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 ($\geq$ 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

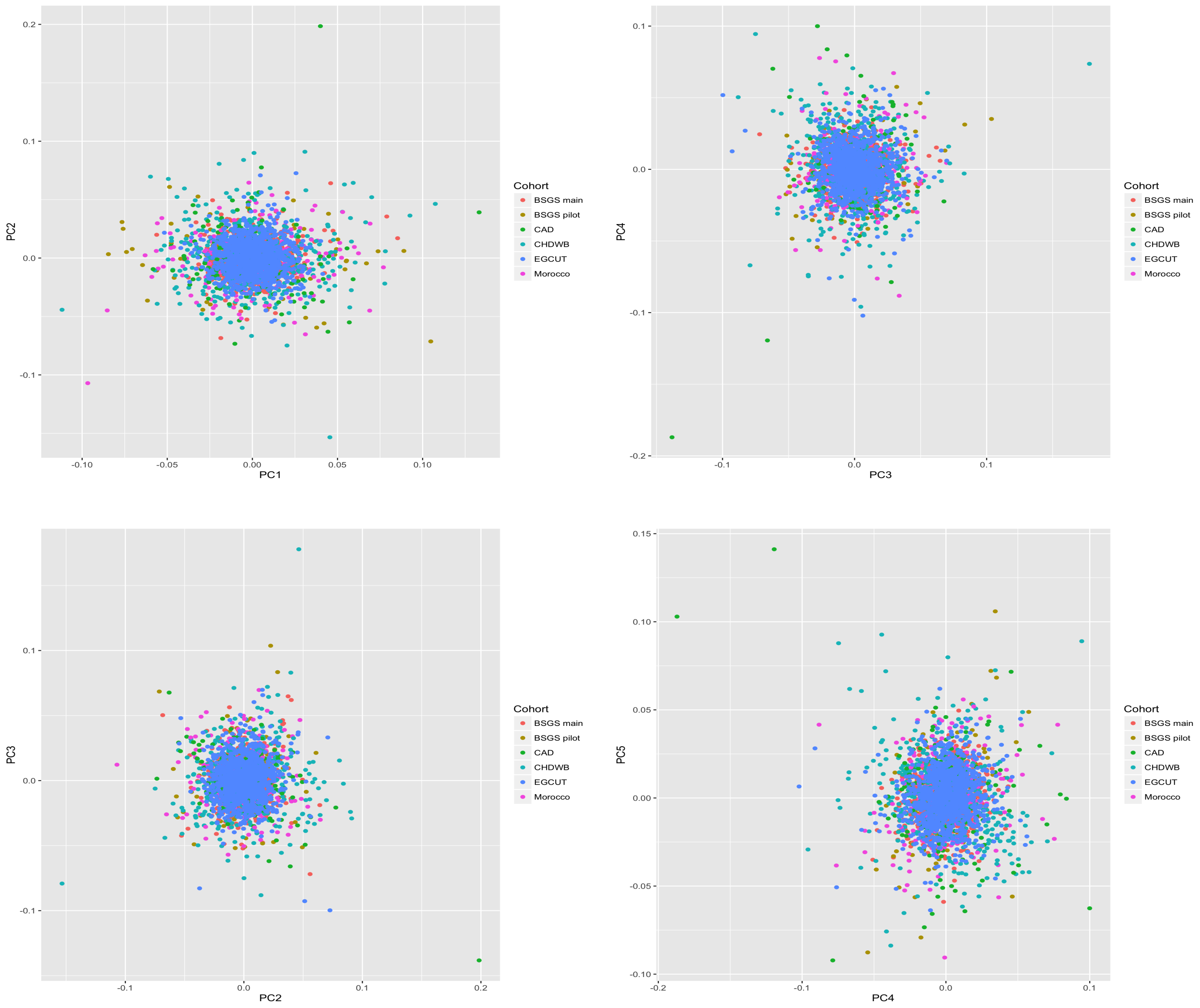

Figure S1 Principal component plots of normalised CAGE expression dataset. Plots depict the first four principal components from a PCA analysis on the whole CAGE expression data set (38,624 probes) after the completion of the normalisation pipeline. Colours indicate the individuals from each cohort and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

