Transcriptome and genome sequencing uncovers functional variation in human populations

Tuuli Lappalainen et al.

Supplementary Material

Index

Supplementary Note

Principal Investigators of the Geuvadis consortium:

Xavier Estivill, Roderic Guigo (Centre for Genomic Regulation, Spain); Emmanouil Dermitzakis, Stylianos Antonarakis (University of Geneva, Switzerland); Thomas Meitinger, Tim M Strom (Helmholtz Zentrum München, Germany), Aarno Palotie (Wellcome Trust Sanger Institute, UK); Jean François Deleuze (Centre National de la Recherche Génomique, France); Ralf Sudbrak, Hans Lerach (Max Planck Institute for Molecular Genetics, Berlin, Germany), Ivo Gut (Centro Nacional d'Anàlisi Genòmica, Spain): Ann-Christine Syvänen, Ulf Gyllensten (Uppsala University, Sweden); Stefan Schreiber, Philip Rosenstiel (Institute of Clinical Molecular Biology, Christian-Albrechts University of Kiel, Germany); Han Brunner, Joris Veltman (Radboud University Nijmegen Medical Centre, the Netherlands); Peter A.C.T Hoen, Gert Jan van Ommen (Leiden University Medical Center, the Netherlands); Angel Carracedo (Universidad de Santiago de Compostela, Spain); Alvis Brazma, Paul Flicek (European Bioinformatics Institute, EMBL-EBI, UK); Anne Cambon-Thomsen (Institut National de la Santé et de la Recherche Médicale (INSERM); Jonathan Mangion (Life Technologies, Germany); David Bentley (Illumina Cambridge Limited, UK); Ada Hamosh (Johns Hopkins University School of Medicine, USA)

Supplementary Figures

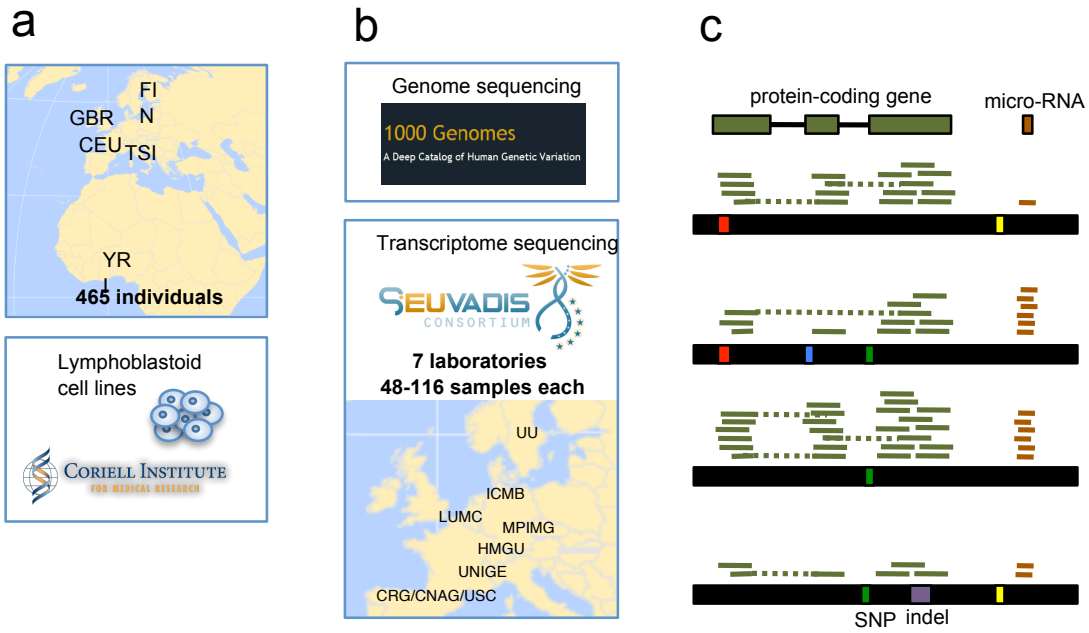

Figure S1: Study design

An illustration of the study design shows the studied populations and samples (a) from which the 1000 Genomes Consortium created genome sequencing and genotype data, and we sequenced mRNA and small RNA in seven European laboratories (b), with the final data set consisting of genotype and RNA-sequencing data from 462 and 452 individuals for mRNA and small RNA, respectively (c).

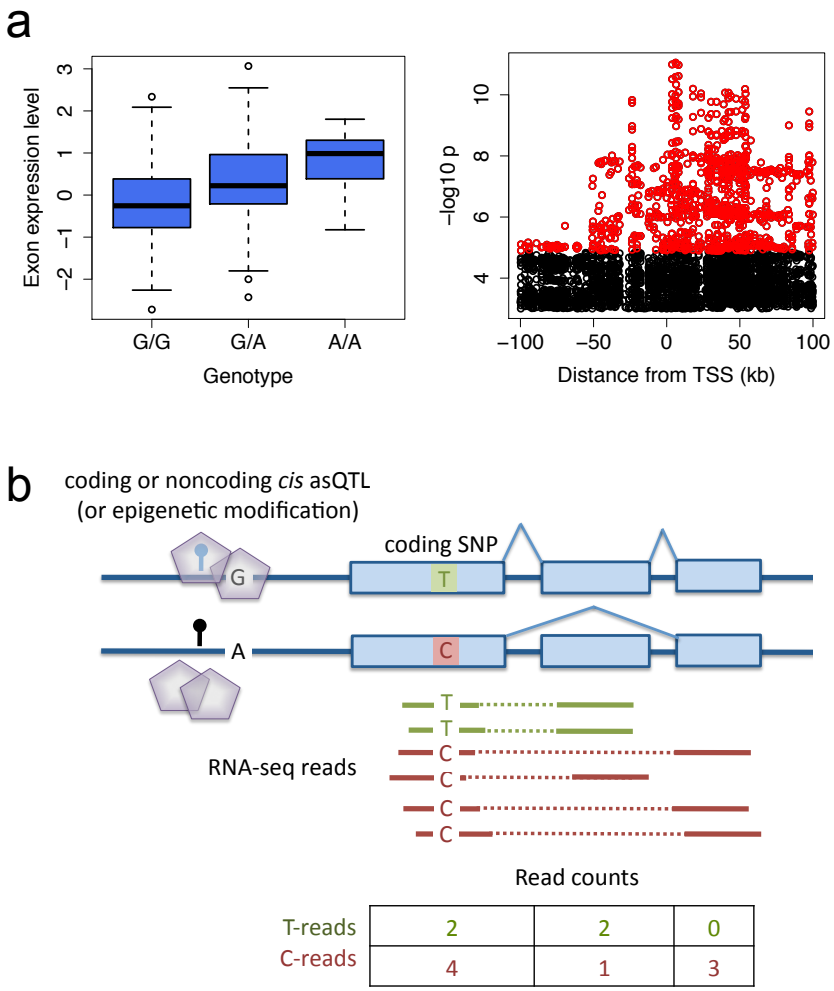

Figure S2: Analysis of regulatory genetic variation

Analytical approaches for studying genetic effects on transcription from RNA-sequencing data: a) transcriptome QTL analysis, where the aim is to find genetic variants that associate to a transcriptome quantitative trait (such as gene expression level) in a population sample. First, association between genotypes and transcriptome quantitative traits is calculated for all variant – transcript feature pairs usually in a genomic window (left panel, where each data point is an individual), and the resulting p-values for each gene can be plotted as a landscape of associations in the reagion flanking the gene (right panel; each data point is a genetic variant with the genome-wide significant associations in red). (b) illustrates allele-specific transcription analysis that aims to identify differences in transcription between the two haplotypes of an individual. In allele-specific expression (ASE) analysis, we search for differences in the ratio of the two alleles, here comparing if the total count of T and C alleles is different from 50-50. In allele-specific transcript structure (ASTS) analysis we ask if the distribution of T- and C-carrying reads or their mates is different across exons (here a 2x4 table).

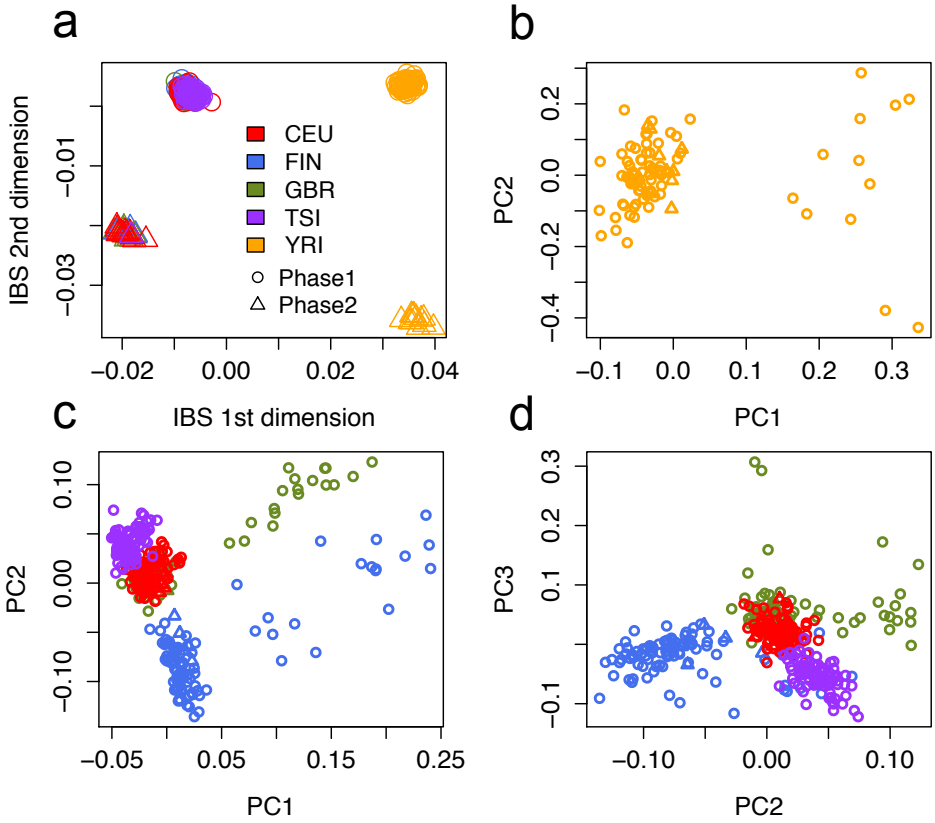

Figure S3. Genotype data quality control

Multidimensional scaling plot of identity-by-state matrix of all the samples shows a clear clustering not only by continent, but also by whether the sample had full genome data or was imputed (a). Principal component analysis within Yoruba (b) and within Europe (c,d) shows population structure especially within Europe. Based on these results, the imputation status and PCs 1-3 for Europeans and PCs 1-2 for Yoruba were included as covariates in the QTL analyses.

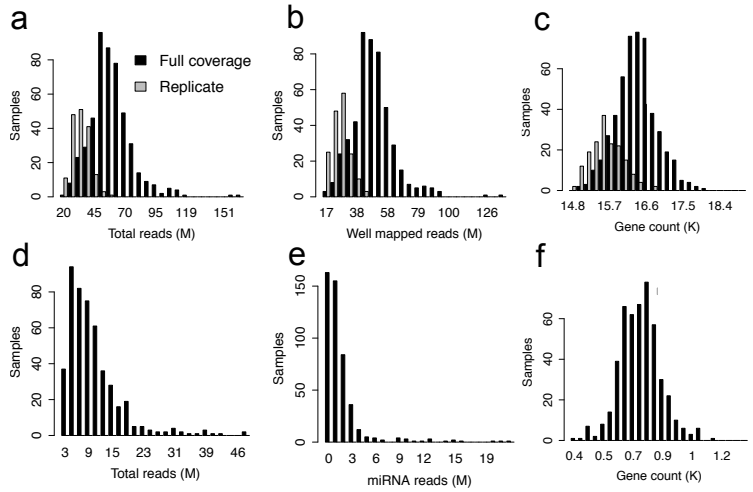

Figure S4. Read and gene count distributions

mRNA statistics per sample of total read counts (a) mapped read counts (MAPQ>150, properly paired, NM<=6) (b), and gene counts (>1 RPKM) (c), and small RNA statistics of total read counts (d) miRNA read counts (e), and miRNA gene (>0) counts (f). These distributions have few outliers, especially in gene counts, demonstrating the uniformity of the raw data. The mRNA replicate samples refer to 168 low-coverage replicate samples that were not used in the analysis.

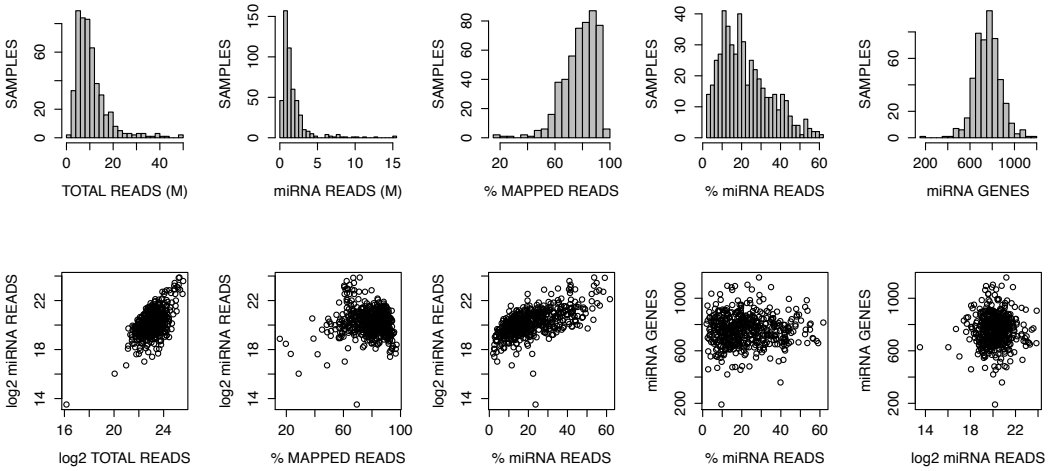

Figure S5. miRNA quality control

Combination of various quality control statistics of small RNA sequencing (total read count, proportion of mapped reads) and miRNA quantification (miRNA read count, proportion of miRNA reads, and number of quantified miRNAs). These plots demonstrate that except for 13 outliers that were excluded from the final data set, the final quantification data is very uniform.

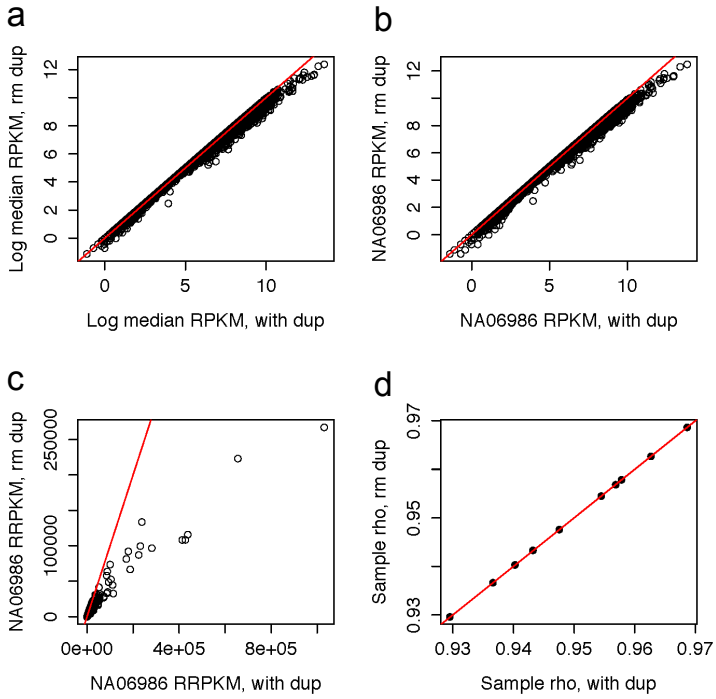

Figure S6. Read duplicates

mRNA exon quantifications with (x-axis) and without (y-axis) read duplicates (RNA-seq reads with identical coordinates) as median of five samples on a log scale (a), in one sample on a log scale (b) and as raw RPKM (c), and Spearman rank correlation between quantifications between all sample pairs of five individuals (d). Other samples look similar to NA06986 shown here. The results show a very similar rank order of quantifications between individuals regardless of if duplicates are removed or not; however, especially (c) demonstrates how highly expressed exons are more affected by duplicate removal. These patterns suggest that the majority of duplicate reads in a high-quality mRNA experiment are due to saturation of the read mapping space driven by real biology of high expression levels, rather than technical artifacts, and that removing duplicates in mRNA-seq data would lead to underestimation of expression levels of highly expressed genes, and to a saturation point where individual variation can no longer be observed. In the analysis of this study, duplicate reads where included in all analyses.

