Supplementary information

Supplementary materials and methods

M.1 Setup of study

This eQTL meta-analysis is based on gene expression intensities measured in whole blood samples. RNA was isolated with either PAXgene Tubes (Becton Dickinson) or Tempus Tubes (Life Technologies). Different Illumina Whole-Genome Expression Beadchips were used: HT12-v3 arrays, HT12-v4 arrays, and HumanRef-8 v2 arrays. Although different identifiers are used across these different platforms, many probe sequences are identical. Meta-analysis could thus be performed if probe-sequences were equal across platforms. Genotypes were harmonized using HapMap2-based imputation. In total, the eQTL meta-analysis was performed on seven independent cohorts (nine datasets), comprising a total of 5,311 unrelated individuals.

M.2 Discovery studies

Fehrmann

The Fehrmann dataset consists of whole peripheral blood samples of 1,469 unrelated individuals from the United Kingdom and the Netherlands[1,2]. Some of these individuals are patients, while others are healthy controls. Individuals were genotyped using Illumina HumanHap300, HumanHap370 or the 610 Quad platform. Genotypes were imputed using Impute v2[3], using the phased genotypes of the CEU subpopulation of HapMap2 release 24 as reference[4]. RNA levels were quantified using both the Illumina H8v2 platform (N = 229) and the HT12v3 platform (N = 1,240), as has been described before. The Fehrmann expression dataset is available at GEO (Gene Expression Omnibus) public repository under the accessions GSE20332 and GSE20142, respectively. As sample mix-up correction was performed prior to the participation in this study, the total number of samples, having both genotype and gene-expression data, equals 1,469.

SHIP-TREND

SHIP (Study of Health in Pomerania, North-East of Germany) is a population-based project consisting of two independent cohorts, SHIP and SHIP-TREND. Study design of SHIP has been previously described in detail[5]. For this eQTL analysis, the SHIP-TREND cohort was used. The SHIP-TREND probands (N=986) were genotyped using the Illumina HumanOmni2.5-Quad arrays. Genotypes were imputed to HapMap v22[4] using IMPUTE[3]. RNA was prepared from whole blood under fasting conditions in PAXgene tubes (BD) using the PAXgene Blood miRNA Kit (Qiagen, Hilden, Germany). For SHIP-TREND this was done on a QIAcube according to protocols provided by the manufacturer (Qiagen). RNA was amplified (Ambion TotalPrep RNA), and hybridized to the Illumina whole-genome Expression BeadChips (HT-12v3). The SHIP-TREND expression dataset is available at GEO (Gene Expression Omnibus) public repository under the accession GSE 36382. After sample mix-up correction both imputed genotypes and whole-blood gene expression data (after sample mix-up correction) were available for a total of 963 SHIP-TREND samples.

Rotterdam Study

The Rotterdam Study (RS) is a large prospective, population based cohort study in the district of Rotterdam, the Netherlands, investigating the prevalence, incidence, and risk factors of various chronic disabling diseases among eldery Caucasians aged 45 years and over. The initial cohort, named Rotterdam Study I (or RS-I) started in 1989, and consisted of 7,983 persons aged 55 years or over, living in the well-defined Ommoord district. In 1999, a second cohort, named Rotterdam Study II (or RS-II) was started and consisted of 3,011 participants who had become 55 years or moved into the study district. In 2006, a further extension of the cohort was initiated in which 3,932 subjects were included, aged 45 years or over, called Rotterdam Study III (RS-III). The Rotterdam Study has been described in detail[6]. Informed consent was obtained from each participant, and the medical ethics committee of the Erasmus Medical Center Rotterdam approved the study.

For this eQTL analysis, the RS-III cohort was used. The RS participants (n=3,054) were genotyped using the Illumina 610K quad arrays, and genotypes were imputed using MACH[7] using the HapMap CEU Phase 2 genotypes (release #22, build 36) as a reference[4]. Whole blood of 768 samples was collected (PAXgene Tubes-Becton Dickinson) and total RNA was isolated (PAXgene Blood RNA kit-Qiagen). RNA was amplified, labelled (Ambion TotalPrep RNA), and hybridized to the Illumina Whole-Genome Expression Beadchips (Human *HT-12v4*). The RS-III expression dataset is available at GEO (Gene Expression Omnibus) public repository under the accession GSE 33828. The total number of RS-III samples with both imputed genotypes and whole-genome expression data equals 768 (before sample mix-up correction). After sample mix-up correction, 762 samples remained.

EGCUT

The Estonian Gene Expression Cohort[8] is composed of 899 samples (Mean age 37 (16.6) years; 50% females) from the Estonian Genome Center, University of Tartu (EGCUT) biobank cohort of 53,000 samples. Genotyping was performed using Illumina Human370CNV arrays (Illumina Inc., San Diego, US), and imputed using Impute v2[3], using the HapMap CEU phase 2[4] genotypes (release #24, build 36). Whole peripheral blood RNA samples were collected using Tempus Blood RNA Tubes (Life Technologies, NY, USA), and RNA was extracted using Tempus Spin RNA Isolation Kit (Life Technologies, NY, USA). Quality was measured by NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, DE, USA) and Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA). Whole-Genome gene-expression levels were obtained by Illumina Human HT12v3 arrays (Illumina Inc, San Diego, US) according manufacturers protocols. After sample mix-up correction, 8 samples were excluded, and 891 samples remained.

DILGOM

The Finnish study samples included a total of 513 unrelated individuals aged 25–74 years from the Helsinki area, recruited during 2007 as part of the Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study, an extension of the FINRISK 2007 study described earlier[9]. Study participants were asked to fast overnight (at least 10 hours) prior to giving a blood sample. DNA was extracted from 10 ml EDTA whole blood samples with salt precipitation method using Autopure (Qiagen GmbH, Hilden, Germany). DNA purity and quantity were assessed with PicoGreen (Invitrogen, Carlsbad, CA, USA) and genotyping used 250 ng of DNA which proceeded on the Illumina 610-Quad SNP array (Illumina Inc., San Diego, CA, USA) using standard protocols. SNPs were imputed

with MACH version 1.0.10[7] using HapMap2 release 24[4] as a reference panel. To obtain stabilized total RNA, we used the PAXgene Blood RNA System (PreAnalytiX GMbH, Hombrechtikon, Switzerland). It included collection of 2.5 ml peripheral blood into PAXgene Blood RNA Tubes (Becton Dickinson and Co., Franklin Lakes, NJ, USA) and total RNA extraction with PAXgene Blood RNA Kit (Qiagen GmbH, Hilden, Germany). Protocol recommended by the manufacturer was used. The integrity and quantity of the RNA samples were evaluated with the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Biotinylated cRNA was produced from 200 ng of total RNA with Ambion Illumina TotalPrep RNA Amplification Kit (Applied Biosystems, Foster City, CA, USA), using the protocol specified by the manufacturer. 750 ng of biotinylated cRNA were hybridized onto Illumina HumanHT-12v3 Expression BeadChips (Illumina Inc., San Diego, CA, USA), using standard protocol (ArrayExpress database: accession number E-TABM-1036). After sample mix-up correction, 509 samples were included for further analysis in this cohort.

InCHIANTI

InCHIANTI[10] is a population-based, prospective study in the Chianti area (Tuscany) of Italy. The participants were enrolled in 1998-2000, and were interviewed and examined every three years. Ethical approval was granted by the Instituto Nazionale Riposo e Cura Anziani institutional review board in Italy. Participants gave informed consent to participate. Genome-wide genotyping was performed using the Illumina Infinium HumanHap550 genotyping chip. We used MACH 1.0.16[7] to impute using the HapMap r22 build-36 reference panel[4]. In the InCHIANTI study, peripheral blood specimens were taken using the PAXgene system (PreAnalytiX GMbH, Hombrechtikon, Switzerland), to preserve transcript expression levels. Samples were collected in 2008/9 (wave 4) from 712 participants and mRNA was extracted using the PAXgene Blood mRNA kit (Qiagen, Crawley, UK) according to the manufacturer's instructions. Whole genome expression profiling of the samples was conducted using the Illumina Human HT-12 v3 microarray (Illumina, San Diego, USA) as previously described[11]. Sample mix-up analysis on 620 samples passing QC and having both genotype and gene-expression data, revealed a total of 9 possible sample mix-ups. The total number of InCHIANTI samples with both imputed genotypes and whole-genome expression data included in this analysis was 611.

HVH

The Heart and Vascular Health[12-14] (HVH) study constitutes a group of population based case control studies of myocardial infarction (MI), stroke, venous thromboembolism (VTE), and atrial fibrillation (AF) conducted among 30-79 year old members of Group Health, a large integrated health care organization in Washington State. Participants of the current study were HVH controls (N=350) for whom expression profiling was done as part of several gene expression pilot studies. Total RNA was extracted using PAXgene Blood RNA Kit (QIAGEN Inc., Valencia, CA); amplified and labeled using Illumina® TotalPrep™-96 RNA Amplification Kit (Life Technologies Corp., Carlsbad, CA); and, hybridized onto Illumina HumanHT-12v3 and v4 Beadchip arrays (Illumina, San Diego, CA). Scanned images of the array chips were imported into Illumina's GenomeStudio Gene Expression Module.