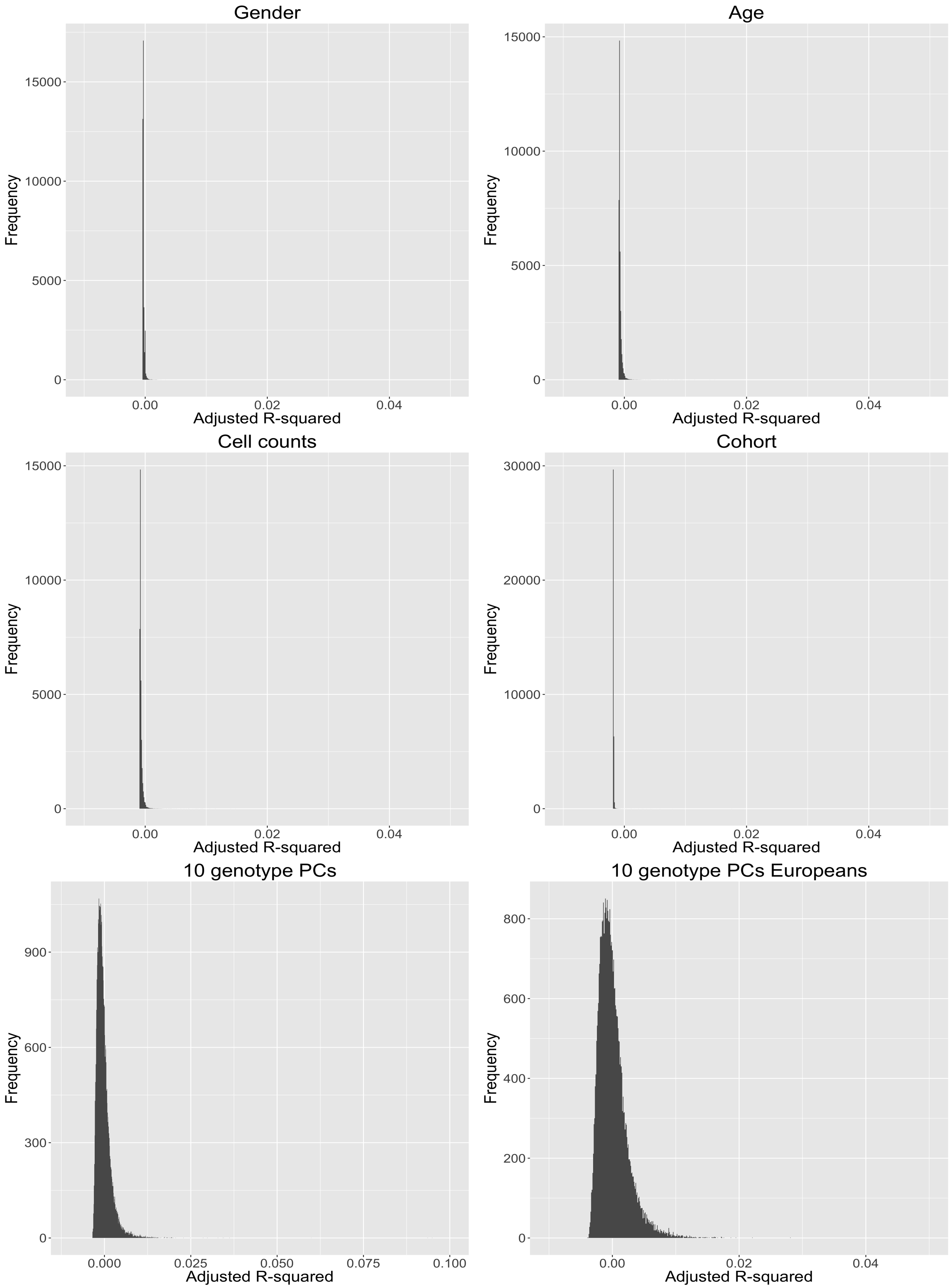

Figure S2 Covariates explaining variation in gene expression. Histograms of adjusted R-squared values from regression of normalised expression measurements of 36,778 probes on covariates gender, age, cell counts, cohort, genotype PCs from all $n = 2,765$ individuals, and genotype PCs from European individuals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

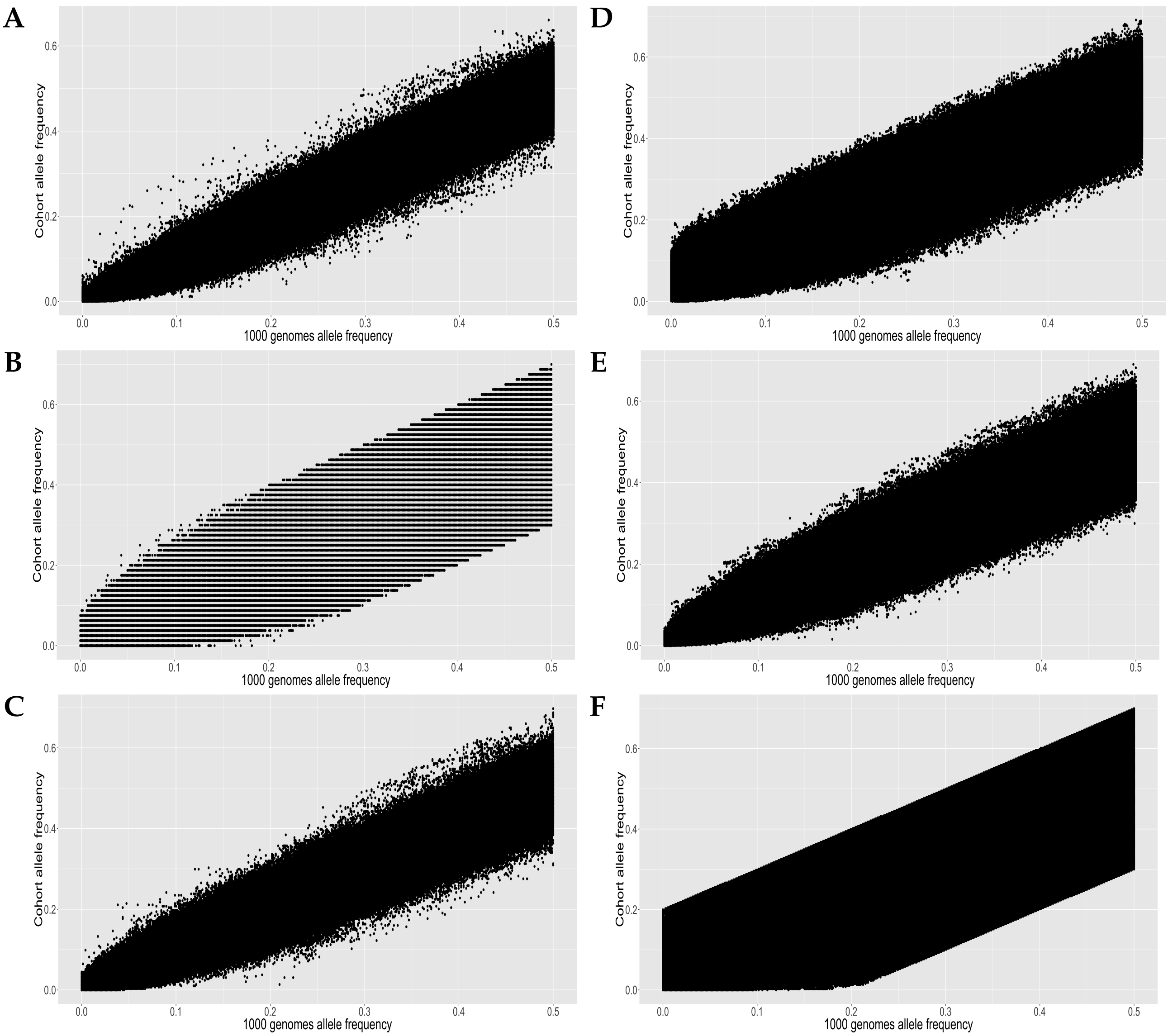

Figure S3 Cohort allele frequency quality control post impuatation. Allele frequency plots of individual cohorts (y-axes) versus the 1000 Genomes Phase 1 Version 3 reference (allele frequencies calculated from European individuals) post removal of SNPs with a frequency difference greater than 0.2 (approximately 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.

