The Genetic Architecture of Gene Expression in Peripheral Blood

Luke R. Lloyd-Jones,[1,2,7,*] Alexander Holloway,[2,7] Allan McRae,[1] Jian Yang,[1,2] Kerrin Small,[3] Jing Zhao,[4] Biao Zeng,[4] Andrew Bakshi,[2] Andres Metspalu,[5] Manolis Dermitzakis,[6] Greg Gibson,[4] Tim Spector,[3] Grant Montgomery,[1] Tonu Esko,[5] Peter M. Visscher,[1,2,7] and Joseph E. Powell[1,2,7,*]

We analyzed the mRNA levels for 36,778 transcript expression traits (probes) from 2,765 individuals to comprehensively investigate the genetic architecture and degree of missing heritability for gene expression in peripheral blood. We identified 11,204 *cis* and 3,791 *trans* independent expression quantitative trait loci (eQTL) by using linear mixed models to perform genome-wide association analyses. Furthermore, using information on both closely and distantly related individuals, heritability was estimated for all expression traits. Of the set of expressed probes (15,966), 10,580 (66%) had an estimated narrow-sense heritability (h^2) greater than zero with a mean (median) value of 0.192 (0.142). Across these probes, on average the proportion of genetic variance explained by all eQTL (h^2_{COJO}) was 31% (0.060/0.192), meaning that 69% is missing, with the sentinel SNP of the largest eQTL explaining 87% (0.052/0.060) of the variance attributed to all identified *cis*- and *trans*-eQTL. For the same set of probes, the genetic variance attributed to genome-wide common (MAF > 0.01) HapMap 3 SNPs (h^2_g) accounted for on average 48% (0.093/0.192) of h^2. Taken together, the evidence suggests that approximately half the genetic variance for gene expression is not tagged by common SNPs, and of the variance that is tagged by common SNPs, a large proportion can be attributed to identifiable eQTL of large effect, typically in *cis*. Finally, we present evidence that, compared with a meta-analysis, using individual-level data results in an increase of approximately 50% in power to detect eQTL.

Introduction

In the past decade, genome-wide association studies (GWASs) have identified thousands of loci for complex traits and diseases. Most associated variants are not located in protein-coding regions and are instead highly enriched for regulatory regions of the genome. Thus, it has been suggested that for many variants, the functional mechanisms by which they affect disease susceptibility is through regulation of gene expression.[1,2] GWA-type approaches have been used to map loci, termed expression quantitative trait loci (eQTL), that influence the expression levels of thousands of transcripts. To date, the majority of identified eQTL are located proximal to their transcript (i.e., *cis*).[3–7] The mean of the estimates of heritability across expressed mRNA transcripts in peripheral blood ranges from 0.14 to 0.24,[7–9] although these studies vary in numerous aspects of their design and methodological approaches. We consider the proportion of transcript narrow-sense heritability not explained by the heritability attributed to identified eQTL as the missing heritability of gene expression.[10–13] On average, the proportion of heritability explained by eQTL across mRNA transcripts, which is largely attributed to *cis* variants, ranges from 0.20 to 0.38,[3,4,7,8] suggesting that to date much of the heritability for gene expression is still unaccounted for.

By using individual-level data, we can investigate some of the hypotheses for missing heritability in more detail. One of the proposed hypotheses is that there is a large contribution from rare variants of large effect. Typically, rare variants are not included on SNP arrays and are not well tagged through imputation to a common reference panel. Another hypothesis is that the majority of missing heritability is due to common variants of small effect that are not detected at the level of genome-wide significance. If the second hypothesis is true, increasing sample size will be more important than extending variant coverage for continued progress in understanding cellular or higher-order complex traits.[14] For gene expression, much of the remaining variation is hypothesized to be hidden in *trans*-eQTLs of small effect.[4,7–9,15]

We use data from the Consortium for the Architecture of Gene Expression (CAGE), which comprises individual-level whole-blood expression and genotype data on 2,765 individuals. For all transcript expression traits (also referred to as probes), we use the method presented in Zaitlen et al.[16] to estimate concurrently the total narrow sense heritability (h^2) and the proportion of phenotypic variance explained by all common SNPs (h^2_g) using a linear mixed model (LMM) that relies on a partitioned identity-by-state (IBS) genetic relationship matrix and takes advantage of both the related and unrelated individuals present in the data. To summarize the extent of missing heritability across expression traits, h^2 and h^2_g are compared to the proportion of genetic variance explained by eQTLs identified from an exhaustive association study.

[1]Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia; [2]Queensland Brain Institute, University of Queensland, Brisbane, QLD 4072, Australia; [3]Department of Twin Research and Genetic Epidemiology, King's College London, London SE1 7EH, UK; [4]School of Biology and Center for Integrative Genomics, Georgia Institute of Technology, Atlanta, GA 30332, USA; [5]Estonian Genome Center, University of Tartu, Tartu 51010, Estonia; [6]Department of Genetic Medicine and Development, University of Geneva, Geneva 1211, Switzerland
[7]These authors contributed equally to this work
*Correspondence: l.lloydjones@uq.edu.au (L.R.L.-J.), joseph.powell@uq.edu.au (J.E.P.)
http://dx.doi.org/10.1016/j.ajhg.2016.12.008.

Furthermore, we investigate the relative power of meta-analyses versus mega-analyses with individual-level data for eQTL detection.

Material and Methods

Consortium for the Architecture of Gene Expression

We investigated the genetic architecture underlying gene expression variation in peripheral blood tissue using data from 2,765 individuals within CAGE (Table S1). For the full details of the cohorts contributing to CAGE and their sample preparation, normalization, and imputation, see the Supplemental Note. In brief, the 2,765 samples consisted of data from five cohorts: BSGS (n = 916),[5,17] CAD (n = 147),[18] CHDWB (n = 449),[19] EGCUT (n = 1,065),[20] and Morocco (n = 188).[21] We conducted the quantification of gene expression for each cohort by isolating RNA from whole blood and then hybridizing RNA to Illumina Whole-Genome Expression BeadChips (HT12 v.3, HT12 v.4). Genotype data were acquired using different genotyping platforms and were imputed to the 1000 Genomes Phase 1 Version 3 reference panel,[22] resulting in 7,763,174 SNPs passing quality control. The gene expression levels in each cohort were initially normalized using variance stabilization,[23] followed by a quantile adjustment to standardize the distribution of expression levels across samples using the software of Ritchie et al.[24] The PEER software[25] was used to concurrently correct for the measured covariates such as age, gender, cell counts, and batch effects, which are known to explain variation in gene expression, and hidden heterogeneous sources of variability. Not all cohorts had measurements for all covariates and thus we relied on the PEER software to correct for these in their absence. For all cohorts we chose the maximum number of relevant factors in the PEER analysis to be 50. The residuals from PEER for each cohort were then standardized to *z*-scores and concatenated across cohorts. We retained only those probes that passed quality control in all cohorts, resulting in 38,624 taken forward. We performed a further PEER correction analysis on the concatenated data with the covariate gender included and then transformed the residuals for each probe using the rank normal transformation of Blom,[26] which alters the distribution of the residuals to be normally distributed with a mean of 0 and a standard deviation of 1. Finally, probes measuring expression levels of genes located on the X and Y chromosomes were removed from the analysis, leaving 36,778 for analysis.

Heritability Estimation

The 2,765 CAGE samples consist of a mix of both highly related individuals and different ancestral groups (Figures S7 and S8). To avoid problems associated with population stratification, we chose to estimate heritability using data from individuals of European ancestry. To investigate ancestry for the 2,765 individuals in CAGE, the relationship between the first two principal components (PCs) of the CAGE genotype matrix relative to the HapMap 3 ancestry cohorts (i.e., projected PCs[27,28]) showed mixed population backgrounds within CAGE (Figure S7). Non-European individuals were defined to be those exceeding the bounds of [lower quartile − 1.5 × IQR, upper quartile + 1.5 × IQR] of the first projected PC[28] (where IQR is the inter-quartile range); this threshold removed 311 individuals leaving 2,454 with European ancestry (see Table S3 for a detailed summary of data subsets used across analyses).

We utilized a method presented by Zaitlen et al.[16] to estimate the narrow-sense heritability (h^2) and the proportion of phenotypic variance explained by genotyped SNPs (h^2_g) via the use of a two-variance component LMM that requires an IBS genetic relationship matrix (GRM) (denoted $\mathbf{K}_{\mathrm{IBS}}$). This method, here termed Big K/Small K, makes use of both the unrelated and related European individuals present in the CAGE dataset by partitioning the phenotypic covariance matrix as $\boldsymbol{\Sigma} = \mathbf{K}_{\mathrm{IBS}>t}\,(h^2_{IBS>t} - h^2_g) + \mathbf{K}_{\mathrm{IBS}}h^2_g + \mathbf{I}(1 - h^2_{IBS>t})$. The $\mathbf{K}_{\mathrm{IBS}>t}$ matrix is estimated by setting the off-diagonal elements of $\mathbf{K}_{\mathrm{IBS}}$ less than the off-diagonal threshold t to zero. The resultant estimate of h^2 is the proportion of phenotypic variance attributed to the sum of the two variance component parameters. The method was implemented in the GCTA software[29] for all European individuals (n = 2,454), with $t = 0.05$ and SNPs common to the HapMap 3 set and the 7.8 M CAGE SNPs (893,626) used to construct the GRM (Figure S9). The first ten PCs of the genotype matrix for the European individuals were included as fixed effects in the REML analysis to control for any residual population stratification in the European individuals. For comparison, the unconstrained and constrained versions of the REML algorithm in GCTA were run. The narrow-sense heritability and proportion of phenotypic variance explained by genotyped SNPs from the unconstrained algorithm are denoted as h^{2*} and h^{2*}_g, respectively, to differentiate from the constrained values.

In order to make inferences regarding the proportion of narrow-sense heritability explained by genome-wide SNPs and identified eQTL, we made comparisons across a set of probes that overlapped with those reported to be expressed in the study of Kirsten et al.[4] This set was chosen because the Kirsten et al.[4] data are completely independent from CAGE, had expression levels determined from peripheral blood, and had a similar data size to CAGE (n = 2,112). The probe list was downloaded from the GEO website and consisted of 18,738 probes that mapped uniquely to the genome and had a probe annotation quality score of at least "good" as per the protocol of Barbosa-Morais et al.[30] Of the set of 18,738 well-expressed probes, 15,966 overlapped with the CAGE data, which formed the comparative set.

eQTL Discovery

BOLT-LMM Association Analysis

We used a LMM, implemented in the BOLT-LMM software,[31] to identify SNP-probe associations across 36,778 mRNA transcript level phenotypes in a computationally efficient manner, while accounting for the population structure present in the data. BOLT-LMM was chosen because it has high computational efficiency, performs LMM analysis, and uses a mixture of two normal distributions for the genetic effects. The standard LMM, referred to as the "the infinitesimal model," implicitly assumes that all variants have an effect that is drawn from independent Gaussian distributions. BOLT-LMM relaxes the assumptions of the infinitesimal model by using a mixture of two Gaussian distributions as the prior on the genetic effects, giving the model greater flexibility to accommodate SNPs of large effect, which are often present for expression traits, while maintaining effective modeling of genome-wide effects (for example, ancestry).[31]

We estimated SNP effects for each combination of 7,763,174 autosomal SNPs against 36,778 probes using data from all 2,765 individuals. To increase computational efficiency while maintaining power and correction for confounding, we used the *modelSnps* option in BOLT-LMM, which requires the specification of a set of linkage disequilibrium (LD) pruned SNPs, and was set to be the HapMap 3 set of SNPs.

COJO Refinement of SNP-Probe Associations

To subset the extensive set of SNP-probe association results generated by BOLT-LMM, we performed a conditional and joint (COJO) stepwise model selection[32] procedure. The method was implemented in the GCTA software and uses the summary statistics generated from the BOLT-LMM analysis. Probes were carried forward for this analysis if they had a SNP-probe association with a p value $< 5 \times 10^{-8}$. To avoid overfitting in the COJO model selection procedure, an initial clumping of the BOLT-LMM association summary statistics was performed for each probe. This analysis was completed with the PLINK 2 software[33] with an LD threshold R^2 of 0.1 and the default clump distance of 250 kb. The clumped summary statistics were then used for the COJO analysis.

The COJO analysis selects SNPs (*cis* and *trans*) on the basis of conditional p values thresholded at $p < 5 \times 10^{-8}$ and then estimates the joint effects of all selected SNPs after the model has been optimized. GCTA allows for the individual-level genotype data to be used in the procedure; thus, we used the CAGE genotype data as an LD reference for the COJO analysis. An estimate of the proportion of phenotypic variance explained by the identified COJO eQTL was calculated for each probe by fitting the selected SNPs in a multiple linear regression model in the R programming language[34] (with ten PCs fitted as fixed effects to correct for population stratification), and the resultant ratio of the genetic variance and the phenotypic variance taken to be the heritability estimate (h^2_{COJO}). The genetic variance was calculated as $\mathrm{Var}(\mathbf{X}\,\widehat{\boldsymbol{\beta}})$, where $\widehat{\boldsymbol{\beta}}$ is the vector of estimated SNP effects from the multiple regression model and $\mathbf{X}$ the corresponding genotypes. Additionally, for the probes that had an identified eQTL, the proportion of phenotypic variance explained by the sentinel SNP (defined to be the SNP with the smallest association p value for each probe) was calculated by fitting the selected SNP in a linear regression model (with ten PCs added to correct for population structure) and estimating the proportion of phenotypic variance explained by that SNP (h^2_S) as above for the COJO set of SNPs.

Power to Detect SNP-Probe Associations: Mega- versus Meta-analysis

We investigated the statistical power for eQTL discovery using individual-level data versus a meta-analysis by comparing association results from using the CAGE data to those presented in Westra et al.[6] In Westra et al.,[6] Spearman's rank correlations were used to measure the association between genotypes and phenotypes for each of the gene expression data cohorts. These correlations were converted to *t* scores, and then, via the inverse normal distribution, to *z* values. For each dataset *i*, the *z* value for each SNP *j* and probe *m* was weighted by the square root of the sample size for the dataset used to calculate the *z* value for the SNP tested in the association test, i.e.,

$$z_{w_{ijm}} = \sqrt{n_{ij}}\, z_{ijm}$$

For each *cis*-eQTL association present after controlling the false discovery rate at 0.05, Westra et al.[6] reported the weighted *z* value $z_{w_{ijm}}$. If at least three cohorts had results for a SNP-probe pair, the combined *z* value was calculated as

$$z_{meta_{jm}} = \frac{1}{\sqrt{n}} \sum_i z_{w_{ijm}},$$

where *n* is the total number of individuals contributing a weighted *z* score; this statistic was then used to calculate the presented p value. To be consistent with the data present in Westra et al.,[6] a set of unrelated European individuals was determined by removing individuals from the subset of 2,454 European individuals in the CAGE dataset via a threshold of 0.05 on the off-diagonals of the genetic relationship matrix (GRM) (Figure S9). This resulted in the removal of a further 706 individuals, leaving n = 1,748 individuals for comparison. We recalculated the $z_{meta_{jm}}$ values from the Westra et al.[6] study using the DILGOM cohort[35] (n = 509) and the largest Fehrmann cohort[36] (n = 1,240), which resulted in n = 1,749 individuals. These cohorts were chosen because they were the largest cohorts that when summed had a similar number of individuals to the set of unrelated Europeans from the CAGE dataset. The resultant *z* values were converted to χ^2 statistics by squaring these values. We preferred to make comparisons between the χ^2 statistics because they are on the scale of the number of individuals and are all positive. Additionally, a comparison between effect sizes was made by estimating $\widehat{\beta}_{jm}$ from the recalculated $z_{meta_{jm}}$ statistics. This required the estimation of an approximate standard error for each $\widehat{\beta}_{jm}$, which was calculated as $\sigma\,(\widehat{\beta}_{jm}) = 1/\sqrt{2p_j(1-p_j)(n + z^2_{meta_{jm}})}$ where p_j is the allele frequency for SNP *j* (obtained from a large independent dataset of unrelated Europeans) and n = 1,749.

To compare the results from the two datasets, the sentinel SNP (from the *cis* set of results in Westra et al.[6]) for each of 3,450 overlapping probes reported in Westra et al.[6] were used. For the 3,450 probes, an association analysis using the BOLT-LMM software was run on the set of unrelated European individuals in CAGE. To provide further comparison, SNP-probe associations for the overlapping sentinel SNPs were investigated using a standard single-SNP linear association analysis performed in the PLINK 2 software, with the first ten PCs of the genotype matrix used as covariates. This analysis was chosen to provide a baseline comparison with a standard analysis performed in the literature and reflected a methodology closer to that used in Westra et al.[6]

We investigated a potential deflation of the test statistics as a function of the amount of variance explained by an individual SNP. BOLT-LMM uses an approximate method that first estimates the variance components of the LMM under the null model (no SNP effect) and then keeps the variance components from the null model fixed when testing the effect of each SNP. This reduces computation time, but the assumption that the variance explained by each SNP is approximately zero is a good approximation only for highly polygenic traits. For eQTL that explain a large proportion of phenotypic variance (up to 60% observed for a single eQTL in the CAGE analysis), this assumption leads to a deflation of the χ^2 statistics by a factor of approximately $1/(1 - R^2)$. For SNPs that explain a large amount of phenotypic variance, an exact test that repeatedly estimates variance components when performing each association is desirable. Zhou and Stephens[37] presented an efficient exact method, referred to as genome-wide efficient mixed-model association (GEMMA), that makes approximations unnecessary in many contexts but is computationally less efficient than BOLT-LMM and thus was not viable for the full CAGE analysis. To provide more exact estimates of χ^2 statistics for reference and comparison, we performed a LMM eQTL analysis using the GEMMA software for the 3,450 overlapping probes.