Genotyping was performed at the General Clinical Research Center's Phenotyping/Genotyping Laboratory at Cedars-Sinai using the Illumina 370CNV BeadChip system. Genotypes were called using the Illumina BeadStudio software. Samples were excluded from analysis for sex mismatch or call rate < 95%.

The following exclusions were applied to identify a final set of 301,321 autosomal SNPs: call rate < 97%, HWE $P < 10^{-5}$, > 2 duplicate errors or Mendelian inconsistencies (for reference CEPH trios), heterozygote frequency = 0, SNP not found in HapMap, inconsistencies across genotyping batches. Imputation was performed using BIMBAM[15] with reference to HapMap CEU using release 22[4], build 36 using one round of imputations and the default expectation-maximization warm-ups and runs.

M.3 Replication studies

KORA F4

KORA F4 (Cooperative Heath Research in the Region of Augsburg) is a follow-up survey (2006-2008) of the population-based KORA S4 survey that was conducted in the region of Augsburg in Southern Germany in 1999-2001. The expression analysis in this study was based on whole blood samples of the KORA F4 participants aged 62 to 81 year[16]. RNA was isolated from whole blood using PAXgene Blood miRNA Kit (Qiagen, Hilden, Germany). Purity and integrity of the RNA was analyzed using the Agilent Bioanalyzer with the 6000 Nano LabChip reagent set (Agilent Technologies, Germany). RNA was reverse transcribed with TotalPrep-96 RNA Amp Kit (Ambion, Germany) and hybridized to the Illumina HumanHT-12 v3 Expression BeadChip[17]. The samples were genotyped on the Affymetrix 6.0 GeneChip array[18]. The SNPs were imputed with MACH (v1.0.15)[7] and the HapMap CEU version 22[3] was used as reference population for calling and imputation. Altogether are 740 samples with gene expression and genotype data available for analysis.

Oxford

Oxford cell-specific eQTL analysis has been previously described[19]. In the initial analysis peripheral blood mononuclear cell fractions were purified from 50ml of freshly collected EDTA anti-coagulated blood from 288 healthy European Volunteers using Ficoll gradients. CD14+ monocytes and CD19+ B-cells were subsequently positively selected from this fraction using magnetic beads (MACS, Miltenyi-Bitotec, Bergisch Gladbach, Germany) with all steps performed on ice as per protocol. Individuals were genotyped at 730,525 markers using Illumina OmniExpress Beadchips and, after controlling for population outliers, 283 individuals were used in the final analysis. For the replication, genotypes were imputed against 1000 genomes phase I integrated variant set from March 2012 using IMPUTE2[3] (available at IMPUTE2 website: http://mathgen.stats.ox.ac.uk/impute/impute_v2.html). The final sample size for B-cells was 282 and the final sample size for monocytes was 283.

BSGS

The Brisbane Systems Genetics (BSGS) eQTL analysis has been previously described[20]. Gene expression data was normalized and corrected for batch effects and population stratification using the following procedure: raw expression levels were quantile-normalized, converted to the $\log_2$ scale and subsequently transformed to have a mean of 0 and standard deviation of 1. Expression data was corrected for possible population stratification by fitting the eigenvalues from the first four PCs generated from SNP genotypes in a linear model. Principal components were generated from the results gene expression data. Each PC was tested against the 528,529 genotyped SNPs using a linear mixed model fitting genotypes and a random effect of the BSGS pedigree calculated using coefficients of co-ancestry (Merlin software[21]). 40 PCs were removed from the gene expression data using a linear model,

excluding those having a significant association (at FDR < 0.05) with SNP genotypes. The resulting normalized gene expression data was used to replicate *trans*-eQTLs, fitting a mixed linear model with a random pedigree effect (assuming additive inheritance only). BSGS comprises of a total of 862 individuals of Northern-European origin, from 274 families, consisting of either monozygotic or dizygotic twin pairs along with their siblings and parents. Expression levels for each individual were measured from whole blood using Illumina HT12-v4.0 microarray chips. Whole genome SNP genotypes were generated using Ilumina 610 Quad-Beadchips and after quality control were imputed to 1000 Genomes Release.

Stranger LCL

Although we have included two cell-type specific cohorts as replication datasets, we also performed replication in lymphobastoid cell-lines, since these cell-lines have been used previously for many *cis*- and *trans*-eQTL studies: we assessed whether our *trans*-eQTLs replicated in LCLs of 608 individuals from HapMap3[22], which were hybridized to Illumina WG6v2 bead chips (ArrayExpress ID: E-MTAB-264). As genotypes, we used HapMap3 release 2 and gene expression measurements were normalized per population, using $\log_2$ transformation, quantile normalization, and principal component (PC) correction. PC correction was limited to 10 PCs because of the small sample size of each individual population. The LCL *trans*-eQTLs were subsequently meta-analyzed over all populations.

M.4 Integration of Illumina platform identifiers.

The different Illumina platforms used by our cohorts share a considerable number of probes with identical 50-mer probe sequences. For many probes, the different platforms use the same probe identifier to identify the same probe sequence. However, some of the platforms use the same probe identifier to identify two or more different probe sequences as we have discussed before[1]. In order to perform the meta-analysis, probes were integrated across HT12-v3, HT12-v4 and HumanRef-8 v2 platforms by determining all unique probe sequences from the different annotation files for each platform. Subsequently we reannotated the probes on the basis of unique probe sequences, by linking these sequences to individual 'Array Address identifiers' that all particitipating cohorts used. Since most of the cohorts used the HT12v3 platform, we limited our analysis to sequences present on this platform. In total, our annotation therfore contained 49,578 unique sequences.

M.5 Initial mapping of Illumina expression probe sequences

Because cross-hybridizing probes, or probes that map with low identity, may give rise to false-positive *cis*- and *trans*-eQTLs, we applied a very stringent mapping procedure prior to our meta-analysis to determine whether the unique probe sequences truly map to a single location on the chomosome. We used three different programs (BLAT[23] v. 34 , SOAPAlign v2.21[24] and BWA 0.5.8c[25]) to map the probe sequences against different sequence indexes created from Ensembl release 54[26] of the human reference genome (HG18 / build 36.3). Probes were mapped against indexes consisting of the autosomes, sex chromosomes, mitochondrial DNA, transcripts, non-coding RNA sequences, exonic sequences and finally against exon-exon boundary sequences. Exon-exon boundary sequences were created by extending the sequence around exon-exon boundaries by 50 basepairs on either side of the boundary. Probes were mapped in using standard parameters for each program. The output files of the

three mapping programs were then merged and mapping positions translated back into chromosomal positions. Subsequently, gapped alignments were only allowed for the alignments that used exon-exon-boundary sequences as a target. Probes were excluded from further analysis if they showed an alignment with multiple genomic locations or low identity mapping (< 96%) by at least one of the three programs applied. In total, 34,061 (69%) probe sequences mapped unambiguously to a single genomic location, reflecting 16,332 genes. We realize that we have been very conservative here, and thus might have missed some true-positive *cis*- or *trans*-eQTL associations, but argue this is justified as this procedure prevents the identification of false-positive eQTLs.

M.6 eQTL mapping procedure

We developed an eQTL Mapping Pipeline (Westra *et al*, manuscript in preparation) which has been used to coordinate and standardize the data collection of results (summary statistics) from the different participating cohort studies. This pipeline was organized in such a way that all participating centers at not a single stage of the meta-analysis needed to share raw genotype or expression data. The pipeline was run by each cohort independently, and the summary statistics were sent to the University Medical Centre Groningen, where the meta-analysis was subsequently performed. Everybody analysed their expression data in exactly the same way, by relying upon the raw, non-normalized and non-background-corrected format, extracted using the Illumina's GenomeStudio V 2010.1 Gene Expression Module. This ensured that the subsequent gene expression normalization within the pipeline was performed in exactly the same way across all cohorts. Every cohort converted imputed dosage and genotype data into TriTyper format[27]. All cohorts subsequently used the eQTL Mapping Pipeline for normalization of the expression data, correction for population stratification, identifying sample mix-ups, correction for confounders and eQTL mapping.

M.6.1 Normalization and Standardization

Per cohort, the raw gene expression data was first quantile normalized to the median distribution and subsequently $\log_2$-transformed. The sample means were then centered to zero, and sample variance was linearly scaled such that each sample had a standard deviation of one (standardization). We corrected the gene expression data for possible population stratification effects by using 4 multidimensional scaling (MDS) vectors as covariates, and used linear regression to obtain the residual gene expression data. The MDS were obtained from the SNP genotype data: SNPs were first pruned on the basis of linkage disequilibrium by PLINK[28] v1.07 (parameters: --indep-pairwise 200 5 0.05), after which MDS was performed (parameters: --cluster --mds-plot 4).

M.6.2 eQTL mapping and meta-analysis

For the *cis*-eQTL mapping, we limited the analysis to combinations of SNPs and probes where the distance between the SNP position and the midpoint of the probe was ≤ 250 kilobases (kb). For the *trans*-eQTL analysis, we included any combination of SNPs and probes, as long as the distance between the SNP position and midpoint of the probe was > 5 megabases (mb). Prior to the *trans*-eQTL analysis, we corrected the gene expression data for all *cis*-eQTL effects to increase statistical power to identify *trans*-eQTLs on genes that also have strong *cis*-eQTL effects (see Supplementary Results).

eQTL association tests were performed using a non-parametric (Spearman's rank) correlation. Since all cohorts used imputed genotype data with the CEU population of the HapMap 2 study as a reference dataset, we performed eQTL mapping on the imputed dosage values which range between zero and two. For the *cis*-eQTL analysis, we limited our analysis to those SNPs with a minor allele frequency (MAF) > 0.05, a Hardy-Weinberg p-value > 0.001, and a call-rate > 95%.
For the *trans*-eQTL analysis, we limited our analysis to 4,542 unique SNPs that are associated to complex traits and diseases ('trait-associated SNPs') as reported by the 'Catalog of Published GWAS studies' (http://www.genome.gov/gwastudies/, accessed July 16th, 2011), and each of which passed quality control in at least three of our datasets.