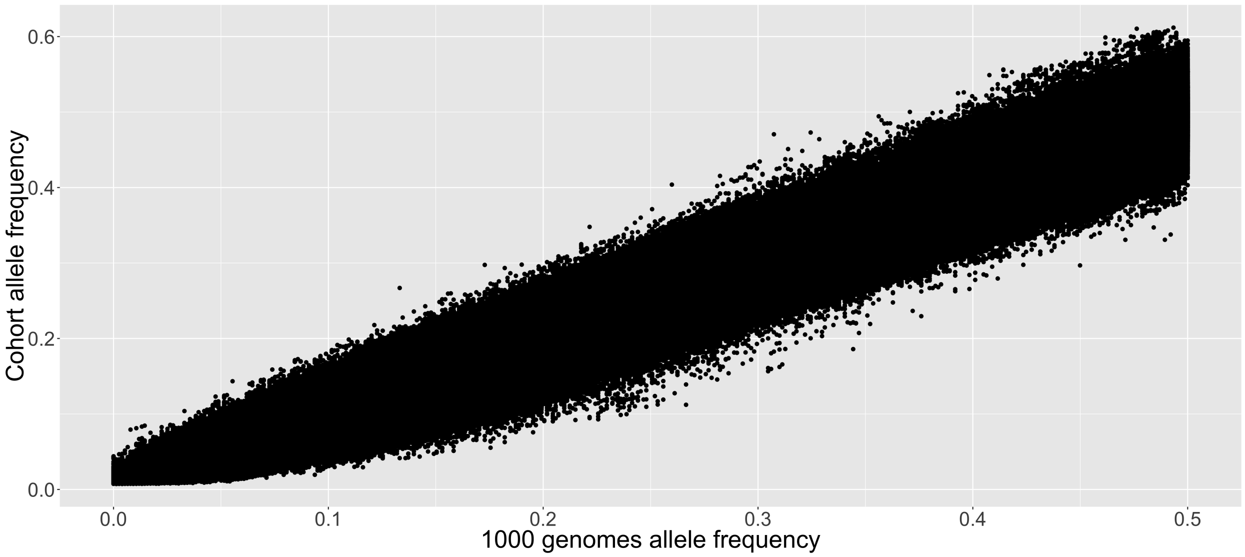

Figure S4 Allele frequency quality control post imputation for the whole CAGE data set. Allele frequency plot of whole CAGE data (y-axis) versus the 1000 Genomes Phase 1 Version 3 reference (allele frequencies calculated from European individuals) post removal of SNPs with a frequency difference greater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

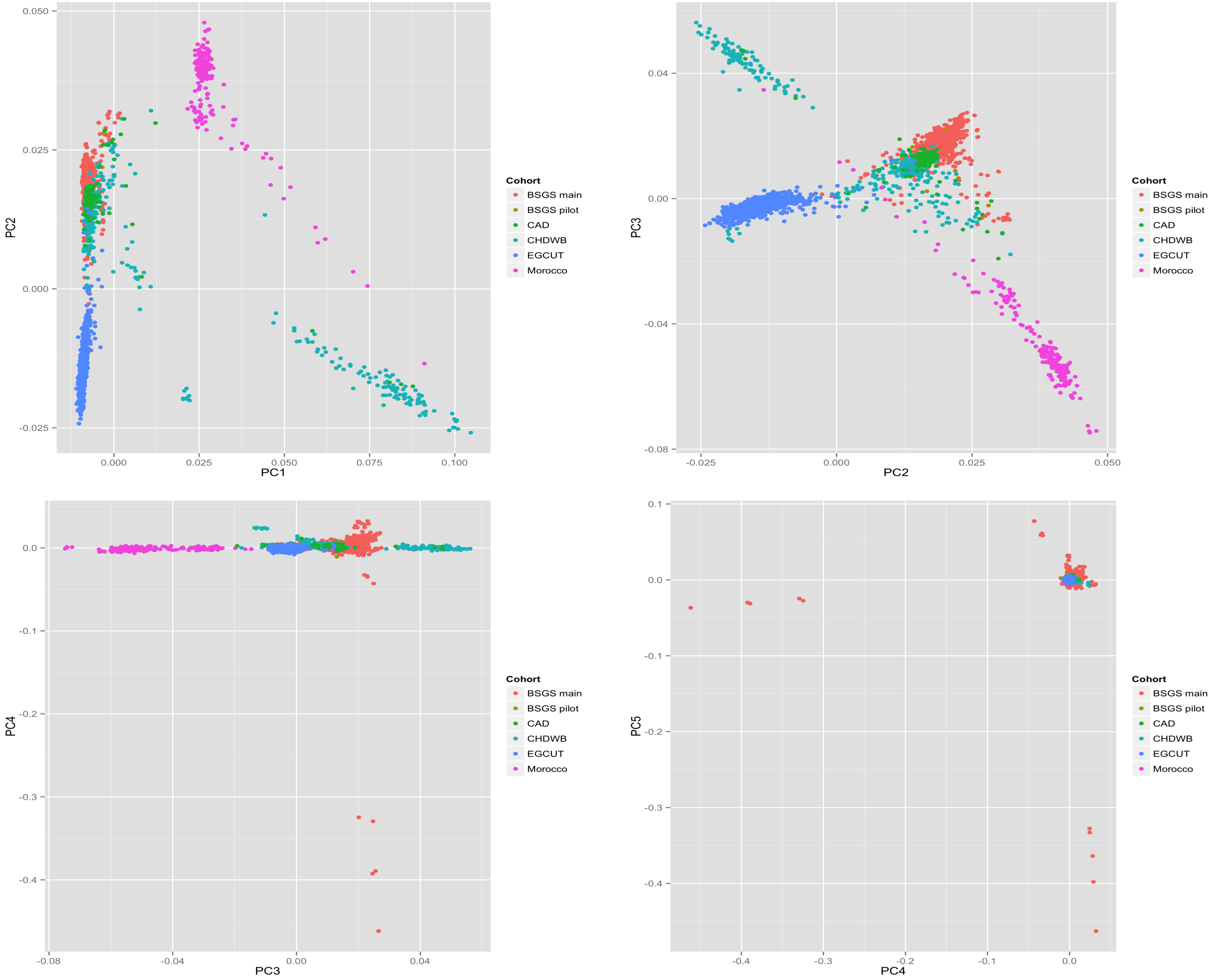

Figure S5 Principal component plots of genotype dataset. Plots depict the first four principal components from a PCA analysis on the whole CAGE genotype data set (7.8 million SNPs) after the completion of the imputation pipeline and merge. Colours indicate the individuals from each cohort and are classified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