To make comparisons between sets of χ^2 statistics for the sentinel SNPs from the different methodologies, a linear model was fitted with no intercept term. Regression slopes were then used to measure whether the χ^2 statistics were on average greater than those calculated in Westra et al.[6]

Table 1. Summary of Identified eQTL

No. eQTL per Probe	Probes	Genes	eQTL	*cis*-eQTL	*trans*-eQTL
≧ 1	9,967	8,080	14,995	11,204	3,791
1	6,617	5,707	6,617	4,692	1,925
2	2,231	2,050	4,462	3,419	1,043
3	754	708	2,262	1,775	487
4	242	232	968	780	188
≧ 5	123	112	686	538	148

Summary of eQTL mapping from the BOLT-LMM and COJO analyses of the whole CAGE dataset. Of the set of 11,829 probes with at least one COJO eQTL, there were 1,862 probes with a genomic annotation quality score of less than "good" as per the protocol of Barbosa-Morais et al.,[30] and thus the results for 9,967 probes are presented. Genes correspond to the number of unique HGNC gene names for each set of probes. *cis*-eQTL were defined to be those associations such that the SNP was located on the same chromosome as the gene and *trans*-eQTL the complement of this.

Results

Expression Quantitative Trait Loci

We performed an eQTL analysis on 2,765 individuals for each of the 36,778 mRNA transcript phenotypes and 7,763,174 SNPs using a LMM implemented in the BOLT-LMM software.[31] A total of 2,733,370 SNP-probe associations were identified at a p value threshold of 5×10^{-8}. Each probe with one or more associations at this threshold was taken forward for clumping using the PLINK 2 software and then for conditional and joint (COJO) analysis.[32] The COJO analysis selects SNPs (*cis* and *trans*) on the basis of conditional p values (thresholded at $p < 5 \times 10^{-8}$) and estimates the joint effects of all selected SNPs after the model has been optimized. The COJO analysis identified a total of 17,608 eQTLs for 11,829 unique probes and 9,190 HGNC genes. Of this set, 2,613 eQTL (1,862 probes) were for probes with a genome annotation quality score of less than "good" as per the protocol of Barbosa-Morais et al.,[30] making them unreliable for classification as *cis* or *trans*. The remaining 14,995 eQTL corresponded to 9,967 probes with 11,204 (75%) located in *cis* and 3,791 (25%) in *trans* (Table 1). *cis*-eQTL were defined to be those associations where the SNP was located on the same chromosome as the gene, and *trans*-eQTL the complement of this. We identified multiple independent eQTLs for 2,306 probes in *cis* and 360 in *trans* (Table S4). All SNP-probe associations below a p value threshold of 1×10^{-6} and the complete set of COJO eQTL are publicly available to download or query using the CAGE Shiny online application (see Web Resources).

Heritability of Gene Expression

For the 36,778 transcripts passing quality control, we estimated narrow-sense heritability (h^2) and the proportion of phenotypic variance explained by genotyped SNPs (h^2_g) via the Big K/Small K method of Zaitlen et al.[16] This analysis was implemented in the GCTA software using both the unconstrained and constrained REML algorithms[29] (see Figure S10 for full distributions of heritability estimates). Poor convergence of the REML algorithm was observed for 6,811 probes in the unconstrained Big K/Small K analysis, and thus to obtain estimates for these probes we used the *–reml-force-converge* option in the GCTA software. The majority of the probes with poor convergence had heritability estimates that were close to 0. As an initial benchmark, we also estimated narrow-sense heritability using just the $\mathbf{K}_{\mathrm{IBS}>t}$ matrix of estimated relatedness and the unconstrained REML algorithm. The unconstrained narrow-sense heritability estimates from this model showed very similar results to the sum of the two variance components estimated using the unconstrained Big K/Small K method (Figure S11A), and thus we focused on the results from the Big K/Small K method.

To make conclusions about the proportion of h^2 explained by genotyped SNPs, COJO eQTL, and the sentinel SNP, we compared means and medians across the set of 15,966 overlapping expressed probes from the study of Kirsten et al.[4] This is in contrast to the COJO eQTL results, which are reported for all probes that had a COJO eQTL. To investigate whether this preselection of probes was reasonable, we calculated the average number of identified COJO eQTL in the overlapping expressed probes from the study of Kirsten et al.[4] and for the complement set of probes (20,812). For the overlapping Kirsten et al.[4] probes, the average number of eQTLs per probe was 0.72 and for the complement the average number was 0.29. Therefore, for the comparative set, we observed a greater than 2-fold enrichment for identified eQTLs, implying that our preselected set was much more likely to contain probes with a genetic contribution to variation. For the set of 15,966 overlapping probes, the mean and median estimates of h^2 from the constrained algorithm were 0.139 and 0.089 (Table 2 and Figure S12). Average standard errors across the 15,966 probes for h^2 and h^{2*} were approximately 0.053 and 0.052, respectively (Figure S13). Of the set of 15,966 probes, 10,580 probes (66%) had a $\widehat{h}^{2*}$ greater than 0, representing 8,842 unique HGNC genes (Table 2). The mean and median from the constrained algorithm for these probes were 0.192 and 0.142, respectively, with smaller estimates from the unconstrained algorithm of 0.158 and 0.103 (Table 2 and Figure 1).

Missing Heritability for Gene Expression

For all probes, estimates of the proportion of variance explained by significant eQTLs (h^2_{COJO}) were summarized to investigate the extent of missing heritability for gene expression. Across the set of 15,966 probes, the sentinel SNP of the largest eQTL for a gene explained on average 88% (0.036/0.041) of the variance attributed to all identified *cis*- and *trans*-eQTL (h^2_{COJO}). Across this same set of probes, h^2_{COJO} explained on average 30% (0.041/0.139) of h^2, suggesting that 70% of the heritability is missing (Table 2). For the set of expressed probes with a h^{2*} estimate greater than zero (10,580 probes), 6,585 (62%) had

Table 2. Summary of Heritability Estimates across Overlapping Probes from the Study of Kirsten et al.[4]

Threshold		h^2	h^{2*}	h_g^2	h_g^{2*}	h_{COJO}^2	h_S^2
Expressed probes (15,966)	mean	0.139	0.089	0.068	0.052	0.041	0.036
	median	0.089	0.042	0.022	0.036	0.000	0.000
$\widehat{h}^{2*} > 0$ (10,580)	mean	0.192	0.158	0.093	0.079	0.060	0.052
	median	0.142	0.103	0.048	0.056	0.018	0.016
$\widehat{h}^{2*} > 0.05$ (7,560)	mean	0.241	0.212	0.116	0.104	0.081	0.070
	median	0.193	0.158	0.074	0.077	0.036	0.029
$\widehat{h}^{2*} > 0.1$ (5,383)	mean	0.294	0.268	0.142	0.136	0.106	0.091
	median	0.245	0.218	0.100	0.100	0.060	0.047
$\widehat{h}^{2*} > 0.2$ (2,987)	mean	0.391	0.368	0.194	0.198	0.158	0.135
	median	0.349	0.329	0.148	0.148	0.117	0.090
$\widehat{h}^{2*} > 0.4$ (997)	mean	0.566	0.538	0.304	0.330	0.273	0.234
	median	0.536	0.512	0.264	0.264	0.258	0.205

Numbers in parentheses indicate the total number of probes used to calculate estimates. For Big K/Small K narrow-sense heritability estimates (h^2 and h^{2*}) and the proportion of phenotypic variance explained by genome-wide HapMap 3 SNPs (h_g^2 and h_g^{2*}), all European individuals in CAGE with varying degrees of relatedness were used (n = 2,454). The asterisk (*) notation refers to the results from the unconstrained variance components REML algorithm implemented in the GCTA software. The parameters h_{COJO}^2 and h_S^2 correspond to the proportion of phenotypic variance explained by COJO eQTL and the sentinel SNPs, respectively.

one or more independent significant eQTL identified from the COJO analysis, leaving 3,995 having no significant eQTL. For those probes with no significant eQTL, h_{COJO}^2 was set to zero when calculating averages across probes, as were all probes without a COJO eQTL across other $\widehat{h}^{2*}$ threshold summaries. For these probes, similar on average proportions were seen, with 87% (0.052/0.060) of h_{COJO}^2 being explained by h_S^2 and 31% (0.060/0.192) of h^2 explained by h_{COJO}^2 (Table 2). For transcripts with a $\widehat{h}^{2*} > 0.4$ (997 probes), on average 48% (0.273/0.566) of h^2 could be attributed to h_{COJO}^2. Of the set of 15,966 probes, a total of 2,634 probes (2,387 unique genes) had an estimate of h_{COJO}^2 that explained greater than 50% of h^2, indicating that their genetic architecture is predominantly driven by a few loci of large effect. We also observed a positive linear relationship between estimates of h^2 and h_{COJO}^2, suggesting that as the heritability of gene expression transcripts increases, so does the proportion of phenotypic variance explained by identified QTLs (Figure 2B).

The ratio of h_{COJO}^2 and h_g^2 gives an indication of the degree of "hiding" heritability, which is most likely due to common variants of small effect.[38] Across the set of 15,966 probes, on average 60% (0.041/0.068) of h_g^2 is explained by h_{COJO}^2, with the proportion increasing to 65% (0.060/0.093) for expressed transcripts with a $\widehat{h}^{2*} > 0$. Average standard errors for h_g^2 and h_g^{2*} across the 15,966 probes were approximately 0.129 and 0.126, respectively (Figure S13). For transcripts with a $\widehat{h}^{2*} > 0.4$, on average 90% (0.273/0.304) of h_g^2 could be attributed to h_{COJO}^2 (Table 2). These results suggest that for more heritable probes there is less hiding heritability.

The ratio of h_g^2 and h^2 represents the "still-missing" heritability, which is most likely due to variants that are poorly tagged by genotyped SNPs, for example due to rare variants. An alternative explanation is that h^2 is biased upward due to confounding by non-additive or non-genetic factors. Across the set of 15,966 probes, on average 49% (0.068/0.139) of h^2 could be attributed to h_g^2, suggesting that 51% is still missing (Table 2). For the set of probes with $\widehat{h}^{2*} > 0$, a similar on average proportion of 48% (0.093/0.192) was observed, which increases to 54% (0.304/0.566) for transcripts with a $\widehat{h}^{2*} > 0.4$. These results suggest that on average approximately half of the narrow-sense heritability is captured by genome-wide HapMap 3 SNPs. This is in contrast to results for human complex traits, where it has been observed across 49 human phenotypes that h_g^2 is approximately one third of h^2.[39] The proportion of hiding and still-missing heritability for each probe is available to download at the CAGE Shiny online application (see Web Resources).

Mega- versus Meta-analysis Chi-Square Statistics

We investigated the relative statistical power to identify eQTL when using individual-level data versus meta-analyzed summary statistics by comparing the results from the analysis of the CAGE data to a published meta-analysis.[6] Association chi-square (χ^2) statistics for 3,450 sentinel SNPs (common to both studies) were compared between the meta-analysis and those obtained by analyzing the CAGE data using a single SNP analysis in PLINK and a LMM fitted with BOLT-LMM. Comparisons between association χ^2 statistics for all common sentinel SNPs were made via regressing the χ^2 statistics generated from CAGE on those obtained in the meta-analysis.

Linear regressions of mega-analysis association χ^2 statistics (CAGE), generated using single-SNP regression in PLINK 2 and a LMM in BOLT-LMM, on meta-analysis χ^2 statistics showed slope coefficients of 1.5 and 0.86,

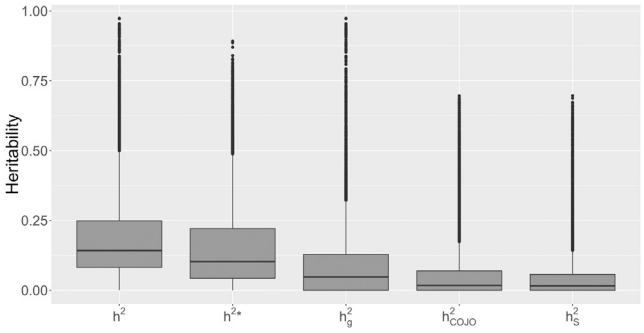

Figure 1. Boxplot Summary of Heritability Estimates
The summarized results are for the set of 10,580 probes that had a $\widehat{h}^{2*}$ greater than 0 from the set of overlapping expressed probes from the Kirsten et al.[4] study. Estimates from the Big K/Small K method are displayed for the narrow-sense heritability from the constrained algorithm (h^2), the narrow-sense heritability from the unconstrained algorithm (h^{2*}), and the proportion of phenotypic variance explained by genome-wide HapMap 3 SNPs (h^2_g) from the constrained REML algorithm, which used European individuals (n = 2,454). The parameters h^2_{COJO} and h^2_S refer to the proportion of phenotypic variance explained by COJO eQTL and the sentinel SNP.

respectively (Figures 3 and S14A). We expected the slopes of the single-SNP regression analysis and the LMM to be approximately the same, but we observed a deflation in the χ^2 statistics from BOLT-LMM relative to the PLINK analysis. Upon investigation, this deflation is expected from theory (see Material and Methods). A deviation between PLINK and BOLT-LMM was seen after a χ^2 statistic of $\approx$100 (Figure S14D), which has little practical consequence for discovery and significance given that such test statistics are large.

The deviation between the BOLT-LMM and GEMMA-LMM statistics for the set of overlapping sentinel SNPs is substantial, with the same parabolic deflation seen as in the comparison of BOLT-LMM and PLINK (Figure S14C). The regression slope from the GEMMA-LMM comparison with the Westra et al.[6] meta-analysis was 1.49 (Figure 3) and thus, the CAGE data have χ^2 statistics for sentinel SNPs across 3,450 probes that are on average approximately 50% greater than the meta-analysis χ^2 statistics. This increase in χ^2 statistics is partially due to an increase in estimated effect sizes. A regression slope of 1.20 was observed when regressing $\widehat{\beta}_{jm}$ statistics from the PLINK and GEMMA-LMM analyses in the CAGE data on those from the approximate effects calculated from the meta-analysis *z* values (Figures S14E and S14F).

Discussion

We have presented results from the examination of the genetic architecture of gene expression in blood tissue from 2,765 individuals. We identified 11,204 *cis*- and 3,791 independent *trans*-eQTLs using a two-step analysis of all 36,778 probes in CAGE, with multiple independent eQTLs detected for 2,306 probes in *cis* and 306 in *trans*. Using information on both closely and distantly related individuals, we estimated heritability for all probes in the CAGE dataset. We showed that across overlapping expressed probes from the study of Kirsten et al.[4] that had a h^{2*} estimate greater than zero (10,580), on average h^2_{COJO} explained 31% (0.060/0.192) of h^2, suggesting that 69% is missing. For this same set of probes, on average 48% (0.093/0.192) of h^2 could be attributed to additive genetic values captured by genome-wide HapMap 3 SNPs (h^2_g), suggesting that approximately half of the heritability of gene expression is "still" missing[38] for these probes. Additionally, 65% (0.060/0.093) of the variance explained by genotyped SNPs (h^2_g) could be detected at a genome-wide significance threshold; this value increased to 90% (0.273/0.304) for transcripts with $\widehat{h}^{2*} > 0.4$. Therefore, for this set of transcripts, approximately half of the variance for gene expression is not tagged by common SNPs, while the majority of variance that is tagged is due to detected eQTL. Additionally, we observed a positive linear relationship between the heritability of probes and the proportion of phenotypic variance that can be explained by COJO-eQTL, implying that, on average, more heritable probes have larger effects. This is in contrast to what is observed for the majority of complex traits and common diseases.[40]

There is the potential for h^2 estimates to be inflated due to effects such as dominance, shared environment, and epistatic variance,[16,41] although there is little evidence that non-additive genetic variation contributes considerably to variation in gene expression.[8] In addition to these sources of bias, we acknowledge that the presented mean Big K/Small K heritability estimates across probes are biased due to sampling variance. The estimates of h^2_{COJO} and h^2_S also contain a contribution from overestimated effects due to the winner's curse, although the contribution to the mean is likely to be small given that the effects are large for the majority of expression traits. Furthermore, the heritability estimates from the constrained REML algorithm are potentially biased due to the bounded variance component parameter space, which is alleviated by the reporting of the estimates from the unconstrained REML algorithm. Schweiger et al.[42] showed that the reported standard errors from the constrained REML algorithm led to the construction of confidence intervals with inaccurate coverage probabilities. However, the reported mean standard error from the constrained REML algorithm is a meaningful measure of the uncertainty in these estimates due to the law of large numbers. Additionally, the array technology used in this study may lack sufficient resolution to identify variation in lowly expressed genes, which may be abated by studying large cohorts with RNA-seq. The ideal set for making conclusions about missing heritability would be the set of probes with a genetic contribution to gene expression variation in peripheral blood. In reality, no selection of probes is perfect for comparison and thus we made a selection based upon external data, where

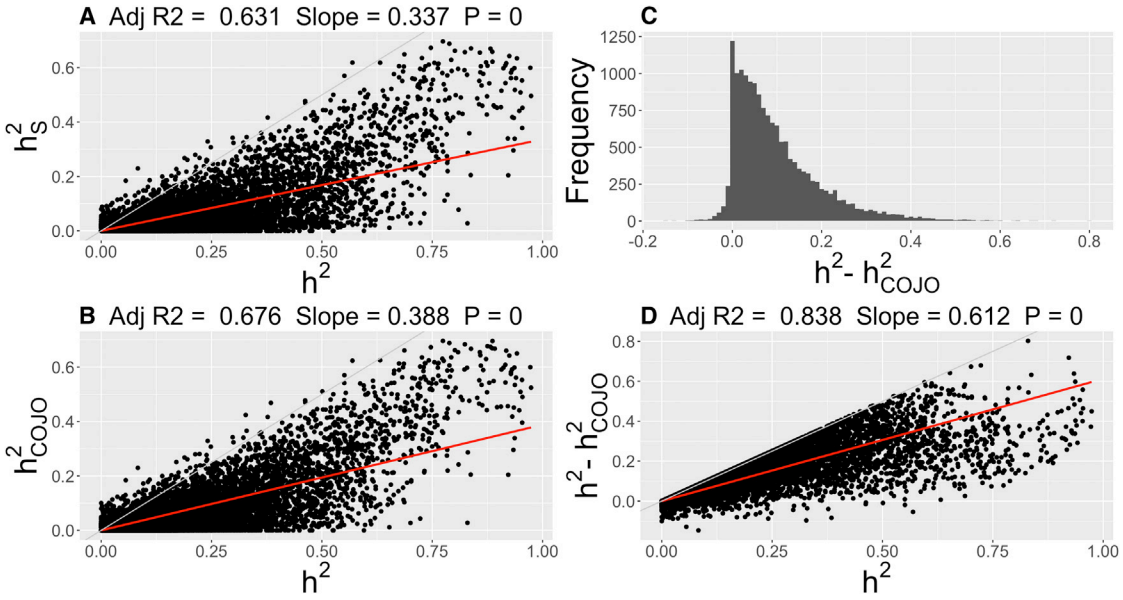

Figure 2. Missing Heritability

Scatterplot and density summaries of narrow-sense heritability estimates (constrained REML algorithm) from the Big K/Small K method (h^2), the proportion of phenotypic variance explained by COJO eQTL (h^2_{COJO}), and the proportion of phenotypic variance explained by the sentinel SNP (h^2_S). Displayed summaries are across 15,966 overlapping expressed probes from the Kirsten et al.[4] study.

