 315:848–53. [PubMed: 17289997]
49. Montgomery SB, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; 464:773–7. [PubMed: 20220756]
50. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–72. [PubMed: 20220758]
51. Price AL, et al. Effects of cis and trans genetic ancestry on gene expression in African Americans. PLoS genetics. 2008; 4:e1000294. [PubMed: 19057673]
52. Spielman RS, et al. Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet. 2007; 39:226–31. [PubMed: 17206142]
53. Gatti DM, Barry WT, Nobel AB, Rusyn I, Wright FA. Heading down the wrong pathway: on the influence of correlation within gene sets. BMC genomics. 2010; 11:574. [PubMed: 20955544]
54. McLaren W, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010; 26:2069–70. [PubMed: 20562413]
55. Sun W, Ibrahim JG, Zou F. Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. Genetics. 2010; 185:349–59. [PubMed: 20157003]
56. Marfil V, et al. Interaction between Hhex and SOX13 modulates Wnt/TCF activity. The Journal of biological chemistry. 2010; 285:5726–37. [PubMed: 20028982]
57. Betancur C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. Brain Res. 2011; 1380:42–77. [PubMed: 21129364]
58. Chiurazzi P, Schwartz CE, Gecz J, Neri G. XLMR genes: update 2007. European journal of human genetics: EJHG. 2008; 16:422–34. [PubMed: 18197188]
59. Inlow JK, Restifo LL. Molecular and comparative genetics of mental retardation. Genetics. 2004; 166:835–81. [PubMed: 15020472]
60. Cooper GM, et al. A copy number variation morbidity map of developmental delay. Nature genetics. 2011; 43:838–46. [PubMed: 21841781]
61. Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nature Reviews Genetics. 2012; 13:537–51.
62. Wise AL, Gyi L, Manolio TA. eXclusion: Toward Integrating the X Chromosome in Genome-wide Association Analyses. American journal of human genetics. 2013; 92:643–7. [PubMed: 23643377]
63. Xavier RJ, Rioux JD. Genome-wide association studies: a new window into immune-mediated diseases. Nature reviews. Immunology. 2008; 8:631–43.

64. Crowley JJ, et al. Pervasive allelic imbalance revealed by allele-specific gene expression in highly divergent mouse crosses. Submitted.
65. Hurst LD, Pal C, Lercher MJ. The evolutionary dynamics of eukaryotic gene order. Nature reviews. Genetics. 2004; 5:299–310.
66. Osborne CS, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. Nature genetics. 2004; 36:1065–71. [PubMed: 15361872]
67. Sproul D, Gilbert N, Bickmore WA. The role of chromatin structure in regulating the expression of clustered genes. Nature reviews. Genetics. 2005; 6:775–81.
68. Hentges KE, Pollock DD, Liu B, Justice MJ. Regional variation in the density of essential genes in mice. PLoS genetics. 2007; 3:e72. [PubMed: 17480122]
69. Cai JJ, Macpherson JM, Sella G, Petrov DA. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS genetics. 2009; 5:e1000336. [PubMed: 19148272]
70. 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. 2009; 10:614. [PubMed: 20015390]
71. Willemsen G, et al. The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. Twin research and human genetics: the official journal of the International Society for Twin Studies. 2010; 13:231–45. [PubMed: 20477721]
72. Penninx B, Beekman A, Smit J. The Netherlands Study of Depression and Anxiety (NESDA): Rationales, Objectives and Methods. International Journal of Methods in Psychiatric Research. 2008; 17:121–40. [PubMed: 18763692]
73. Boomsma DI, et al. Netherlands Twin Register: from twins to twin families. Twin Res Hum Genet. 2006; 9:849–57. [PubMed: 17254420]
74. Boomsma DI, et al. Genome-wide association of major depression: Description of samples for the GAIN major depressive disorder study: NTR and NESDA Biobank Projects. European Journal of Human Genetics. 2008; 16:335–42. [PubMed: 18197199]
75. Sullivan PF, et al. Genomewide association for major depressive disorder: a possible role for the presynaptic protein piccolo. Molecular Psychiatry. 2009; 14:359–75. [PubMed: 19065144]
76. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]
77. Altshuler DM, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–8. [PubMed: 20811451]
78. Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–73. [PubMed: 20981092]
79. Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. Nature genetics. 2012; 44:603–8. [PubMed: 22484626]
80. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007; 3:1724–35. [PubMed: 17907809]
81. Falconer, DS.; Mackay, TFC. Introduction to Quantitative Genetics. Longman Group Ltd; London: 1996.
82. Neale, MC.; Cardon, LR. Methodology for the Study of Twins and Families. Kluwer Academic Publisher Group; Dordrecht, the Netherlands: 1992.
83. Wang X, Guo X, He M, Zhang H. Statistical inference in mixed models and analysis of twin and family data. Biometrics. 2011; 67:987–95. [PubMed: 21306354]
84. Visscher PM, et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS genetics. 2006; 2:e41. [PubMed: 16565746]
85. Wright FA. The phenotypic difference discards sib-pair QTL linkage information. American journal of human genetics. 1997; 60:740–2. [PubMed: 9042938]
86. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002; 30:97–101. [PubMed: 11731797]
87. Barry WT, Nobel AB, Wright FA. A statistical framework for testing functional categories in microarray data. Annals of Applied Statistics. 2008; 2:286–315.

88. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012; 28:1353–8. [PubMed: 22492648]

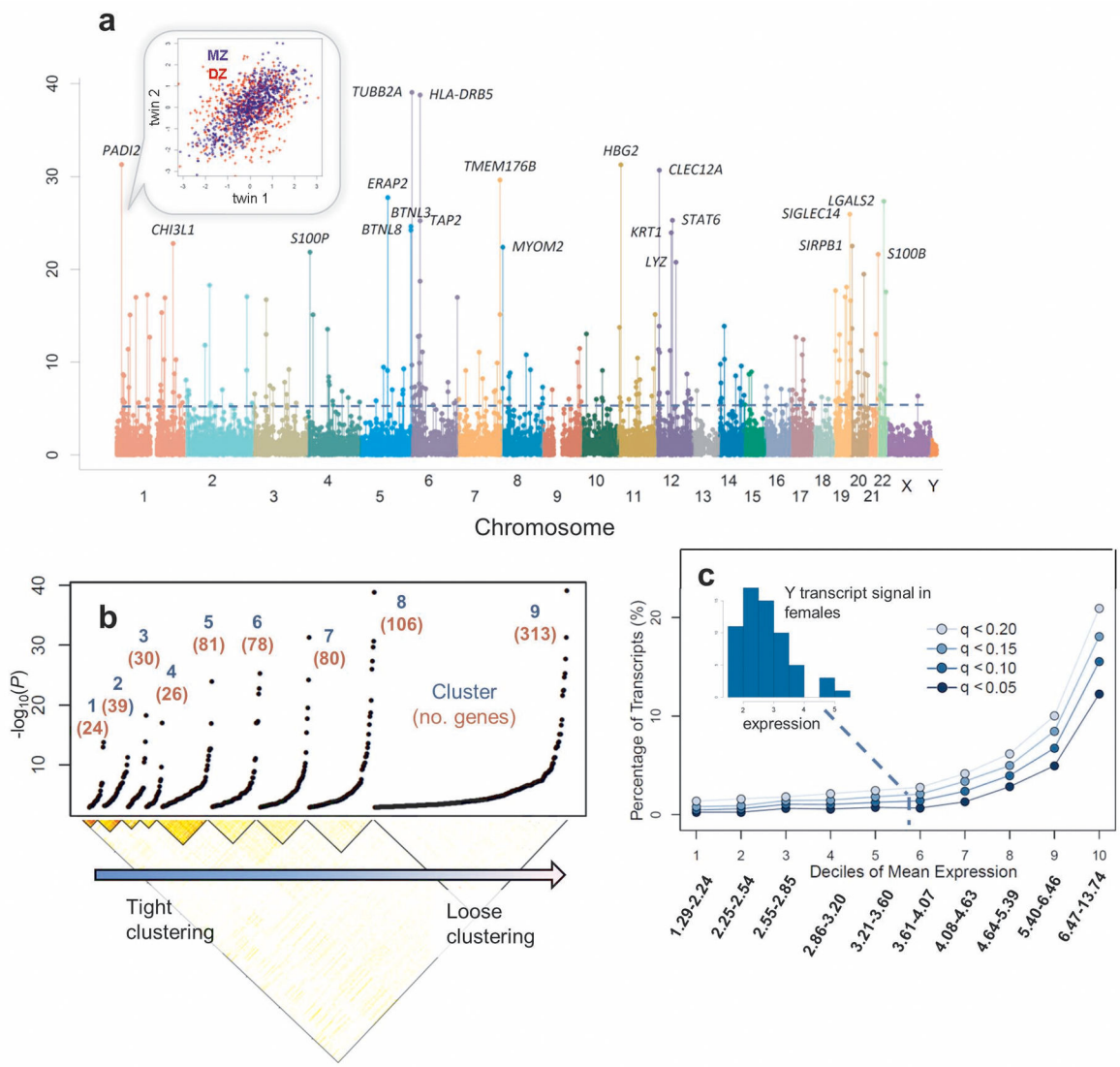

Figure 1.