Spearman's rank correlations were converted to Z-scores using the t-distribution for each dataset:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Where r is the correlation coefficient, and n is the sample size of the dataset. Via the inverse normal distribution, these t-values were consequently converted to Z-scores. This Z-score was then weighted for the square root of the sample size:

$$Z_{weighted_{Dataset_i SNP_j Probe_k}} = \sqrt{n_{Dataset_i SNP_j}} * Z_{Dataset_i SNP_j Probe_k}$$

When a particular SNP-probe pair was present in at least three of the cohorts, we calculated a joint p-value by summing up the weighted Z-scores over all datasets:

$$Z_{sum_{SNP_j Probe_k}} = \sum_{Dataset=0}^{Dataset=i} Z_{weighted_{Dataset_i SNP_j Probe_k}}$$

Finally, the sum of Z-scores was weighted for the sum of samples across datasets which had both genotype and gene-expression data (N):

$$Z_{meta_{SNP_j Probe_k}} = \frac{Z_{sum_{SNP_j Probe_k}}}{\sqrt{N}}$$

From this joint-Z-score, the final meta-analysis p-value was calculated.

M.6.3 Correction for multiple testing

As we performed many statistical tests (especially in the *trans*-eQTL analysis) we had to properly correct for multiple testing. We used the false discovery rate (FDR) procedure to determine what should be the significance threshold in order to end up with a list of significant eQTLs of which a predefined percentage is false-positives: for instance, if we aim to control the FDR at 0.05, among the eQTLs that we ultimately declare significant, 5% will be false-positives.

For this purpose, we needed to obtain the distribution of p-values that we would expect by chance, when conducting exactly the same eQTL analysis, but using randomized phenotypes. Although this would result in an uniform distribution when tests are independent, this is not the case when mapping eQTLs, as many of the statistical tests that we have conducted are correlated, due to structure in both the genotype data (LD) and structure in the gene expression data (due to co-expression). Therefore, we shuffled the sample identifiers in the gene expression data, in order to ensure this structure remained intact, although effectively breaking the correct link between genotypes and gene expression data. All individual cohorts performed ten of such permutations, and sent these permuted eQTL results for meta-analysis, resulting in 10 meta-analyses of permuted eQTL results. With these permutation results we could then determine what the significance threshold in the real data should be, in order to get only a limited fraction of false-positives, determined by using that same significance threshold in the 10 permutations[29].

The obtained FDR estimates are 'probe-level' FDR estimates: for both the real and 10 permuted meta-analysis p-value distributions, we only used the most significant SNP per probe to determine the FDR (one p-value per probe). We applied the following reasoning for the probe-level FDR estimate: if in the real analysis there is a *cis*-eQTL within a locus that has extensive LD and contains many SNPs (e.g. the HLA region), many of these SNPs are likely to show association with this particular *cis*-probe. If we would subsequently estimate the FDR based on the distribution of all p-values (thus all combinations of SNPs and probes), many of the most significant p-values of this distribution will correspond to the aformenentioned probe that is correlated with many HLA SNPs. When we subsequently would use an FDR of 0.05 to determine the significance threshold, we would observe many significant SNP-probe combinations, many of these pertaining to the aforementioned *cis*-probe. Subsequently, although the number of false-positives among all combinations of SNPs and probes will then be 5%, the number of unique probes with a significant *cis*-eQTL effect, that are actually false-positive, will likely be higher than 5%. Due to the existence of correlations in significant p-values in the real analysis (i.e. the HLA SNPs that are in strong LD that are all correlated with the aforementioned probe), this is likely to happen when a disproportianate number of many real *cis*-eQTLs exist that map within regions of extensive LD. In summary, to overcome this potential issue, we determined the FDR significance by using only the most significant SNP per probe in both the real and permuted meta-analysis results. We realize this is a somewhat conservative approach compared, for example, to a method that would determine the FDR by comparing the number of unique SNP-probe combinations in the real analysis that pass a certain significance threshold with the number of unique SNP-probe combinations detected at that significance threshold in the permutations. Through our approach we for instance do not acknowledge the potential presence of secondary, independent SNPs that also affect the same probes.

We stress that we determined the significance thresholds for the *cis*- and *trans*-eQTL analysis completely seperately: the significance threshold for the *cis*-eQTL analysis is substantially less stringent than the significance threshold of the *trans*-eQTL analysis, since I) we conducted many more tests in the *trans*-eQTL analysis, II) *cis*-eQTL effects are more numerous and III) often have a larger effect-size than *trans*-eQTLs. For the *trans*-eQTL analysis we only included SNP-probe pairs that were mapping at least 5 Mb,

to ensure we did not accidentally include long-range *cis*-eQTLs, that otherwise would lead to a small inflation of the number of significantly identified *trans*-eQTLs.

For the *trans*-eQTL meta-analysis, we assessed the stability of our false discovery threshold by running ten additional permuted eQTL meta-analyses: we systematically ascertained which *trans*-eQTL significance threshold corresponds to a false-discovery rate of 0.05 depending on the number of permutations used. We observed that after five permutations the estimated significance threshold was already very stable: the significance threshold that corresponds to an FDR of 0.05 when using only five permutations resulted in identification of 1,513 significant SNP-probe *trans*-eQTLs, identical to what we found in the real analysis using 10 permutations. Increasing the number of permutations to 20 did not alter the FDR threshold: the meta-analysis using 20 permutations also identified 1,513 significant SNP-probe *trans*-eQTLs (Supplementary Figure 16).

M.6.4 Sample mix-up correction

We have previously shown that mislabeling of samples in functional genomics datasets may decrease the power to detect small genetic effects on gene expression[30]. Such sample mix-ups may arise from either mislabeling of genotype or gene expression arrays (including missing arrays and sample swaps) and on average affect 3% of the samples. To correct for these sample mix-ups, we applied our *MixupMapper*[30] methodology to each of the datasets independently. The concept behind this method is straightforward: *cis*-eQTLs define a relationship between the SNP genotype and gene expression levels. As such gene expression levels can be used as a predictor for the associated SNP genotypes. For any pair of genotype and gene expression arrays, we can thus check the concordance between predicted genotype based on strong *cis*-eQTLs and the actual genotype of the associated SNP. After sample mix-up correction by *MixupMapper*, these detected mis-labeled samples were excluded from further analysis.

M.6.5 Principal component adjustment

In order to increase the number of detectable *cis*- and *trans*-eQTLs, principal component analysis (PCA) was subsequently applied on the sample correlation matrix, which was calculated from the quantile normalized, $\log_2$ transformed and standardized gene expression data. We argued that the dominant components capture the majority of the variation within the sample correlation matrix[1]. This variation could either reflect differences between samples caused by environmental, technical (e.g. batch effects) and physiological, or eQTL effects. Therefore, we treated each component as a quantitative trait and correlated each component to SNP genotypes using Spearman's rank correlation to determine whether individual components captured genetic variation. To correct for multiple testing, we conducted 10 permutations, and denoted what the most significant p-value was in these permutations. We used this significance threshold to identify those components that were under genetic control. Once we had identified these components, we corrected the gene expression data, by removing up to 50 PCs from the gene expression data using linear regression, although we did not remove those PCs that were under genetic control. To determine the optimal number of PCs to remove, we iteratively performed *cis*- and *trans*-eQTL mapping on the residual gene expression data after correction for an inceasing amount of PCs (increments of 5 PCs per analysis were used). In order to determine this maximum quickly we limited the analysis to 300,000 SNPs present on the Illumina Human-Hap300 platform (known to capture HapMap2 variation generally well), while for the *trans*-eQTL analysis, we tested 4,542 SNPs from the

Catalog of Published GWAS studies. In each of the datasets that comprised more than 300 samples, the highest numbers of *cis*- and *trans*-eQTLs were observed when we corrected for 40 PCs. This procedure was performed for all datasets, except for HVH-v3 and HVH-v4, where sample size was limited we respectively only corrected for the first five and ten PCs in those two cases.

M.6.6 Identification of false eQTLs due to probe sequence polymorphisms and cross-hybridization

Sequence polymorphisms, which may lead to altered hybridization efficiency of mRNA sequences with the gene expression probe, can cause many false-positive *cis*-eQTL effects[31]. Therefore, we removed SNP-probe combinations from the *cis*-eQTL analysis where the SNP was in LD ($r^2 > 0.2$) with any known SNP (as reported in dbSNP[32] build 130) located within the probe's sequence. To calculate the LD between the *cis*-eQTL SNPs and the SNP in the probe's region, we used the subpopulations of central European descent of the 1000 genomes project[33,34] (2011-05-21 release, 286 individuals, excluding Finnish individuals). If either the *cis*-eQTL SNP or the SNP within the probe region was not present in 1000 genomes, we used the CEU subpopulation of HapMap2[4] release 24 to detect a perfect proxy. For those cases where the LD could not be calculated between the *cis*-eQTL SNP and the SNP in the probe region (for example because there was no apparent proxy in the HapMap data), we excluded the SNP-probe pair from further analysis.