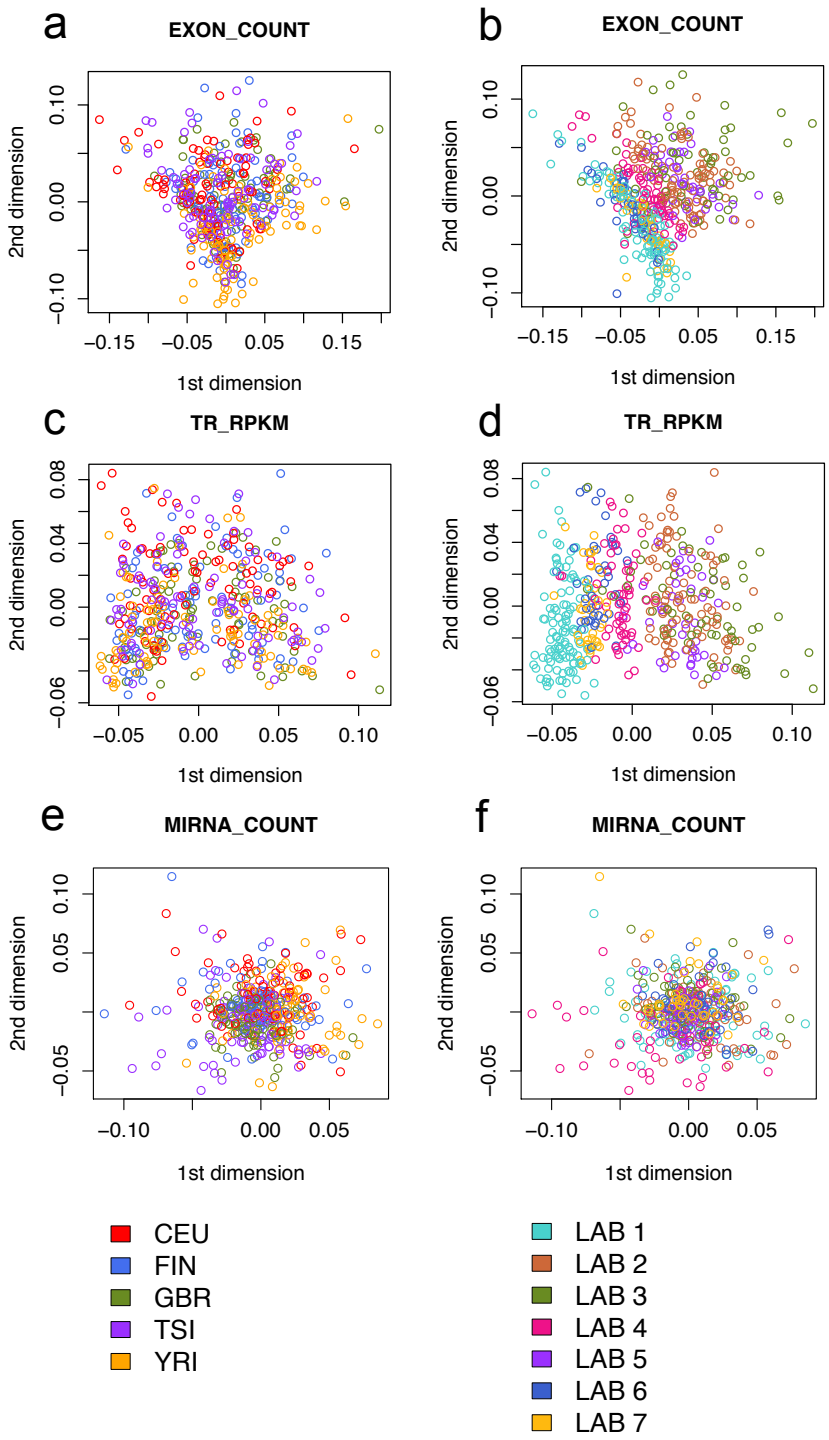

Figure S7. Sample clustering

Multidimensional scaling of pairwise sample correlations based on exon (a, b), transcript (c,d) and miRNA (e,f) quantifications normalized only for the total number of mapped reads. The same data is shown colored by population (a, c, e) and by sequencing laboratory (b, d, f).

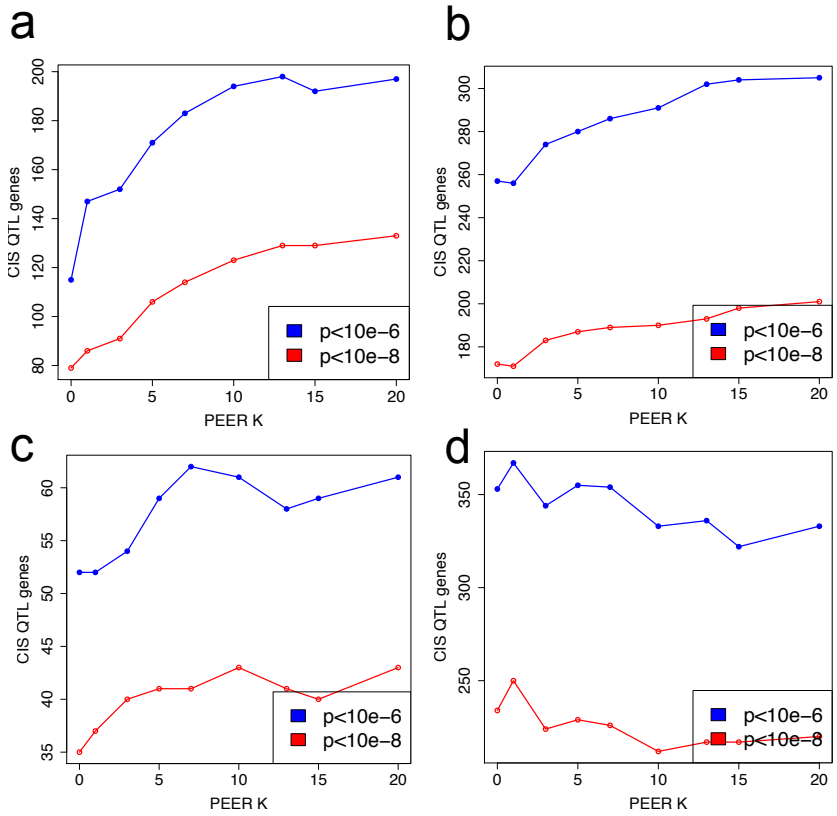

Figure S8. PEER covariate analysis

The number of cis-eQTLs in a small test data set was used to evaluate the performance of PEER normalization for all quantifications; here it is shown as a function of the number of corrected covariates for mRNA (a), transcript (b), miRNA (c) and repeat (d) quantifications. For the final analysis, the data was normalized with K=10 except for repeat quantifications for which PEER normalization was not done.

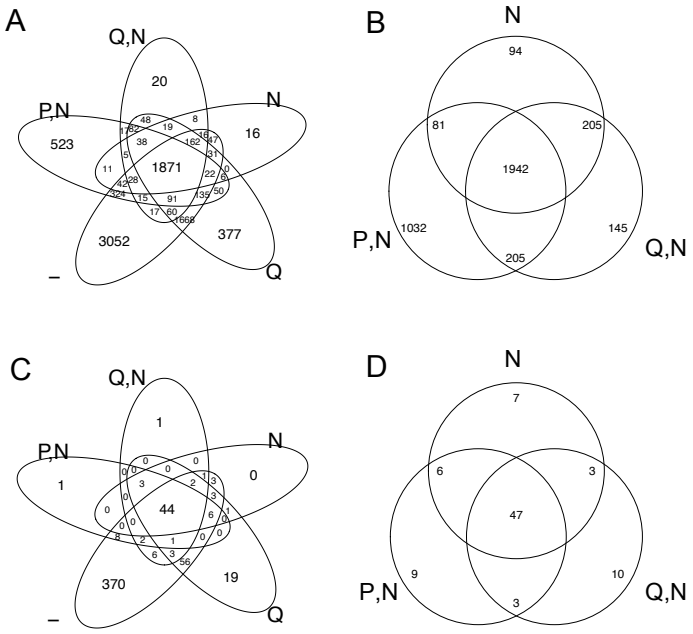

Figure S9. eQTL discovery with different normalizations

Number of discovered eQTL genes (a,b) and miRNAs (c,d) with linear regression in Europeans after different normalizations of the data set. B) and D) are subsets of the categories in A) and C). The abbreviations are: P,N (PEER + standard normal transformation, as in the main analysis in this paper), Q,N (quantile normalization + standard normal transformation), N (standard normal transformation on library size corrected raw quantifications), Q (quantile normalization on library size corrected raw quantifications), - (no normalization except for library size). PEER and quantile normalization scale the distributions of different samples, and standard normal transformation is applied to each gene/miRNA across samples. Not using standard normal transformation violates the assumptions of linear regression, and leads to an extremely high number of likely false eQTLs likely due to outliers that often occur in RNA-seq data (much more than in expression array data due to a wider dynamic range). In b) and d), i.e. only in categories where the values have been appropriately transformed to standard normal, PEER normalization yields the highest number of eQTL discoveries. Of all gene and miRNA eQTLs, 83% and 76%, respectively, of the union of all three categories are discovered by PEER, indicating high reproducibility (see also Fig. S8, and 't Hoen et al. submitted). In miRNA analysis the advantage of PEER correction is less pronounced due to the smaller amount of expression data that is used to calculate the factor for correction.

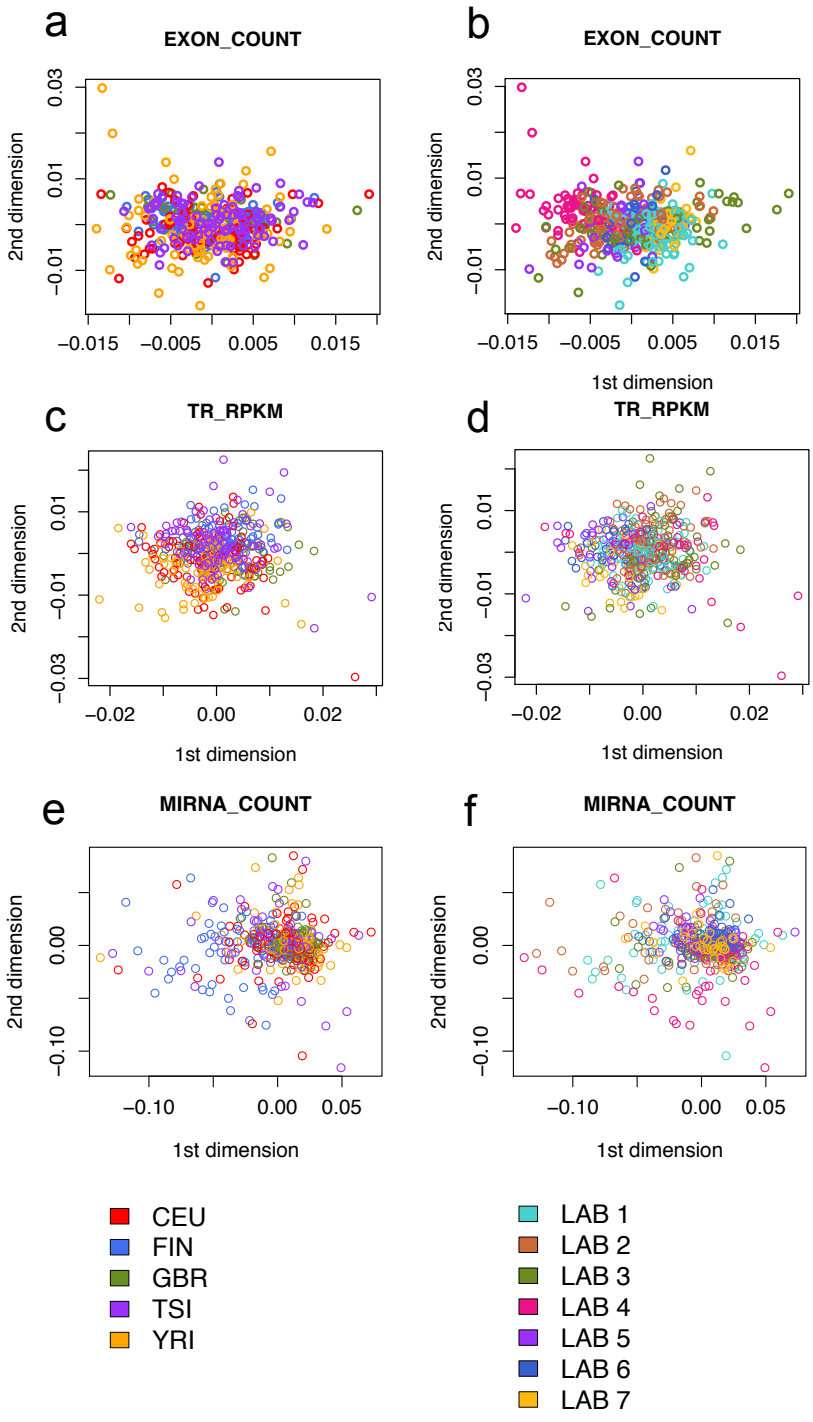

Figure S10. Sample clustering after normalization

Multidimensional scaling of pairwise sample correlations based on exon (a, b), transcript (c,d) and miRNA (e,f) quantifications after PEER normalization. The same data is shown colored by population (a, c, e) and by sequencing laboratory (b, d, f).

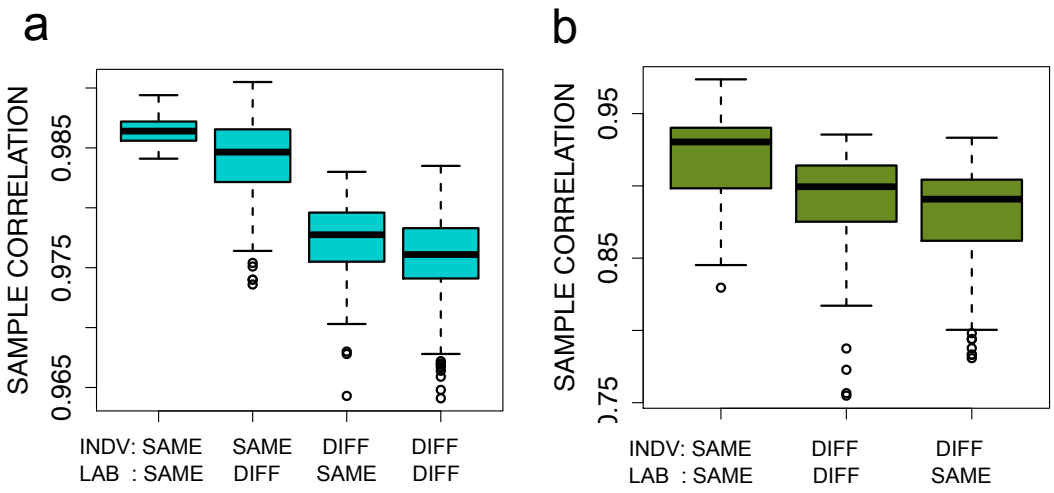

Figure S11. Replicate correlation after normalization

Correlation of the five replicate samples based on mRNA exon (a) and miRNA (b) quantifications after PEER normalization, partitioned by lab and individual. In mRNA sequencing, the same samples were sequenced twice in one lab. See Fig. 1a for similar analysis before normalization. While the normalization does not eliminate all lab effects, the lab effects are considerably smaller than biological differences between individuals.