(A) Scatterplot of Big K/Small K heritability estimates versus the proportion of phenotypic variance explained by the sentinel SNP.

(B) Scatterplot of Big K/Small K heritability estimates versus the proportion of phenotypic variance explained by the COJO eQTL.

(C) Histogram of the difference between Big K/Small K heritability estimates and the proportion of phenotypic variance explained by the COJO eQTL.

(D) Scatterplot of Big K/Small K heritability estimates versus the difference from (C).

For (A), (B), and (D), the fitted regression line (red) and 95% confidence interval (shaded) is plotted with the key statistics of this regression (no intercept term fitted) displayed at the top of the panels. The light gray line represents the $y = x$ line. The p value is with regard to the regression slope.

each probe had evidence for variation of which additive genetic variation could be a potential contributor. The set of probes chosen showed a greater than 2-fold enrichment for identified eQTLs, which reinforced our preselection of this set of probes.

The estimated value of h^2_g is an upper bound on the proportion of variation that can be attributed to all SNPs on a given genotyping platform and is almost entirely made up of common variation. One potential reason for the differences between h^2_g and h^2 is that rare variation accounts for a significant fraction of the total narrow-sense heritability. Recently, Zhao et al.[43] showed that an excess of rare variants contributed to both the high and low expression levels of many genes in blood. It is important to recognize that blood is a heterogenous tissue made up of multiple cell types, and although it is likely that *cis* effects will be shared across cell types,[9] we expect some variability in average heritability estimates for expression transcripts across blood cell types, meaning that our estimates are likely to reflect averaged effects. This heterogeneity may be particularly evident for immune-specific cells, where Brodin et al.[44] showed that for many of the component parts of the immune system, a considerable amount of the variation in humans is driven by non-heritable factors.

The individual-level data of the CAGE resource allowed for a genome-wide eQTL analysis to be performed using a LMM, which accounts for population stratification and cryptic relatedness and improves statistical power due to joint modeling of all genotyped markers. Additionally, the LMM methodology used has increased flexibility to model SNPs of large effect, which are often present for gene expression phenotypes. The results from the COJO-eQTL analysis allowed for a characterization of independent eQTL signals with 17,608 eQTLs identified for 11,829 transcripts (9,190 unique genes). The majority of the identified eQTL are located in *cis* with 25% of the identified eQTL being in *trans*. A similar percentage (29%) of genes were identified as being *trans*-regulated (relative to all genes with an eQTL) in the study of Kirsten et al.[4] While the majority of COJO eQTLs are likely to tag independent causal variants, there is the possibility that multiple eQTLs could be in LD with a single causal variant of very large effect.[32] The meta-analysis comparison also showed that linear mixed model methods that reduce computational burden by assuming that the variance components estimated under the null model of no effect at the candidate marker,[45] or the variance explained by a single SNP is small, may not be adequate for gene expression traits because many loci can explain a large amount (>10%) of the phenotypic variance. We demonstrated that using individual-level data can increase the χ^2 statistics for eQTLs on average, with a 50% increase in χ^2 statistics compared

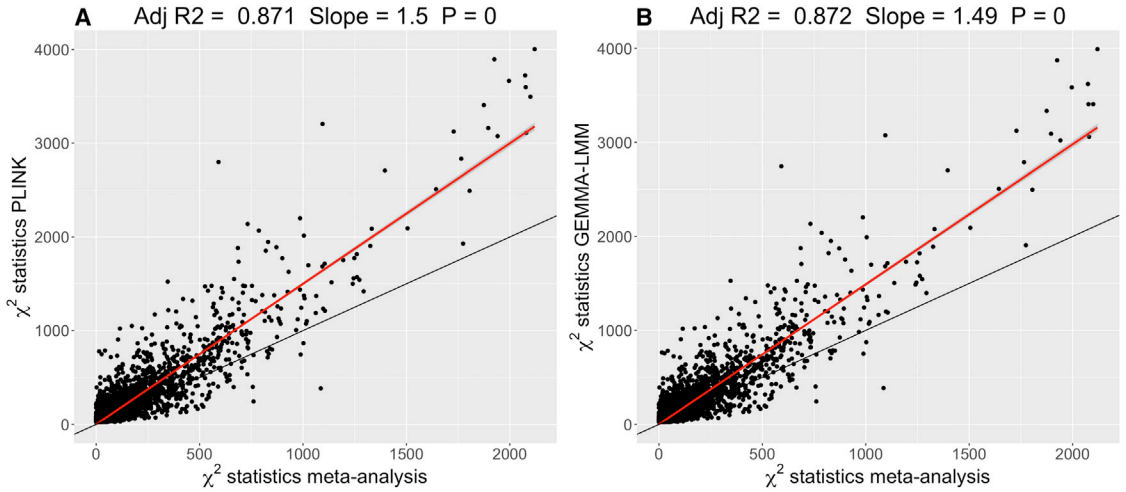

Figure 3. Mega- versus Meta-analysis Chi-Square Statistics
Comparison of association χ^2 statistics for the sentinel SNP from the top 3,450 *cis* probes generated from a subset of the meta-analysis of Westra et al.[6] (n = 1,749) and analyses of CAGE data using European unrelated individuals (n = 1,748).
(A) Comparison of the set of association χ^2 statistics generated using a linear model analysis of sentinel SNPs from the CAGE dataset (analyzed in PLINK and corrected for ten PCs) versus those from the meta-analysis.
(B) Comparison of the association χ^2 statistics for sentinel SNPs from the GEMMA-LMM analysis (GRM generated from HapMap 3 SNPs) and the meta-analysis. All panels include the fitted regression line (red) and its 95% confidence interval (shaded) with the key statistics of this regression (no intercept term fitted) displayed at the top of each panel. The p value is with regard to the regression slope. Additionally, the $y = x$ line (black) line is plotted for reference.

with a meta-analysis. However, it is important to note that the meta-analysis of Westra et al.[6] is more powerful given its larger sample size. The information differences shown here may be caused by the difficulties inherent in sharing summary statistics and the heterogeneity caused in cohort processing.[46] A final additional benefit of using raw-level data is the ability to employ a variety of data normalization pipelines and more complex analyses such as the LMM, to account for cryptic relatedness and population structure, and conditional single SNP modeling.

This resource has allowed for an exhaustive eQTL analysis and has characterized the heritability of gene expression by studying thousands of mRNA profiles using contrasting methods. Our eQTL results are a valuable resource to explore the relevance of SNPs identified in current as well as future GWASs. These results and data will form the basis of further study into the genetic basis of gene expression with the dataset opening the door to explore questions, such as multivariate modeling of joint *cis* effects of SNPs on gene expression variation, genetic co-regulation of mRNA transcripts within peripheral blood across all probes, and sexual dimorphism in gene expression.

Supplemental Data

Supplemental Data include 14 figures, 4 tables, and a supplemental note and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.12.008.

Acknowledgments

This work was supported by the Australian National Health and Medical Research Council (NHMRC) grants (1046880, 1083405, 1107599, 1083656, 1078037, 1078399, 1107599) and the Sylvia and Charles Viertel Charitable Foundation.

Received: June 7, 2016
Accepted: December 14, 2016
Published: January 5, 2017; corrected online February 2, 2017

Web Resources

CAGE Shiny, http://cnsgenomics.com/shiny/CAGE/
GEO, http://www.ncbi.nlm.nih.gov/geo/
International HapMap Project, ftp://ftp.ncbi.nlm.nih.gov/hapmap/

References

1. Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. *16*, 197–212.
2. Edwards, S.L., Beesley, J., French, J.D., and Dunning, A.M. (2013). Beyond GWASs: illuminating the dark road from association to function. Am. J. Hum. Genet. *93*, 779–797.
3. Grundberg, E., Small, K.S., Hedman, Å.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.-P., Meduri, E., Barrett, A., et al.; Multiple Tissue Human Expression Resource (MuTHER) Consortium (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat. Genet. *44*, 1084–1089.
4. Kirsten, H., Al-Hasani, H., Holdt, L., Gross, A., Beutner, F., Krohn, K., Horn, K., Ahnert, P., Burkhardt, R., Reiche, K., et al. (2015). Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci. Hum. Mol. Genet. *24*, 4746–4763.

5. Powell, J.E., Henders, A.K., McRae, A.F., Wright, M.J., Martin, N.G., Dermitzakis, E.T., Montgomery, G.W., and Visscher, P.M. (2012b). Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. Genome Res. *22*, 456–466.
6. Westra, H.-J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. *45*, 1238–1243.
7. Wright, F.A., Sullivan, P.F., Brooks, A.I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y.-H., et al. (2014). Heritability and genomics of gene expression in peripheral blood. Nat. Genet. *46*, 430–437.
8. Powell, J.E., Henders, A.K., McRae, A.F., Kim, J., Hemani, G., Martin, N.G., Dermitzakis, E.T., Gibson, G., Montgomery, G.W., and Visscher, P.M. (2013). Congruence of additive and non-additive effects on gene expression estimated from pedigree and SNP data. PLoS Genet. *9*, e1003502.
9. Price, A.L., Helgason, A., Thorleifsson, G., McCarroll, S.A., Kong, A., and Stefansson, K. (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet. *7*, e1001317.
10. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.
11. Hill, W.G., Goddard, M.E., and Visscher, P.M. (2008). Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. *4*, e1000008.
12. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
13. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.
14. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A., Lee, S.H., Robinson, M.R., Perry, J.R., Nolte, I.M., van Vliet-Ostaptchouk, J.V., et al.; LifeLines Cohort Study (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. *47*, 1114–1120.
15. Gaffney, D.J. (2013). Global properties and functional complexity of human gene regulatory variation. PLoS Genet. *9*, e1003501.
16. Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., and Price, A.L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genet. *9*, e1003520.
17. Powell, J.E., Henders, A.K., McRae, A.F., Caracella, A., Smith, S., Wright, M.J., Whitfield, J.B., Dermitzakis, E.T., Martin, N.G., Visscher, P.M., and Montgomery, G.W. (2012a). The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. PLoS ONE *7*, e35430.
18. Kim, J., Ghasemzadeh, N., Eapen, D.J., Chung, N.C., Storey, J.D., Quyyumi, A.A., and Gibson, G. (2014). Gene expression profiles associated with acute myocardial infarction and risk of cardiovascular death. Genome Med. *6*, 40.
19. Preininger, M., Arafat, D., Kim, J., Nath, A.P., Idaghdour, Y., Brigham, K.L., and Gibson, G. (2013). Blood-informative transcripts define nine common axes of peripheral blood gene expression. PLoS Genet. *9*, e1003362.
20. Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L., et al. (2015). Cohort profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int. J. Epidemiol. *44*, 1137–1147.
21. Idaghdour, Y., Czika, W., Shianna, K.V., Lee, S.H., Visscher, P.M., Martin, H.C., Miclaus, K., Jadallah, S.J., Goldstein, D.B., Wolfinger, R.D., and Gibson, G. (2010). Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nat. Genet. *42*, 62–67.
22. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.; and 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
23. Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics *18* (*Suppl 1*), S96–S104.
24. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). *limma* powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.
25. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.
26. Blom, G. (1958). Statistical Estimates and Transformed Beta-Variables (New York: Wiley).
27. Chen, C.-Y., Pollack, S., Hunter, D.J., Hirschhorn, J.N., Kraft, P., and Price, A.L. (2013). Improved ancestry inference using weights from external reference panels. Bioinformatics *29*, 1399–1406.
28. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
29. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
30. Barbosa-Morais, N.L., Dunning, M.J., Samarajiwa, S.A., Darot, J.F., Ritchie, M.E., Lynch, A.G., and Tavaré, S. (2010). A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. Nucleic Acids Res. *38*, e17.
31. Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.
32. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J., et al.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. *44*, 369–375, S1–S3.
33. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

34. R Core Team (2015). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Founda- tion for Statistical Computing).
35. Inouye, M., Silander, K., Hamalainen, E., Salomaa, V., Harald, K., Jousilahti, P., Männistö, S., Eriksson, J.G., Saarela, J., Ripatti, S., et al. (2010). An immune response network associated with blood lipid levels. PLoS Genet. *6*, e1001113.
36. Fehrmann, R.S., Jansen, R.C., Veldink, J.H., Westra, H.-J., Arends, D., Bonder, M.J., Fu, J., Deelen, P., Groen, H.J., Smolonska, A., et al. (2011). 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.
37. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.
38. Witte, J.S., Visscher, P.M., and Wray, N.R. (2014). The contribution of genetic variants to disease depends on the ruler. Nat. Rev. Genet. *15*, 765–776.
39. Yang, J., Lee, T., Kim, J., Cho, M.C., Han, B.G., Lee, J.Y., Lee, H.J., Cho, S., and Kim, H. (2013). Ubiquitous polygenicity of human complex traits: genome-wide analysis of 49 traits in Koreans. PLoS Genet. *9*, e1003355.
40. Robinson, M.R., Wray, N.R., and Visscher, P.M. (2014). Explaining additional genetic variation in complex traits. Trends Genet. *30*, 124–132.
41. Lynch, M., and Walsh, B. (1998). Genetics and Analysis of Quantitative Traits, *Volume 1* (Massachusetts: Sinauer Sunderland).
42. Schweiger, R., Kaufman, S., Laaksonen, R., Kleber, M.E., März, W., Eskin, E., Rosset, S., and Halperin, E. (2016). Fast and accurate construction of confidence intervals for heritability. Am. J. Hum. Genet. *98*, 1181–1192.
43. Zhao, J., Akinsanmi, I., Arafat, D., Cradick, T.J., Lee, C.M., Banskota, S., Marigorta, U.M., Bao, G., and Gibson, G. (2016). A burden of rare variants associated with extremes of gene expression in human peripheral blood. Am. J. Hum. Genet. *98*, 299–309.
44. Brodin, P., Jojic, V., Gao, T., Bhattacharya, S., Angel, C.J.L., Furman, D., Shen-Orr, S., Dekker, C.L., Swan, G.E., Butte, A.J., et al. (2015). Variation in the human immune system is largely driven by non-heritable influences. Cell *160*, 37–47.
45. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348–354.
46. Panagiotou, O.A., Willer, C.J., Hirschhorn, J.N., and Ioannidis, J.P. (2013). The power of meta-analysis in genome-wide association studies. Annu. Rev. Genomics Hum. Genet. *14*, 441–465.

The American Journal of Human Genetics, Volume *100*

Supplemental Data

The Genetic Architecture of Gene Expression in Peripheral Blood

Luke R. Lloyd-Jones, Alexander Holloway, Allan McRae, Jian Yang, Kerrin Small, Jing Zhao, Biao Zeng, Andrew Bakshi, Andres Metspalu, Manolis Dermitzakis, Greg Gibson, Tim Spector, Grant Montgomery, Tonu Esko, Peter M. Visscher, and Joseph E. Powell

Supplemental note

Gene expression normalisation

Data summary

The initial (Phase 1) CAGE dataset contains expression data from seven unique cohorts: The Brisbane Systems Genetics Study (BSGS) main and pilot studies[20,21](Gene Expression Omnibus number GSE33321); Coronary Artery Disease (CAD)[15] (GSE49925); The Centre for Health Discovery and Well-Being (CHDWB)[22] (GSE35846); The Estonian Genome Centre - University of Tartu (EGCUT)[18] (GSE48348); Morocco[13] (GSE17065); and The Multiple Tissue Human Expression Resource Consortium (MuTHER)[9] (ArrayExpress archive under accession E-TABM-1140).