However, for *trans*-eQTL effects, false-positives may be caused by cross-hybridizing probes rather than polymorphisms within the probe sequence. Although we initially used a stringent mapping methodology to map our probe sequences, it may be very well possible that the detected *trans*-eQTLs are actually caused by cross-hybridizations (e.g. hybridization of small portions of the probe sequence to a transcript located near the SNP). Therefore, we tried to falsify each of the detected *trans*-eQTLs by mapping the probe sequences to a region of 5 mb surrounding the *trans*-eQTL SNP. For this purpose, we used SHRiMP[35] v2.2.2, which uses both local and global methodologies to perform alignment, with very relaxed settings (accepting the fact that we might accidentally also remove genuine *trans*-eQTLs): we used a match score of 10, a mismatch score of 0, a gap open penalty of 250, a gap extension penalty of 100 and a minimal Smith-Waterman score of 30% (parameters: -m 10 -i 0 -q -250 -f -100 -h 30%). SNP-probe combinations where the probe mapped with at least 15 bases of identity within the 5 mb region surrounding the SNP were deemed potential false-positive and removed from any subsequent analysis.

After the potential false positive *cis*- and *trans*-eQTL effects had been removed from the real, non-permuted data, we repeated the FDR calculation (again controlling the FDR at 0.05 and 0.5), to ensure these false positives did not accidentally inflate the number of significant eQTL at a certain FDR.

M.7 Functional annotation of eQTLs

To analyze the relationship between overall gene expression levels and the number of detected eQTLs we detected, we ranked all tested probes according to their average expression level in the Fehrmann-HT12v3 dataset, and subsequently binned them in groups of 2,000 probes each. Afterwards, we determined the number of *cis*- and *trans*-eQTLs per bin. To perform functional annotation of *cis*- and *trans*-eQTL SNPs, we used the online tools HaploReg[36], SNPInfo[37] and SNPNexus[38,39]. Since SNPs may tag other variants due to of LD structure, we also determined perfect proxies ($r^2 = 1$) for those SNPs using

the CEU subpopulations in 1000 genomes[40] (excluding Finnish individuals). This set of *cis*- and *trans*-eQTL SNPs was then used as input for SNPInfo and SNPNexus. For the *trans*-eQTL SNP enrichment analyses, we limited this analysis to those SNPs that were associated with complex traits at genome-wide significance levels ('trait associated SNPs', $P < 5 \times 10^{-8}$). These SNPs were subsequently pruned using the 'clump' command found in PLINK[28] ($r^2 < 0.2$, within a window of 1Mb). Default settings were used for both tools. Both tools use database lookups in an integrated database consisting of multiple data sources, to annotate SNPs on human genome build 18.

In order to get accurate and realistic null-distributions for each of these tools, we used the following procedure: we repeated the analysis using each of the three tools, by taking an equal number of SNPs as in the real analysis, but chosen based on the top hits out of permuted runs, thus selecting random SNPs that still adhere to the same criteria as the SNPs that we used in the real analysis (i.e. only those SNPs that map within 250kb from tested probes for the *cis*-eQTL analysis, and only trait-associated SNPs for the *trans*-eQTL analysis). For the *trans*-eQTL SNPs, we selected only those SNPs that did not show a significant *trans*-eQTL effect in our discovery meta-analysis (including SNPs linked to these *trans*-eQTL SNPs, using $r^2 > 0.2$). We then combined the results from SNPInfo and SNPNexus, and tested the difference between real and permuted sets of SNPs for each category of annotation using a Fisher's exact test. For the *cis*-eQTL SNPs, we corrected for multiple testing using the Bonferroni correction, accounting for the number of functional categories tested (0.05/15 = 0.003). We then determined the fold change for each category by dividing the fraction of real data SNPs, by the fraction of permuted data SNPs.

For HaploReg, we used a different approach: this tool calculates the enrichment for enhancer regions in 9 ENCODE project cell-lines, using a binomial test. To determine the null-distribution, we used the same approach as we used for SNPInfo and SNPNexus. Because of the large number of observed *cis*-eQTL effects, we used bins of 2,000 probes (probes were ranked by effect size) and then selected their accompanying top SNPs (and perfect proxies) in both real and permuted data. Then, we performed the enrichment analysis on each of these sets. Finally, we determined the average enrichment over all cell-lines per bin (and over all permutations) in order to establish the relationship between enhancer annotation and eQTL effect size.

M.8 Correlations with endophenotypes

With the help of the *cis*- and *trans*-eQTL analyses we identified sets of genes that are regulated by single SNPs. By correlating gene expression levels of significant eQTL-probes directly to related phenotypes (such as cholesterol levels, BMI, weight, and bloodpressure), we gained insight how these downstream gene are related to endo-phenotypes. We selected the most significant eQTL probes, and adjusted the fully normalized gene expression levels for age and gender, both adjusted for 40 PCs (not adjusting for PCs under genetic control) and non-adjusted for 40PCs. All (endo) phenotypes were also adjusted for age and gender. Correlations between probes and phenotypes were subsequently calculated using a Spearman's rank correlation. We corrected for multiple testing by applying a Bonferroni correction.

M.9 Correlations with blood cell-counts and age

A number of trait-associated SNPs have been associated with changes in cellular composition within blood, which may increase expression for genes specific for these cell types. As a consequence, some *trans*-eQTLs may reflect differences in cell-types rather than genetic effects on gene expression. Although our PCA based normalization strategy captures[41] and thus to some extent compensates for differences in cellular composition, residual effects may be still present after PC correction. Therefore, we compared the effect of cellular composition on gene expression with the *trans*-eQTL effect. Because not all endophenotypes were available for each cohort, we chose to meta-analyse those endophenotypes that were shared between cohorts. Consequently, we selected 19 phenotypic measurements that were available for at least 1,500 samples over all cohorts (including age and different cellular composition measurements). Each cohort individually correlated residual gene expression values (after PC correction) with the phenotypic measurements using Spearman's rank correlation. Summary statistics were then meta-analyzed using a Z-score method, weighted for the sample size. Using the inverse Student's T-distribution, the obtained meta-analysis p-values were then converted to correlation coefficients to determine the variance explained by each phenotype. Similarly, the meta-analysis p-values obtained from the *trans*-eQTL mapping were used to calculate the variance explained by each *trans*-eQTL. If the *trans*-eQTL effect size is larger than the endophenotype effect size, this means the *trans*-eQTL effect cannot be solely explained by endophenotypes.

We also repeated the *trans*-eQTL analysis on the EGCUT cohort, which had measurements for all 18 blood cell parameters and age. We multiple linear regression to adjust the gene expression data for each of these components and then reran the *trans*-eQTL analysis, comparing the effect sizes of the *trans*-eQTLs before and after endophenotype correction (see Supplementary results).

M.10 SLE IKZF1 ENCODE ChIP-seq Analysis

We used ENCODE ChIP-seq signal data for *IKZF1*[42]. For every human gene we determined the average signal (corrected for gene size), corrected for GC-content bias, and performed a Wilcoxon Mann-Whitney test to test determine the down regulated genes (*MX1, TNFRSF21, IFIT1/LIPA, HERC5, CLEC4C, IFI6*) showed a higher ChIP-seq signal, compared to all other human genes.

Supplementary results

R.1 *Cis*-eQTL Mapping

We performed *cis*-eQTL mapping (SNP-probe's midpoint distance < 250 kb) in 5,311 unrelated peripheral blood samples, and observed significant *cis*-eQTL effects for 397,310 SNPs, 8,228 probes which mapped to 6,418 unique genes (44% of all tested genes) and 4,690 unique genes when using a more stringent Bonferroni multiple testing correction. Top-associated SNPs per probe (FDR < 0.5) are listed in Supplementary Table 1 and their positions are shown in Supplementary Figure 4.

R.1.1 *Cis*-eQTL replication

We compared the results of our current meta-analysis (excluding the Fehrmann datasets) with the results we had published before on the both Fehrmann datasets[1]: out of 57,102 SNP-probe pairs reported as significant at FDR < 0.05 in our previous study, 48,071 pairs (84%) were also significant in the meta-analysis excluding the Fehrmann samples. Reassuringly, only 40 (0.08%) of these *cis*-eQTLs showed an opposite allelic effect, indicating that the new approach does not confound our results (Supplementary Figure 7).

Finally, we determined whether the effect directions in each of the datasets were in concordance with the final meta-analysis Z-score (using all datasets). We observed uniform directionality for a majority of the tested *cis*-eQTLs across all datasets, where on average 94% of the eQTLs shared between each of the datasets and the meta-analysis showed an identical effect direction (Supplementary Figure 3).