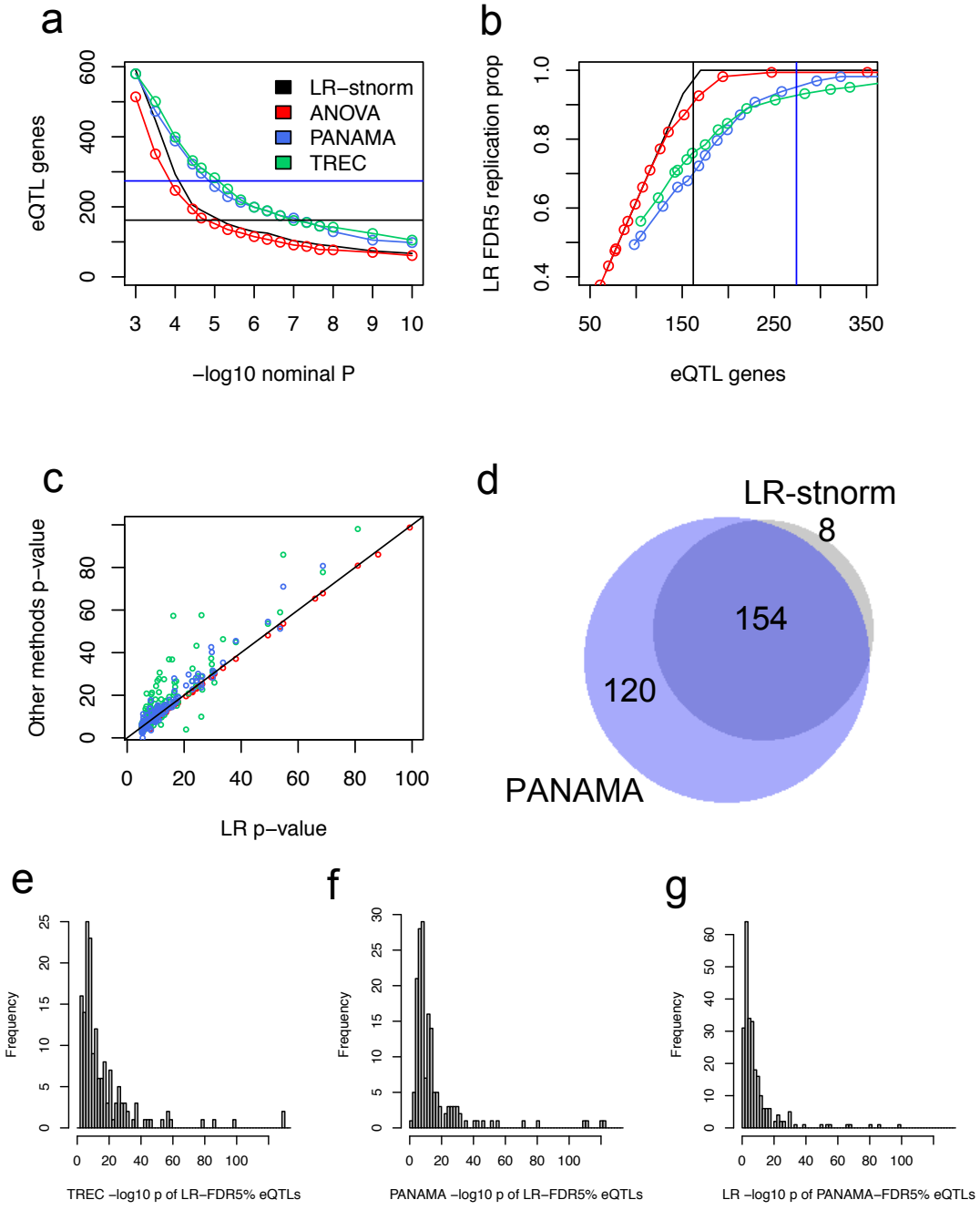

Figure S12. Comparison of eQTL methods

We compared eQTL discovery with linear regression (LR+stnorm) that was used in the main analysis of this paper to three alternative eQTL methods: ANOVA, TReC that is based on RNA-seq count data, and PANAMA that jointly detects and corrects for hidden covariates in eQTL analysis (see Supplementary Methods section 11.1). The analysis was run for 645 gene quantifications of chr7 of European samples. For LR and ANOVA, the gene RPKM values were normalized by PEER and transformed to standard normal distribution. TReC was run on read counts with total number of reads per sample and PEER factors as

covariates, and PANAMA was ran after Anscombe transformation with population label and sequencing lab as additional covariates (see Supplementary Methods for additional details). For linear regression and PANAMA, the significant eQTLs at 5% FDR are shown by vertical and horizontal lines. A) shows the number of genes below different nominal p-value thresholds, b) shows the proportion of LR eQTLs at FDR 5% (used in the main analysis) that are replicated with each method as a function of different significance thresholds that yield different number of eQTL genes. The p-values of the LR FDR5% significant genes in other methods are plotted in (c), showing a high concordance in p-values, especially between LR and ANOVA. Comparison of the significant eQTLs at FDR 5% in PANAMA and LR in (d) shows that almost all LR eQTLs are replicated by PANAMA that finds a substantial number of additional eQTLs. The –log10 p-value distributions of LR FDR5% significant eQTLs in TReC (e) and PANAMA (f), and PANAMA FDR5% eQTLs in LR (g) show that these eQTLs tend to have good p-values also with alternative methods. In fact, the p-value distribution is so heavily skewed towards low values that calculating pi1 was not possible. Altogether, these results indicate a high replication rate of our linear regression eQTLs with different statistical methods. ANOVA results are nearly identical, TREC also reproduces the eQTL signals well despite using the data in a very different manner. PANAMA replicates nearly all LR eQTLs and finds a substantial number of new ones, likely due to more efficient control of confounding factors. PANAMA is an attractive eQTL method for future use; however, since it has not yet been tested and validated extensively by several earlier studies, in this paper we preferred to use the more established linear regression method that we here show to be very robust.

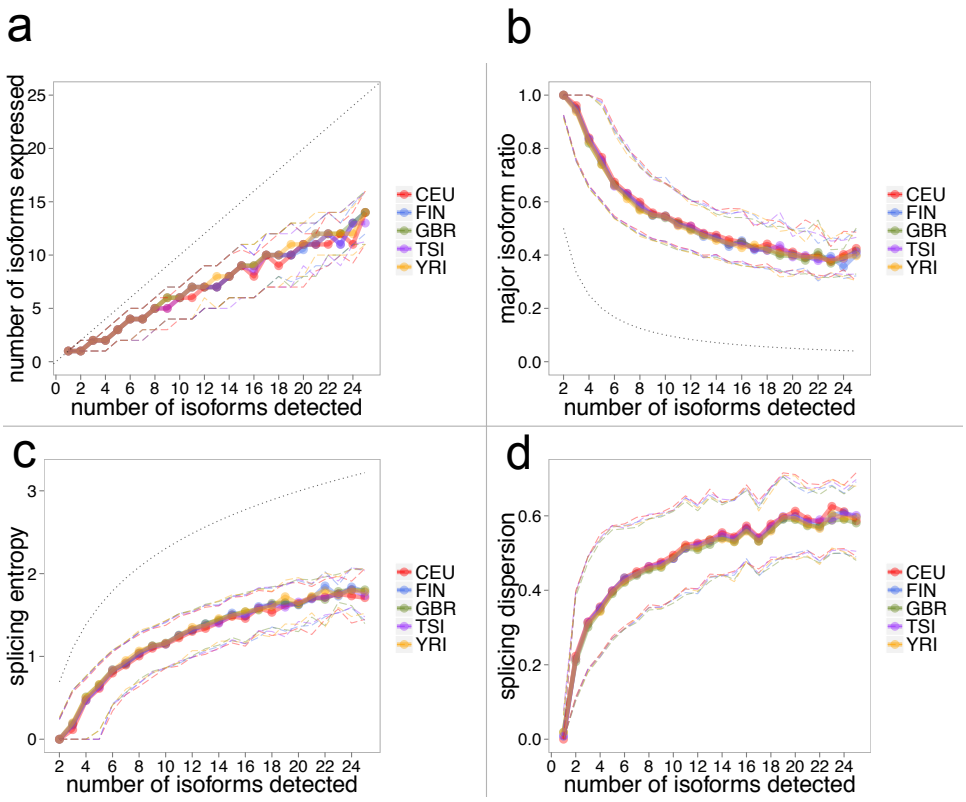

Figure S13. Transcript quantification statistics

Various transcript quantification metrics were computed at gene-level to evaluate the quality of the quantifications. The plot is based on values calculated for all individuals, plotted as population medians (points) and first/third quartiles (dashed lines). We show the number of transcripts expressed (>0.1 RMPK) in each population (a), ratio of the major transcript of the total per gene (b), splicing entropy, where one transcript expressed would result in low entropy and all transcripts expressed equally would give a high value (c), and splicing dispersion which represents the variability in the space of the splicing ratios. (d). The genes are grouped according to the number of different transcripts detected when pooling all the samples together (x-axis), as the theoretical boundaries of the metrics depend on the number of transcripts observed (shown as lines in (a) and (c).

The general distribution of these metrics indicates good quality transcript quantifications with expected patterns – not all possible transcripts should be detected in a sample derived from one individual and one cell line, and having one transcript per gene with higher expression levels than others rather than equal quantifications of all is also a pattern that is believed to reflect the underlying biology of transcription. These basic metrics are also highly consistent across the populations.

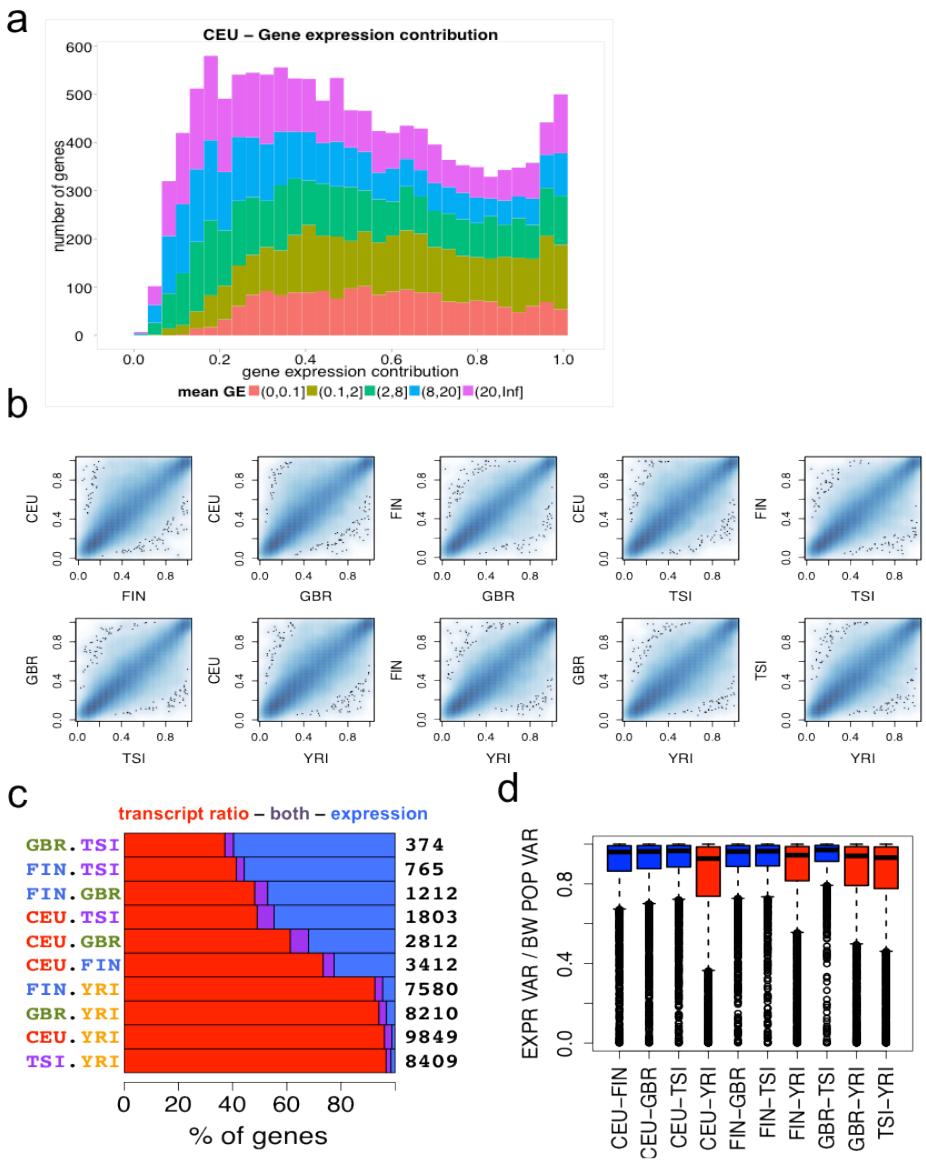

Figure S14. Transcript variation between population pairs

For each gene, we can calculate the proportion of expression level variation (as opposed to splicing) of the total expression variation between individuals in each population (see Supplementary Methods). (A) shows the general distribution of this statistic (similarly to Fig. 1b), here only for CEU but separating genes by expression level. This shows that lowly expressed genes where the statistic might be noisier do not cause major shifts in the distribution. Other populations show a similar pattern (data not shown). In (b), we compare the proportion of expression variation per gene between all population pairs. The consistency between populations indicates that each gene has a characteristic pattern of how much expression versus splicing variation it allows.

Furthermore, for each gene, a small proportion of total gene expression variation is explained by difference between population pairs, and we characterized these differences using several methods. (C) shows the proportion

of differentially expressed and differentially spliced genes between population pairs using the DEXSeq method. Additionally, in (d), we calculated how much of between-population variation is explained by variation in expression levels, and show this distribution for population pairs for genes with high level of population differentiation (between-population variation >2.5% of total). We observe that differences between the African YRI and European populations have less contribution of differential expression levels than European pairs, suggesting a bigger contribution of splicing variation between continental populations.

Together with the results from differential expression/transcript usage analysis (Fig. 1c), all three analysis methods indicate higher contribution of splicing differences in between-continent transcript variation. Cell line age may contribute to the variation seen between European population pairs (Ferreira et al. submitted), but since the population with a distinct pattern is YRI rather than the oldest CEU cell lines, and the YRI-EUR comparisons are all very similar, it appears unlikely that the trend would be solely due to cell line batch effects.

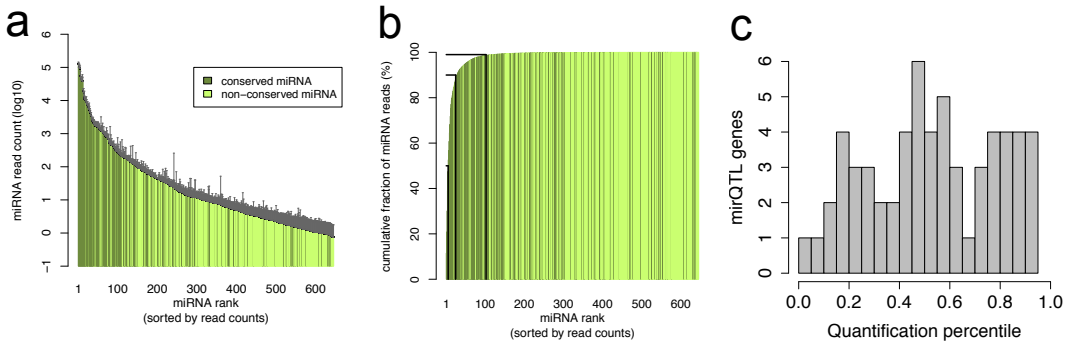

Figure S15. miRNA quantification statistics

Expression of 644 autosomal miRNAs detected in 452 individuals is shown in (a) with mean and s.d. of normalized read counts. (b) shows the cumulative fraction of miRNA reads explained by miRNAs of decreasing abundance. The black lines indicate the 50% fraction (explained by the 6 most abundant miRNAs), 90% fraction (explained by the 29 most abundant miRNAs) and the 99% fraction (explained by the 122 most abundant out of 644 total miRNAs). These plots indicate that the highest expressed miRNAs account for the vast majority of the total miRNA pool of the cell. (c) shows the quantification distribution of miRNAs with mirQTLs, which are found relatively evenly in lowly and highly expressed miRNAs.