Transcriptome-wide estimates of heritability, based on *n*=2752 twins. **(a)** Manhattan plot of h^2 *P*-values for the highest h^2 transcript for each of 18,392 genes. The inset (showing *PADI2*) illustrates that the evidence for heritability is based on higher a correlation between MZ pairs (blue) than between DZ pairs (red). **(b)** Clustering of 777 genes with h^2 $q < 0.05$. The most heritable genes belong to the cluster with lowest inter-gene correlation, but many significant genes belong to clusters with high inter-gene correlation. **(c)** Among 43,628 transcripts, the significant proportion (in terms of false discovery q-value) is dependent on mean transcript expression, increasing rapidly for transcripts above an approximate detection threshold (expression ≥ 3.584, determined as the 90th percentile of chrY expression in females).

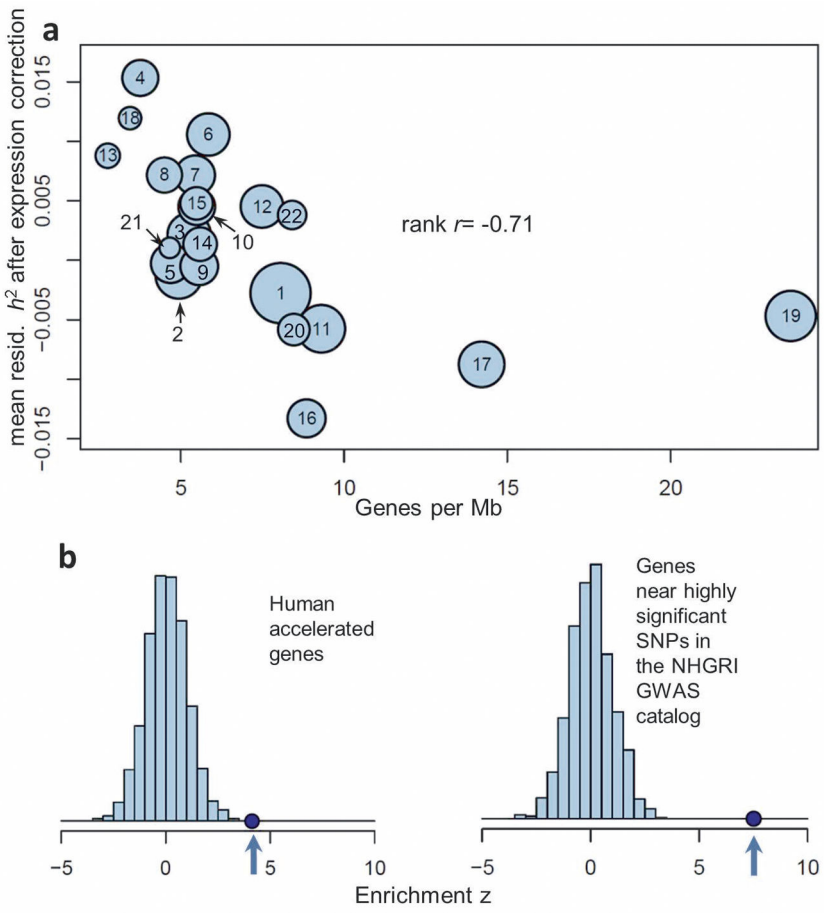

Figure 2.

Gene density and other predictors of heritability, using *n*=2616 paired co-twins and 18,392 genes. **(a)** Mean h^2 (corrected for gene expression level) vs. density of protein coding genes per autosome, showing that heritability is considerably higher for gene-poor chromosomes. Plot symbol area is proportional to number of array genes per chromosome. **(b)** Histograms of the permuted enrichment z-statistics for two predictors listed in Table 2. Observed values (blue dots) are extreme compared to the permutations.