R.1.2 *Cis*-eQTLs are enriched in highly expressed genes

We examined whether various SNP and gene characteristics of our *cis*-eQTLs reflected those of previously published *cis*-eQTLs[1,17,43-46]. For example, we determined the distance between the SNP and the transcription start site for those probes mapping to a gene in Ensembl release 54. We observed that for 66% of the significant *cis*-eQTLs, the SNP showing the highest association was within 50 kb of the transcription start site (TSS) of the relevant *cis*-eQTL gene (Supplementary Figure 1b). For the majority of the *cis*-eQTLs (97%) the SNP showing the highest association mapped within 250 kb of the TSS. Since we mapped *cis*-eQTLs on the basis of a SNP-probe midpoint distance within 250 kb, some *cis*-eQTLs (3%) show a distance between SNP and TSS that is larger than 250 kb. Such cases could be interesting for follow-up because they imply long range regulation of transcription, or possibly regulation of transcription via external factors such as miRNA binding. Furthermore, we observed that the meta-analysis Z-score shows a dependence upon the SNP-TSS distance (Supplementary Figure 1a), with smaller distances showing a larger effect size in general, and small effect *cis*-eQTLs showing a high variance in SNP-TSS distance. Finally, we observed a clear relationship between the average gene expression level of probes and their likelihood of showing *cis*-eQTL effects (Supplementary Figure 1c): we first ranked all tested gene expression probes on the basis of their average gene expression level (in bins of 2,000 probes). We observed that 84% of the top-2000 most abundantly expressed probes show a *cis*-eQTL effect. For the probes in this bin that did not show a *cis*-eQTL effect, we observed a strong overrepresentation of genes involved in mRNA splicing (Wilcoxon-Mann-Whitney test; GO biological process 'RNA Splicing' term p-value = 6×10^{-9}, KEGG 'Spliceosome' p = 1×10^{-12}). Since these functions

are fundamental to the maintenance for the cell, this may explain why we are unable to detect *cis*-eQTL effects on these probes, irrespective of their expression level and variance.

R.1.2 *Cis*-eQTL SNPs are enriched for functional regions

We investigated whether *cis*-eQTL SNPs were enriched for functional regions: we annotated the *cis*-eQTL SNPs and their proxies (r^2 = 1.0, estimated from European populations in 1000 genomes[40]) showing the highest association per *cis*-eQTL probe, in bins of 1,000 probes, using the online tools SNPInfo[37] and SNPNexus[38,39] (annotating SNPs for functional categories using the ENCODE-project data[42]). Since we had permuted the data ten times (to establish the *cis*-eQTL FDR threshold), we repeated this analysis for each permutation, allowing us to determine the enrichment statistics these methods provide under a null-hypothesis. We observed that the *cis*-eQTL SNPs were enriched for multiple functional categories (Supplementary Figure 2a, Fisher exact p-values: copy number polymorphism: $p = 9 \times 10^{-33}$; miRNA binding sites (from miRanda database): $p = 1 \times 10^{-22}$; transcription factor binding sites: $p = 5 \times 10^{-17}$; 3' UTR: $p = 5 \times 10^{-15}$; miRNA binding sites (from Sanger): $p = 3 \times 10^{-10}$; splice enhancers: $p = 2 \times 10^{-7}$; variants that abolish splice sites: 2×10^{-5}; upstream variants: 1×10^{-5}; downstream variants: 3×10^{-5}). The miRNA binding site and 3' UTR enrichment are of particular interest, since many miRNAs have been reported to exert their function on the 3' UTR regions of genes[47,48], which may explain the effect of some of the associated SNPs. To investigate whether our *cis*-eQTL SNPs were enriched for enhancer regions, we submitted the same set of SNPs (permuted and real data) to the online tool HaploReg[36], which uses the ENCODE pilot ChIP-Seq data to calculate enrichment for enhancer signals in nine different cell types (Supplementary Figure 2b). We observed that the enrichment for enhancer sequences increases with *cis*-eQTL effect size as compared to the permuted SNPs. This is especially pronounced for K562 and GM12878 cell-types, which are myeloid and lymphoblastoid cell-lines, respectively. These enhancer and functional enrichment results indicate that among the *cis*-eQTL SNPs there is enrichment of causal variants.

R.2 *Trans*-eQTL mapping

We performed a focused *trans*-eQTL analysis on 4,542 SNPs obtained from the Catalog of published GWAS (http://www.genome.gov/gwastudies/, accessed July 16th, 2011): all SNPs have previously been implicated in complex traits at various levels of significance. We observed *trans*-eQTLs for 346 SNPs, acting on in total 430 unique genes. The *trans*-eQTL effect sizes are typically very small: 95% of the 1,513 *trans*-eQTLs have an explained variance smaller than 3% (Supplementary Figure 11b). The 346 SNPs for which we have identified a *trans*-eQTL effect represent 176 unique loci (using the 'clump' function in PLINK, with < 1Mb and r^2 < 0.2 setting). A more stringent Bonferroni correction revealed 643 significant *trans*-eQTLs, including 200 unique SNPs and 223 different genes). Significant SNP-probe combinations (FDR < 0.5) are listed in Supplementary Table 2 and their genomic positions are shown in Supplementary Figure 4.

For 26 *trans*-eQTL genes the eQTL SNP affected multiple probes within these genes (Supplementary Table 3), always with consistent allelic directions, suggesting that our probe filtering procedure was effective in preventing false-positive *trans*-eQTLs.

We observed that for 88 out of 1,513 (6%) *trans*-eQTL associations (FDR < 0.05), the probe was located on the same chromosome as the SNP. The average distance between SNP and probe for these *trans*-eQTLs was 50.5Mb. For 36 of these *trans*-eQTLs, the probe was located within 10Mb of the SNP. We therefore conclude that the far majority of the reported *trans*-eQTLs cannot be explained by (long-range) *cis*-regulatory effects.

To ensure that our choice for the removal of 40 PCs did not lead to overfitting, we repeated the meta-analysis, removing either 35 or 45 PCs in each of the cohorts, using 10 permutations in each meta-analysis, in order to establish an FDR < 0.05. For the meta-analysis where we removed 35 PCs in the cohorts with larger sample sizes, we removed 0 and 5 PCs in the HVH-HT12v3 and HVH-HT12v4 cohorts, respectively. For the meta-analysis where we removed 45 PCs in the cohorts with larger sample sizes, we removed 10 and 15 PCs in the HVH-HT12v3 and HVH-HT12v4 cohorts, respectively. Upon correcting the expression for 35 PCs, we observed that 1,433 out of the 1,513 (95%) *trans*-eQTLs that we had detected when correcting for 40 PCs, were significant at FDR < 0.05, all with an identical direction of effect. When correcting for 45 PCs, 1,358 out of 1,513 (90%) *trans*-eQTLs were significant at FDR < 0.05, all with an identical direction of effect (Supplementary Figure 14). This indicates that although a few *trans*-eQTLs disappear when removing different numbers of components, the far majority of identified *trans*-eQTLs are not sensitive to the number of PCs removed.

R.2.1 Correcting for *cis*-eQTL effects increases power to detect *trans*-eQTLs

Our rationale for regressing out *cis*-eQTL effects prior to *trans*-eQTL mapping was that *cis*-eQTL effects (often having fairly strong effects) may somewhat obscure the detectability of *trans*-eQTLs that also act on such genes. Therefore, by correcting for *cis*-eQTL effects, we gain statistical power to detect *trans*-eQTL effects. Therefore, we also performed the *trans*-eQTL meta-analysis without prior correction for the identified *cis*-eQTLs. In this analysis, we identified 1,335 significant SNP-probe *trans*-eQTLs at FDR < 0.05, as compared to the 1,513 that we identified in the original analysis where we had explicitly corrected for *cis*-eQTLs (Supplementary Figure 15), yielding a 12% increase of the number of significant *trans*-eQTLs.

R.2.2 Highly expressed genes are more often *trans*-eQTLs

As with the *cis*-eQTL results, we investigated whether there was a relationship between the average level of expression and the number of *trans*-eQTLs. Gene expression probes were ranked according to their average gene expression level, and divided over bins of 2,000 probes. We then determined the number of probes in each bin constituting a *trans*-eQTL effect. We observed that among the 4000 highest expressed probes a majority (54%) of *trans*-eQTL probes exist (Supplementary Figure 11a).

R.2.3 *Trans*-eQTL replication

To determine the impact of the changes we had made in our normalization strategy, compared to our previously used normalization strategy[1] (i.e.: not removing PCs under genetic control, controlling for population stratification, different probe mapping strategy), we compared the results of our current meta-analysis (excluding the Fehrmann datasets) with the results we had published before on the Fehrmann datasets[1] (Supplementary Figure 7). Out of 226 SNP-probe pairs previously reported as significant at FDR < 0.05, we tested 113 pairs in our new study (50%), of which 71 were significant at FDR

< 0.05 in our new meta-analysis. We did not test the other 113 SNP-probe combinations in our current meta-analysis because: I) we have mapped all probe sequences prior to the start of the meta-analysis using a different strategy as in our previous study, and II) we excluded those SNP-probe pairs when there were less than three datasets that interrogated the specific SNP-probe *trans*-eQTL pair. Reassuringly, irrespective of the significance threshold in the current analysis, none of these *trans*-eQTLs showed an opposite allelic effect (Supplementary Figure 7), indicating that we can replicate our previously reported *trans*-eQTLs, even though the normalization procedures were slightly different.

When we compared the eQTL Z-score directions across the different datasets in the meta-analysis, we observed uniform directionality for the majority of the tested *trans*-eQTLs across all datasets: on average 90% of the eQTLs that were shared between each of the datasets and the meta-analysis showed an identical effect direction (Supplementary Figure 5).