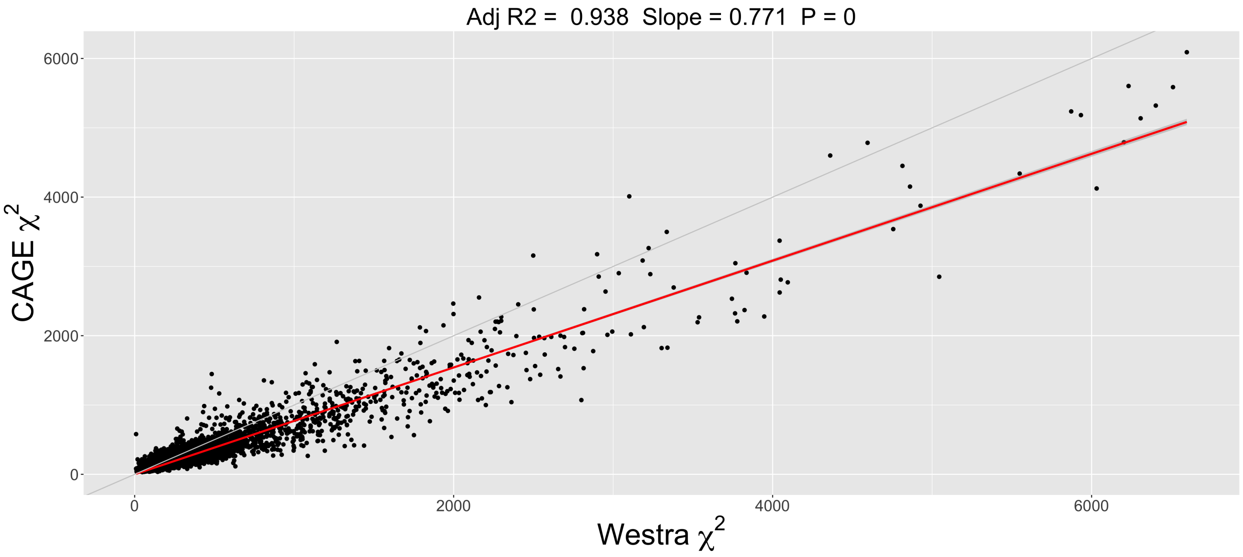

Figure S6 Meta-analysis chi-squared statistics comparison. Scatterplot of chi-squared statistics for 3,202 sentinel SNPs (*cis*) from the Westra *et al.*[29] study versus chi-squared statistics from CAGE data (all individuals $n = 2,765$) generated using a linear model in PLINK with 10 PCs fitted as 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.

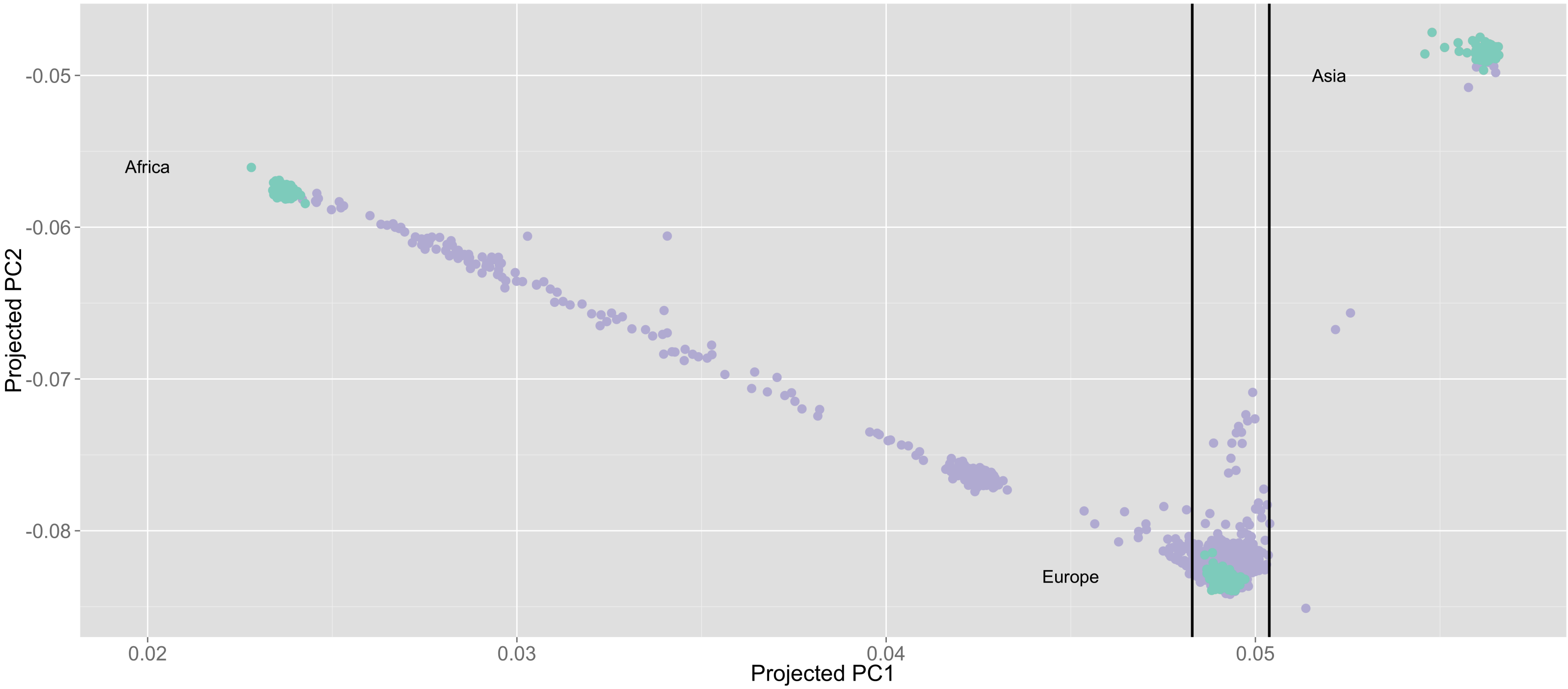

Figure S7 Ancestry investigation. Projected principal component (PPC) plot (PPC1 versus PPC2) of Hap Map 3 cohorts (green) and CAGE data ($n = 2,765$) (purple). The Utah residents of northern and western European ancestry (CEU) cohort from Hap Map 3 formed the European sample, the Yoruba trios from Ibadan, 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).