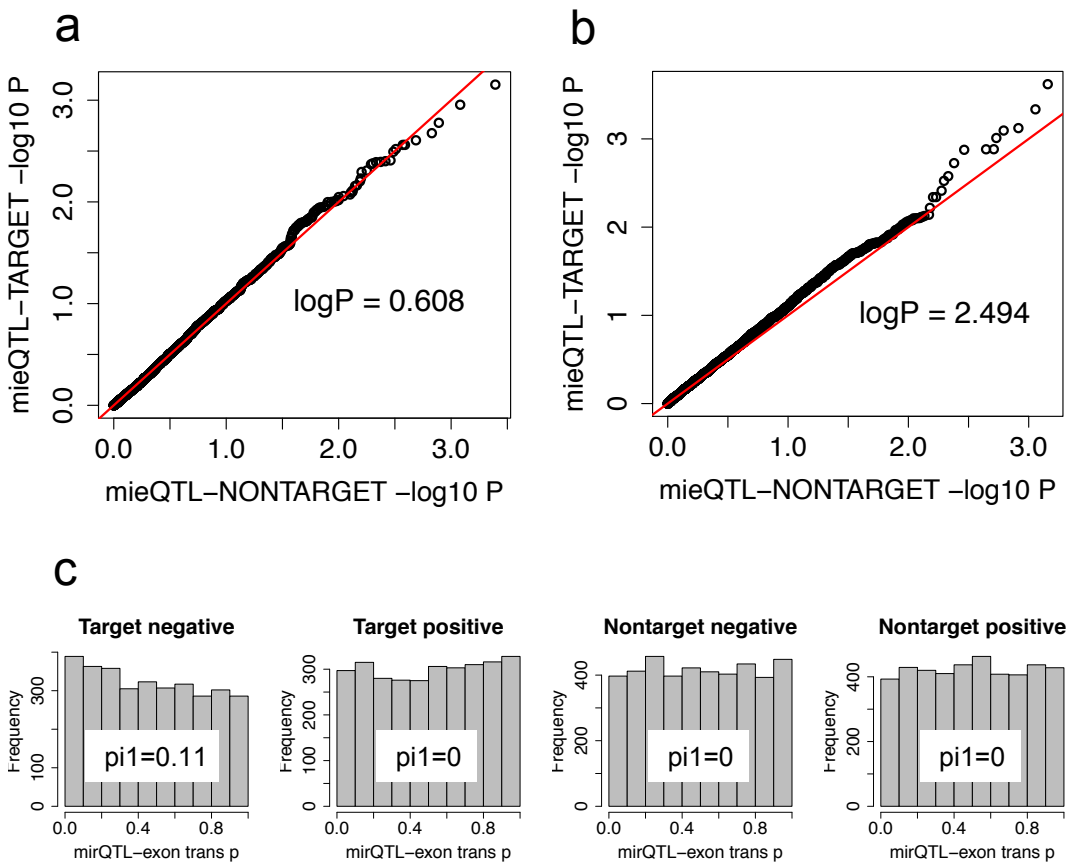

Figure S16. Trans-effects of mirQTLs

Variants that associate to miRNA expression levels (mirQTLs) can potentially be trans-eQTLs for the target genes of these miRNAs. We sought for this effect in (a,b) by comparing trans-eQTL p-value distributions of cis-mirQTLs to the predicted target exons of the affected miRNA (y-axis) to a null distribution to non-target exons. This comparison was done separately for positive (a) and negative associations (b), i.e. those where the cis-mirQTL allele increasing the miRNA expression has positive or negative correlation to the exon, respectively. The p-values are from a KS test, and indicate a small excess of low p-values for negative associations. Furthermore, (c) shows histograms of the p-value distributions, with the skew towards low p-values quantified by the pi1 statistic (1-pi0), again indicating enrichment of low p-values only in negative associations with real targets. The slightly lower p-values for negative associations makes biological sense, since miRNAs are usually believed to downregulate their targets. While we do not find a long tail of p-values indicating cis-mirQTLs being highly significant trans-eQTLs, the overall shift of the p-value distribution can be a sign of small effects of genetic variants affecting miRNA levels leading to downstream effects on the miRNA targets.

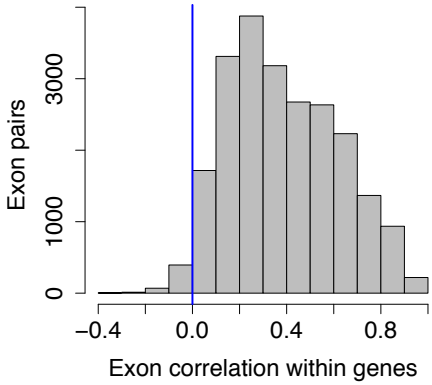

Figure S17. Coexpression of exons of the same gene

Correlation between quantifications of exons from the same gene for chr20 in the European data set. For many exon pairs, the correlation is not very high, indicating frequent splicing variation within genes, consistently with the large number of independent eQTL signals for different exons of the same gene.

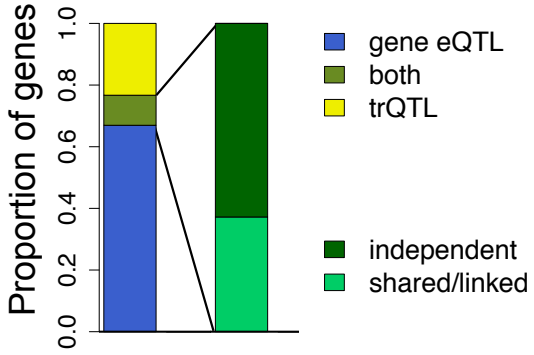

Figure S18. eQTL-trQTL sharing

We analyzed the overlap of gene eQTLs affecting total gene expression levels, and trQTLs affecting transcript ratios. The barplot shows the proportion of genes with each type of QTL (left), and of the genes with both gene eQTL and trQTLs, how many are driven by independent genetic effects, versus same or linked causal variants (right). The sharing is measured by correcting for the best trQTL variant in eQTL analysis, and observing if the significant gene eQTL signal remains. This analysis is possible due to the measurement of gene expression levels and transcript ratios being independent – a biological change in only one does not change the statistical measure of the other, thus any correlation is likely to be biological rather than technical. For exon eQTLs this would not true, since they can be affected both by changes in splicing and expression.

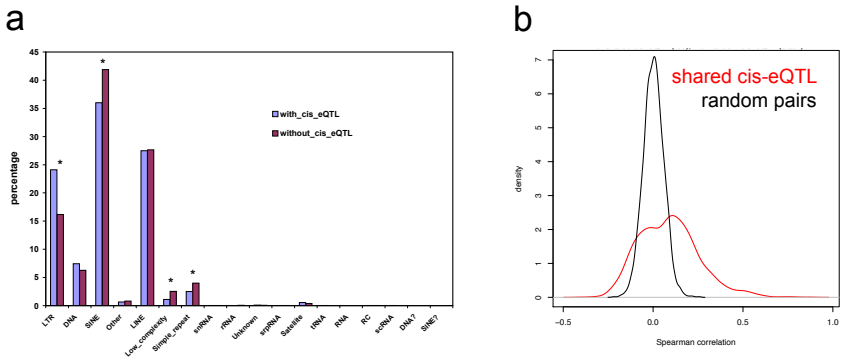

Figure S19. Transcribed repeat eQTLs

Out of the 43,875 repeat elements that were expressed in at least 50% of the samples, 5,763 repeat elements had significant cis-eQTLs in EUR. The class of LTR (long terminal repeat retrotransposons) was overrepresented among the repeats with significant cis-eQTLs (Fisher P<0.001; a). (b) Distributions of correlation of quantifications between repeats and exons for pairs that share an eQTL or for random pairs. This analysis was done after summing up quantifications of flanking repeats with a shared eQTL. This suggests that transcribed repeats can be controlled by the same regulatory machinery as nearby genes and affected by same regulatory variants (KS $p<2.2 \times 10^{-16}$)

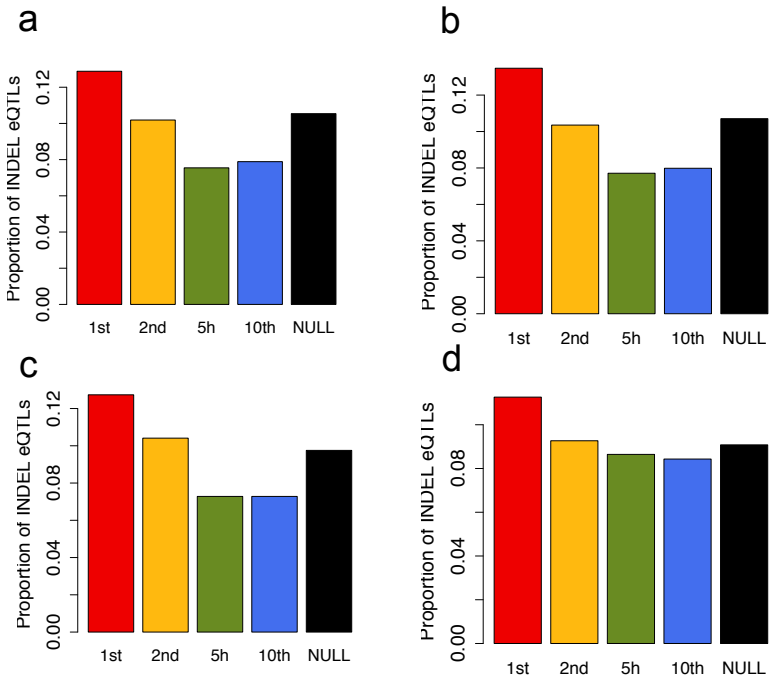

Figure S20. Indel enrichment in eQTL variants

We calculated the proportion of indels among the 1st, 2nd , 5th and 10th best QTL variants and in the matched null for eQTLs in EUR, p= 0.00019 (a), eQTLs in EUR using only noncoding variants, p= 2.707e-05 (b), trQTLs in EUR, p=0.17 (c), eQTLs in YRI , p= 0.0019 (d). We see a clear significant overrepresentation of indels in eQTL and trQTL variants. We analyzed only noncoding-variants in (b) to confirm that the enrichment is not driven by mapping bias. The higher number of mismatches in mapping of indel reads could potentially lead to allele-specific bias in quantifications, and a false eQTL signal for the variant causing the mapping bias. However, the similar result for only noncoding variants implies that this is unlikely, since RNA-seq reads rarely map to noncoding regions and eQTLs there should not be affected by this bias.

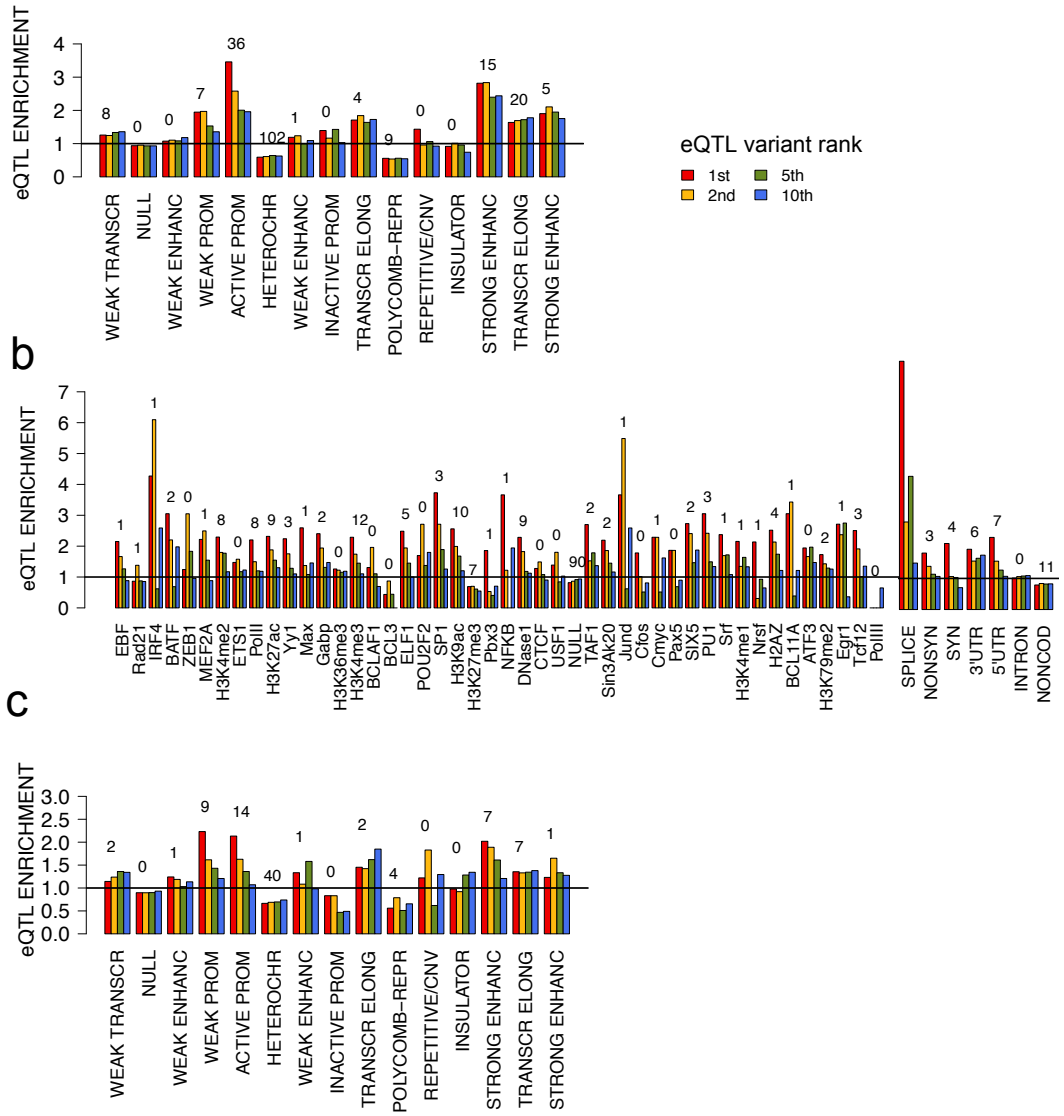

Figure S21. Functional annotation of eQTLs

Enrichment of eQTL variants in functional annotations relative to a matched null distribution for the 1st (most significant) eQTL variant as well as 2nd, 5th and 10th best variants for EUR eQTLs in chromatin states (a), and YRI eQTLs in Ensembl Regulatory Build and coding annotations (b) and in chromatin states (c). See Figure 2a for the figure of EUR eQTLs corresponding to (b) and (c). The numbers above the bars denote –log10 p-values of a Fisher test between 1st eQTL variants and the null for each category. There is an overall high enrichment of eQTLs in functional elements, especially for the 1st variant.

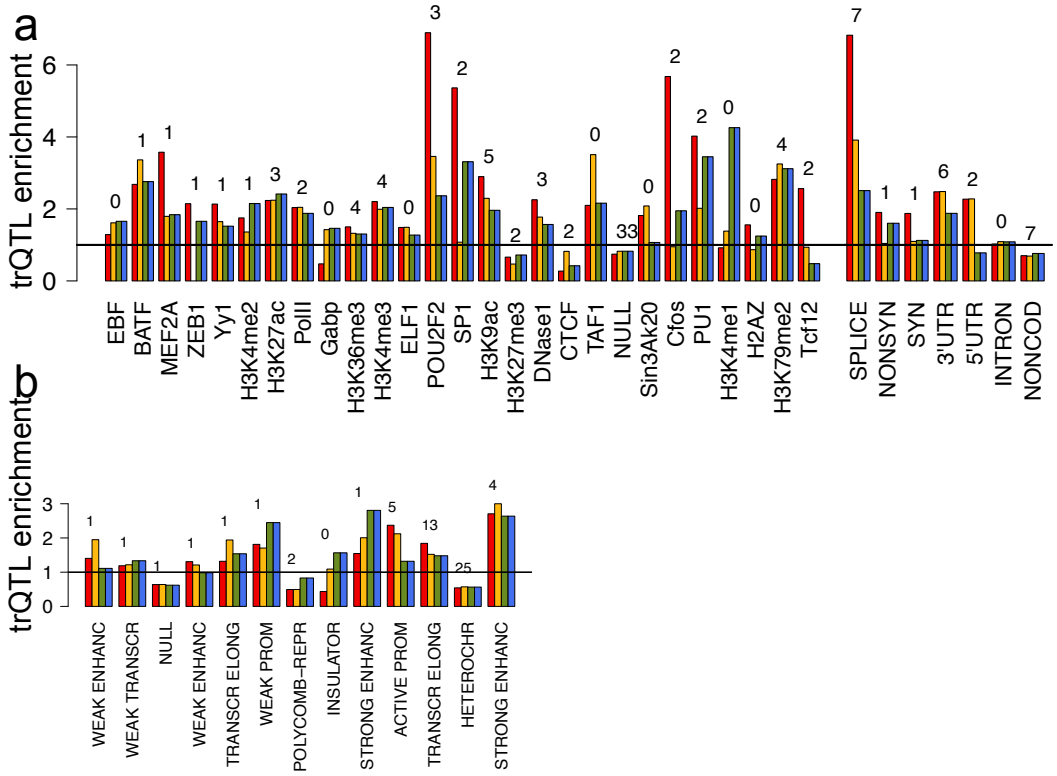

Figure S22. Functional annotation of trQTLs

Enrichment of EUR trQTL variants in functional annotations relative to a matched null distribution for the 1st (most significant) trQTL variant as well as 2nd, 5th and 10th best variants in the Ensembl Regulatory Build and coding annotations (a) and in chromatin states (b). The numbers above the bars denote –log10 p-values of a Fisher test between the 1st the best eQTL and the null for each category. Several annotations are significantly enriched for trQTLs. This analysis is not shown for YRI due to the low number of variants.