We subsequently attempted replication of our *trans*-eQTLs (1,513 SNP-probe pairs) in two independent datasets peripheral blood datasets, and three cell-type specific datasets.

The cell type specific datasets consisted of a cohort of B-cells, monocytes (Oxford[19], N = 282 and 283 individuals, respectively) and lymphoblastoid cell lines (LCL[22], N=608). As peripheral blood replication datasets, KORA F4[17] (N = 740) and BSGS (N = 862) were available. For the KORA F4 dataset, irrespective of the significance threshold, 1,483 pairs were tested, of which 131 showed an opposite effect direction (9%). Controlling the FDR at 0.05, 771 *trans*-eQTLs were significant. Of these pairs, 15 showed an opposite effect (Supplementary Figure 8). For the BSGS[20] dataset 1,490 pairs were tested, of which 109 showed an opposite effect direction (7%). When controlling the FDR at 0.05, 1,239 *trans*-eQTLs were significant. Of these pairs, 65 showed an opposite effect (Supplementary Figure 8). We subsequently performed a meta-analysis on the two peripheral blood replication studies. Using this meta-analysis, 1,472 out of the 1,512 tested *trans*-eQTLs showed an identical direction of effect. Controlling the FDR at 0.05, 1,346 *trans*-eQTLs were significant (89%), of which 4 had an opposite direction of effect (Supplementary Figure 8). In the B-cell and monocyte replication studies, 1,364 pairs were tested in the datasets, of which 621 (46%) and 609 (45%) showed an opposite effect direction, respectively. Controlling FDR at 0.05, 57 SNP-probe pairs were significant, of which 4 showed an opposite effect direction in B-cells, and 101 were significant in monocytes, of which 12 showed an opposite effect direction (Supplementary Figure 9). For the lymphoblastoid cell-lines, irrespective of the significance threshold, 792 SNP-probe pairs were tested, 359 (45%) of which showed an opposite effect direction. Controlling the FDR at 0.05, 31 SNP-probe pairs replicated significantly, all of which showed an identical direction of effect (Supplementary Figure 10).

R.2.4 *Trans*-eQTLs are enriched for affecting transcription factor genes in *cis*

We ascertained whether the trait-associated SNPs that showed *trans*-eQTL effects at FDR < 0.05 were affecting transcription factors in *cis*, which could be indicative that the *trans*-eQTLs are driven by these transcription factors. We used Ensembl GO annotation (GO:0003700, “DNA binding transcription factor activity”) to define a set of known transcription factors. For each trait-associated SNP we took the most significant *cis*-gene (at FDR < 0.05). We limited this analysis to 254 *trans*-eQTL SNPs that have been previously associated with complex traits at genome-wide significant levels (‘trait-associated SNPs’, p < 5

x 10^{-8} as reported in the Catalog of Published GWAS studies). We subsequently pruned these SNPs using the 'clump' command in PLINK (using $r^2 < 0.2$ and distance < 1Mb as settings).

We observed that the trait-associated SNPs that cause *trans*-eQTLs were also affecting 68 unique genes in *cis*, of which 8 (12%) were transcription factor genes (*C1orf85*, *ZFP90*, *SPI1*, *HHEX*, *IKZF1*, *GATA2*, *PKNOX1*, *RREB1*. We compared this to the tested trait-associated SNPs that did not show *trans*-eQTLs effects, and observed that these SNPs affected 428 unique genes in *cis*, of which 24 (6%) were transcription factor genes, , indicating that the trait-associated *trans*-eQTL SNPs were more often affecting transcription factor genes in *cis* as compared to trait-associated SNPs not affecting genes in *trans* (one-tailed Fisher's exact P = 0.032). We also performed this analysis while using a background based on all SNPs that give *cis*-eQTLs: out of the 6,418 unique *cis*-regulated genes, 293 were transcription factor genes (5%). Using the all *cis*-eQTL genes as background, the significance of the enrichment further increased (Fisher's exact one-tailed test P = 0.006).

As only a few trait-associated SNPs are non-synonymous or in strong LD with non-synonymous proxies, we lacked statistical power to show that SNPs that change protein structure more often cause *trans*-eQTLs than SNPs that do not change protein structure. We did not observe any differences when stratifying the *trans*-eQTL SNPs in groups of SNPs that have multiple *trans*-targets and SNPs that only affect a single gene in *trans*.

R.2.5 SNPs associated with traits at genome-wide significance levels are more often *trans*-eQTL SNPs compared to other SNPs with weaker associations

The Catalog of Published GWAS studies contains many SNPs that are in strong LD with each other. We therefore pruned the 4,542 SNPs that we had tested for *trans*-eQTLs (using the 'clump' function in PLINK, with < 1Mb and $r^2 < 0.2$ setting), yielding 3,363 independent SNPs. We subsequently tested these SNPs for *trans*-eQTLs (using all 5,311 samples), resulting in 145 significant *trans*-eQTL SNPs (594 *trans*-eQTLs, FDR < 0.05 using 10 permutations, Supplementary Figure 12).

Many of the SNPs in the Catalog of Published GWAS studies are not associated with complex traits at a genome-wide significant level ($P < 5 \times 10^{-8}$): out the 4,542 tested SNPs, 2,082 (46%) were reported to have a genome-wide significant association. We stratified the 4,542 SNPs for genome-wide significance, and pruned the SNPs (using the 'clump' function in PLINK, with < 1Mb and $r^2 < 0.2$ setting), resulting in two sets of unlinked SNPs: 1,320 SNPs having a genome-wide trait-association, and 2,280 SNPs having a nominal association with disease. We then tested both sets for *trans*-eQTLs and observed 122 significant *trans*-eQTL SNPs (9,2%; 591 SNP-probe pairs, FDR < 0.05, 10 permutations) for the 1,320 trait-associated SNPs, and 47 significant *trans*-eQTL SNPs (2.1%) for the other set. These results indicate that SNPs that have been found associated with disease at a genome-wide significance more often are also *trans*-eQTL SNP (Fisher's exact $P = 8 \times 10^{-22}$), as compared to SNPs for which no genome-wide significant disease association has been reported (Supplementary Figure 12).

R.2.6 Trait-associated SNPs are more often *cis*- and *trans*-eQTL SNPs compared to random SNPs

We previously showed that trait-associated SNPs ($p < 5 \times 10^{-8}$ in the Catalog of Published GWAS studies; 2,082 out of the 4,542 tested SNPs) are more likely to be *trans*-eQTLs compared to randomly selected

SNPs[1]. However, in order to show this to be the case in this larger study, we have now repeated that analysis. Before performing this analysis, we identified the number of independent SNPs within these 2,082 trait-associated SNPs, by clumping SNPs within 1Mb that show at least an r^2 of 0.2 (using the HapMap2 CEU population and the PLINK 'clump' command), yielding 1,320 independent trait-associated loci. Subsequently, we randomly selected 100 sets of 1,320 unlinked ($r^2 < 0.2$ in the HapMap CEU population) SNPs from all SNPs that passed QC in our *cis*-eQTL analysis. We ensured the SNPs in each of these sets were matched to the trait-associated SNPs (in terms of distance to the transcription start site of the nearest gene and minor allele frequency). Due to computational constraints, we could only test each of the random sets and the set of 1,320 trait-associated loci for showing *trans*-eQTL effects in four cohorts (EGCUT, Fehrmann-HT12v3, Fehrmann-H8v2 and the Rotterdam Study; n=3,122 samples). We subsequently determined the number of SNPs showing significant (FDR < 0.05) *trans*-eQTL effect for each set of SNPs, and observed a six-fold increase in *trans*-eQTLs for the trait-associated loci, as compared to the random sets ($P = 5.6 \times 10^{-49}$, Supplementary Figure 13B). These results thus clearly indicate that SNPs that cause *trans*-eQTLs are also more likely to cause disease.

We also ascertained whether trait-associated SNPs were enriched for *cis*-eQTLs, as compared to random SNPs: we determined how many SNPs affected at least one gene in *cis*, for the clumped set of 1,320 trait-associated SNPs and the randomly selected sets. We observed that 601 trait-associated SNPs (46%) were *cis*-eQTL SNPs (1,431 *cis*-eQTLs in total), which is significantly higher than in the random sets ($P = 1.4 \times 10^{-16}$, Supplementary Figure 13A), where on average 434 SNPs showed a *cis*-eQTL effect. For 1,090 out of the 1,431 *cis*-eQTLs, the trait-associated SNP was the strongest effect for the *cis*-gene, or in strong LD ($r^2 > 0.8$) with the SNP that was the top effect for the *cis*-gene.

R.2.7 Trait-associated SNPs are often the top *trans*-eQTL SNP

We have performed a fine-mapping approach for each of the significant *trans*-eQTL SNPs that also showed a genome-wide significant association ($P < 5 \times 10^{-8}$) with complex traits in the Catalog of Published GWAS studies (1,340 out of the 1,513 significant *trans*-eQTLs): for each *trans*-eQTL, we tested SNPs within 250kb of the *trans*-eQTL SNP (using all 9 discovery studies). We then compared the explained variance between each of those `fine-mapping` *trans*-eQTLs and the trait-associated SNP *trans*-eQTLs and observed that for 671 out of the 1,340 *trans*-eQTLs (50%), the trait-associated SNP was either the top effect, linked to the top effect ($R^2 \geq 0.8$), or independent ($R^2 \leq 0.2$) from the top *trans*-eQTL in that locus (Supplementary Table 6).