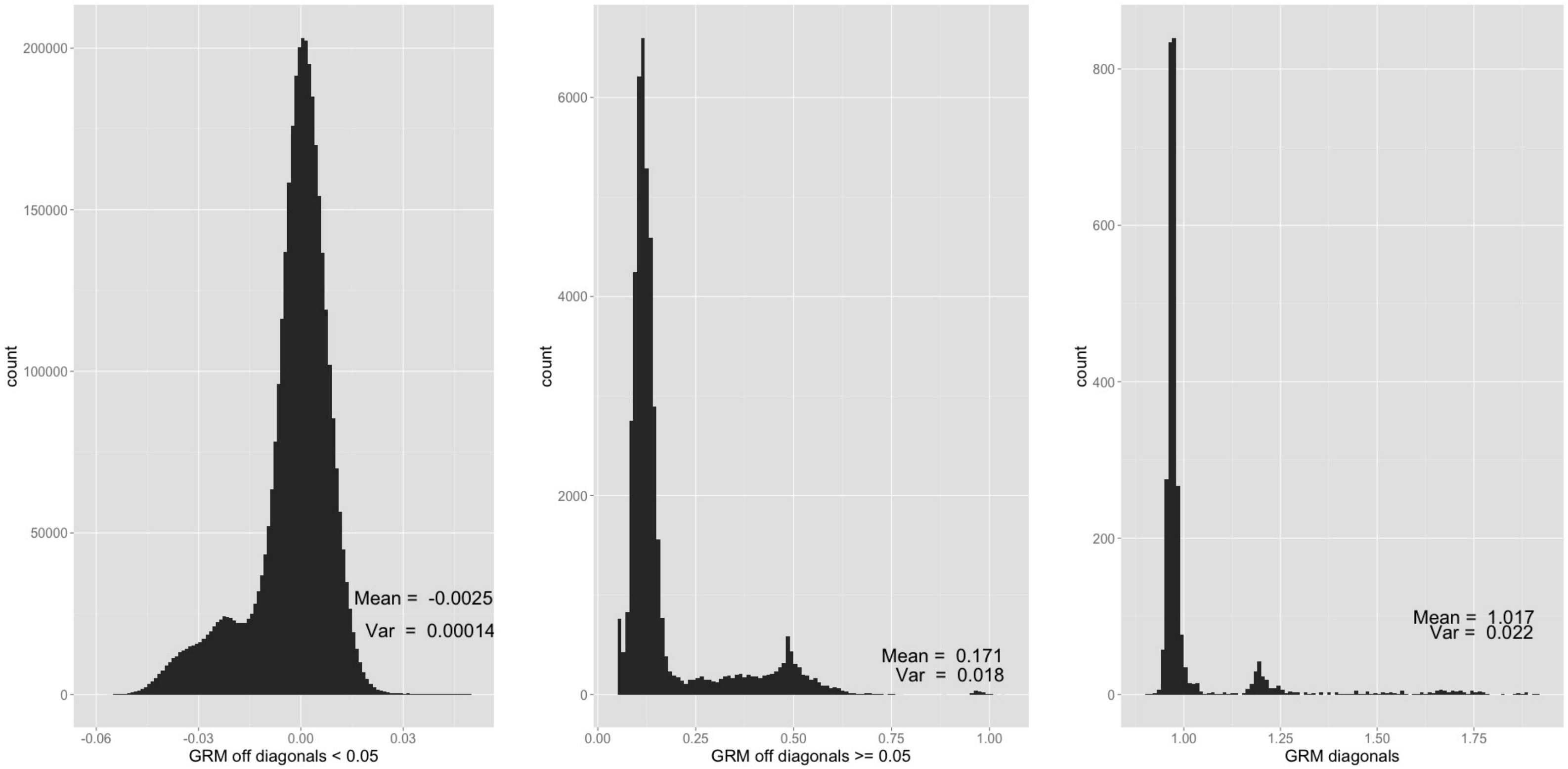

Figure S8 Genetic relationship matrix for all of CAGE. Summary of elements of the genetic relationship matrix (GRM) built using overlapping Hap Map 3 SNPs (893,626) and all individuals ($n = 2,765$) in CAGE. Means and variances are summarised for the histogram displayed. The GRM off diagonals are partitioned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

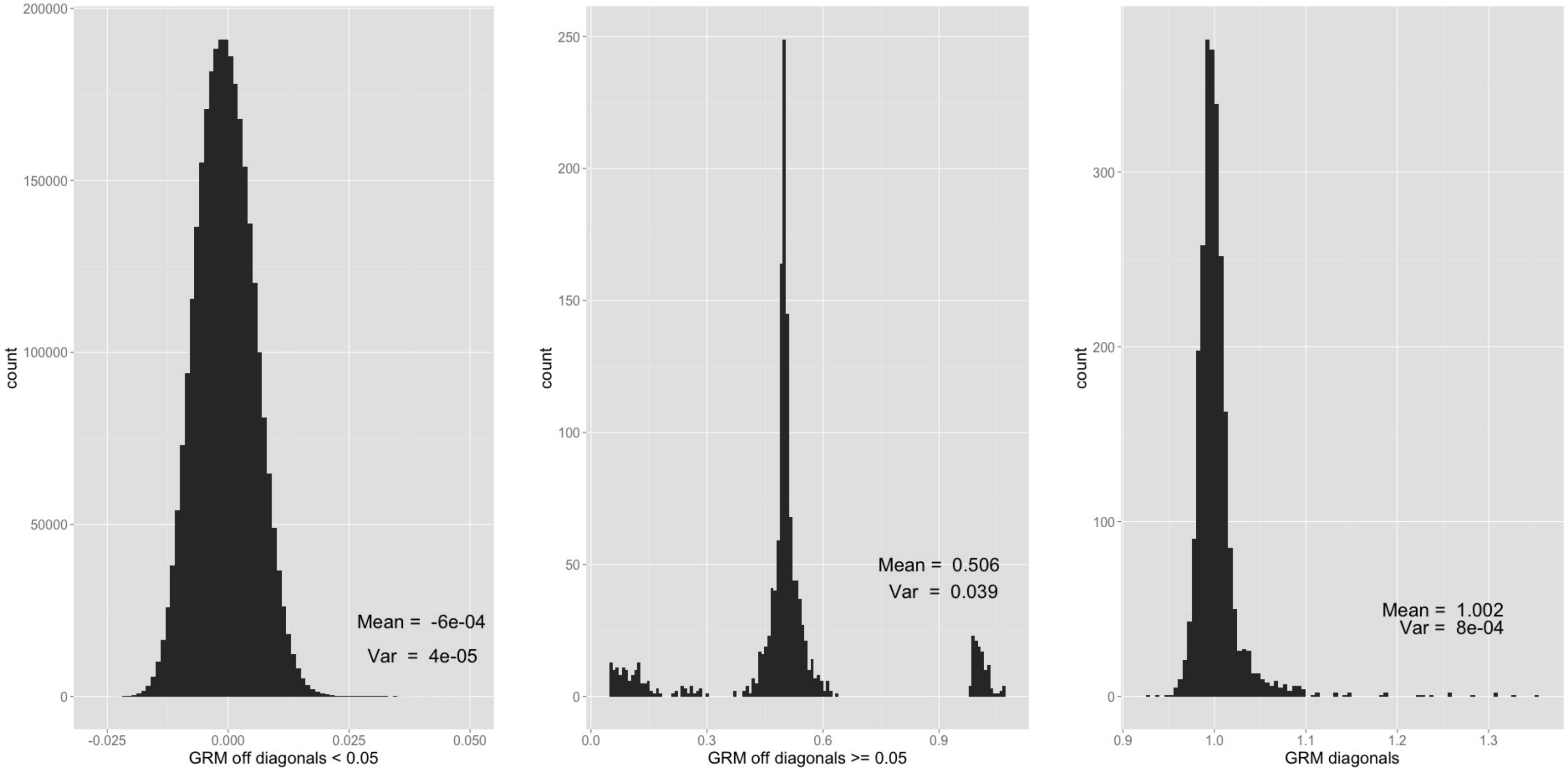

Figure S9 Genetic relationship matrix for European individuals. Summary of elements of the genetic relationship matrix (GRM) built using overlapping Hap Map 3 SNPs (893,626) and European individuals (n = 2,454). Means and variances are summarised for the histogram displayed. The GRM off diagonals are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