A summary of the original, uncombined data from these cohorts is given in Table S1. The MuTHER and BSGS pilot LCL cohorts were not taken forward, as the expression levels were not measured from whole blood. Genotype data were not available for the CAD batch 2 cohort, and thus these samples were excluded from further analysis.

Quality Control and Normalisation

The CAGE data set comprises multiple cohorts with gene expression levels measured in whole blood. Due to variation in microarray gene assaying processes such as sample treatment, labelling, dye hybridisation and detection, the gene expression levels (measured in array fluorescence intensities) cannot, in general, be compared directly without first performing normalisation steps. Most approaches to normalising gene expression levels from microarray data assume that the overall distribution of mRNA does not vary much between samples. This seems reasonable for most laboratory treatments, however, within and between laboratories large systemic error effects may arise—*i*.e. between laboratory batch effects. The expression normalisation method implemented here consists of six steps, with a subset of the steps carried out on the individual data cohorts (Table S1), followed by concatenation into a single dataset and subsequent final normalisation.

- Variance stabilisation – an alternative to $\log_2$ transformation that more adequately corrects for the fact that the variance of microarray measured spot intensities increases with mean signal intensity
- Quantile normalisation – coerces the intensity values for all probes on a chip to a single common distribution
- Age, cell counts and batch effect correction along with correction for other unobserved heterogeneous sources of variability using the PEER[27] software
- PEER residual phenotypes standardised to z-scores within cohort and concatenation of all cohorts to a final matrix
- PEER and gender correction of final concatenated residual matrix
- Rank normal transformation of PEER residuals to a normal distribution with mean 0 and variance 1

All of the expression normalisation steps were carried out in the statistical computing software, R[24], using a combination of native functions, the PEER[27] program, and functions made available by Bioconductor packages[11].

Variance stabilisation

It is common practice to transform microarray data to a logarithmic (usually base 2) scale. This transformation collapses the original range of the signal and, moreover, it decouples a random multiplicative error term from the true signal intensity. This is desirable because it is well known that the variance of microarray signal intensities increases with the mean signal intensity[17]. However, this transformation assumes a multiplicative model which predicts that measurement error vanishes for very small signals, whereas microarray data will always contain background noise. Thus, the logarithmic transform does not adequately adjust the variance for low-intensity signals with the post transformation variances being larger than expected. A more realistic model allows for both an additive and a multiplicative error term.

The method of Huber *et al.*[12] includes both an additive and a multiplicative error term, and has been shown to be more successful at decoupling the signal variance and signal

mean intensity in real data. As an alternative to performing $\log_2$ transformation, we used the method of Huber *et al.* [12] as implemented in the `vsn` package in Bioconductor.

Quantile normalisation

In order to allow for a fair comparison of intensities between probes, the distribution of expression intensities are mapped to a standard distribution (generated from the data) via a process known as quantile normalisation[3]. This procedure explicitly assumes that the distribution of gene expression measures does not change across samples. We used the function `normalizeBetweenArrays`, from the `limma` package[26] to implement this method. While quantile normalisation is a fast solution, one potential problem is that the genes in the upper range of intensity are forced into a common distribution shape, leading to a reduction in both biological and technical variation[25].

PEER correction analysis

Age, gender, cell counts and batch effects are known to be large sources of variation in gene expression array data[8]. Not all cohorts had recorded values for age, cell counts and batch information such as Illumina Sentrix ID, Sentrix position, and extraction date. Therefore, we utilised the PEER software[27] to account for such sources of variation in the absence of these measurements. The algorithm used by the PEER software reduces overfitting by estimating a suitable number of factors that explain a broad amount of the variation. The software also allows for known covariates, such as age, gender, cell counts and batch effects, to be included in the variance correction analysis concurrently. Relevant covariate measurements available for some cohorts, included age, gender, cell counts for basophils, eosinophils, neutrophils, lymphocytes, monocytes, and array scan date, scan order, Sentrix ID, and Sentrix position. If any such covariates were available for an individual cohort they were included in the PEER correction analysis. Correction for hidden sources of variation via principal components analysis (PCA) is less effective than PEER in the sense that the number of unobserved factors is often pre-specified, whereas PEER uses automatic relevance determination to choose a suitable effective number of factors[27]. Hence, the

number of factors initially specified for the PEER analysis only needs to be sufficiently large. For all cohorts we chose the maximum number of relevant factors to be 50. The PEER correction analysis was performed on all cohorts separately with residuals from the analysis standardised to z-scores across individuals to form the new within cohort expression phenotypes.

Concatenation, final PEER correction and rank normal transformation

Residual phenotypes for each cohort from the previous step were concatenated to form a large expression matrix with $n = 2,765$ individuals. To create a combined gene expression matrix, it was necessary to retain only those probes that are common to all cohorts. In the case of blood samples, this meant reducing the total number of examined probes from approximately $47,000$ to $38,624$.

Post concatenation, the expression matrix was again PEER corrected, using a potential of 50 factors and gender as a covariate. Gender was included at this stage of the analysis because it was the only covariate measured on all individuals in CAGE. The residuals for each probe from this final PEER analysis were transformed using the rank normal transformation of Blom[2], which alters the distribution of scores to be normally distributed with a mean of 0 and a standard deviation of 1.

Removal of probes on sex chromosomes

Probes measuring expression levels of genes located on the X and Y chromosomes were removed from the analysis. The analysis was restricted to autosomal probes because of the difficulties in adequately modelling the potential sex biases in gene expression, which are primarily driven by escape from X chromosome inactivation and male-only expression on the Y chromosome. Illumina probe identifiers were mapped to a genomic location using the re-annotated Illumina Human HT12v4 probe sequences in the Bioconductor illuminaHumanv4.db database[6], and if they mapped to the X and Y chromosomes they were removed. Of the 38,624 probes present after cohort concatenation, 1,846 were mapped to positions on the sex chromosomes leaving 36,778 for analysis.

Expression matrix quality control

To verify the performance of the normalisation steps, and to identify any cohorts that contained irregularities, PCA was performed on the final normalised expression matrix. The results of the analysis for the first four PCs can be seen in Figure S1, where all of the samples are distributed with no unique patterns across cohorts, implying that the main sources of variation are not generated by cohort differences. This check is qualitative in the sense that if individual within cohorts are seen to cluster, it would indicate between cohort differences in variance structure. The same pattern was observed for all combinations of PCs 1-20 (figures not shown), suggesting that no single cohort has a unique variance structure across probes for the first 20 PCs.

To verify the correction for covariates within the PEER analysis, we performed linear regression (in the R programming language) of the normalised gene expression measurements for all 36,778 probes on the covariates age, cohort, gender, cell counts, the first 10 principal components (multiple regression) of the genotype matrix from all individuals, and the first 10 principal components (multiple regression) from the genotype matrix of European individuals (defined in Supporting Material). The regression for age, and cell counts was only performed on those individuals that had these measurements (age - $n = 1,164$, cell counts - $n = 793$). The adjusted R-squared values from these 36,778 regressions were visualised as a histogram for each covariate (Figure S2). These analyses indicate that the PEER analysis has adequately adjusted for age, gender, cell counts and cohort differences with means and medians across all probes for these covariates being 0 (Figure S2). The first 10 PCs of the genotype matrix have an on average adjusted R-squared greater than 0 and thus when performing genetic analyses we used a combination of linear mixed models and genotype PC adjustment to account for population stratification.

Genotype imputation and quality control

In addition to the imputation process itself, it was necessary to perform quality control steps on both pre- and post-imputation data, for example, filtering on data features such as minor allele frequency (MAF), genotype missing rate, and Hardy-Weinberg equilibrium. The entire imputation process, and its associated quality control steps were performed using the following publicly available pipeline <https://github.com/CNSGenomics/impute-pipe>.

The imputation pipeline comprised the following steps:

- Pre-imputation quality control, and data-consistency checks
- Imputation to reference panel
- Post-imputation quality control – filtering
- Merging datasets on common SNPs

Pre-Imputation quality control, and imputation to the reference panel

In order to perform imputation, it was necessary to supply a "strand file" for the genotype chip used on each cohort, in order to correctly align alleles to a common strand (*i.e.* positive or negative). In cases where this information had been supplied by the data providers, the necessary strand file was taken from <http://www.well.ox.ac.uk/~wrayner/strand/>. This process ensures that the strand from the 1000 Genomes reference panel and the data set being imputed are the same.

For each dataset, a strand summary table with key statistics on SNP allele alignment with the 1000 Genomes Phase 1 Version 3[4] imputed (in house) Health and Retirement Study (HRS) data set used as a reference (dbGaP Study Accession: phs000428.v1.p1) was produced. Strand alignment was checked using the Genotype Harmoniser software Deelen *et al.*[5].

Once the pre-imputation quality control was completed, imputation was performed as per the protocol outlined at <https://github.com/CNSGenomics/impute-pipe>. The reference panel used was the 1000 Genomes Phase 1 Version 3.

Imputed data merging

After imputation each cohort contained approximately 38 million SNPs. A post imputation check for an adequate proportion of SNPs with high 'info' score was conducted for each cohort; the prior expectation for this proportion was driven by previous experience with imputation. The info score is a quality metric output by IMPUTE2[10] (a component of the imputation pipeline) that ranges between 0 and 1 – where a higher value indicates greater certainty of imputation. To merge these datasets it was necessary to identify the subset of SNPs that were common to all cohorts. To reduce the computational cost of this process, we applied initial filtering on two info score thresholds: 0.9 and 0.3. Two thresholds allow for more flexibility in downstream analyses. Matching over common SNPs yielded approximately 5.4 millions SNPs for the 0.9 threshold, and 8.2 million SNPs for the 0.3 threshold.

Once the common SNP lists were determined, we used PLINK[23] to merge the datasets to form the final genotype dataset. During this process approximately 500 SNPs were removed due to multi-allelic differences between cohorts. These are likely to be a mix of true multi-allelic SNPs and so-called "palindromic SNPs" that were not flipped correctly during the imputation process.

The BSGS and EGCUT cohorts consisted of multiple data sets and were found to contain some duplicates IDs (89 in total). BSGS contained 10 duplicate IDs between the main and pilot studies. For BSGS, the genotype data were subsetted to the duplicate individuals and a subset of 10,000 SNPs; correlations between the genotypes of the individuals with duplicate IDs across these SNPs showed that these individuals were either monozygotic (MZ) twins or the same individual (i.e., they had correlations across the 10k SNPs of $>$ 0.95). To differentiate these samples further, we performed a correlation analysis of the gene expression data across all common probes for the duplicate ID individuals. The BSGS main and pilot data were generated from distinct samples at two time instances with procedural and microarray differences. The gene expression correlation results showed that these individuals had a high correlation (average of approximately 0.9). The values

were lower than expected for a duplicate individual but this could be accounted for by the differences in procedure between the main and pilot studies. Further investigation of the empirical distribution of correlations generated by comparing all individuals across these two data sets was carried out; given this distribution we could not conclude with certainty that these individuals were the same. Further correspondence with the laboratory established that approximately 10 individuals were duplicated across the main and pilot studies. Given this evidence, we decided to retain one set of these individuals with the genotype and expression data kept from the main study. The main study was chosen because it was a more recent study, from a larger cohort, and from the more recent array (Table S1)

For EGCUT, the data provided consisted of one expression data set containing 1,065 individuals, and two sets of genotypes containing 1,144 total (non-unique) individuals (Tables S1 and S2). A total of 79 duplicate IDs were identified between the two genotype datasets, accounting for the difference in total individuals observed between the expression and genotype data. A similar correlation study (to BSGS above) was carried out for the genotype data and again we concluded that these individuals were either MZ twins or the same individuals. As no expression duplicates IDs were found, we concluded that these individuals were very likely to have been duplicated across the two data sets and thus we carried forward the genotype data from the newer chip (*i.e.* the HumanOmniExpress 12v1).

Post-merge quality control

Post merging of the genotype data, allele frequency checks were performed within cohort (by subsetting the merged genotype matrix) to remove any potential SNPs with large allele frequencies differences from the 1000 Genomes reference. This analysis was performed by comparing the allele frequencies for all SNPs in the merged CAGE data with European allele frequencies ($n = 379$) in the 1000 Genomes Phase 1 Version 3[4]. These analyses were performed on the 8.2 million SNPs for the 0.3 threshold data set as the 0.9 (info score threshold) set of SNPs was a subset of the 0.3 set. To make these comparisons, the allele used to calculate the allele frequency was updated for each cohort to the allele in the 1000 Genomes using the GCTA software[30], to ensure comparison of allele frequencies for the same allele. If SNP allele frequencies within cohort differed by more than 0.2 (absolute value) from those in the 1000 Genomes then they were removed from the CAGE genotype data set using the PLINK 2 software. The choice of a 0.2 allele frequency difference cutoff was based on the standard used for the Haplotype Reference Consortium's[19] data preparation toolbox. The BSGS, CAD, CHDWB, and EGCUT cohorts contained individuals of predominantly European ancestry, and therefore the variation in allele frequencies in these cohorts relative to the 1000 Genomes European reference was smaller than that of the Moroccan cohort (Figure S3). As the Moroccan cohort was relatively small ($n = 188$) and is ancestrally diverged from Europe there was greater variation in the allele frequencies relative to the 1000 Genomes European reference. This led to many more SNPs being removed due to allele frequency differences in the Moroccan cohort (Figure S3F) . Approximately 300,000 SNPs were remove from the CAGE genotype data set due to allele frequency differences across all the cohorts, with nearly all of these removed due to the Moroccan cohort. The 0.2 allele frequency threshold was kept for the Moroccan cohort for consistency, and although a large number of SNPs were removed it was a relatively small number of the 8.2 million available. Post removal of allele frequency outlier SNPs, a final check of allele frequencies versus the 1000 Genomes in the whole CAGE data set was performed. No allele frequency outliers were detected with this comparison (Figure S4).

Final quality control on the genotype matrix was implemented, with a minor allele frequency threshold of 0.01, a Hardy-Weinberg equilibrium p-value threshold of 1×10^{-6}, and a genotype call rate threshold of 99% applied to the genotype datasets using the PLINK 2 software. The two final CAGE genotype datasets contained 2,765 individuals with 5,083,862 SNPs for info score threshold 0.9, and 7,763,174 SNPs for info score threshold 0.3.

Post merge we conduced final checks to investigate the quality of the imputed data. To investigate cohort differences in the merged genotype matrix, we generated the first 20 principal components of the genotype matrix using PLINK. These were visualised by plotting successive pairs of PCs against each other. For the 0.3 threshold data the cohorts separate on the first three principal components plots and by the fourth-versus-fifth comparison, separation is reduced (Figure S5). This trend of reduced separation is observed in the remaining PC plots. These plots show that much of the variation in the genotype data can be explained by differences between cohorts. Depending on their research objectives, it will be up to the analysts using these data to decide whether to correct for these differences or not.

Matching expression and genotype data

The final stage of the data preparation process was to match samples between the normalised gene expression and the imputed genotype files. Ensuring the samples' IDs match correctly is vital to ensuring the integrity of downstream analyses. This required three main steps:

- Encode the merged and normalised gene expression matrix with unique CAGE sample identifiers
- Map CAGE sample identifiers to their respective genotype entries, stripping expression samples that lack genotype data
- Verify correctness of identifier mapping

Step one was achieved by simply generating a six-digit, zero-padded numeric identifier

for each unique sample ID in the merged gene expression matrix. This identifier was then prepended with the prefix "CAGE", and appended with an abbreviated dataset code (the inclusion of which simplifies the process of tracing a CAGE-encoded sample back to its parent dataset). The resulting identifiers are of the form `CAGE000123_BSGS_M`—where `BSGS_M` is the abbreviated code for the main BSGS cohort.

The second step was performed by using PLINK to recode (`-update-ids`) the family information of individuals in each of the imputed datasets. A plaintext file was used to map the original sample identifiers to their respective CAGE identifiers, thus creating a list of IDs for PLINK to update. In the cases where a sample did not have a unique family identifier (*i*.e. their individual ID and family ID were the same in PLINK's `.fam` file), it was assigned as the sample's original ID – again, in an attempt to keep the recoding process transparent. Genotyped samples lacking a unique CAGE identifier – indicating they had no associated expression data – were also found during this process, and were dropped via PLINK's `-remove` option.

Finally, it was necessary to determine whether the expression and genotype sample identifiers still mapped individuals correctly. In order to perform this check, we made use of a software tool, MixupMapper[28]. MixupMapper makes use of known eQTL in combination with the genotypic information of each sample in the supplied data to calculate the expected expression level for a number of genes. These estimates are then compared against the observed gene expression levels, and discordance between the two values is taken to be indicative of a "mixup"—*i*.e. an individual whose label in the genotype data does not match the expression data entry of the same label.

The output of MixupMapper is a plaintext report, with one row for each individual in the supplied dataset. Each individual's original expression and genotype IDs are listed, with a score describing their relationship, the ID of the "best-matched" sample in the supplied dataset, and its score. If the best-matched ID aligns with the original genotype ID, the mixup verdict "false" is reported—otherwise, the verdict is "true", suggesting that the samples are mislabelled.

The final report from MixupMapper gave very few 'true' results suggesting that only a small subset were potentially mixed up. Upon investigation these were found to be the monozygotic twins from the BSGS pilot study.

Replication of eQTLs from Westra *et al.*[29]

As a final check that the genotype and expression data have been aligned well throughout the quality control processes, we attempted to replicate the top 3,202 sentinel SNPs (SNP with the greatest evidence for association for each probe) from Westra *et al.*[29] study. This was done for the whole CAGE blood dataset, as well as for the individual cohorts to help diagnose if any individual cohorts had errors. For each of the sentinel SNP-probe combinations regression analysis was performed using the PLINK2 software, with 10 PCs of the genotype matrix fitted. Chi-squared statistics were calculated from the summary statistics provided from the study of Westra *et al.*[29] and the CAGE analysis and compared via a scatter plot (Figure S6).