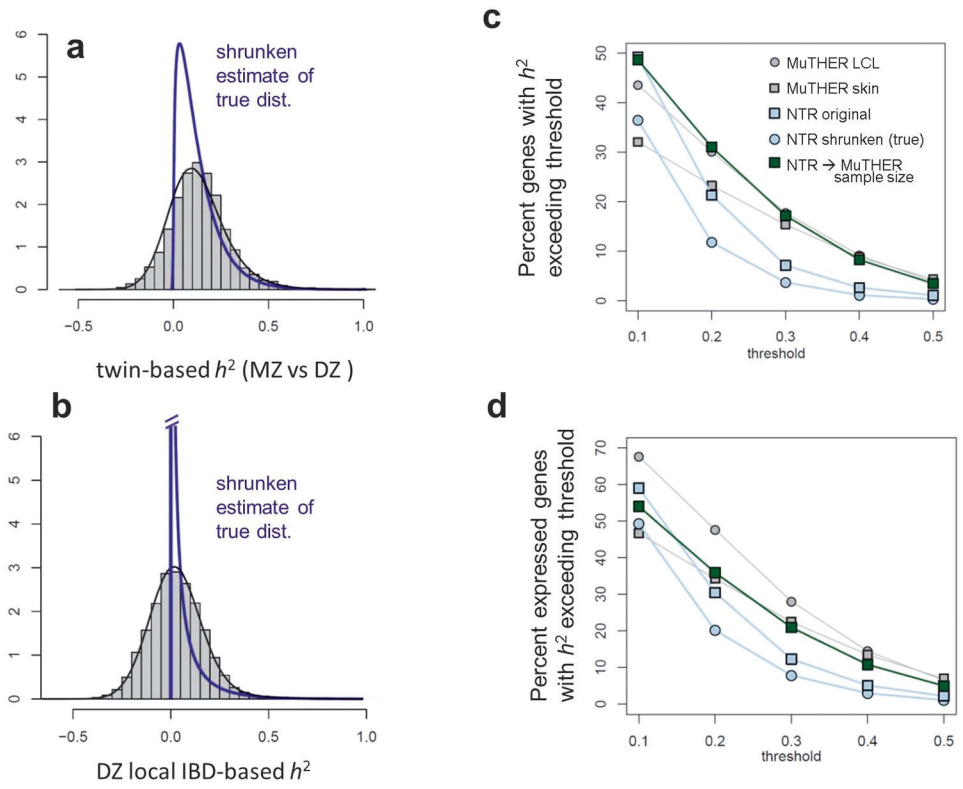

Figure 3.

Apparent heritability and local IBD effects vs. true underlying distributions. **(a)** For the twin-based h^2 estimates (n=2752, 8818 expressed genes shown), subtracting the effects of sampling variation produces an estimated true distribution (blue). Re-simulating from the fitted true assumed distribution closely approximates the observed h^2 (black curve). **(b)** The 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)** The analogous figure using only expressed genes from both studies.