R.3 Endophenotype correlations

R.3.1 Principal component adjustment corrects for endophenotype effects

We correlated gene expression levels of the significant *trans*-eQTL probes (167 unique probes at FDR < 0.05) to their original GWAS phenotypes and other related phenotypes in both the Rotterdam Study and the EGCUT study (see Supplementary Table 7 for all tested *trans*-gene / endophenotype combinations). The *trans*-gene / endophenotype correlations are very consistent across the two studies. We identified significant correlations with the mean corpuscular volume (MCV), different blood cell counts (number of granulocytes, number of lymphocytes, number of erythrocytes, and number of platelets), body mass index (BMI), body weight, diastolic and systolic blood pressure, cholesterol levels, triglyceride levels, and

hemoglobin values. We used the Bonferroni method to correct for the number of tests (see Supplementary Table 7 for significance levels per phenotype).

One nice example is the GWAS hit found for blood pressure, rs653178, located in SH2B3 locus on chromosome 12. This SNP affects the *MYADM* gene located on chromosome 19 in *trans*. Gene expression levels of *MYADM* are significantly positively correlated with both systolic and diastolic blood pressure ($p = 7.8 \times 10^{-9}$ and $p = 5.3 \times 10^{-7}$ in the Rotterdam Study, $p = 1.5 \times 10^{-5}$ and $p = 2.8 \times 10^{-3}$ in the EGCUT study, respectively). This indicates that rs653178 may act as a modulator for *MYADM*, thereby influencing both systolic and diastolic blood pressure.

Interestingly, most of the significant correlations with blood cell counts disappeared by adjusting the normalized gene expression levels for the first 40 Principal Components (PCs). For example, gene expression levels of the gene *IDS* are significantly correlated with the number of granulocytes ($p = 1.5 \times 10^{-24}$ in the Rotterdam Study, $p = 3.3 \times 10^{-30}$ in the EGCUT study). When we adjust for the first 40 PCs, the correlations are not significant anymore ($p = 0.31$ in the Rotterdam Study, $p = 0.49$ in the EGCUT study). This indicates that through the removal of PCs, some differences that pertain to differences in proportions of specific cell-types diminish, and other, weaker gene-endophenotype relationships become more readily detectable.

R.3.2 *Trans*-eQTLs are not confounded by blood-cell counts

Although the gene expression data was corrected for 40 PCs, some of which capture differences in cell-counts across individuals[41], ideally, cell-counts for each of the different cell-types should be used as covariate to correct the expression data. However, because cell-count measurements were not available for all cohorts, we were not able to correct the gene expression data for possible cell-count effects. Therefore, we conducted various analyses to ascertain whether the identified *trans*-eQTLs might be due to differences in cellular composition:

For each gene that constitutes a *trans*-eQTL at FDR < 0.05 (1,513 SNP-probe combinations), we first assessed whether the eQTL effect size was larger than any correlation with blood cell-count information, which would suggest that such a *trans*-eQTL effect cannot solely appear due to differences in cellular composition. To do so, we first performed a meta-analysis across our cohorts, correlating *trans*-gene expression levels with age (known to have an effect on cellular blood composition) and 18 different blood count parameters (lymphocyte count and percentage, monocyte count and percentage, basophil count and percentage, eosinophil count and percentage, neutrophil count and percentage, white blood cell count, red blood cell count, platelet count, hematocrit, hemoglobin, mean corpuscular volume, mean corpuscular hemoglobin and mean corpuscular hemoglobin concentration, although not all of these parameters were available for each of the cohorts). For each *trans*-eQTL we compared the *trans*-eQTL effect-size (proportion explained variance, R^2) with the effect-size (R^2) of the correlation of the *trans*-gene expression with each of the 18 different blood count parameters. We observed that for the far majority (80.3%) of *trans*-eQTLs, the SNP effect was larger than any of the 18 blood cell-count parameters, which indicates that the majority of *trans*-eQTLs cannot be fully explained by differences in cellular composition (Supplementary Table 4). For example, for age, we observed that for 1,474 out of 1,513 *trans*-eQTL unique SNP-probe pairs, the *trans*-eQTL SNP explained a higher amount of gene

expression variance than age, which indicates that very few *trans*-eQTLs might in theory be fully explained by age.

To test this formally, we concentrated on the EGCUT cohort for which both age and each of the 18 blood count parameters were available. We first corrected the expression data for these 19 parameters (treating these as covariates), and subsequently tested whether some of the *trans*-eQTLs (at FDR < 0.05) became less significant after covariate correction. We did not observe this to be case: upon comparison of the *trans*-eQTL P-Value distributions, before and after correction for these 19 parameters, there was no evidence that associations became less significant (Wilcoxon P-Value = 0.99, Supplementary Figure 6).

Captions

Supplementary Figure 1 - *Cis*-eQTL characteristics

a) *Cis*-eQTL effect size is dependent upon the distance between the SNP position and the gene start. b) *Cis*-eQTL SNPs are generally located close to the *trans*cription start site: 97% of the detected *cis*-eQTL SNPs are within 250 kb of the TSS. c) *Cis*-eQTL probes are generally highly expressed with 84% of the top-2000 tested gene expression probes (ranked by average expression) having a *cis*-eQTL effect.

Supplementary Figure 2 - *Cis*-eQTL annotation

a) *Cis*-eQTL SNPs are significantly enriched for regulatory sequences such as miRNA binding sites, transcription factor binding sites, CpG-islands, nonsynonymous SNPs, splice enhancers and silencers, and 3'-UTRs. Enrichment and significance is with respect to top *cis*-eQTL SNPs for each probe in the permuted data, using Fisher's exact test. Data obtained from SNPInfo (FuncPred) and SNPNexus. c) *Cis*-eQTL SNPs with a large effect size show enrichment for enhancer sequences. Especially enhancer sequences in blood related cell-types (K562 and GM12878) are enriched for high effect *cis*-eQTL SNPs. Data obtained using HaploReg.

Supplementary Figure 3 - Comparison of *cis*-eQTL effect sizes between studies within the discovery meta-analysis

We performed a pair-wise comparison of Z-scores between studies in our discovery meta-analysis and with the discovery meta-analysis Z-score. We observed that *cis*-eQTL Z-scores are highly concordant between studies and the meta-analysis: on average, the majority (94%) of the SNP-probe pairs shared with the meta-analysis showed an identical direction of effect.

Supplementary Figure 4 - Position of detected *cis*- and *trans*-eQTLs

Cis- and *trans*-eQTL mapping in our current discovery meta-analysis revealed 8,228 and 1,513 eQTLs, respectively. Our current meta-analysis is capable of detecting many novel *trans*-eQTLs, compared to our previously published results. Sizes of dots reflect their respective significances.

Supplementary Figure 5 - Comparison of *trans*-eQTL effect sizes between cohorts within the meta-analysis

We performed a pair-wise comparison of Z-scores between studies in our discovery meta-analysis and with the meta-analysis Z-score. We observed that *trans*-eQTL Z-scores are highly concordant between studies and the meta-analysis: on average, the majority (90%) of the SNP-probe pairs shared with the meta-analysis showed an identical direction of effect.

Supplementary Figure 6 - Endophenotype correction in EGCUT

In order to determine the effect of age and differences in cellular composition of individuals on *trans*-eQTL effect size, we performed *trans*-eQTL mapping on the EGCUT dataset before and after adjusting for age and 18 different blood cell-count parameters. We observed that the effect of these different cell-count parameters and age on *trans*-eQTL effect size is minimal.

Supplementary Figure 7 - Detectability of previously reported *cis*- and *trans*-eQTLs in new meta-analysis

We compared previously identified *cis*- and *trans*-eQTLs (Fehrmann et al, PLoS Genetics, 2011) with our current meta-analysis. As the samples from the previous study are also part of this study, we redid the meta-analysis, excluding the samples the previous study (1,469 samples). We investigated all previously reported *cis*- and *trans*-eQTLs, comparing both the significance and allelic direction (Z-scores). The majority of previously reported *cis*- and *trans*-eQTL are also highly significant in the new meta-analysis. (Some previously reported eQTLs were not tested, as we used some additional, new quality control measures in the current meta-analysis).

Supplementary Figure 8 – Replication of *trans*-eQTLs in KORA F4 and BSGS

We attempted replication of our identified *trans*-eQTLs using peripheral blood samples from the KORA F4 and BSGS cohorts. We also performed a meta-analysis of both peripheral blood replication cohorts. We observed a high concordance in the direction of the *trans*-eQTL effect Z-scores, compared to our discovery meta-analysis.

Supplementary Figure 9 – Replication of *trans*-eQTLs in B-cells and monocytes

We attempted replication of our identified *trans*-eQTLs using B-cells and monocyte samples from the Oxford cohort. For the significant effects in both datasets, we observed a high concordance in the direction of the *trans*-eQTL effect Z-scores, compared to our discovery meta-analysis. However, some *trans*-eQTLs were significantly replicated in monocytes, but not in B-cells, indicating that some *trans*-eQTL SNPs may exert cell-type specific effects.

Supplementary Figure 10 – Replication of *trans*-eQTLs in lymphoblastoid cell lines

We attempted replication of our identified *trans*-eQTLs using HapMap3 lymphoblastoid cell lines (LCL). At an FDR < 0.05, we observed a high concordance in the direction of the *trans*-eQTL effect Z-scores, compared to our meta-analysis.