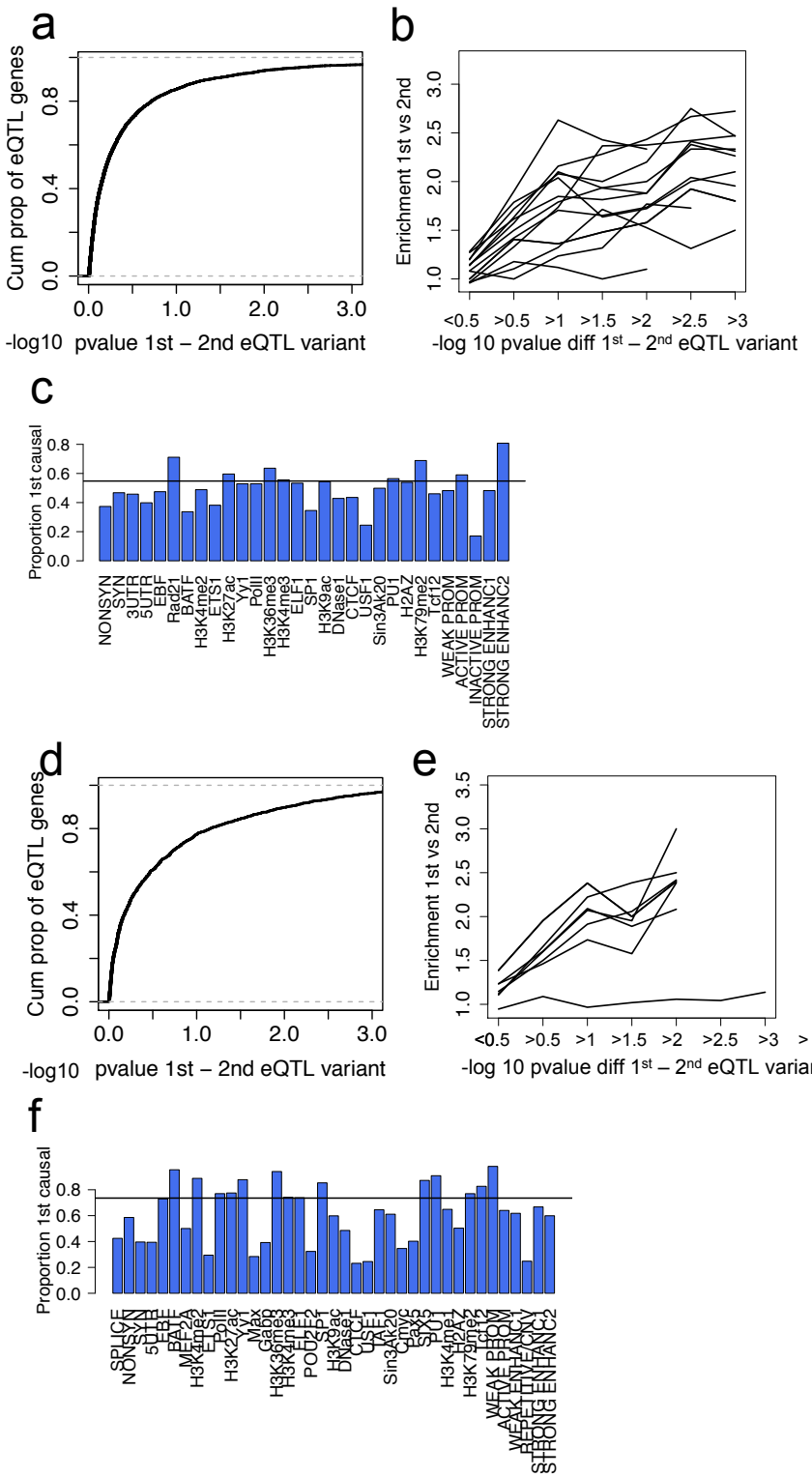
a
Cum prop of eQTL genes
0.0
0.4
0.8
0.0
1.0
2.0
3.0
-log10 pvalue 1st – 2nd eQTL variant
b
Enrichment 1st vs 2nd
1.0
1.5
2.0
2.5
3.0
<0.5
>0.5
>1
>1.5
>2
>2.5
>3
-log 10 pvalue diff 1st – 2nd eQTL variant
c
Proportion 1st causal
0.0
0.2
0.4
0.6
0.8
NONSYN
SYN
3UTR
5UTR
EBF
Rad21
BATF
H3K4me2
ETS1
H3K27ac
Yy1
PolII
H3K36me3
H3K4me3
ELF1
SP1
H3K9ac
DNase1
CTCF
USF1
Sin3Ak20
PU1
H2AZ
H3K79me2
Tcf12
WEAK PROM
ACTIVE PROM
INACTIVE PROM
STRONG ENHANC1
STRONG ENHANC2
d
Cum prop of eQTL genes
0.0
0.4
0.8
0.0
1.0
2.0
3.0
-log10 pvalue 1st – 2nd eQTL variant
e
Enrichment 1st vs 2nd
1.0
1.5
2.0
2.5
3.0
3.5
<0.5
>0.5
>1
>1.5
>2
>2.5
>3
>
-log 10 pvalue diff 1st – 2nd eQTL variant
f
Proportion 1st causal
0.0
0.2
0.4
0.6
0.8

Figure S23. Causal eQTL variants

(From the previous page:) We estimated the probability of the best eQTL variant being the causal regulatory variant by comparing the annotation enrichment (relative to the null) of the best eQTLs variants of all loci to that on loci where the p-value distribution indicates that the best eQTL variant is very likely to be causal. (a-c) show analysis of EUR eQTLs, and the corresponding figures for YRI eQTLs are in panels (d-e). In the majority of eQTLs the p-value difference between the 1st and the 2nd variant (Δp) is small (a,d) due to strong LD between the variants, however, there are also large numbers of eQTLs where the 1st variant association is orders of magnitude more significant than for the 2nd variant, and in such cases the first variant is very likely to be the causal one. We calculated the annotation enrichment of the 1st variant relative to the 2nd variant for different classes of Δp (b,e), and based on the modest increase from 1-1.5 log10 Δp onwards, we chose log10 Δp >1.5 as the main limit above which we can assume that the first variant is causal, and Δp >2.5 as a more conservative estimate. Then, for these causal variants, we calculated the enrichment of the best variants relative to the null, and comparing the same enrichment of all the variants to this number (c,f) gave us an estimate of the proportion of all variants where the 1st variant is causal, with the weighted median (horizontal line) based on the proportion and frequency of each annotation class. Using a log10 Δp >2.5 as a more conservative estimate gave us proportions of the first variant being causal in 34% for EUR and 41% YRI (plots similar to c and f not shown).

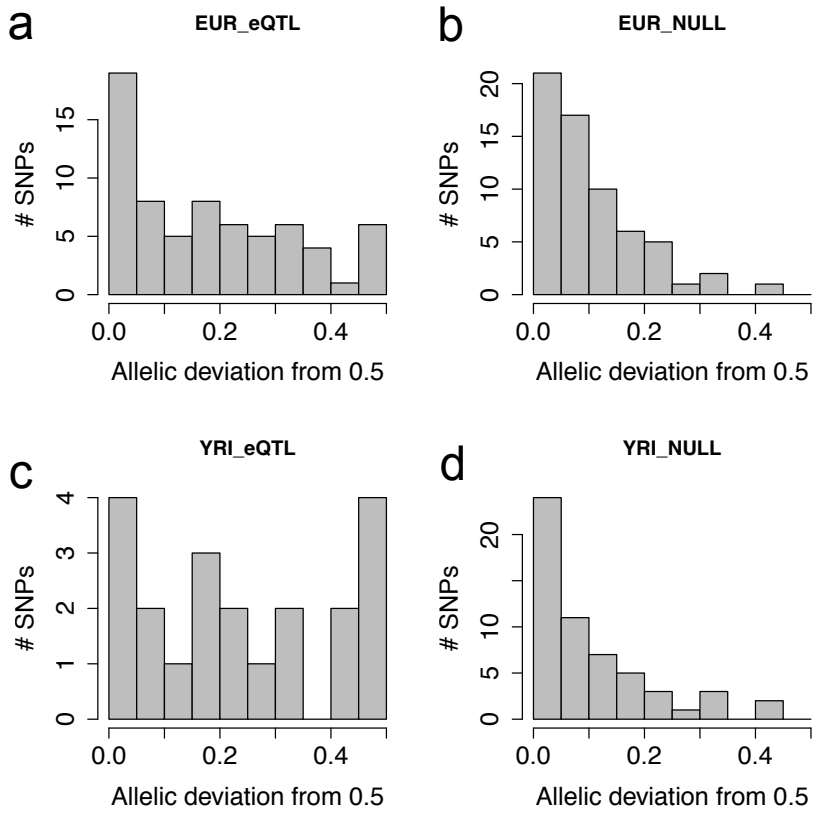

Figure S24. Allele-specific binding of CTCF in eQTLs

In order to validate the functional effects of our best eQTL SNPs using additional cellular assays, we measured allele-specific binding in Chip-seq data of CTCF from 6 individuals, using best eQTL SNPs (a,c) as well as a matched null set of variants (b,d) from EUR (a,b) and YRI (c,d) data, analyzing sites that were heterozygous in these samples and covered by Chip-seq reads. We observe a significant enrichment of allele-specific binding in eQTL sites (p=0.0020 for EUR, 0.0023 for YRI, Mann-Whitney between allelic ratios for eQTLs and null), suggesting that these sites can indeed be causal, with a change in CTCF binding being the putative cellular mechanism underlying the expression level change.

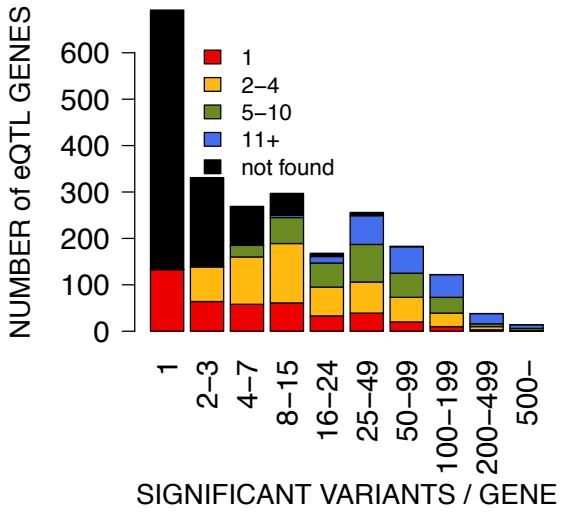

Figure S25. Overlap of eQTLs with Omni 2.5M SNPs

The rank of the best Omni2.5M SNP among the significant YRI eQTL variants per gene, in bins on the x-axis according to the total number of significant variants. See Figure 2 for the plot of European eQTLs.

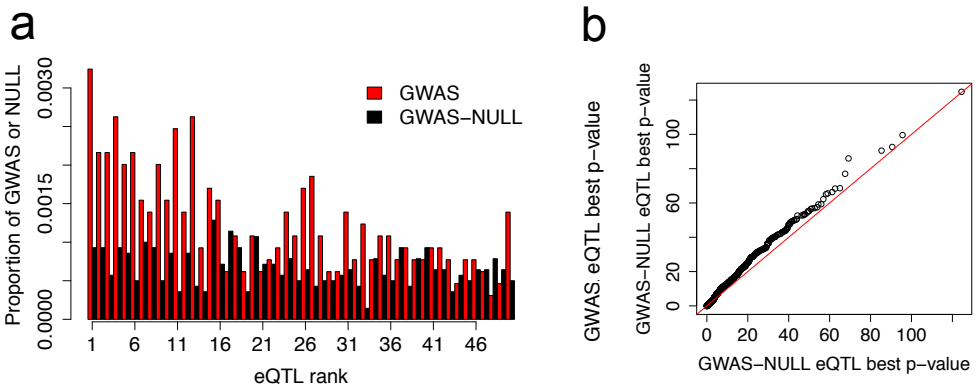

Figure S26. GWAS signal of eQTLs

We analyzed the overlap of GWAS SNPs and EUR eQTL signals. First, we calculated for each GWAS variant its rank among eQTLs (a) – i.e. if the SNP is the best associating eQTL of a gene, it gets a value of one – and this repeated for a null distribution of variants matched to the GWAS minor allele frequencies. GWAS variants are clearly enriched around the peak of eQTL associations compared to the null, suggesting that eQTL variants are the causal variants for many GWAS associations. The distribution is truncated at the rank of 50. (B) shows the best cis-eQTL p-value of GWAS SNPs and the matched null variants plotted as a qq-plot, indicating that GWAS variants are more likely to be eQTLs.

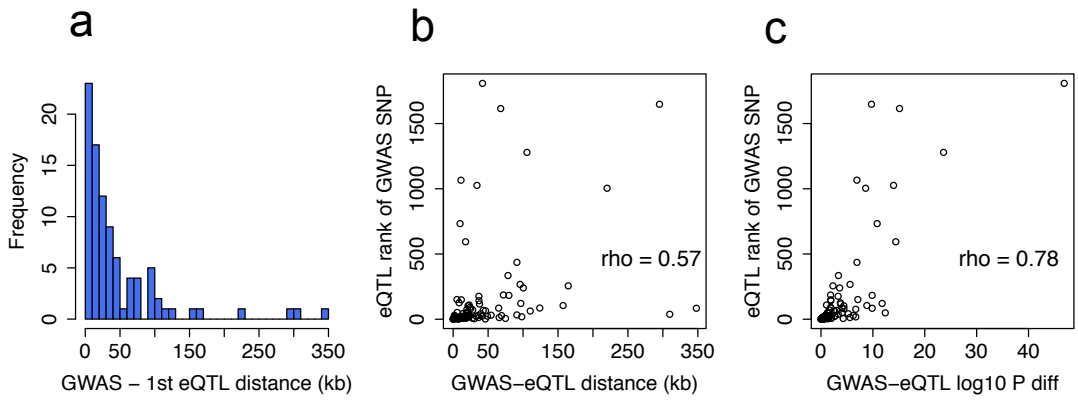

Figure S27. Causal GWAS variants prediction

For the 91 GWAS eQTLs that have been assessed to share a causal variant (see Supplementary Methods), we assessed how close the GWAS variant is to our best EUR eQTL – the most likely causal variant – in terms of distance (a), eQTL rank versus distance (b), and eQTL rank versus p value difference between the GWAS SNP and the best eQTL (c). 75% of GWAS variants are >10kb away from the most likely causal variant or region, and in the 1000 Genomes data there is a median of 31 variants with better eQTL p-value than the GWAS variant. The correlations in (b) and (c) are significant with p= 4.676e-09 and p< 2.2e-16, respectively.