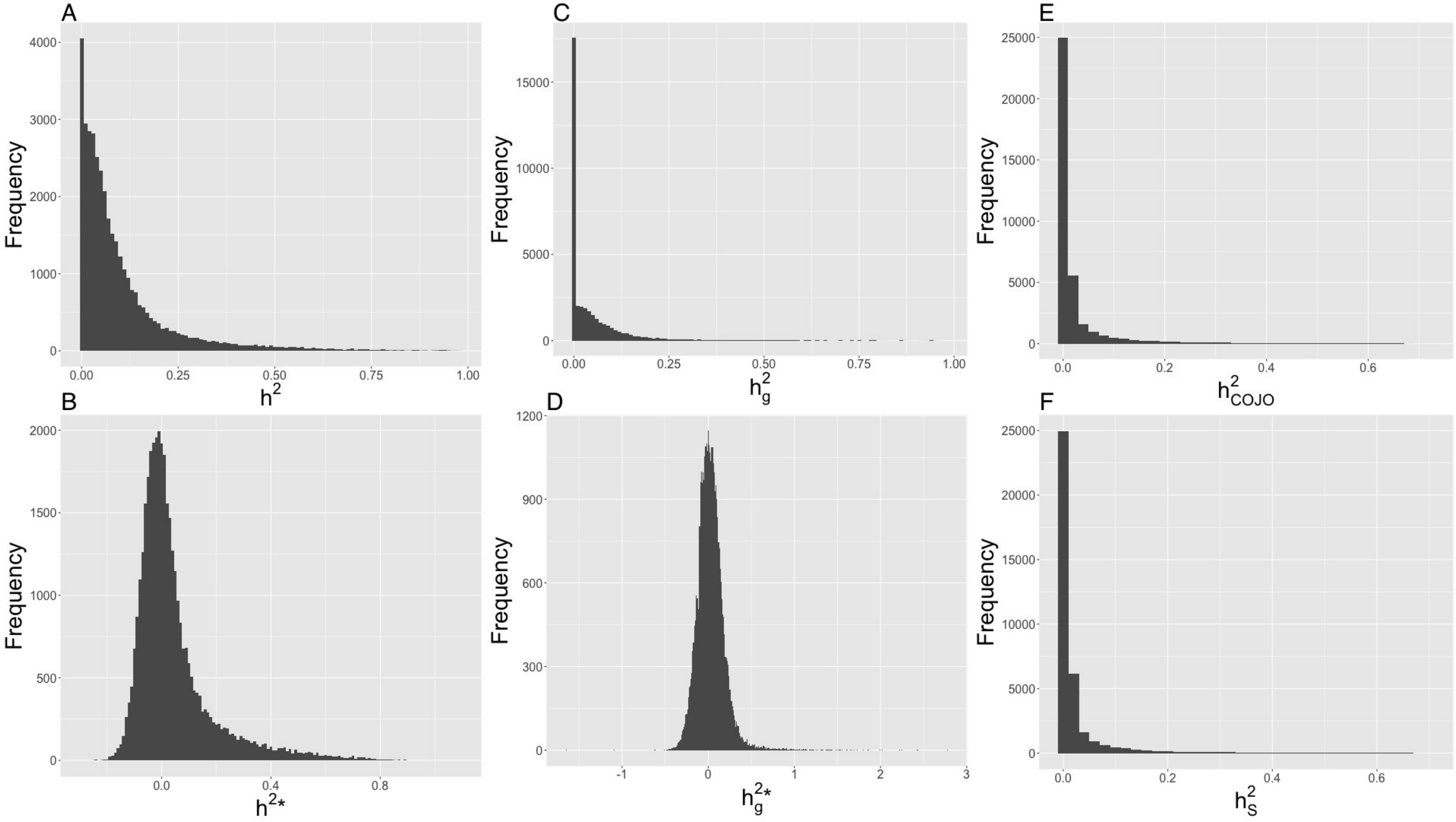

Figure S10 **Distributions of heritability estimates for all probes**. Histogram of heritability estimates across 36,778 probes generated using the Big K/Small K method, and estimates of h^2_{COJO} and h^2_S. Panels (A) and (B) display histogram summaries of the narrow-sense heritability estimates using 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.