For the combined individual data, the Westra *et al.*[29] sentinel SNPs replicated well with chi-squared statistics nearing those in the Westra *et al.*[29] study. Given that the Westra *et al.*[29] study contained 5,311 individuals, which is nearly two times those in CAGE, the chi-squared statistics across these probes suggest that the CAGE data have more power per individual.

The final CAGE blood dataset consists of expression and genotypes for 2,765 individuals, has 36,778 expression probes, and 7,763,174 or 5,083,862 SNPs (dependent upon info score filtering). These data form CAGE release 2.0.

Annotation of Illumina HT12 v4 array probes to the genome

Entrez gene identifiers were taken from the Bioconductor illuminaHumanv4.db_1.26.0 data base, which follows the probe remapping protocols of Barbosa-Morais *et al.*[1] and were based on gene data from NCBI from 17 March 2015. Transcription start and stop site information was retrieved for each of the Entrez gene identifiers from the Bioconductor org.Hs.eg.db data base, which was built on data from NCBI from 27 September 2015.

Genomic location mappings were based on data provided from UCSC Genome Bioinformatics (Homo sapiens) on hg19 coordinates. Mappings based on the illuminaHumanv4.db database were only accepted if the chromosome of the probe was on the same chromosome as that of TSS/TES information provided from the org.Hs.eg.db data base. Each probe maps to multiple transcripts and thus the median of the transcription start and stop site was used as a summary measure. Of the the 36,778 probes present in the CAGE data set 31,690 had Entrez gene identifiers, which corresponded to 19,505 genes. These mappings are available to download from <http://cnsgenomics.com/shiny/CAGE/>.

For those CAGE probes that had a COJO eQTL, probe quality was determined as per the re-annotated results in the Bioconductor illuminaHumanv4.db database, which follows the protocols of Barbosa-Morais *et al.*[1]. Under the protocols of Barbosa-Morais *et al.*[1], a probe is considered specific if all its transcriptomic matches align to a single genomic location, regardless of the number of isoforms for the targeted genes and differences between gene model sources. These probes are given a quality score of "good" to "perfect" (please see Barbosa-Morais *et al.*[1] for stricter definitions). Probes are deemed "bad" if the probe matches repeat sequences, intergenic or intronic regions, or if probes target multiple (≥ 3) transcripts from different locations in the genome. The "no match" score is given to a probe if it does not significantly match any transcript or genomic location[1]. We tested for genomic location "match back"[1] for these probes, and identified 40 that did not map to a known genomic location. A further 1,822 probes had a genomic annotation score of "bad" and were not included in the presentation of eQTL results. Probes with "good" or "perfect" quality score were deemed reliable. All "bad" and "no match" probes are still reported in the nominal association database and COJO eQTL results but do not contain information on probe genomic location or transcript start and stop sites.

Supplemental Figures and Tables

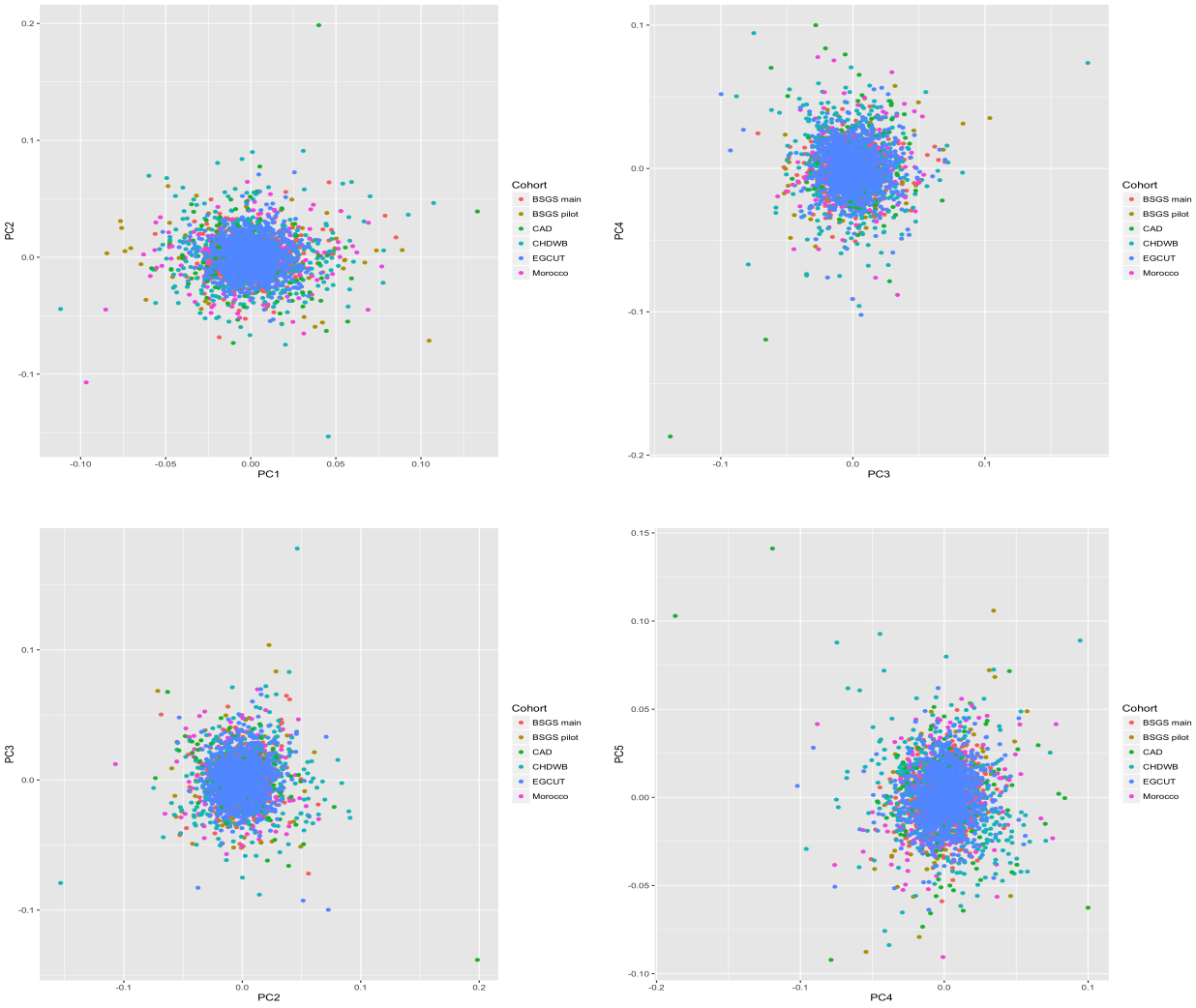

Figure S1 Principal component plots of normalised CAGE expression dataset. Plots depict the first four principal components from a PCA analysis on the whole CAGE expression data set (38,624 probes) after the completion of the normalisation pipeline. Colours indicate the individuals from each cohort and are classified in the legend.

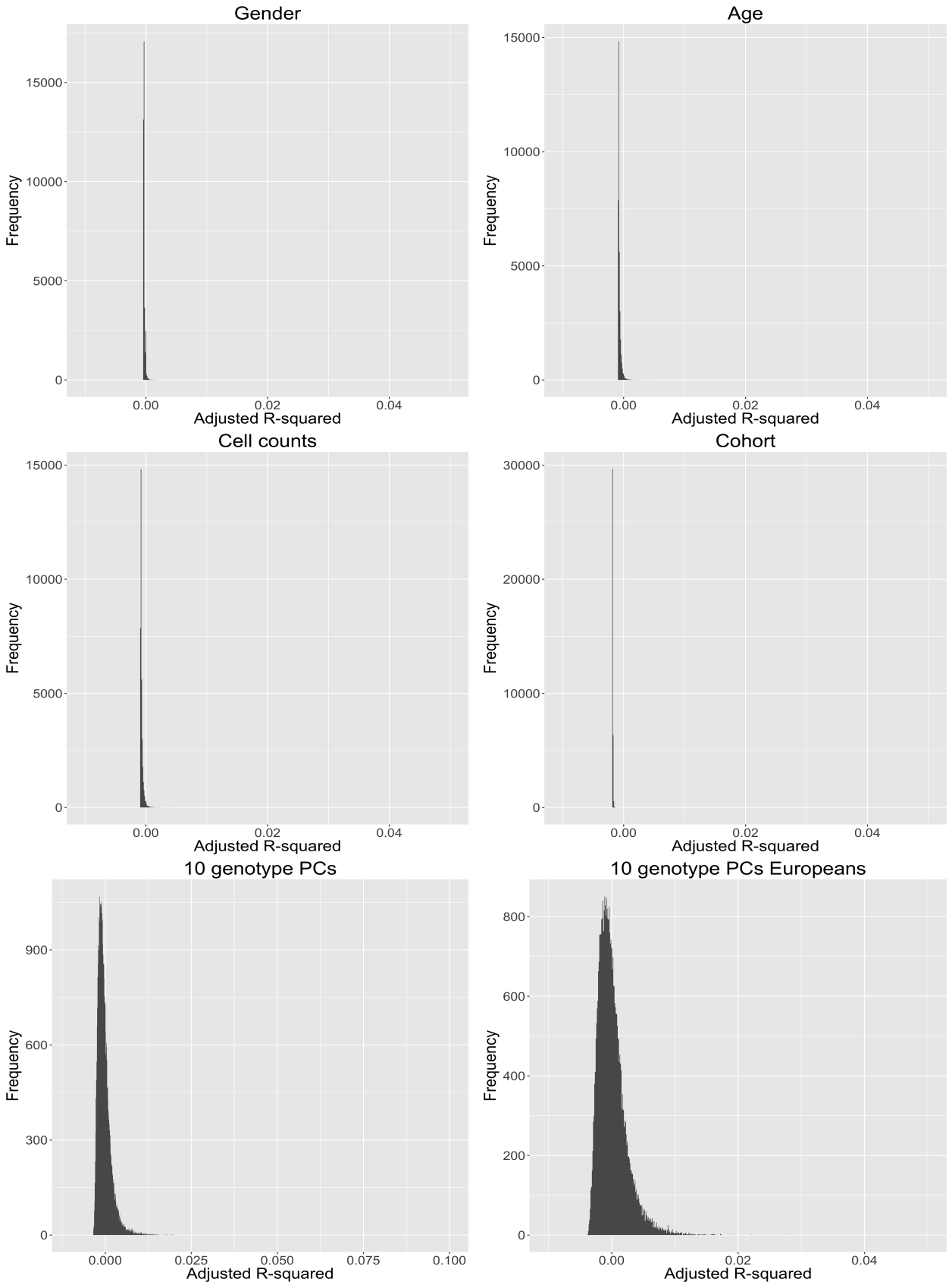

Figure S2 Covariates explaining variation in gene expression. Histograms of adjusted R-squared values from regression of normalised expression measurements of 36,778 probes on covariates gender, age, cell counts, cohort, genotype PCs from all $n = 2,765$ individuals, and genotype PCs from European individuals $n = 2,454$.

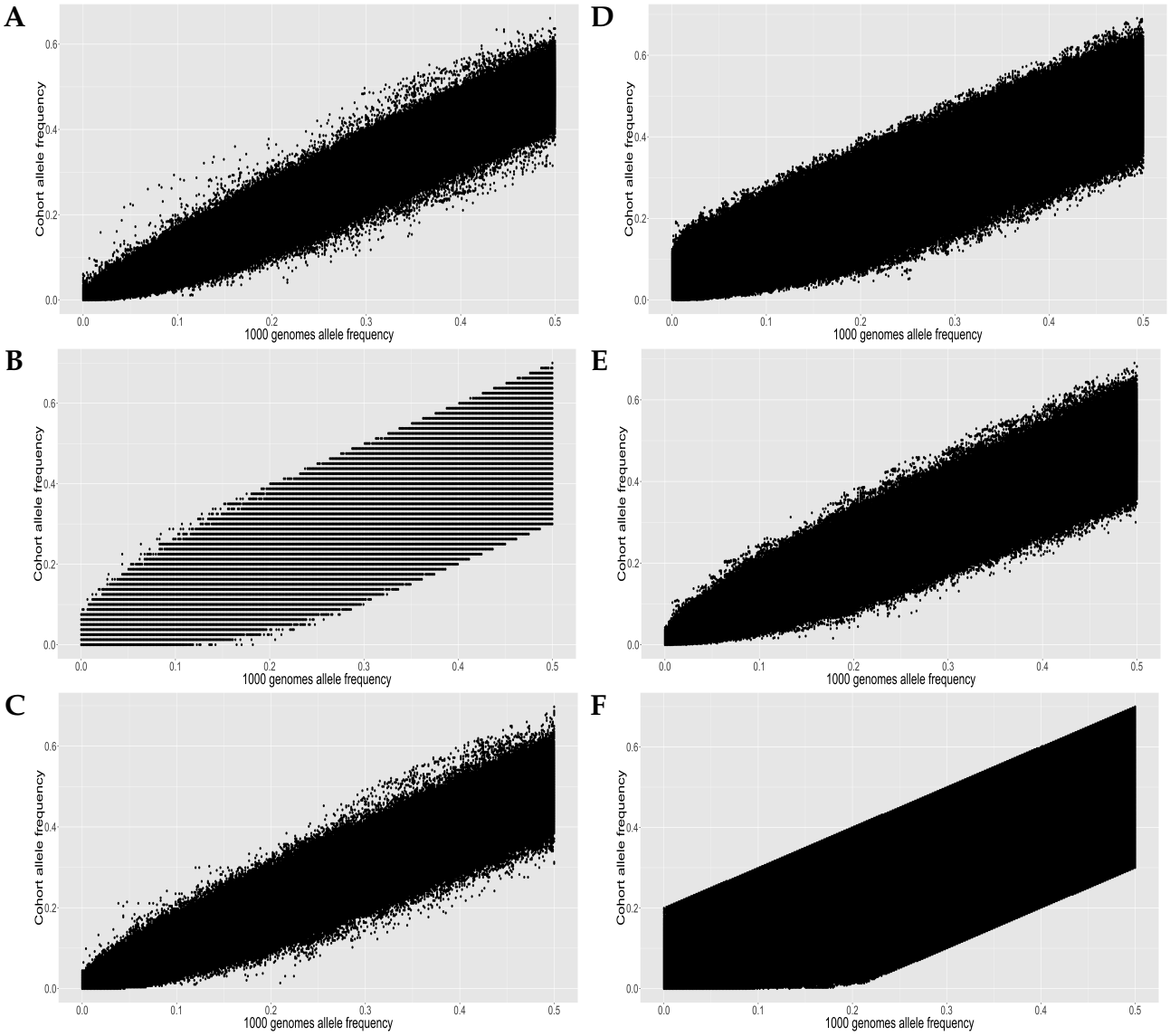

Figure S3 Cohort allele frequency quality control post impuatation. Allele frequency plots of individual cohorts (y-axes) versus the 1000 Genomes Phase 1 Version 3 reference (allele frequencies calculated from European individuals) post removal of SNPs with a frequency difference greater than 0.2 (approximately 7.8 million SNPs plotted). Panel (A) depicts the BSGS main cohort, (B) BSGS pilot, (C) CAD, (D) CHDWB, (E) EGCUT, and panel (F) depicts the Moroccan cohort.

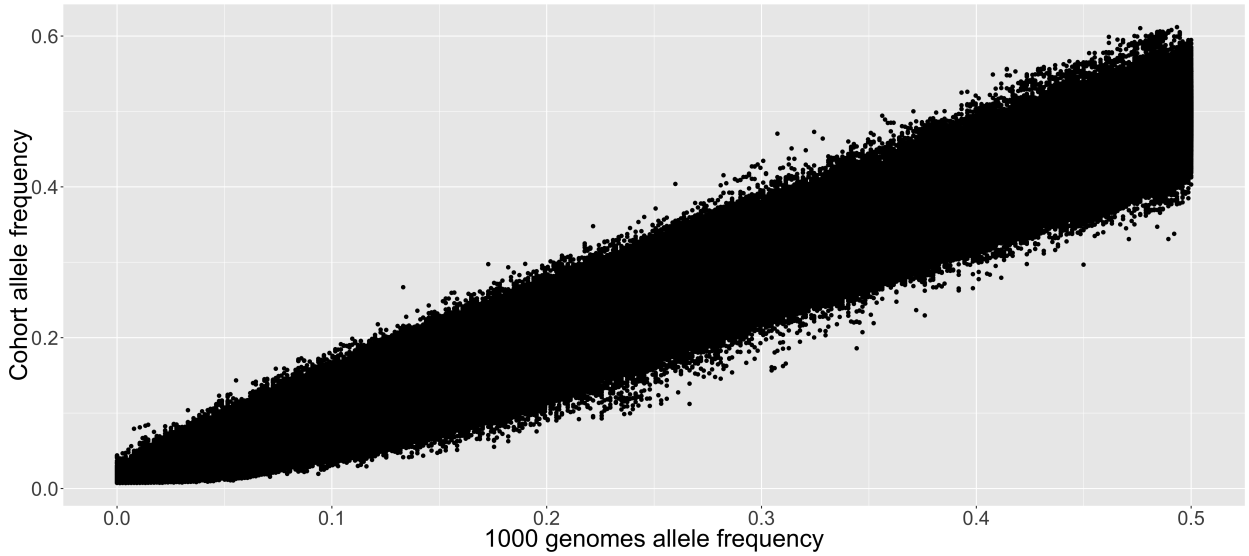

Figure S4 Allele frequency quality control post imputation for the whole CAGE data set. Allele frequency plot of whole CAGE data (y-axis) versus the 1000 Genomes Phase 1 Version 3 reference (allele frequencies calculated from European individuals) post removal of SNPs with a frequency difference greater than 0.2 (approximately 7.8 million SNPs plotted).