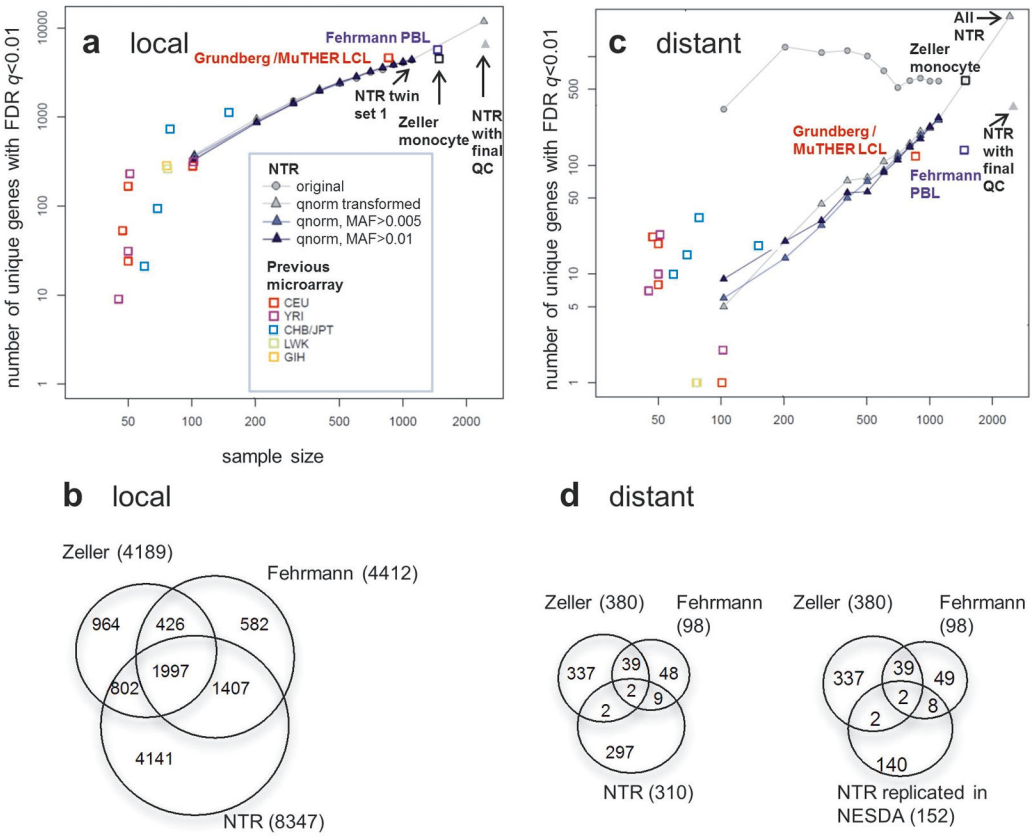

Figure 4.

Comparison and replication of eQTL results. **(a)** Number of unique genes with evidence of local association ($q < 0.01$, SNP ± 1 Mb window of gene), depicted for published leukocyte eQTL studies (LCLs, monocytes, and PBLs), as well as subsampling of NTR data (PBLs) using only genotyped markers and moderate QC (n=2494, 43,628 transcripts examined). Sample sizes are corrected for the number of covariates used. The "NTR with final QC" value applies q<0.001. **(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 (greater than 1Mb) association. The implausible non-monotone pattern for NTR on original expression values illustrates the importance of robust association methods. Using the final QC 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 QC is essential. **(d)** Overlap of distant eQTL findings (q<0.001) with previous studies (within 1 Mb).