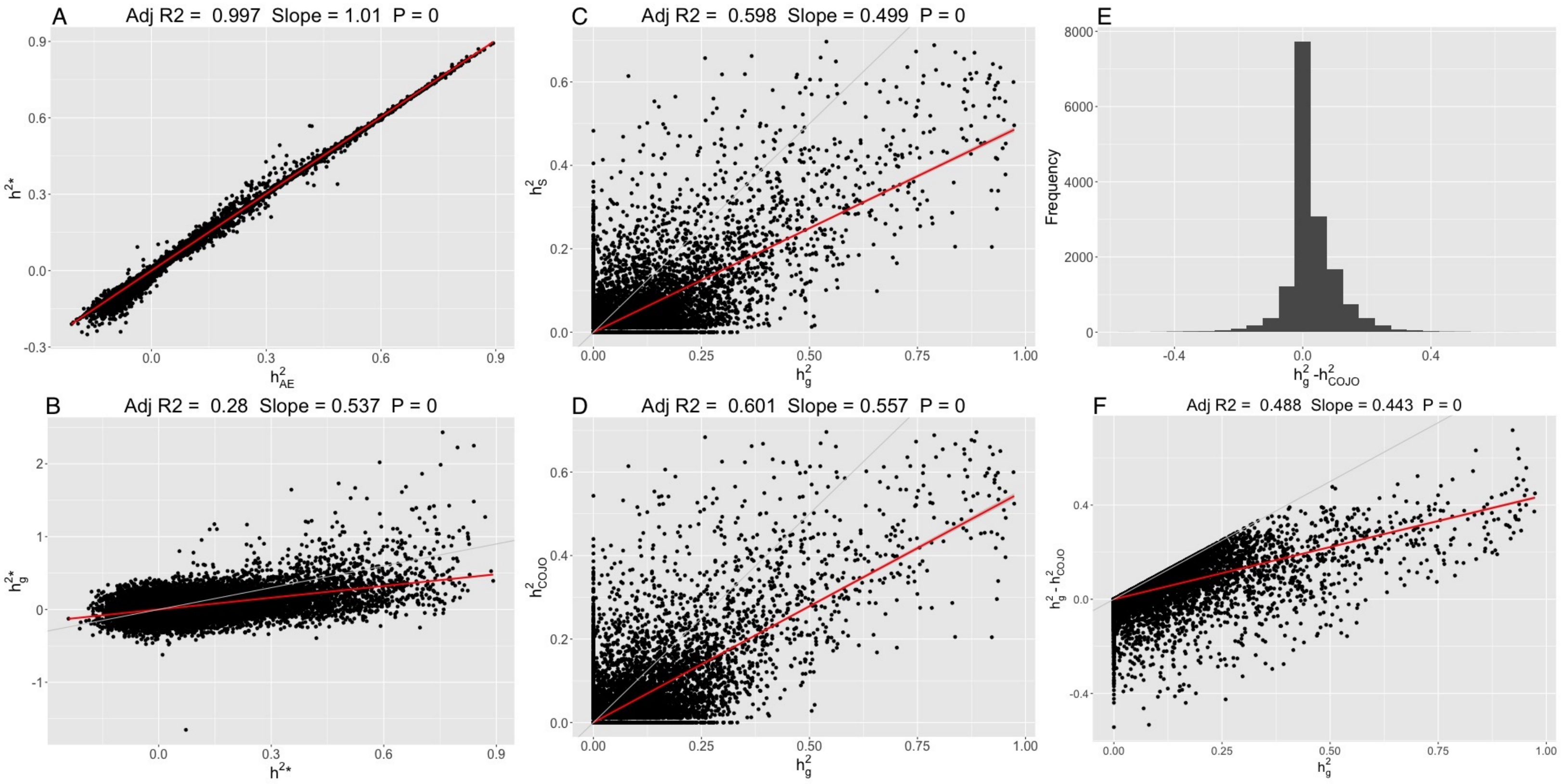

Figure S11 **Comparison of heritability estimates for expressed probes**. Summary of heritability estimates (unconstrained) using only the $\mathbf{K}_{\mathrm{IBS}>t}$ matrix of estimated relatedness (h^2_{AE}), h^2_g (constrained) and h^{2*}_g (unconstrained) of Big K/Small K method, proportion of 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.

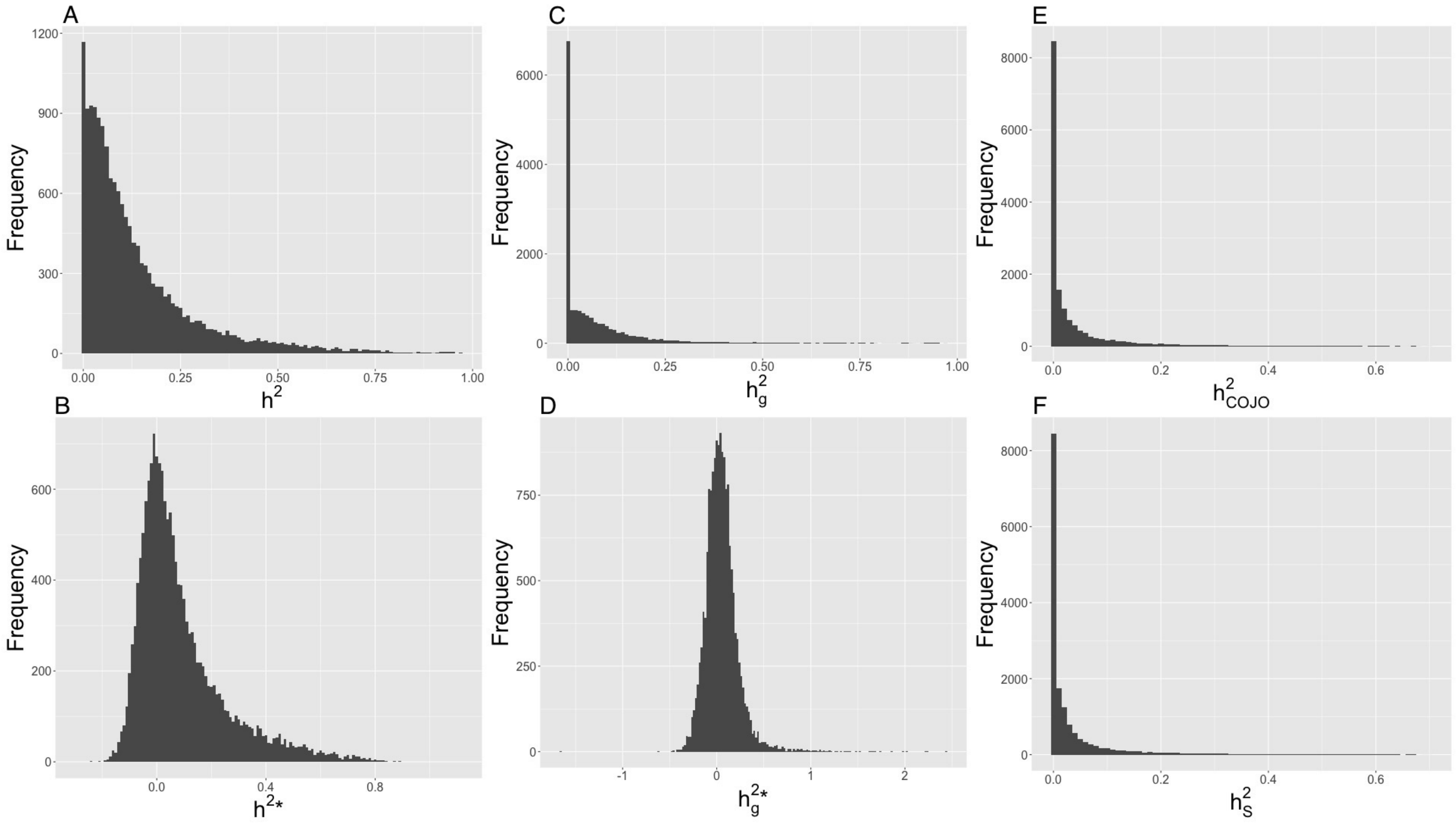

Figure S12 **Distributions of heritability estimates for expressed probes**. Histogram of heritability estimates across 15,966 expressed probes generated using the Big K/Small K method, and estimates of h^2_{COJO} and h^2_S. Panels (A) and (B) display histogram summaries of the narrow-sense heritability 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.