Supplementary Figure 11 – *Trans*-eQTL characteristics

a) *trans*-eQTL probes are generally highly expressed with 54% of the *trans*-eQTL probes residing in the top-4000 tested gene expression probes (ranked by average expression). b) The distribution of *trans*-eQTL effect sizes indicates that most *trans*-eQTLs have a small effect size.

Supplementary Figure 12 - Genome wide significant trait-associated SNPs are more often *trans*-eQTL SNPs

In order to determine whether genome-wide significant trait-associated SNPs are enriched for *trans*-eQTL effects, we first pruned the SNPs present in the Catalog of Published GWAS studies: the first set contained all SNPs (tested in our *trans*-eQTL meta-analysis) of the Catalog of Published GWAS studies, irrespective of GWAS association p-value. The second set contained all SNPs with a genome-wide significant association with complex traits ($P < 5 \times 10^{-8}$), and the third set all SNPs that were not genome-wide significant according to the Catalog of Published GWAS studies. We then performed *trans*-eQTL

analyses on each set (after pruning), using all nine cohorts, and observed that genome-wide significant trait-associated SNPs within the Catalog of Published GWAS studies are more often *trans*-eQTL SNPs, as compared to SNPs that are not genome-wide significant.

Supplementary Figure 13 - Trait-associated SNPs are enriched for *cis*- and *trans*-eQTLs

In order to determine whether trait-associated SNPs are enriched for eQTL effects, we first pruned the SNPs present in the GWAS catalog. We then randomly selected 100, equally sized, unlinked sets of SNPs, which were matched to the pruned GWAS catalog SNPs, for minor allele frequency and distance to genes. A) Using these sets of SNPs, we determined how many SNPs affected a gene in *cis*. We observed that trait-associated SNPs are enriched for *cis*-eQTL effects, as compared to randomly selected SNPs ($P = 1.4 \times 10^{-16}$). B) We performed *trans*-eQTL mapping on 3,122 individuals (EGCUT, Fehrmann HT12v3, Fehrmann H8v2 and the Rotterdam Study). We observed that trait-associated SNPs show a sixfold increase in the number of significant *trans*-eQTL SNPs ($P = 5.4 \times 10^{-49}$) as compared to random SNPs.

Supplementary Figure 14 – Differences in Z-scores when removing 35 or 45 PCs

To ensure that our choice for the removal of 40 PCs did not lead to overfitting, we repeated the discovery *trans*-eQTL meta-analysis, removing either 35 or 45 PCs in each of the cohorts. We observed that the majority of the *trans*-eQTLs are unaffected by the number of PCs chosen.

Supplementary Figure 15 - Effect of *cis*-eQTL regression on *trans*-eQTL effect size

In our final discovery meta-analysis, we corrected gene expression values for the presence of *cis*-eQTLs, in order to gain power to detect *trans*-eQTLs. We repeated the meta-analysis without removal of *cis*-eQTL effects, and observed that the majority of *trans*-eQTLs are not affected by *cis*-eQTL removal, although we observed a 12% increase in the number of *trans*-eQTL effects after correcting for *cis*-eQTLs.

Supplementary Figure 16 - FDR stability

In our meta-analysis, we used permutations to ascertain which p-value threshold corresponds to a false discovery rate of 0.05. However, this estimate may be dependent upon the number of permutations performed. Therefore, we have performed 20 meta-analyses, where we used an increasing amount of permutations to determine the FDR threshold. We observed that the FDR threshold estimate is quite stable after 5 permutations, with 1,513 significant *trans*-eQTL SNP-probe combinations, which is equal to the number found in our meta-analysis (using 10 permutations). Adding additional permutations does not greatly change the significance threshold.

Supplementary Table 1
Significant *cis*-eQTLs effects (top effect per probe), having an FDR < 0.5.

Supplementary Table 2
Significant *trans*-eQTL SNP-probe combinations, having an FDR < 0.5.

Supplementary Table 3
For some genes, the Illumina HT12v3 platform has multiple probes. This table lists the significant *trans*-eQTL SNP-probe pairs for such genes (FDR < 0.05).

Supplementary Table 4
Results of the endophenotype correlation meta-analysis, comparison to *trans*-eQTL effect size, and effect of endophenotype correction in EGCUT cohort.

Supplementary Table 5
Summary of replication results, listing p-values and Z-scores for *trans*-eQTL SNP-probe combinations that are significant in the meta-analysis (FDR < 0.05).

Supplementary Table 6
Results of the fine-mapping approach for each of the 1,340 *trans*-eQTL SNP-probe combinations (FDR < 0.05), showing the top *trans*-eQTL association per trait-associated SNP locus.

Supplementary Table 7
Correlations between *trans*-eQTL genes and several endophenotype measurements in the EGCUT and Rotterdam Study cohorts.

Supplementary Table 8
Convergent *cis*- and *trans*-eQTL effects per trait (FDR < 0.05), where two independent trait-associated SNPs are affecting the same gene.

References

1. Fehrmann, R.S.N. *et al.* Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet* **7**, e1002197 (2011).
2. Dubois, P.C. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* **42**, 295-302 (2010).
3. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
4. The International HapMap Project. *Nature* **426**, 789-96 (2003).
5. Volzke, H. *et al.* Cohort profile: the study of health in Pomerania. *Int J Epidemiol* **40**, 294-307 (2011).
6. Hofman, A. *et al.* The Rotterdam Study: 2012 objectives and design update. *Eur J Epidemiol* **26**, 657-86 (2011).
7. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-34 (2010).
8. Metspalu, A. The Estonian Genome Project. *Drug Development Research* **62**, 97-101 (2004).
9. Inouye, M. *et al.* An immune response network associated with blood lipid levels. *PLoS Genet* **6**(2010).
10. Tanaka, T. *et al.* Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. *PLoS Genet* **5**, e1000338 (2009).
11. Gibbs, J.R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* **6**, e1000952 (2010).
12. Heckbert, S.R. *et al.* Antihypertensive treatment with ACE inhibitors or beta-blockers and risk of incident atrial fibrillation in a general hypertensive population. *Am J Hypertens* **22**, 538-44 (2009).
13. Psaty, B.M. *et al.* The risk of myocardial infarction associated with antihypertensive drug therapies. *JAMA* **274**, 620-5 (1995).
14. Smith, N.L. *et al.* Esterified estrogens and conjugated equine estrogens and the risk of venous thrombosis. *JAMA* **292**, 1581-7 (2004).
15. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**, 629-44 (2006).

16. Rathmann, W. *et al.* Incidence of Type 2 diabetes in the elderly German population and the effect of clinical and lifestyle risk factors: KORA S4/F4 cohort study. *Diabet Med* **26**, 1212-9 (2009).
17. Mehta, D. *et al.* Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur J Hum Genet* (2012).
18. Marzi, C. *et al.* Genome-wide association study identifies two novel regions at 11p15.5-p13 and 1p31 with major impact on acute-phase serum amyloid A. *PLoS Genet* **6**, e1001213 (2010).
19. Fairfax, B.P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet* **44**, 502-10 (2012).
20. Powell, J.E. *et al.* The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. *PLoS One* **7**, e35430 (2012).
21. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97-101 (2002).
22. Stranger, B.E. *et al.* Patterns of *cis* regulatory variation in diverse human populations. *PLoS Genet* **8**, e1002639 (2012).
23. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).
24. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-7 (2009).
25. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).
26. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res* **30**, 38-41 (2002).
27. Franke, L. *et al.* Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. *Am J Hum Genet* **82**, 1316-33 (2008).
28. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
29. Breitling, R. *et al.* Genetical genomics: spotlight on QTL hotspots. *PLoS Genet* **4**, e1000232 (2008).
30. Westra, H.J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27**, 2104-11 (2011).
31. Alberts, R. *et al.* Sequence polymorphisms cause many false *cis* eQTLs. *PLoS One* **2**, e622 (2007).

32. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-11 (2001).
33. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
34. Patterson, K. 1000 genomes: a world of variation. *Circ Res* **108**, 534-6 (2011).
35. David, M., Dzamba, M., Lister, D., Ilie, L. & Brudno, M. SHRiMP2: sensitive yet practical SHort Read Mapping. *Bioinformatics* **27**, 1011-2 (2011).
36. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-4 (2012).
37. Xu, Z. & Taylor, J.A. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res* **37**, W600-5 (2009).
38. Chelala, C., Khan, A. & Lemoine, N.R. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics* **25**, 655-61 (2009).
39. Dayem Ullah, A.Z., Lemoine, N.R. & Chelala, C. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic Acids Res* **40**, W65-70 (2012).
40. Siva, N. 1000 Genomes project. *Nat Biotechnol* **26**, 256 (2008).
41. Schurmann, C. *et al.* Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. *PLoS One* **7**, e50938 (2012).
42. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
43. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423-8 (2008).
44. Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet* **8**, e1002431 (2012).
45. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat Genet* **39**, 1217-24 (2007).
46. Nica, A.C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet* **7**, e1002003 (2011).
47. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338-45 (2005).

48. Baek, D. *et al.* The impact of microRNAs on protein output. *Nature* **455**, 64-71 (2008).
49. Berland, R. & Wortis, H.H. Origins and functions of B-1 cells with notes on the role of CD5. *Annu Rev Immunol* **20**, 253-300 (2002).