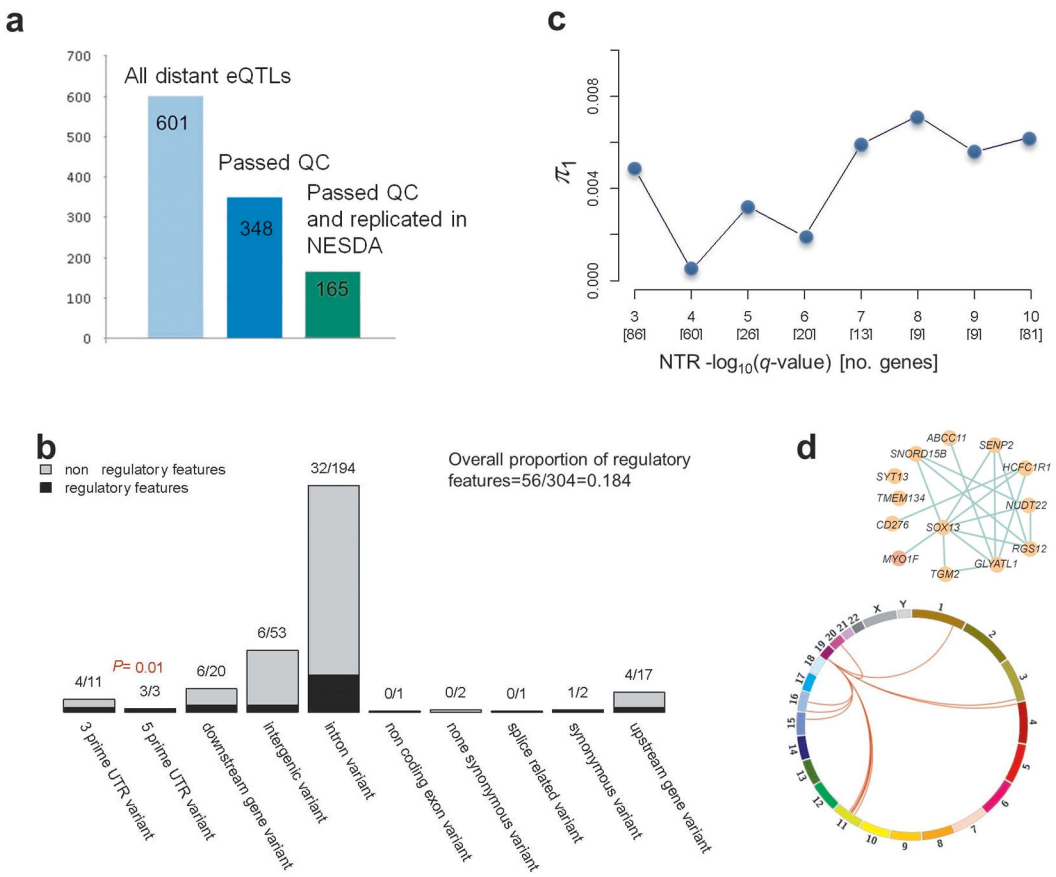

Figure 5.

Properties of distant eQTLs. **(a)** 348 eQTLs (gene-SNP pairs) were significant ($q < 0.001$) and passed the QC procedures and, of these, 165 replicated ($q < 0.1$) in 1895 NESDA individuals. **(b)** The 304 SNPs in significant eQTLs were examined for overlap with regulatory features, including DNase/FAIRE and transfactor binding sites, using Variant Effect Predictor (version 2.8) of Ensembl. [54] Most features were not enriched, although the 3 SNPs annotated as 5′ UTR variants all overlap with regulatory features, representing a significant enrichment compared to the total 18.4% overlap of distant eQTL SNPs with regulatory features representing a significant enrichment compared to the total 18.4% overlap of distant eQTL SNPs with regulatory features. **(c)** The π_1 value represents the estimated proportion of the transcriptome influenced by the 304 QC-passing SNPs in significant eQTLs. Across all significant bins the cumulative proportion is only ~3%. **(d)** A distant eQTL hotspot on chr19 was associated with the expression of 12 distant genes, and one 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*.

Table 1

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

Variable	Median (IQR)[a] or proportion
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 0700–1100	0.940
Fasting at time of blood draw	0.947
Current smoker	0.216
Alcohol user (12 drinks/year)	0.771

[a] IQR=inter-quartile range.

Table 2

Predictors of high heritability expression levels.

Predictor	Mean h^2 change	Enrichment z	*P*	Expr -corrected Enrichment z	*P*
Mean expression		**11.25**	**2.43×10^{-29}**	--	--
Variance of expression		**14.14**	**2.23×10^{-45}**	**14.89**	**4.02×10^{-50}**
GC content, +5kb of TSS		−1.42	0.155	**−5.33**	**9.60×10^{-8}**
GC content, −5kb of TSS		−0.72	0.471	**−5.00**	**5.73×10^{-7}**
DNase I hypersens. site (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 (2343) [l]	**0.018**	**7.42**	**1.14×10^{-13}**	**7.52**	**5.53×10^{-14}**
NHGRI, chr6 genes removed (2142)	**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 (1623)	**0.011**	**3.71**	**0.0002**	**4.88**	**1.03×10^{-6}**
OMIM disease entries (3089) [n]	**0.018**	**8.87**	**7.63×10^{-19}**	**7.54**	**4.81×10^{-14}**
NHGRI + OMIM (4809)	**0.019**	**10.81**	**2.96×10^{-27}**	**9.84**	**7.81×10^{-23}**

Abbreviations: TSS=transcription start site, NHGRI=National Human Genome Research Institute, GWAS=genome-wide association study, OMIM=Online Mendelian Inheritance in Man. Values in boldface correspond to $P < 0.0022$, for Bonferroni significance at α=0.05 for the 23 tests in each of uncorrected and corrected analyses. Values in blue depict significant negative associations.

[a] From the Encode Duke UCSC tracks.

[b] Defined as the reversed rank of the variance of bp position of gene and two flanking genes.

[c] End transcription bp minus start transcription bp.

[d] Decode sex-averaged standardized recombination maps at http://www.decode.com/addendum/ in 10kb bins.

[e] LD block boundaries as described in Supplementary Methods.

[f] NCBI HomoloGene (build 66) score, defined as the ratio of number of appearances in other organisms to the total of 21.

[g] Reference 36, genes with the property is shown in parentheses.

[h] References 36–38.

[i] Reference 39.

[j] Reference 35

[k] Reference 41. Reference 1, for SNPs with $P<5\times10^{-8}$.

[m] Following classification in Reference 42.

[n] Reference 43.