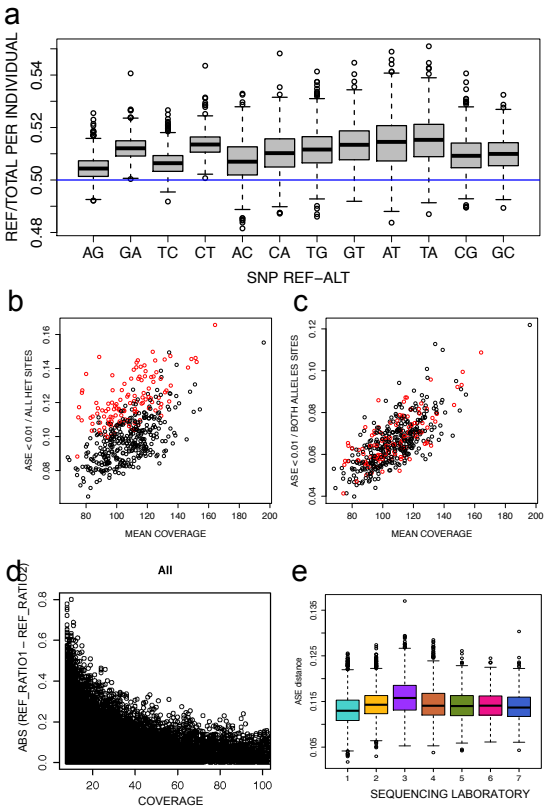

Figure S28. Quality control of ASE data

The expected allelic ratio in ASE analysis (a) is not strictly 50-50 as might be expected for heterozygous sites – there is a slight genome-wide bias favoring the reference allele, and there is also a nucleotide bias favoring G and C, shown in (a) where the genome-wide allelic ratios for each individual are plotted. These ratios are used as the expected ratio in the calculation of binomial probability of ASE as an overall correction of these biases.

B) and c) demonstrate the effects of slight variation in genotype quality in ASE data. Here, each dot is an individual, with median coverage of ASE sites plotted on the x-axis, and proportion of ASE ($p<0.01$) on the y-axis. The general correlation between these two is expected, due to higher power when coverage is high. The red individuals are the lowest 25% of DNA-RNA genotype concordance of heterozygous sites. Discordance may be due to allelic expression (the average amount of which should, however, be similar across individuals), and some due to false genotype calls (which may vary). Excluding sites with only one allele seen (c) leads to more stable ASE proportions between individuals, and for the majority of analyses, we used only such sites.

The effect of coverage of the ASE site is demonstrated in (d) using replicate samples: difference in the allelic ratio of the same site is plotted as a function of coverage. Consistency is good after 30-40 reads.

While ASE is less sensitive to laboratory effects than quantifications, ASE distance (see Supplementary Methods) between individuals still shows some laboratory effects (e).

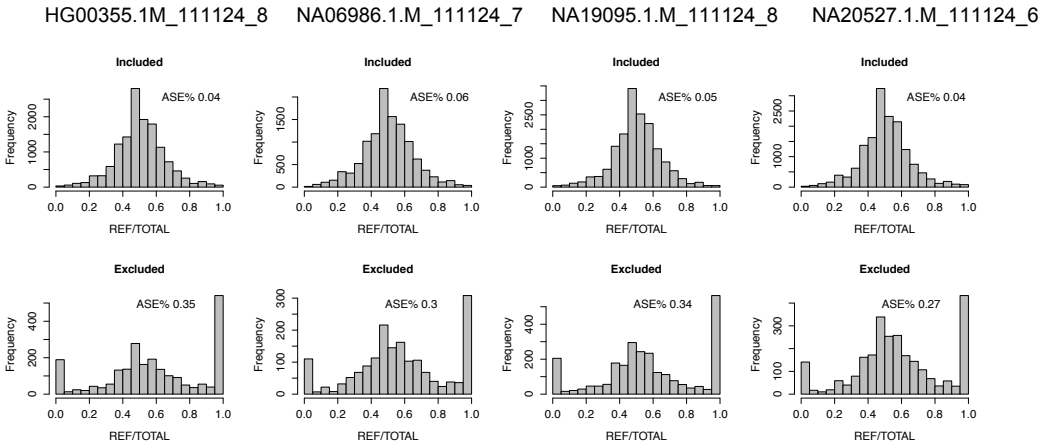

Figure S29. Filters for allelic mapping bias

Allelic ratio in four samples for SNPs that were kept in ASE analysis (top row) and in SNPs that were excluded due to increased risk of mapping bias based on simulations, genomic mapability estimates, or having only one allele observed (REF/TOTAL < 0.02 or > 0.98) (bottom row; see Supplementary Methods for details). The numbers denote the proportion of sites with significant ($p<0.005$) ASE, showing that excluded sites have clearly elevated ASE signal that is likely due to mapping problems.

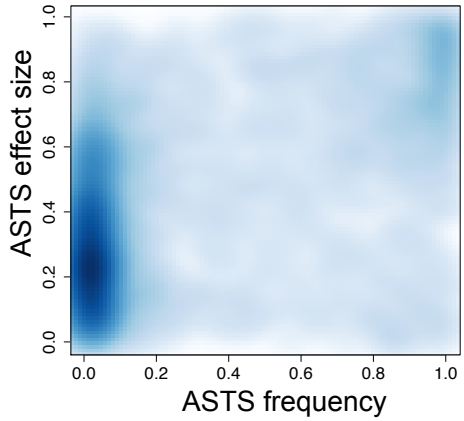

Figure S30. Frequency spectrum of allele-specific transcript structure

ASTS effect size (maximum allelic ratio distance of exons from the total ratio) as a function of frequency of the ASTS effect in the population, calculated for sites with >=20 ASTS measurements. This is analogous to Fig. 3b of ASE, and shows that the majority of ASTS effects are rare in the population.

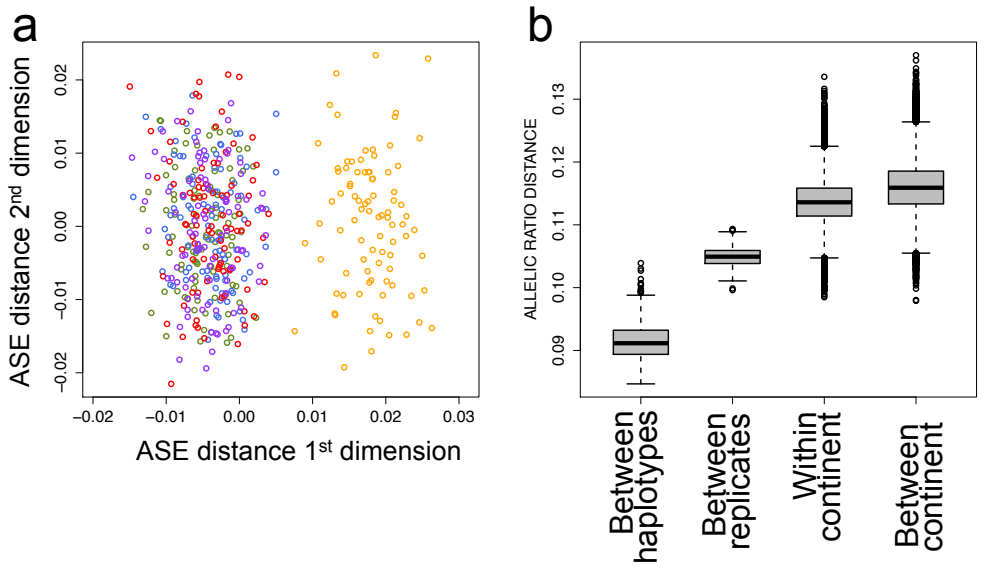

Figure S31. Population variation in ASE

Multidimensional scaling of a matrix of allelic ratio distance between individual pairs (see Supplementary Methods) shows a clear clustering to African and European individuals (a). In (b), we further dissected allelic ratio distances to differences between two haplotypes of an individual (abs(0.5-REF_RATIO)), and allelic ratio distances for replicate samples from the same individual, two individuals from the continent, or from a different continent.

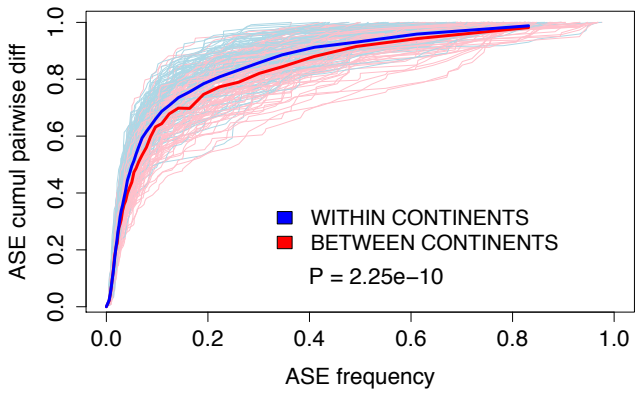

Figure S32. Population variation across ASE frequency spectrum

We used ASE data to partition how much of phenotypic discordance between two individuals come from rare and common events in the population, based on the idea that ASE is a proxy for regulatory variation.

To this end, for each individual pair we took ASE SNPs where the individuals are discordant (ASE $p < 0.005$ & ASE $p > 0.1$). Here, we used only sites present in >=15 individuals in the data set sampled to a coverage of 30, and individual pairs with >=70 sites that were measured in both. For all these sites per individual pair, we calculated the sum of differences in allelic ratios as a measure of total phenotypic difference. Additionally, for each site we calculated how frequent the ASE effect (present in only one of the individuals of the pair) is in the entire sample, which is a proxy for the frequency of the regulatory variant driving this effect.

The plot shows the relationship between the two: the population frequency of ASE (~frequency of the regulatory variant) on the x-axis, and the cumulative proportion of how much of the total allelic ratio (~phenotypic effect) difference between two individuals explained by each event. Each pair of individuals is represented by a thin line (randomly selected 100 pairs of each class), colors are according to whether the individuals of the pair come from same or different continents. The thick lines represent medians. The p-value is from a Mann-Whitney test of ASE frequencies in within-continent vs between-continent pairs.

We can see that most differences between two individuals are caused by regulatory effects that are rare in the population, which is consistent with the frequency spectrum in Fig. 3b. Importantly, differences between individuals of the same continent are relatively more often caused by rare effects. This is consistent with what we know of population sharing of genetic variants: rare variants are very population specific; thus a relatively larger proportion of differences within populations are expected to be driven by rare variants, whereas different continents differ in terms both rare and common variation and so they can contribute more equally to individual differences.

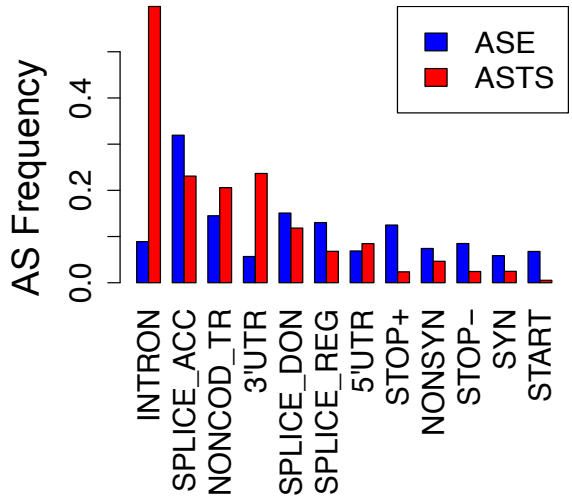

Figure S33. Likelihood of significant allelic effects by annotation class

Frequency of ASE and ASTS by annotation class of the AS variant shows a clear enrichment of allelic imbalance in loss-of-function sites. The likelihood of AS signal is affected by the transcriptome impact (such as splicing change or NMD) that a variant may have, and also the consequences that this has on the coverage of the AS site, since our analysis is naturally only based on the reads that remain in the data and cover the site in question. For example, >=16 reads of coverage over fully intronic sites is rather unusual, and can be due to rare intron retention effects or unannotated exons, both with unusual splicing easily leading to an ASTS signal. The ASTS signal in the 3' end is expected knowing the widespread variation in 3' ends of transcripts. Splice site variants are an interesting case: for example, let us consider a nonreference allele in a splice acceptor site that decreases splicing efficiency of the exon by 60%. In this situation, 60% of nonreference transcripts would skip the exon or undergo NMD, leading to a strong ASE signal over the splice site – however, the remaining 40% of the nonreference reads over the splice site that we would use to calculate ASTS would have the normal splicing pattern, without ASTS over this site.

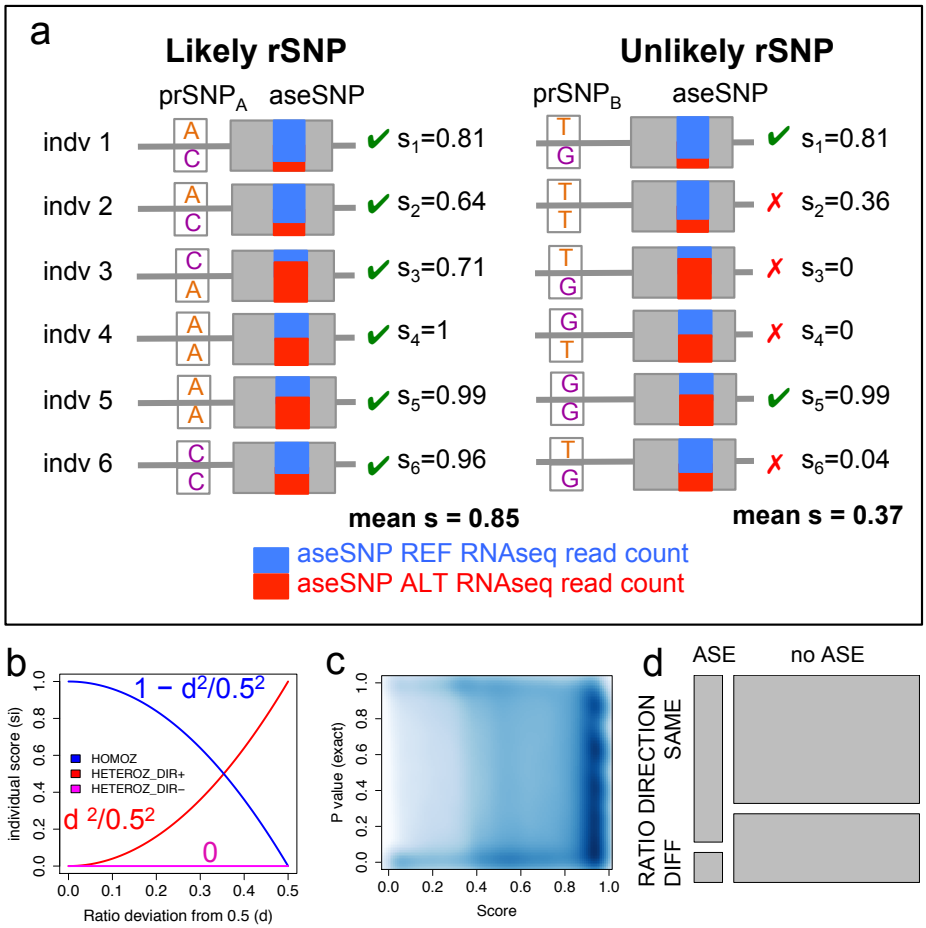

Figure S34. Mapping putative regulatory SNPs (prSNPs) with ASE data

Differential expression of the two haplotypes of an individual, or allele-specific expression, is believed to be often driven by the individual being heterozygous for a regulatory variant elsewhere in the *cis*-regulatory region. (A) illustrates the basic principle of the method, trying finding maximal concordance between the allelic ratios of an aseSNP, and the genotypes of the prSNPs in the surrounding region: for a true rSNP, we would expect heterozygote individuals to have large deviation of the null allelic ratio of 0.5, whereas homozygotes would be expected to have ratio close to 0.5. This situation is illustrated on the left side of (a), whereas right side illustrates a situation with poor concordance.

To quantify the concordance, we calculated for each prSNP–aseSNP pair in each individual *i* a concordance score s_i according to the equations in (b), where d_i = abs(0.5 – ref_count$_i$ / total_count$_i$), separating prSNP homozygotes and heterozygotes (blue versus red/magenta). Here, the closer s_i is to 1, the better the concordance. Having phased data, we also took allelic direction into account in prSNP heterozygotes as a true rSNP should have one allele consistently higher expressed: we assigned the majority allelic direction based on the data (see Supplementary Methods section 12.3) and penalized individuals showing the

opposite direction by assigning $s_i = 0$ (magenta). For each prSNP-aseSNP pair, we then calculated score *s* as the average of s_i.

Having a score *s* for each prSNP-aseSNP pair, we next evaluated how likely it is to obtain such a score by random combination of the allelic ratios over this particular prSNP-aseSNP pair – thus, we permuted the prSNP genotype labels as many times as there are unique combinations, up to 1000 times, and recalculated the score on each round. From this random distribution of *s*, we obtained an empirical p-value for *s*. (C) shows that the absolute value of s is not correlated to its p-value as it depends on genotype frequencies and aseSNP allelic ratios, and even relatively low scores can be much higher than expected by chance. Thus, in the selection of most likely true rSNPs from all the tested prSNPs, the absolute value of *s* is not informative; we selected prSNPs that had a permuted p-value of 0 with no higher permuted scores than the observed one. Since each prSNP-aseSNP pair has [100,1000] permutations, this corresponds to p-value [0.01,0.001] .