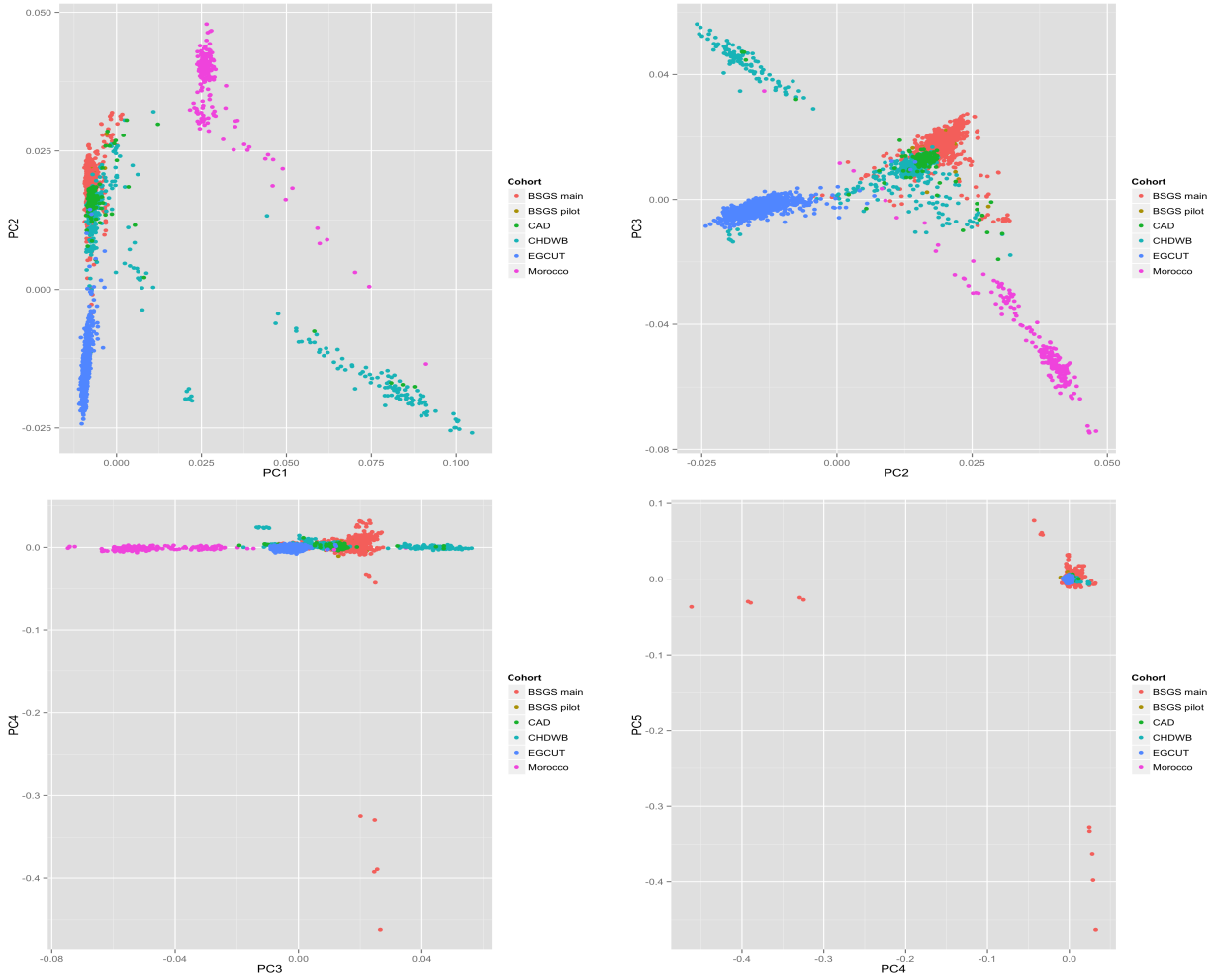

Figure S5 Principal component plots of genotype dataset. Plots depict the first four principal components from a PCA analysis on the whole CAGE genotype data set (7.8 million SNPs) after the completion of the imputation pipeline and merge. Colours indicate the individuals from each cohort and are classified in the legend.

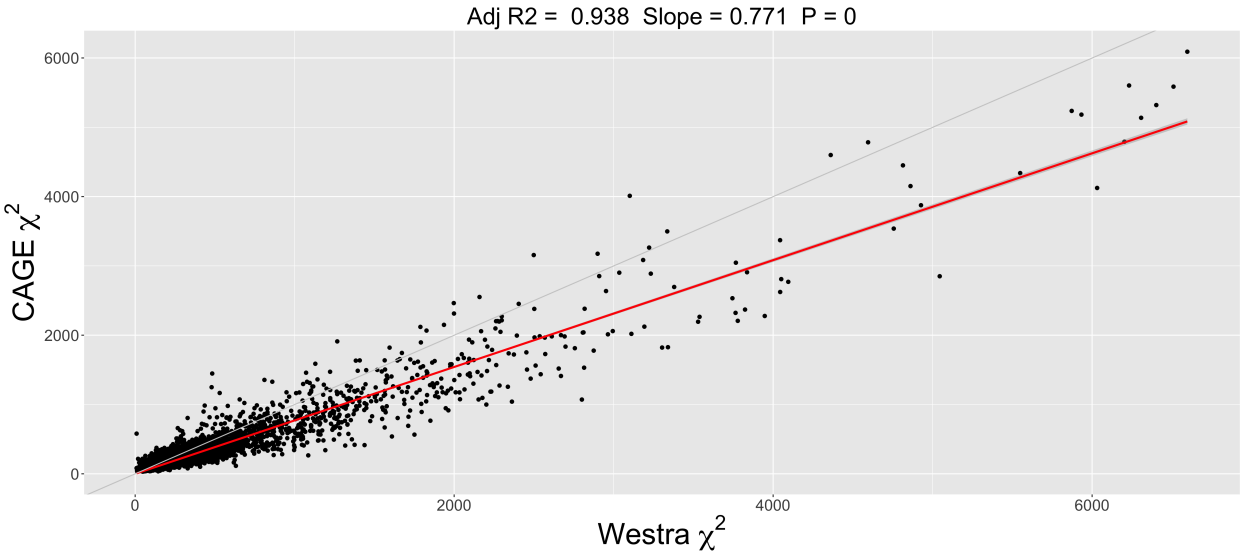

Figure S6 Meta-analysis chi-squared statistics comparison. Scatterplot of chi-squared statistics for 3,202 sentinel SNPs (*cis*) from the Westra *et al.*[29] study versus chi-squared statistics from CAGE data (all individuals $n = 2,765$) generated using a linear model in PLINK with 10 PCs fitted as additional fixed effects. The fitted regression line (red) is plotted with the key statistics of this regression (no intercept term fitted) is displayed at the top of panels. The light grey line represents the $y = x$ line. The p-value is with regard to the regression slope.

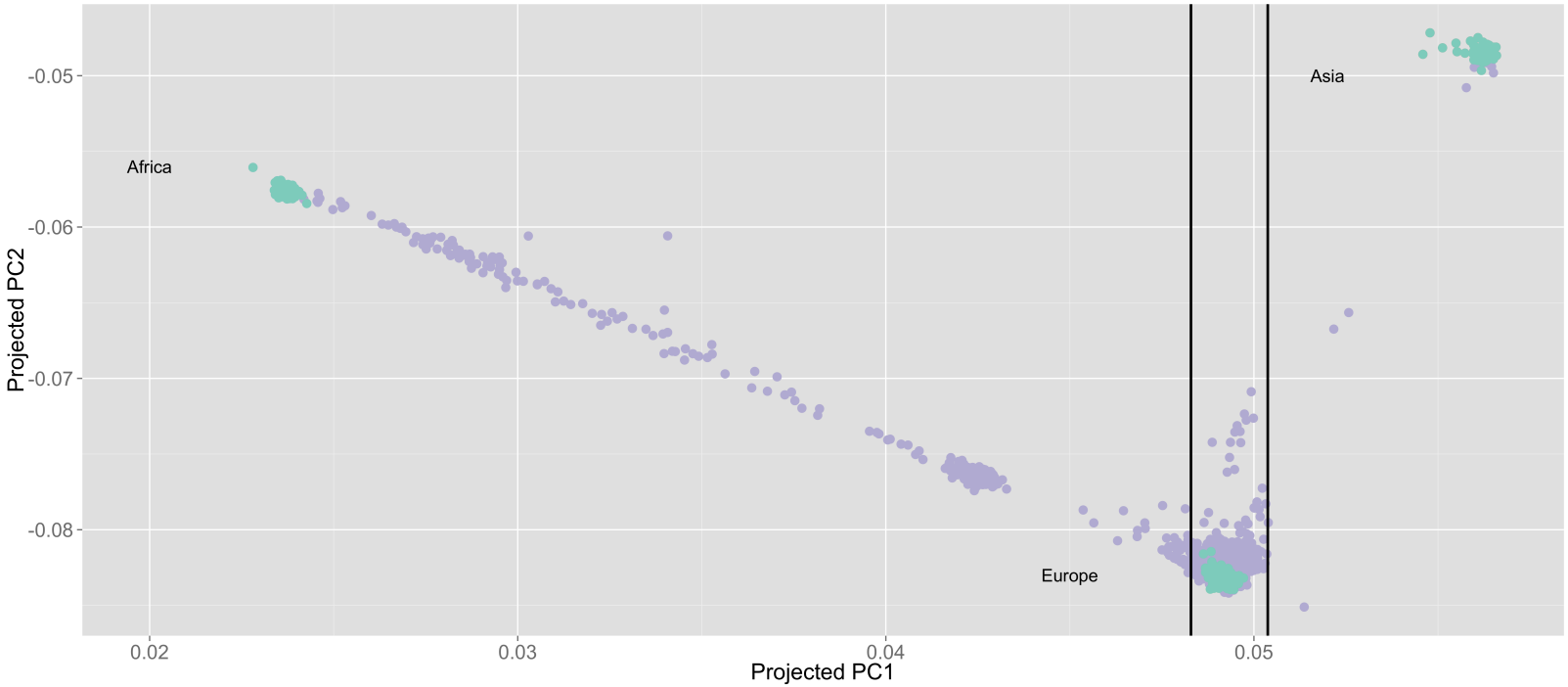

Figure S7 Ancestry investigation. Projected principal component (PPC) plot (PPC1 versus PPC2) of Hap Map 3 cohorts (green) and CAGE data ($n = 2,765$) (purple). The Utah residents of northern and western European ancestry (CEU) cohort from Hap Map 3 formed the European sample, the Yoruba trios from Ibadan, Nigeria (YRI) formed the African cohort, and the Han Chinese individuals from Beijing, China were used for the Asian cohort. Solid vertical lines indicate the bounds for removing European ancestry outliers. The bounds were [lower quartile - 1.5× IQR, upper quartile + 1.5× IQR] of the first projected PC (where IQR is the inter-quartile range).

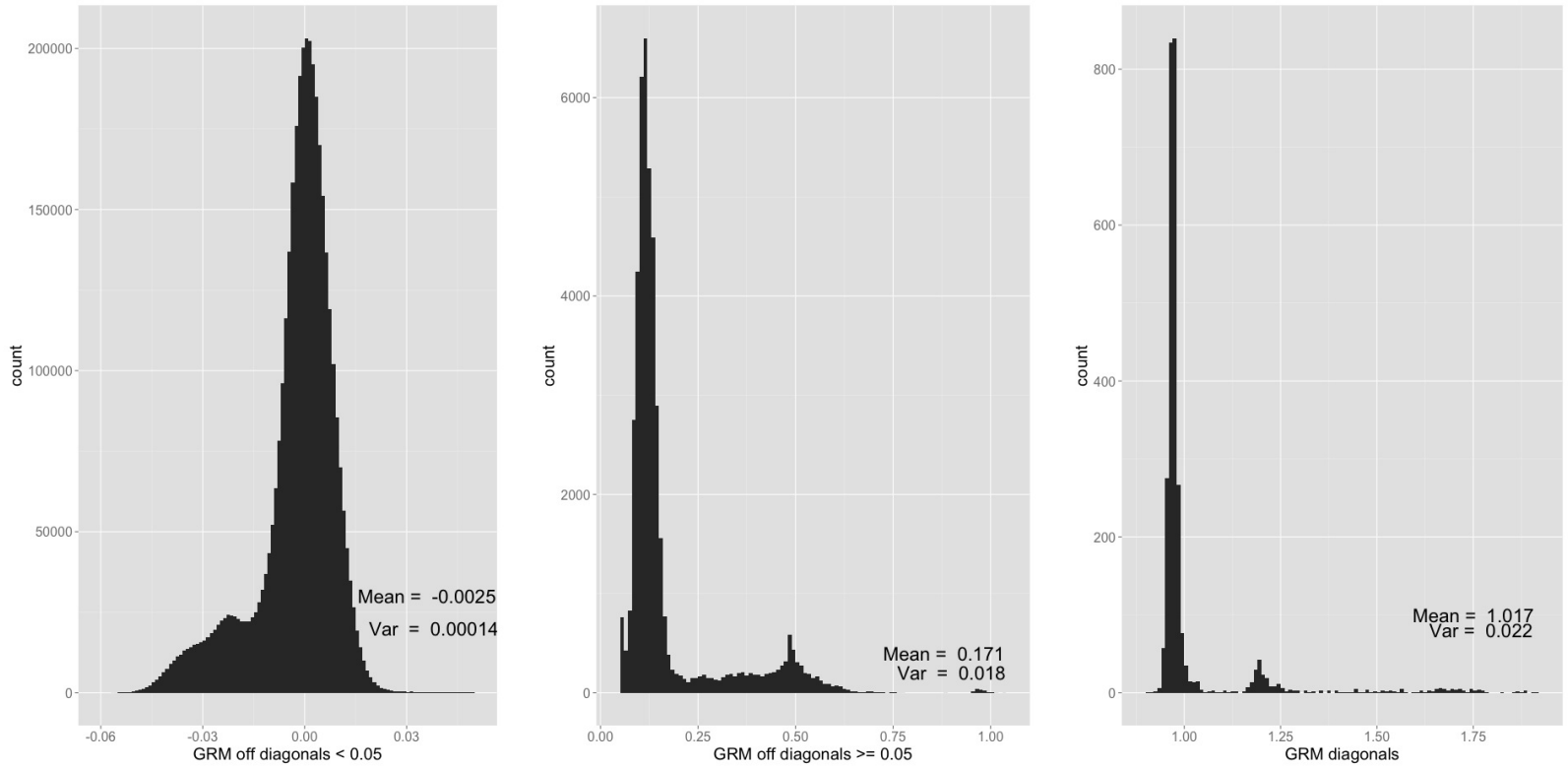

Figure S8 Genetic relationship matrix for all of CAGE. Summary of elements of the genetic relationship matrix (GRM) built using overlapping Hap Map 3 SNPs (893,626) and all individuals ($n = 2,765$) in CAGE. Means and variances are summarised for the histogram displayed. The GRM off diagonals are partitioned into those elements greater and less than 0.05 for ease of interpretation.

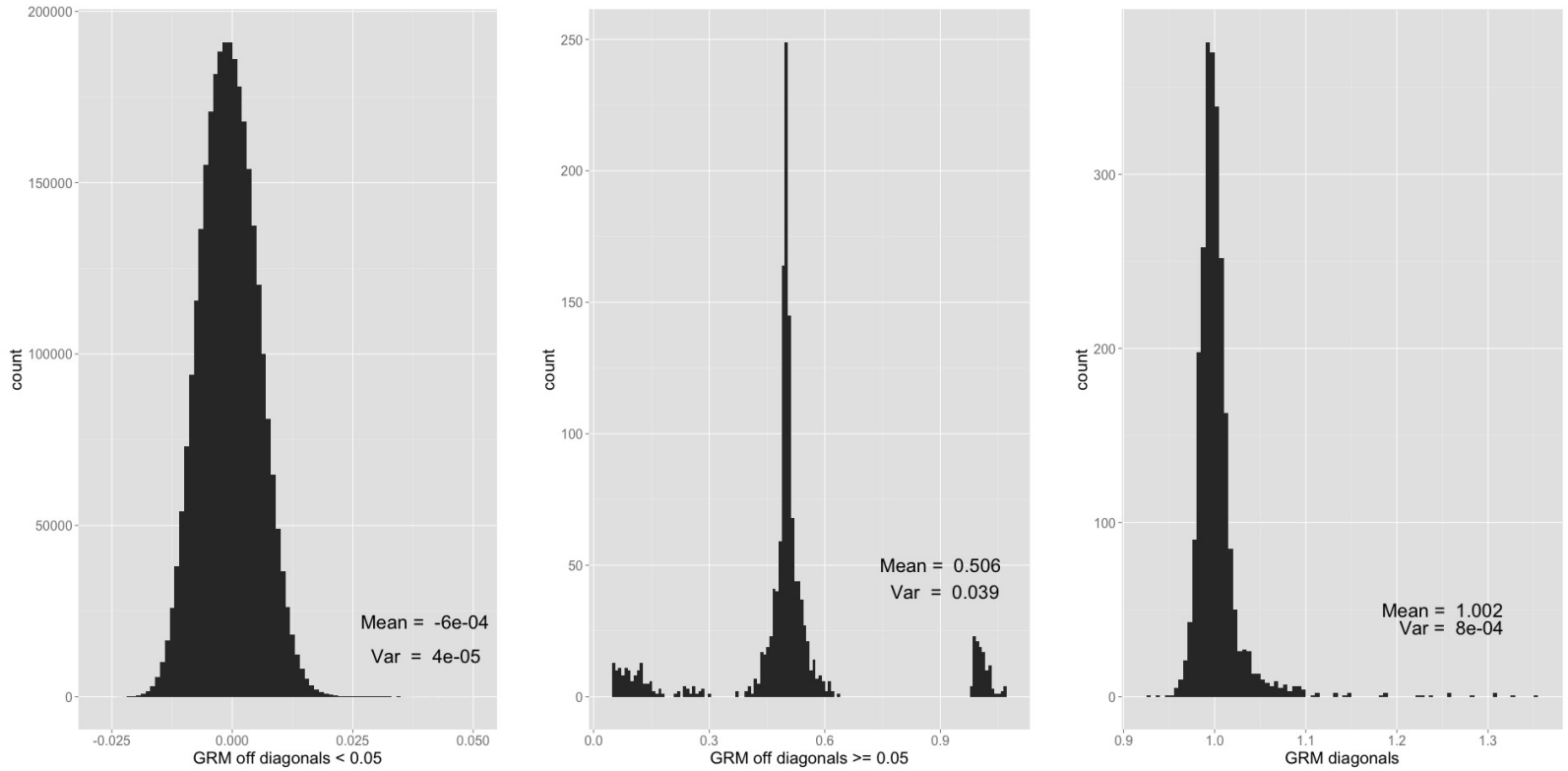

Figure S9 Genetic relationship matrix for European individuals. Summary of elements of the genetic relationship matrix (GRM) built using overlapping Hap Map 3 SNPs (893,626) and European individuals (n = 2,454). Means and variances are summarised for the histogram displayed. The GRM off diagonals are partitioned into those elements greater (or equal to) and less than 0.05 for ease of interpretation.