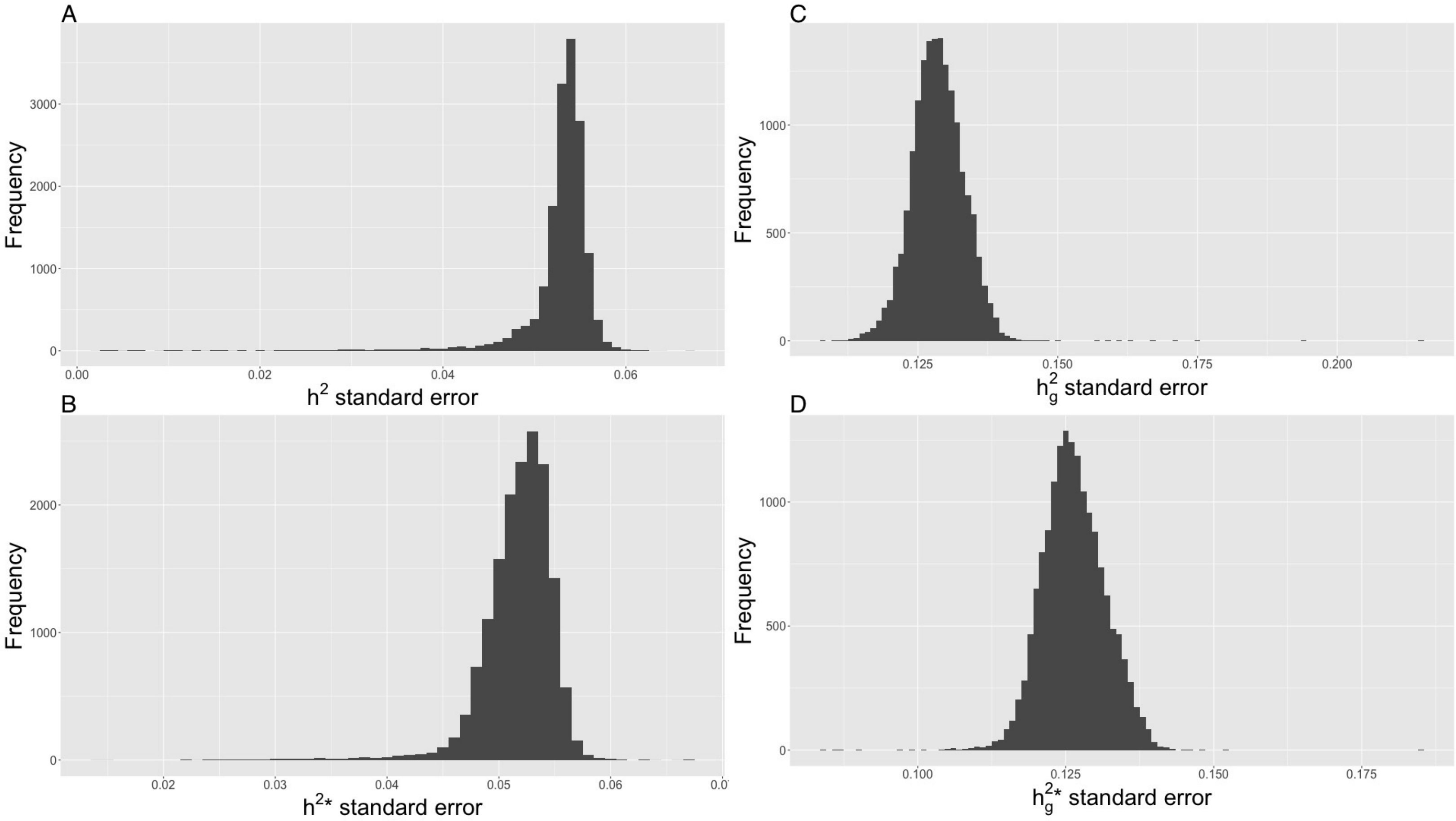

Figure S13 Distributions of standard errors of heritability estimates for expressed probes. Histogram of standard errors of heritability estimates across 15,966 probes generated using the Big K/Small K method. Panels (A) and (B) display histogram summaries of the standard errors for narrow-sense heritability 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.

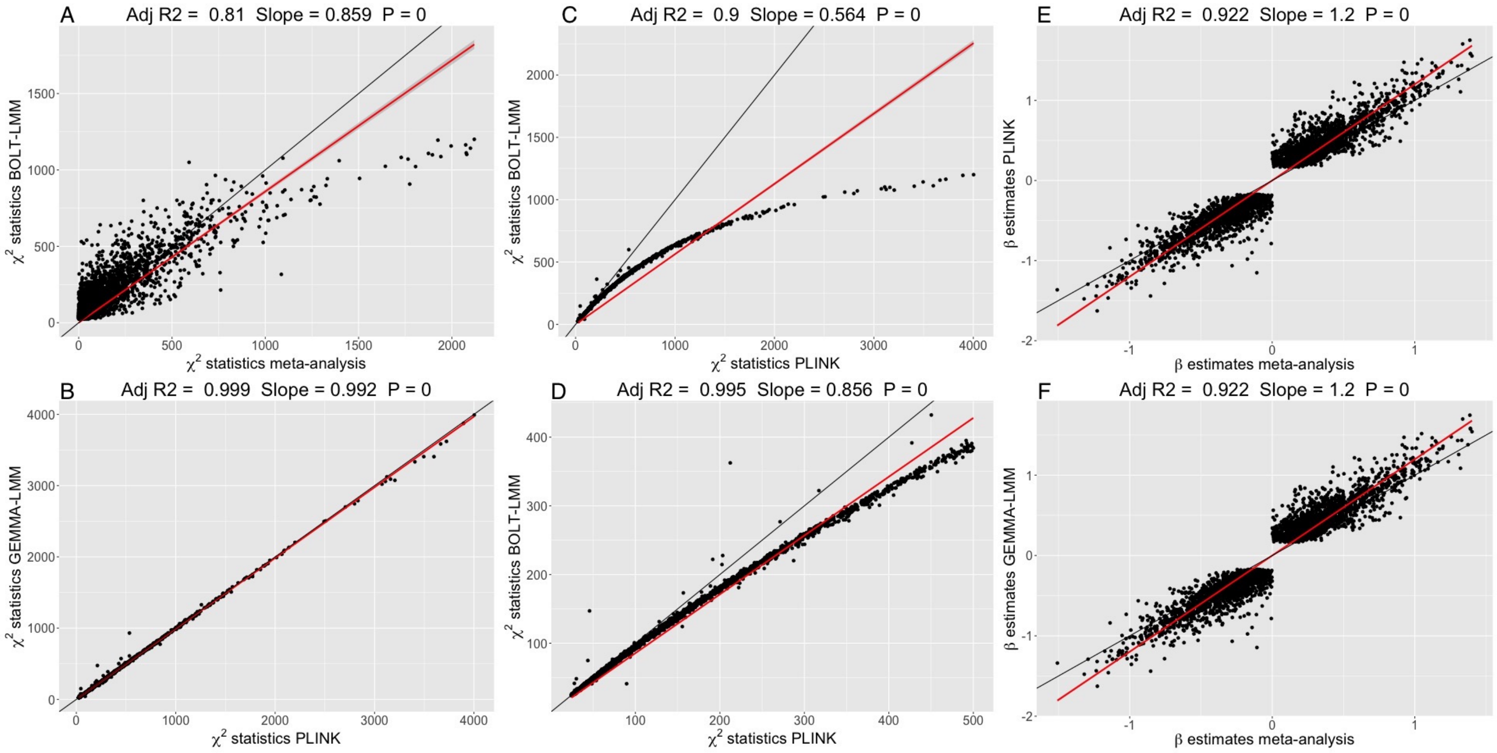

Figure S14 Comparison of mega- versus meta-analysis chi-squared statistics and effect sizes. Comparison of χ^2 statistics for the sentinel SNPs of 3,450 *cis* probes generated from the meta-analysis of Westra *et al.*[29] (n = 1,749) and an eQTL analysis using European unrelated individuals (n = 1,748) 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 prephasing. 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.