One gene can have several aseSNPs that can capture the same rSNP effect when it is driven by regulatory variant that affects both sites, or they can be independent e.g. in the presence of splicing variation. In (d), we measured how similar ASE signals are within the same gene in a single individual: We took a aseSNP within an individual with significant ASE ($p<0.005$), and found cases where the individual has another aseSNP in the same gene as the first one. The mosaic plot shows the proportion of the second aseSNPs that are significant (like the first one) and that have allelic ratio to the same direction as the first one. The results indicate that while there is some excess of shared signal of the two aseSNPs, the signals appear mostly independent. This is consistent with much of ASE overlapping with allele-specific transcript structure signals (Fig 3a) as well as high degree of independence of exon eQTL signals of the same gene (see eQTL section and Fig S17). Consequently, trying to collapse ASE signals from the same gene for prSNP analysis would be likely to lead to loss rather than gain of power, and we decided to analyze aseSNPs from the same gene independently.

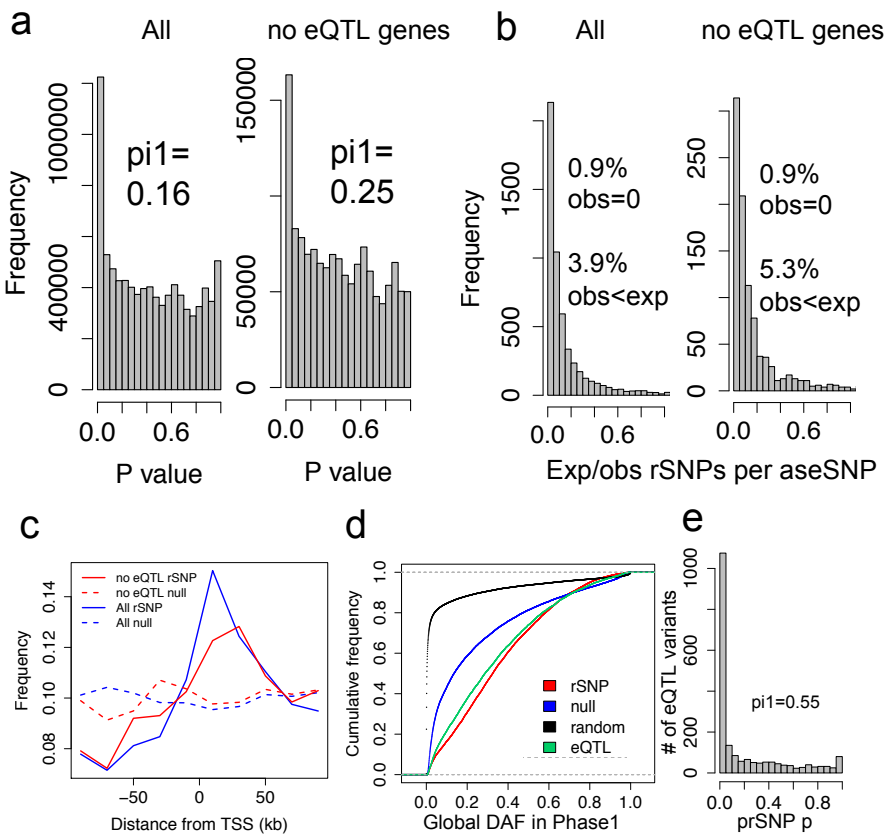

Figure S35. rSNP characteristics

(A) shows the p-value distribution of all tested prSNP-aseSNP pairs (left) and of those aseSNPs that are not in genes with an exon eQTL (right). Both distributions show a clear enrichment of low p-values. From these data, we selected the most likely set of regulatory SNPs (rSNPs) and a null with high p-values (see Supplementary Methods section 12). In (b), for each aseSNP separately, we calculated how many rSNPs we would expect to find by chance based on the p-values of the rSNPs and the number of tested prSNPs, finding that in the vast majority of cases we find much more than expected by chance. The numbers denote statistics not visible in the plots: the cases where we either find no rSNPs for an aseSNP, or that the number of rSNPs is smaller than expected by chance (see also Table S6).

From all the rSNPs, we sampled 5 rSNPs per aseSNP to investigate their properties. Distance of these variants from TSS is shown in (c), with a peak close to TSS that is characteristic for cis-regulatory variants. In (d), we compare allele frequency distributions of rSNPs, eQTLs, the null selected from nonsignificant prSNPs, and random SNPs sampled from the genotype data. In order to make a fair comparison, we use only rSNPs that were included in eQTL analysis (have MAF >5% in EUR or YRI), and the plot shows the derived allele frequency in the whole 1000 Genomes Phase 1 sample. The allele frequency spectra of significant eQTLs and rSNPs are very similar. (E) shows the prSNP p-value for tested eQTL SNPs, indicating the concordance of genotype-based and ASE-based detection of regulatory effects.

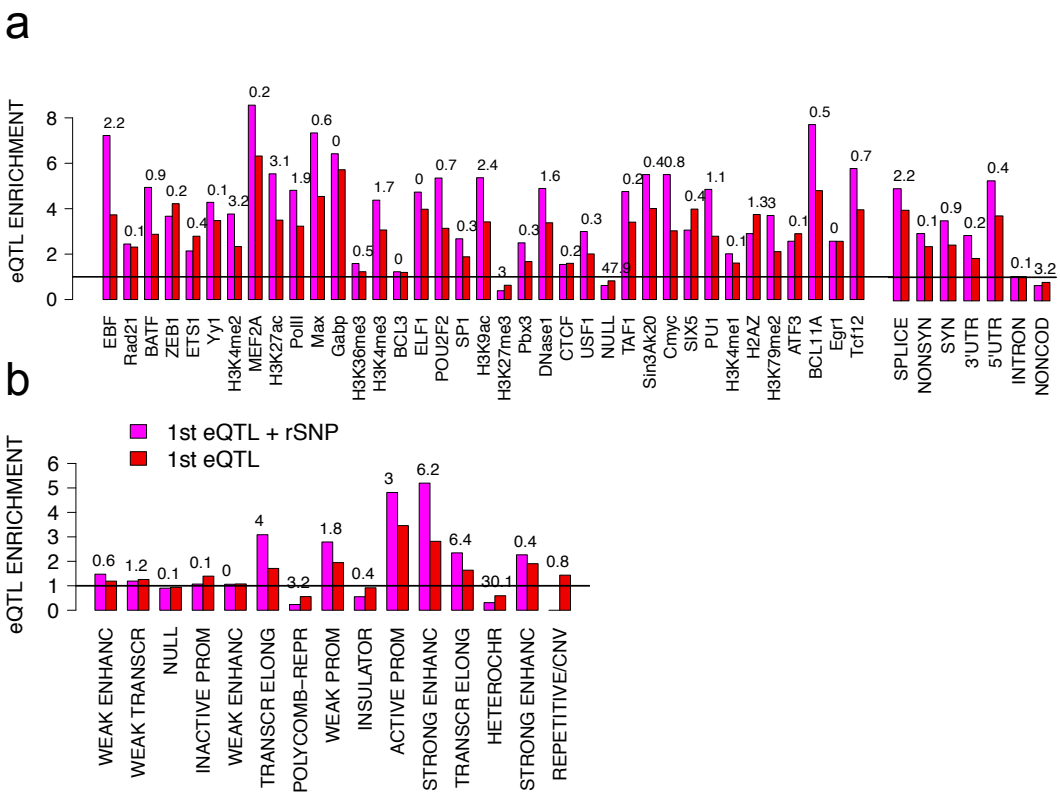

Figure S36. rSNP and eQTL annotation overlap

Enrichment of variants in functional annotations relative to a matched null distribution for all the most significant eQTL variants (red; similarly to Fig. 2a, S21), and to the subset of these that are also rSNPs (magenta). A subset of (a) of the categories with most data is shown in Fig. 3c. (A) shows Ensembl Regulatory Build and coding annotations, and (b) chromatin states. The null is the same as that used in eQTL analysis, matched to eQTL allele frequency and distance from TSS. The numbers above the bars denote –log10 p-values of a Fisher test between the eQTL & prSNP variants and only eQTL variants. There is an increased enrichment of eQTLs in functional elements when they are also prSNPs.

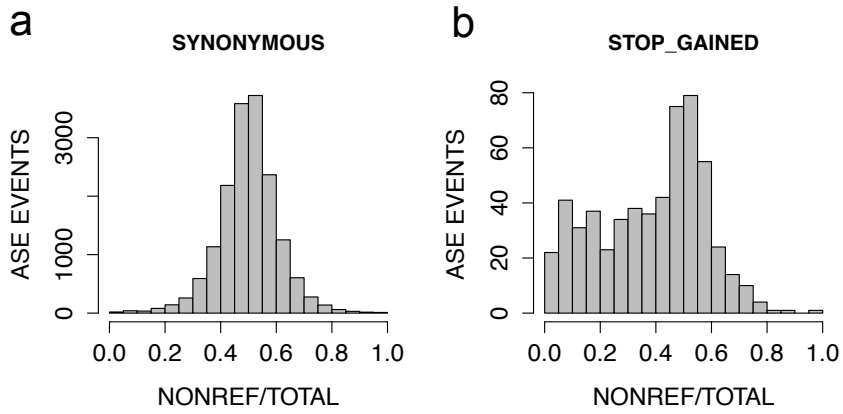

Figure S37. Nonsense-mediated decay

Alternative allele ratio in RNA-sequencing read counts for a random set of heterozygous synonymous variants in (a) and for stop-gained variants (b), from one random individual per site. The synonymous variants show the basic null distribution of allelic ratios, with most sites being close to 0.5, with some variation likely due to cis-regulatory variation. Premature stop variants, however, show a substantial decay in the counts of the nonreference (stop) allele, suggesting frequent nonsense-mediated decay.

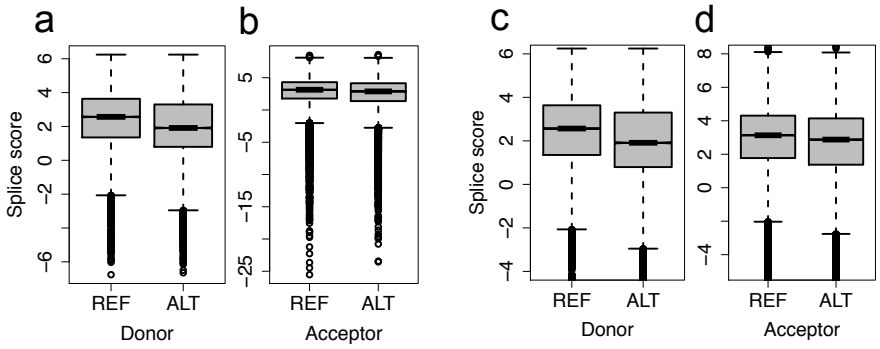

Figure S38 . Splice scores

For variants overlapping an annotated splice motif, we calculated splicing scores to predict the efficiency of splicing of the splice site. Distribution of scores for reference and alternative alleles for donor and acceptor sites is illustrated, with the plots showing the full distribution for donor (a) and acceptor (b) and the same distributions with zoomed y-axis in (c) and (d). The difference between reference and alternative allele distributions are significant for both donor and acceptor sites ($p < 2.2e\text{-}16$).

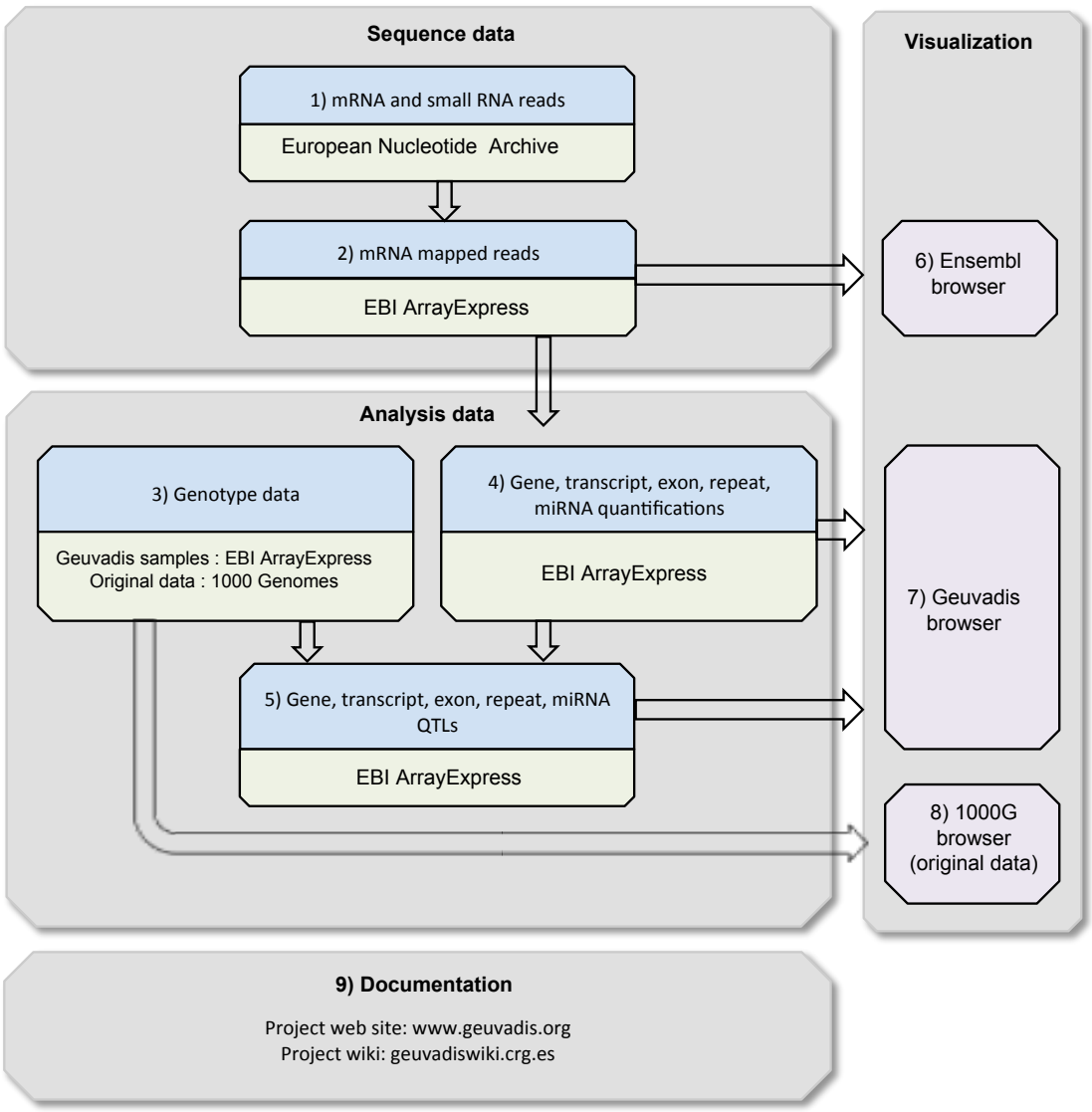

Figure S39. Data Access Schema

The figure illustrates the Geuvadis data sets that are available, all with open access. Full links to all sites can be found in the project website www.geuvadis.org. The main accession site to the data created and analyzed by the Geuvadis RNA-sequencing project is EBI ArrayExpress, where the data is stored under three accessions: E-GEUV-1 for mRNA post-QC samples used in analyses of this paper, E-GEUV-2 for small RNA post-QC samples, and E-GEUV-3 for all the sequenced data.