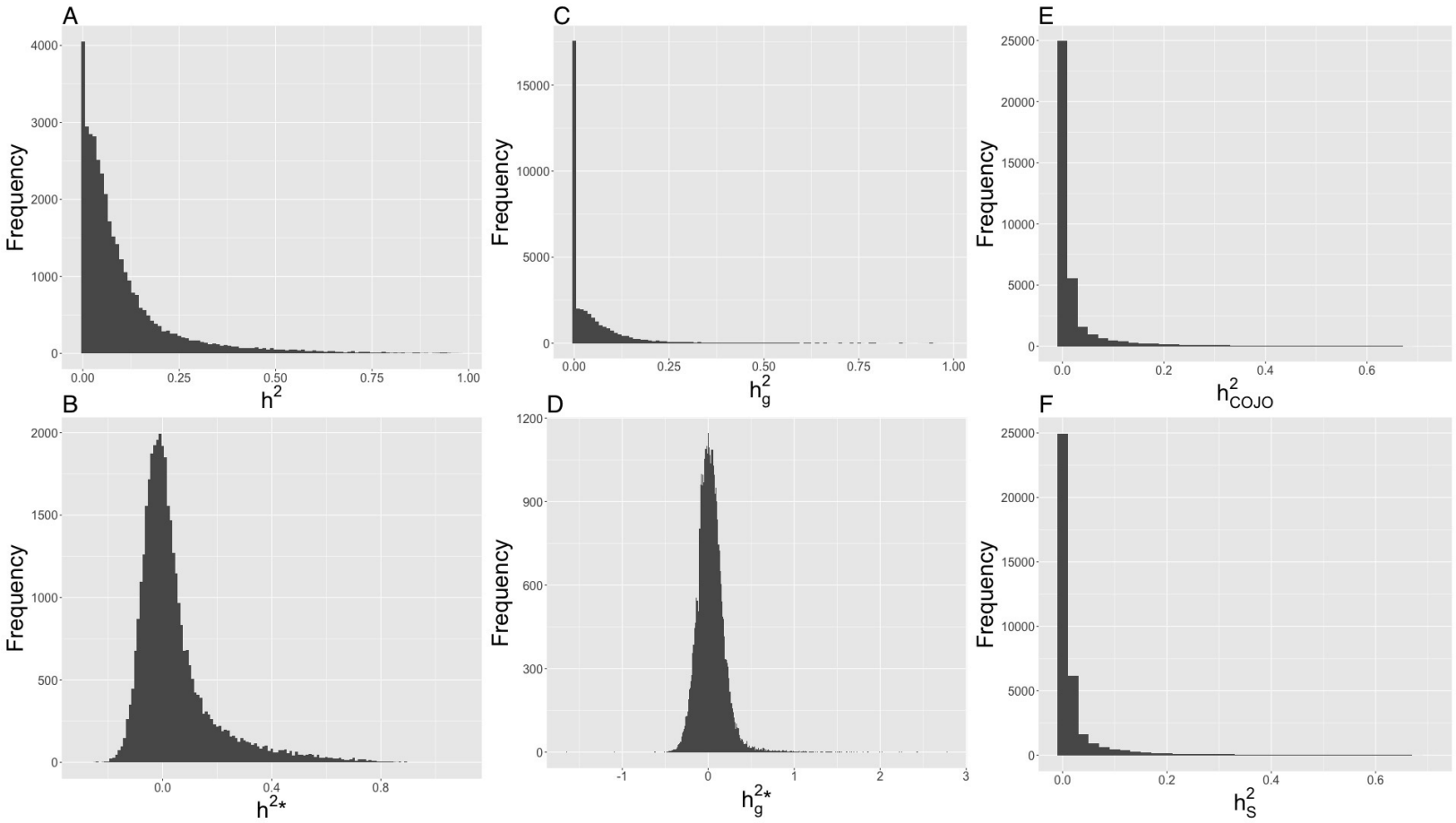

Figure S10 Distributions of heritability estimates for all probes. Histogram of heritability estimates across 36,778 probes generated using the Big K/Small K method, and estimates of h^2_{COJO} and h^2_S. Panels (A) and (B) display histogram summaries of the narrow-sense heritability estimates using the constrained and unconstrained REML algorithms respectively. Panels (C) and (D) display histogram summaries of the heritability estimates of the proportion of phenotypic variance explained by genome-wide Hap Map 3 SNPs using the constrained and unconstrained REML algorithms respectively. Panels (E) and (F) display histogram summaries of the estimates of the proportion of phenotypic variance explained by COJO eQTL (h^2_{COJO}) and the sentinel SNP (h^2_S) respectively.

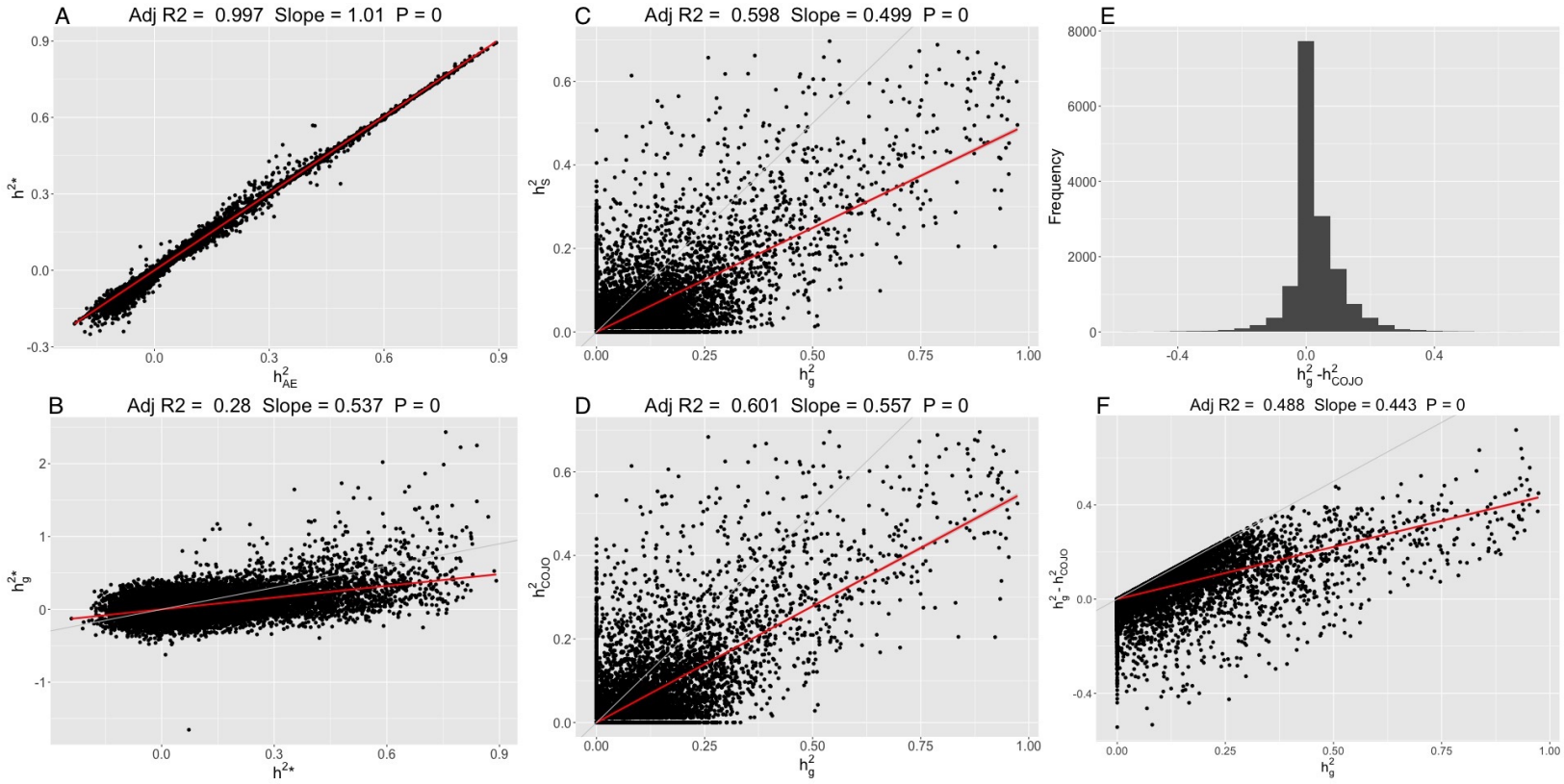

Figure S11 Comparison of heritability estimates for expressed probes. Summary of heritability estimates (unconstrained) using only the $\mathbf{K}_{\text{IBS}>t}$ matrix of estimated relatedness (h^2_{AE}), h^2_g (constrained) and h^{2*}_g (unconstrained) of Big K/Small K method, proportion of phenotypic variance explained by COJO SNPs (h^2_{COJO}), and the proportion of phenotypic variance explained by the sentinel SNP (h^2_S). Displayed summaries are across 15,966 overlapping probes from the study of Kirsten *et al.* [16], except for panel (A), which displays estimates for all 36,778 probes. Panel (A) displays the scatter plot of the AE model estimates of narrow-sense heritability versus Big K/Small K heritability estimates using the unconstrained REML algorithm. Panel (B) is a scatter plot of Big K/Small K heritability estimates of h^{2*}_g versus Big K/Small K heritability estimates of h^{2*}. Panel (C) is a scatter plot of h^2_g estimates versus the proportion of phenotypic variance explained by the sentinel SNP. Panel (D) is a scatter plot of h^2_g estimates versus the proportion of phenotypic variance explained by COJO eQTL. Panel (E) displays a histogram plot of the difference between h^2_g estimates and the proportion of phenotypic variance explained by COJO eQTL. Panel (F) displays a scatterplot of h^2_g estimates versus the difference between h^2_g and the proportion of phenotypic variance explained by the COJO SNPs. For panels (A), (B), (C) (D) and (F), the fitted regression line (red) and 95% confidence interval (shaded) is plotted with the key statistics of this regression displayed at the top of panels. The p-value is with respect to the regression slope.

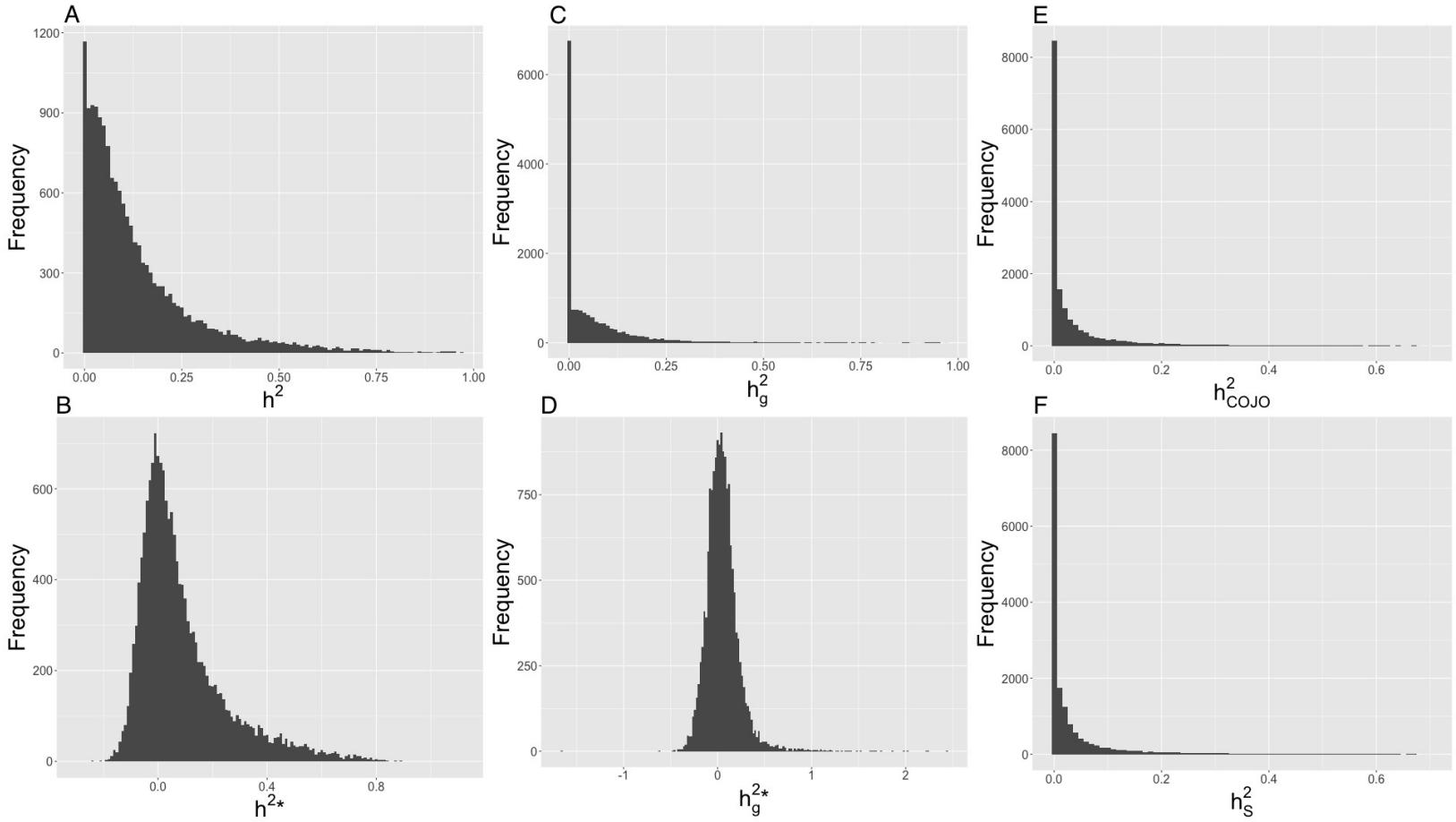

Figure S12 Distributions of heritability estimates for expressed probes. Histogram of heritability estimates across 15,966 expressed probes generated using the Big K/Small K method, and estimates of h^2_{COJO} and h^2_S. Panels (A) and (B) display histogram summaries of the narrow-sense heritability estimates using the constrained and unconstrained REML algorithms respectively. Panels (C) and (D) display histogram summaries of the heritability estimates of the proportion of phenotypic variance explained by genome-wide Hap Map 3 SNPs using the constrained and unconstrained REML algorithms respectively. Panels (E) and F) display histogram summaries of the estimates of the proportion of phenotypic variance explained by COJO eQTL (h^2_{COJO}) and the sentinel SNP (h^2_S) respectively.

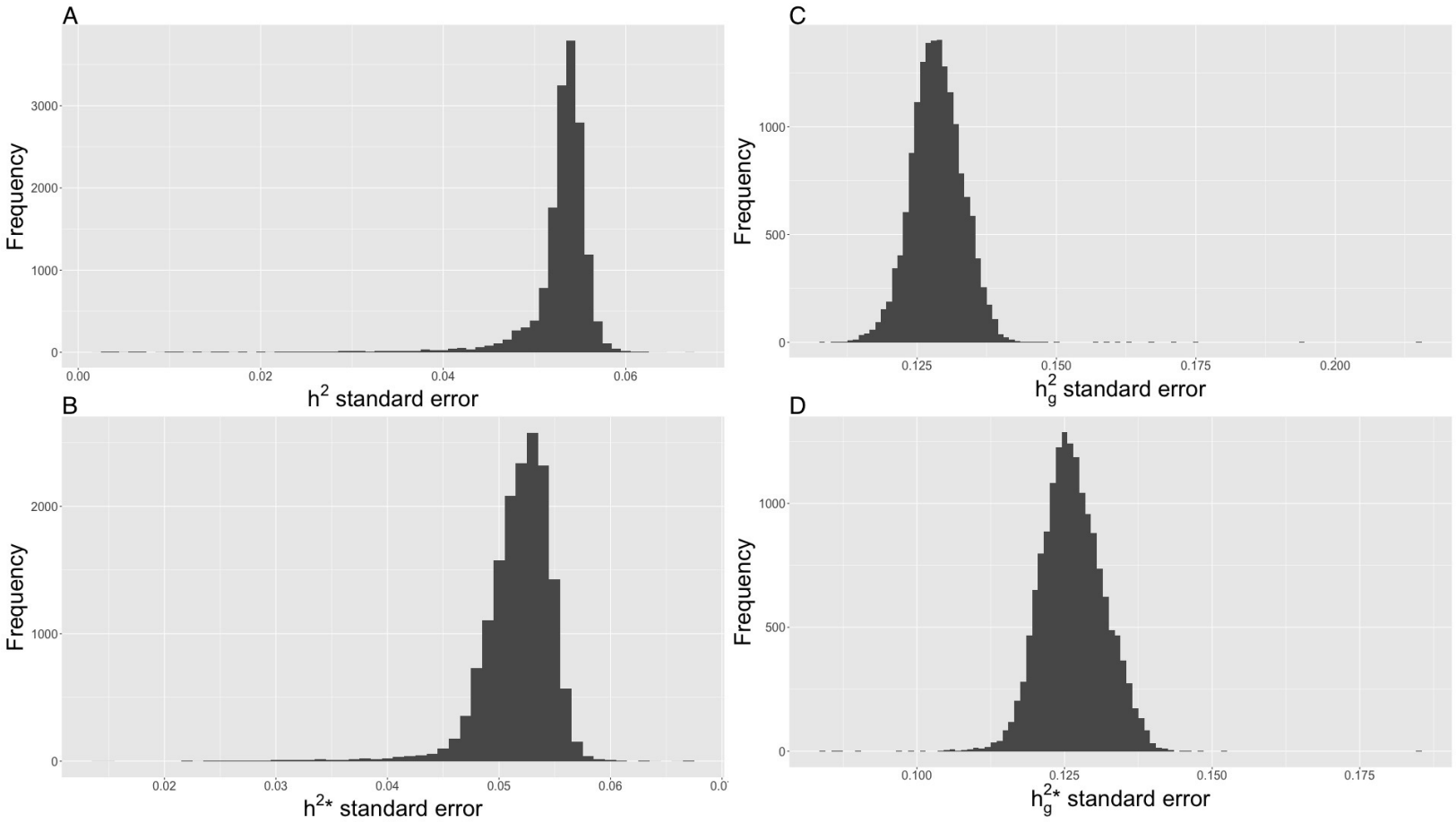

Figure S13 Distributions of standard errors of heritability estimates for expressed probes. Histogram of standard errors of heritability estimates across 15,966 probes generated using the Big K/Small K method. Panels (A) and (B) display histogram summaries of the standard errors for narrow-sense heritability estimates using the constrained and unconstrained REML algorithms respectively. Panels (C) and (D) display histogram summaries of the standard errors for heritability estimates of the proportion of phenotypic variance explained by genome-wide Hap Map 3 SNPs using the constrained and unconstrained REML algorithms respectively.

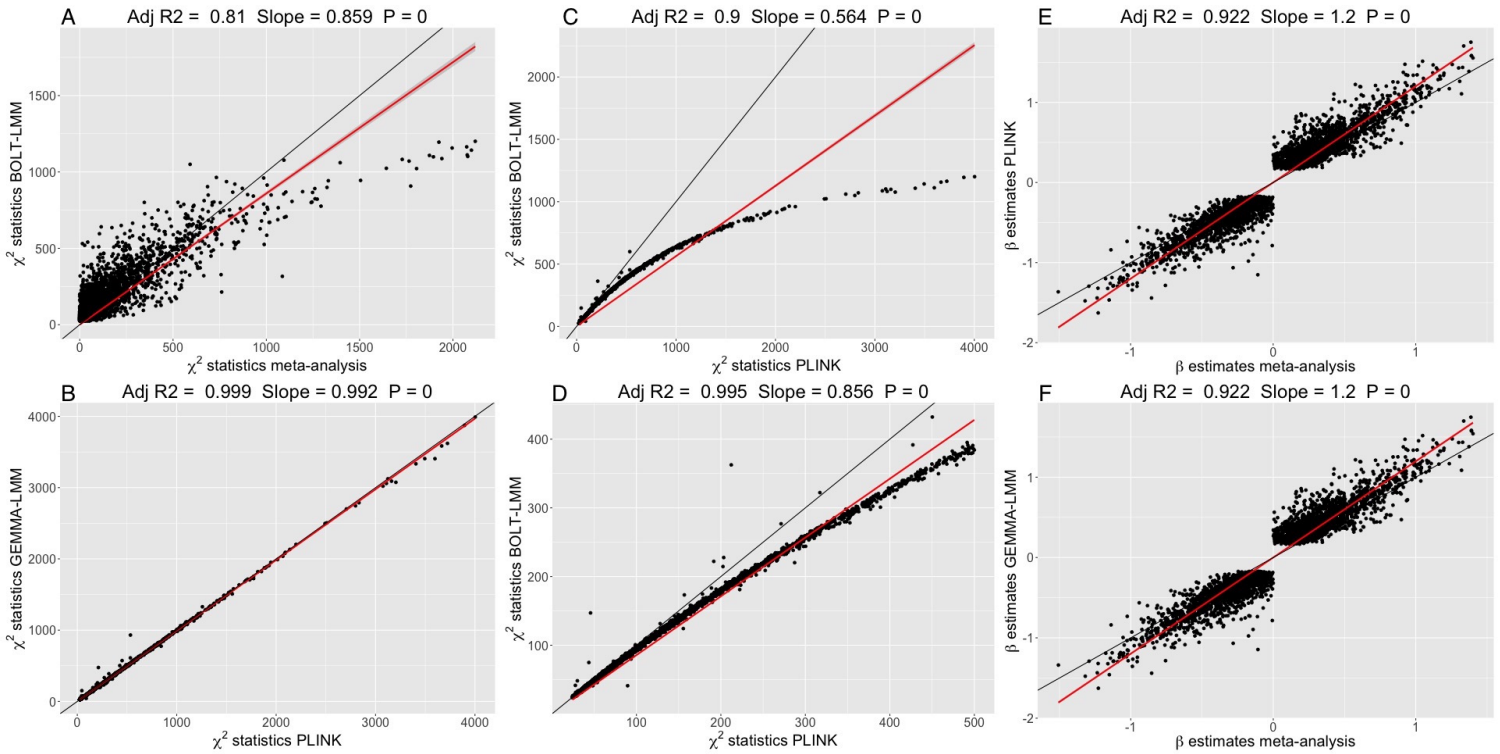

Figure S14 Comparison of mega- versus meta-analysis chi-squared statistics and effect sizes. Comparison of χ^2 statistics for the sentinel SNPs of 3,450 *cis* probes generated from the meta-analysis of Westra *et al.*[29] (n = 1,749) and an eQTL analysis using European unrelated individuals (n = 1,748) from CAGE. Panels (A), (B), and (C) compare the same set of sentinel SNP χ^2 statistics generated using a single SNP analysis in PLINK corrected for 10 PCs (PLINK), eQTL analysis in BOLT-LMM (HapMap 3 SNPs used as model SNPs), eQTL results from GEMMA (GRM generated from Hap Map 3 SNPs), and the meta-analysis χ^2 statistics. Panel (D) displays a zoomed view of panel (C) to investigate the point at which the χ^2 statistics from the PLINK analysis deviated from those from the BOLT-LMM analysis. Panels (E) and (F) show the approximate effects sizes from the meta-analysis versus those generated using PLINK and GEMMA-LMM. All panels include the fitted regression line (red) and 95% confidence interval (shaded) is plotted and $y = x$ line (black) for reference with the key statistics of this regression (no intercept term fitted) displayed at the top of each panel. The p-value is with respect to the regression slope.

Cohort	Probes	Individuals	Array
BLOOD			
BSGS main	47323	846	Illumina HumanHT-12 v4.0
BSGS pilot	48760	80	Illumina HumanHT-12 v3.0
CAD (batch 1)	47231	147	Illumina HumanHT-12
CAD (batch 2)	46331	163	Illumina HumanHT-12
CHDWB (batch 1)	46328	176	Illumina HumanHT-12
CHDWB (batch 2)	46328	141	Illumina HumanHT-12
CHDWB (batch 3)	46328	132	Illumina HumanHT-12
EGCUT	48803	1065	Illumina HumanHT-12 v3.0
Morocco	48803	188	Illumina HumanHT-12
LYMPHOBLASTOID CELL LINES			
BSGS pilot (LCL)	48760	95	Illumina HumanHT-12 v3.0
MuTHER (LCL)	48638	825	Illumina HumanHT-12 v3.0
FAT			
MuTHER	48638	826	Illumina HumanHT-12 v3.0
SKIN			
MuTHER	48646	705	Illumina HumanHT-12 v3.0
Total		5302	

Table S1 CAGE cohort sizes and expression arrays. Summary of gene expression data sets in phase 1 of CAGE. Array versions were not available for all cohorts; array information was gathered from the relevant citations.

Dataset	Individuals
BSGS main	846
BSGS pilot	80
CAD (batch 1)	147
CHDWB (batch 1)	176
CHDWB (batch 2)	141
CHDWB (batch 3)	132
EGCUT-CNV	982
EGCUT-Omni	162
Morocco	188
Total	2,854
Duplicates	89
Total post-merge	2,765

Table S2 Contributing individuals to CAGE peripheral blood data set. Summary of CAGE cohort data dimensions post imputation and merge

Individuals			
	Number	Analyses	Description
	2,765	BOLT-LMM	Total number of individuals in CAGE with expression and genotypes across contributing cohort data sets
	2,454	Big K/Small K	Set of individuals with European ancestry, which includes both related and unrelated individuals. Non-Europeans were excluded via an outlier analysis of projected PC 1.
	1,748	Westra *et al.*[29] comparison	Set of unrelated European individuals. Unrelated status was determined via a relatedness threshold of 0.05 on the genetic relationship matrix off diagonals
Probes			
	Number	Analyses	Description
	36,778	BOLT-LMM/GREML	Total number of overlapping probes passing quality control across contributing cohort data sets used for eQTL analysis
	11,829	COJO	Number of probes with a SNP-probe association (BOLT-LMM) p-value $< 5 \times 10^{-8}$ carried forward for COJO analysis
	15,966	h^2 comparison	Number of overlapping expressed probes from the set of 18,738 probes from the study of Kirsten et al. 2015 that mapped uniquely to the genome and had a probe annotation quality score of at least 'good' as per the protocol of Barbosa-Morais *et al.*[1] 2010
	3,450	Mega vs Meta	Subset of overlapping probes with *cis*-eQTLs from Westra *et al.*[29] with z-values contributing from both the DILGOM cohort[14] ($n = 509$) and Fehrmann cohorts[7] ($n = 1,240$)

Table S3 Summary of data subsets and thresholds used in CAGE analysis. Summary of the number of individuals and probes used for different analyses. Descriptions outline the reasons or thresholds used to come to this number of individuals or probes.

Multiple	1	2	3	4	5	6	7	8	9	≥ 10
No. probes (all)	6,617	2,231	754	242	78	27	12	5	0	1
No. *cis* probes	5,551	1,588	503	148	42	16	4	4	0	1
No. *trans* probes	2,978	289	52	17	1	1	0	0	0	0

Table S4 Multiple eQTL. Summary of the number (No.) of probes with a particular multiple of COJO eQTLs for 9,967 probes (excluding probes with a genomic annotation quality score of less than 'good'). *Cis* and *trans*-eQTL probes were separated if the SNP and gene were located on different chromosomes. Column sums of *cis* and *trans* do not sum to equal the 'all' row value because, for example, if a probe has 3 *cis*-eQTL and 1 *trans*-eQTL then the count would be incremented in the three column for *cis*, the one column for *trans*, and the four column for 'all'.

Literature Cited

[1] Barbosa-Morais, N. L., M. J. Dunning, S. A. Samarajiwa, J. F. Darot, M. E. Ritchie, A. G. Lynch, and S. Tavaré, 2010 A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. Nucleic Acids Research **38**: e17–e17.

[2] Blom, G., 1958 *Statistical estimates and transformed beta-variables*. Wiley; New York.

[3] Bolstad, B. M., R. A. Irizarry, M. Åstrand, and T. P. Speed, 2003 A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19**: 185–193.

[4] Consortium, . G. P. *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature **491**: 56–65.

[5] Deelen, P., M. J. Bonder, K. J. van der Velde, H.-J. Westra, E. Winder, D. Hendriksen, L. Franke, and M. A. Swertz, 2014 Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC research notes **7**: 901.

[6] Dunning, M., A. Lynch, and M. Eldridge, 2015 *illuminaHumanv4.db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4)*. R package version 1.26.0.

[7] Fehrmann, R. S., R. C. Jansen, J. H. Veldink, H.-J. Westra, D. Arends, M. J. Bonder, J. Fu, P. Deelen, H. J. Groen, A. Smolonska, *et al.*, 2011 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 **7**: e1002197.

[8] Goldinger, A., A. K. Henders, A. F. McRae, N. G. Martin, G. Gibson, G. W. Montgomery, P. M. Visscher, and J. E. Powell, 2013 Genetic and nongenetic variation revealed for the principal components of human gene expression. Genetics **195**: 1117–1128.

[9] Grundberg, E., K. S. Small, Å. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T.-P. Yang, E. Meduri, A. Barrett, *et al.*, 2012 Mapping cis-and trans-regulatory effects across multiple tissues in twins. Nature Genetics **44**: 1084–1089.

[10] Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis, 2012 Fast

and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics **44**: 955–959.

[11] Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Ole's, A. K., Pag'es, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., Morgan, and M., 2015 Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods **12**: 115–121.

[12] Huber, W., A. Von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron, 2002 Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics **18**: S96–S104.

[13] Idaghdour, Y., W. Czika, K. V. Shianna, S. H. Lee, P. M. Visscher, H. C. Martin, K. Miclaus, S. J. Jadallah, D. B. Goldstein, R. D. Wolfinger, *et al.*, 2010 Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nature Genetics **42**: 62–67.

[14] Inouye, M., K. Silander, E. Hamalainen, V. Salomaa, K. Harald, P. Jousilahti, S. Männistö, J. G. Eriksson, J. Saarela, S. Ripatti, *et al.*, 2010 An immune response network associated with blood lipid levels. PLoS Genetics **6**: e1001113.

[15] Kim, J., N. Ghasemzadeh, D. J. Eapen, N. C. Chung, J. D. Storey, A. A. Quyyumi, and G. Gibson, 2014 Gene expression profiles associated with acute myocardial infarction and risk of cardiovascular death. Genome Medicine **6**: 40.

[16] Kirsten, H., H. Al-Hasani, L. Holdt, A. Gross, F. Beutner, K. Krohn, K. Horn, P. Ahnert, R. Burkhardt, K. Reiche, *et al.*, 2015 Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci. Human Molecular Genetics .

[17] Kreil, D. P. and R. R. Russell, 2005 Tutorial section: There is no silver bullet–a guide to low-level data transforms and normalisation methods for microarray data. Briefings in Bioinformatics **6**: 86–97.

[18] Leitsalu, L., T. Haller, T. Esko, M.-L. Tammesoo, H. Alavere, H. Snieder, M. Perola, P. C. Ng, R. Mägi, L. Milani, *et al.*, 2014 Cohort profile: Estonian biobank of the Estonian Genome center, University of Tartu. International Journal of Epidemiology p. dyt268.

[19] McCarthy, S., S. Das, W. Kretzschmar, et al., R. Durbin, G. Abecasis, and J. Marchini, 2016 A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics **48**: 1279–1283.

[20] Powell, J. E., A. K. Henders, A. F. McRae, A. Caracella, S. Smith, M. J. Wright, J. B. Whitfield, E. T. Dermitzakis, N. G. Martin, P. M. Visscher, *et al.*, 2012a The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. PLoS One **7**: e35430.

[21] Powell, J. E., A. K. Henders, A. F. McRae, M. J. Wright, N. G. Martin, E. T. Dermitzakis, G. W. Montgomery, and P. M. Visscher, 2012b Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. Genome Research **22**: 456–466.

[22] Preininger, M., D. Arafat, J. Kim, A. P. Nath, Y. Idaghdour, K. L. Brigham, and G. Gibson, 2013 Blood-informative transcripts define nine common axes of peripheral blood gene expression. PLoS Genetics **9**.

[23] Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics **81**: 559–575.

[24] R Core Team, 2015 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[25] Reimers, M., 2010 Making Informed Choices about Microarray Data Analysis. PLoS Comput Biol **6**: e1000786+.

[26] Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, 2015 *limma* powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research **43**: e47.

[27] Stegle, O., L. Parts, M. Piipari, J. Winn, and R. Durbin, 2012 Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature Protocols **7**: 500–507.

[28] Westra, H.-J., R. C. Jansen, R. S. Fehrmann, G. J. te Meerman, D. Van Heel, C. Wijmenga, and L. Franke, 2011 MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics **27**: 2104–2111.

[29] Westra, H.-J., M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M. W. Christiansen, B. P. Fairfax, K. Schramm, J. E. Powell, *et al.*, 2013 Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature Genetics **45**: 1238–1243.

[30] Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics **88**: 76–82.