1) Raw reads in the form of fastq files are stored in ENA under the accession ERP001942 and ERP001941, accessible also through ArrayExpress (the ENA and FASTQ columns)

2) mRNA mapped reads are stored and accessible from EBI ArrayExpress, where the "Processed" column contains downloadable bam files. Files of mapped small RNA reads are not provided due to the more complex nature of mapping to different references for different analytical purposes and the large number of multimapping reads making file sizes very large.
3) Genotype data that have been used in Geuvadis data analysis are available from EBI ArrayExpress site under accession E-GEUV-1, and the vcf files include also a functional reannotation of all the variants. The original data created by 1000 Genomes Project are available in the 1000 Genomes web site.
4 and 5) Geuvadis analysis results for gene, transcript, exon and repeat quantifications and QTLs are available from EBI ArrayExpress site under accession E-GEUV-1, and corresponding data for miRNA is under accession E-GEUV-2.
6) mRNA mapping results per sample down to the level of individual reads can be visualized using Ensembl Genome Browser using the links from ArrayExpress (the Ensembl icon)
7) Geuvadis data browser (http://www.ebi.ac.uk/Tools/geuvadis-das/) was created specially for the Geuvadis RNA-seq project to visualize quantification and QTL results, and allows searching by variant ID, gene and region, as well as download of quantification and QTL data by region. Links to the archives with raw and mapped data as well links to the analysis results and genotypes are available from the browser as well. See also Fig. S40.
8) Original genotype data can be viewed and downloaded in 1000 Genomes Browser: http://browser.1000genomes.org/index.html.
9) Documentation of the files and links to all the sites are in www.geuvadis.org as well as in readme files. The project wiki in geuvadiswiki.crg.es has been made openly accessible, and contains additional analysis results and method descriptions.

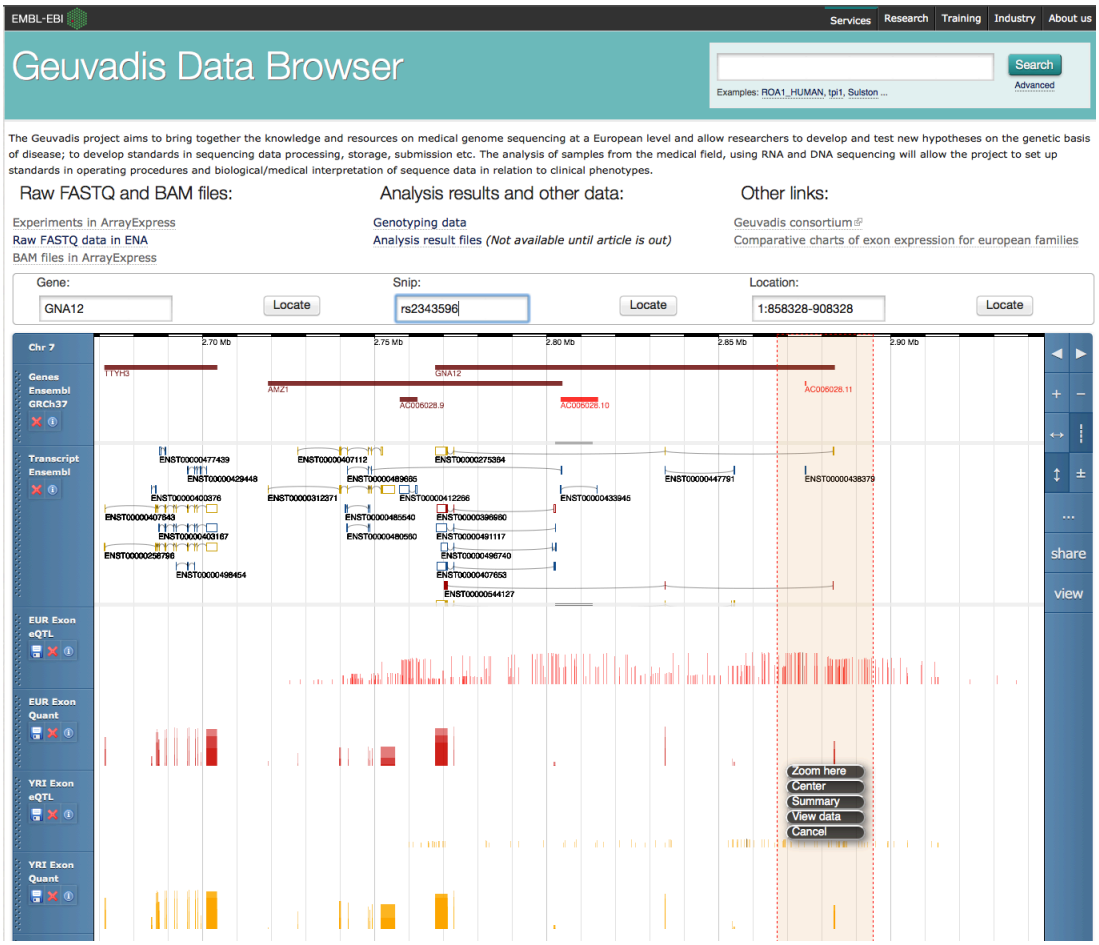

Figure S40. The Geuvadis Data Browser

For the visualisation of RNA-sequencing analysis we created the Geuvadis Data Browser (www.ebi.ac.uk/Tools/geuvadis-das) for viewing and downloading exon, transcript and miRNA quantifications as well as exon eQTLs, transcript ratio QTLs and mirQTLs for EUR and YRI.

Supplementary Methods

1. Study design (Fig. S1, Table S1)

Transcriptome sequencing was performed in seven European laboratories, each processing 48-116 randomly assigned samples. Five samples were sequenced in replicate in each of the labs for both mRNA and miRNA, and twice at the University of Geneva (UNIGE) for mRNA. Additionally, 168 samples, also sequenced in other laboratories, were mRNA-sequenced at the University of Geneva, at 2/3 of the standard coverage. Of the replicate samples, the one with the highest coverage was used in the main analysis of unique samples.

2. RNA-sequencing data production

2.1. Cell line processing

EVB transformed lymphoblastoid cell lines (LCLs) directly from Coriell Cell Repositories (GBR, FIN, TSI) or originally from Coriell but grown at the University of Geneva (CEU, YRI) were shipped to ECACC (European Collection of Cell Cultures) as live cultures, in batches of ~30 samples from Coriell (GBR/FIN/TSI somewhat randomized) and 2 x ~90 samples (by population) from Geneva.

In ECACC, these cell lines were cultured to approximately 1.2 x 10e8 cells. These cultures were split to produce 8 x cell banks of the samples, and a snap frozen pellet of 2 x 10e7 cells from a proliferating culture. The cell pellets were shipped from ECACC to University of Geneva in three batches, the first batch consisting of CEU/GBR/FIN/TSI samples, and the second and third batch with YRI and the rest of CEU samples.

2.2. RNA extraction

RNA was extracted in Geneva about 14 samples at a time, first extracting 2/3 of the first shipping batch with full randomization, then adding the second batch and randomizing among that and the remaining 1/3 of the first batch, and finally extracting the third batch.

Total RNA was extracted from cell pellets using the TRIzol Reagent (Ambion). The pellets had been frozen at ECACC without any additives like RNAlater or TRIzol. In Geneva they were thawed, 1mL of TRIzol was added in each sample, and the samples were transferred to eppendorf tubes. The rest of the protocol followed the manufacturer's guidelines. No DNAse treatment was done to the RNA samples.

RNA quality was assessed by Agilent Bioanalyzer RNA 6000 Nano Kit according to the manufacturer's instructions. RNA quantity was measured by Qubit 2.0 (Invitrogen) using the RNA Broad range kit according to the manufacturer's instructions.

2.3. RNA sequencing

Each of the sequencing laboratories were sent a minimum of 4 ug of total RNA of the samples allocated to them, and RNA Bioanalyzer was ran for 10-20% of the RNA samples before library preparation to confirm sample quality after shipping. No further purification steps were done to the RNA samples other than that specified in the sequencing protocols. Library preps were done in random order in every laboratory.

mRNA sequencing was done on the Illumina HiSeq2000 platform with 75 bp paired-end sequencing with fragment size of ~280 bp – some laboratories sequenced 100bp reads, which were trimmed to 75bp. TruSeq RNA Sample Prep Kit v2 (the high-throughput protocol) was used for library preparation, TruSeq PE Cluster Kit v3 for cluster generation, and TruSeq SBS Kit v3 for sequencing. The laboratories were allowed to choose freely how to pool the samples to get the desired minimum of 10M mapped and properly paired read pairs from any standard mapper, without filtering for mapping quality.

Small RNA sequencing was done on the Illumina HiSeq2000 platform with 36 bp single-end sequencing with fragment size of 145-160 bp. Some laboratories sequenced 50bp reads which were trimmed to 36bp. TruSeq Sm RNA Sample Prep kit was used for library preparation, TruSeq SR Cluster Kit v3 for cluster generation, and TruSeq SBS Kit v3 for sequencing. The laboratories were allowed to choose freely how to pool the samples to get the desired minimum of 3M total reads.

Extensive information of sample processing was collected from all the laboratories for both mRNA and miRNAseq in order to enable control of batch effects.

2.4. Raw data processing

Each lab submitted one demultiplexed fastq file per sample per mRNA and miRNAseq, produced by CASAVA 1.8 or 1.8.2 allowing one mismatch in the index. Reads failing Illumina quality filtering were removed. The fastq files are named as: SAMPLE_ID.SeqLabNumber.M/MI_YYMMDD_Lane_Read.fastq.gz, where M/MI stands for mRNA or miRNA sequencing, and YYMMDD is the sequencing date. All the data were submitted and initially stored in the project ftp site. Samtools[28] was used for general data processing throughout the project.

3. Genotype data

We used 1000 Genomes Phase 1 data set that is based on 1092 individuals , with whole-genome sequencing of to an average depth of 5x and exome sequencing to depth of 80x, and high genotype quality[29]. Variant identifiers follow dbSNP v137, with the variants lacking rs-identifier named as follows: SNPs had an identifier of type snp_chr_pos (e.g. snp_21_357682), and indels and structural variants were of type indel/sv:lengthI/D_chr_startpos (indel:3D_1_10523).

3.1. Variant annotation (Table S2)

The Variant Effect Predictor (VEP v2.5; http://useast.ensembl.org/info/docs/variation/vep/vep_script.html) tool from Ensembl was modified to produce custom annotation tags and additional loss of function (LoF) annotations. The additional LoF annotation was applied to variants that were annotated as STOP_GAINED, SPLICE_DONOR_VARIANT, SPLICE_ACCEPTOR_VARIANT, and FRAME_SHIFT and flagged if any filters failed. A LoF variant is predicted as high confidence (HC) if there is at least one transcript that passes all filters, otherwise it is predicted as low confidence (LC). This modified version of VEP was applied to the 1000 Genomes Phase1 data using the Gencode v12 annotation. To this, we added information of overlap with chromatin states[30], Ensembl Regulatory Build elements, miRNA targets from TargetScan[31], and miRBase v18[32] mature and hairpin miRNA loci. Annotation information is stored in the vcf file info field as ordered lists. Detailed documentation is provided together with the vcf files.

3.2. Imputation

For 421 samples of the project, we used the 1000 Genomes Phase1 release v3. Genotype imputation was done for 42 samples from 1000 Genomes project Phase 2 with Omni 2.5M genotype data, using the IMPUTE2 software[33]. As the reference panel we used the entire Phase 1 v3 release, and for a study panel we took Omni Shapeit haplotypes for the whole Phase 2 sample set, and extracted our 42 samples from the imputation results. These were merged to a single vcf file together with the Phase 1 samples.

Since IMPUTE2 did not handle multiallelic genotypes well, we kept only biallelic genotypes for the analysis. Additionally, the genotype calls of imputed genotypes with posterior probability <0.9 were marked as missing.

3.3. Quality control (Fig. S3)

First, we calculated an IBS matrix of genotype data of chr20, which showed clear clustering to Phase1 and Phase2 individuals (Fig. S3), even though we verified that all variants had consistent allele frequencies. Furthermore, PCA[34] showed a clear clustering to populations, as expected. To make sure that our QTL associations are not driven by biases from imputation or from population structure, we included the imputation status (0|1) and principal components 1-3 for Europeans and 1-2 for Yoruba as covariates in QTL analyses.

In QTL analyses, we used variants with >5% MAF in either EUR or YRI, which gave us 10,785,347 variants in total, of which 9,836,718 are SNPs, 945,987 are indels, and 2642 are SVs. QTL analysis was done with genotype dosage values.

4. mRNA read mapping

We employed the JIP pipeline (Griebel & Sammeth submitted) to map RNA-Seq reads and to quantify mRNA transcripts. For alignment to the human reference genome sequence (GRCh37, autosomes + X + Y + M), we used the GEM mapping

suite (v1.349 which corresponds to publicly available pre-release 2)[35] to first map (max. mismatches = 4%, max. edit distance = 20%, min. decoded strata = 2 and strata after best = 1) and subsequently to split-map (max.mismatches = 4%, Gencode v12 and *de novo* junctions) all reads that did not map entirely. Both mapping steps are repeated for reads trimmed 20 nucleotides from their 3'-end, and then for reads trimmed 5 nucleotides from their 5'-end additionally to earlier 3'-trimming—each time considering exclusively reads that have not been mapped in earlier iterations. Finally, all read mappings were assessed with respect to the mate pair information: valid mapping pairs are formed up to a max. insert size of 100,000 bp, extension trigger = 0.999 and minimum decoded strata = 1. The mapping pipeline and settings is described below, and can also be found in http://github.com/gemtools, where the code as well as an example pipeline are hosted.

The GEM output format was converted to bam format, with following mapping quality scores and flags:

1. Matches which are unique, and do not have any subdominant match: 251 >= MAPQ >= 255, XT=U
2. Matches which are unique, and have subdominant matches but a different score: 175 >= MAPQ >= 181, XT=U
3. Matches which are putatively unique (not unique, but distinguishable by score): 119 >= MAPQ >= 127, XT=U
4. Matches which are a perfect tie: 78 >= MAPQ >= 90, XT=R.

Furthermore, the NM flag contains the number of total mismatches (read1+read2). In analysis, we used reads in categories 1 and 2 and with NM<=6.

4.1. Analysis of allelic mapping bias (Fig. S29)

In RNAseq mapping, it is possible that reads from a locus with a genetic variant map differently to the reference genome depending on whether the read carries the reference or nonreference allele. Such allelic mapping bias can be problematic especially in analysis of allele-specific expression (ASE) comparing RNAseq read counts of the two alleles in heterozygous individuals. Furthermore, also quantifications of exons or other units could be affected by mapping error – analogously to SNPs in probes in expression array studies – although this is less likely due to quantifications being based on a larger genomic region rather than on a single site.

We addressed this concern by simulating RNAseq reads over all SNP and indel variants in 1000 Genomes Phase 1 release that are polymorphic in the populations of this study. For each locus, we created all possible 75bp reads overlapping the site in all haplotype combinations present in the 1000 Genomes data – reads carrying the reference allele, reads carrying the nonreference allele, and additional sets of reads according to the haplotype phase when there were other variants <75bp from the variant in question. The simulated reads were constructed based on the genomic sequence without taking transcript structure or paired-end sequencing into account due to the extremely large number of combinations that would be created by considering these factors. We then mapped the simulated reads to the reference genome with GEM, and calculated

for each variant site the ratio of correct mapping of reads carrying the reference or nonreference allele.

Based on these results, we excluded from ASE and ASTS analysis 2,810,388 sites with >5% bias in simulations. Additionally, we excluded sites in regions with <1 genomic mapability score (based on the UCSC mapability track, 50bp segments) and sites in collapsed repeat elements[36], and altogether filtered out approximately 12% of sites in the ASE analysis of each individual (with slight variation according to which sites are covered by enough reads to analyze ASE; Fig S29).

Using the results from simulations, we also analyzed the effect of potentially biased reads on exon quantifications, observing that this had an effect only in a very small proportion of exons, and only a handful of eQTLs out of thousands appeared to be affected by mapping bias (Panousis et al. in preparation; see also Fig. S20). Thus, we decided not to account for the bias in our eQTL analysis of mRNA data. However, in small RNA sequencing the shorter read length and higher degree of homology makes allelic mapping bias much more likely. Thus, our small RNA reads were mapped to miRNA stemloop sequences containing not only the standard sequences but also alternative allele versions of all the stemloop sequences that overlap a variant (see section 6.3). This guarantees that miRNA sequences containing nonreference alleles are not lost in mapping, which would easily lead to false mirQTL associations.

4.2. Duplicate reads (Fig. S6)

PCR in library preparation can sometimes amplify fragments such that they become overrepresented among sequencing reads, resulting in a stack of reads with exactly same coordinates, often called (PCR) duplicate reads. In genome sequencing where equal coverage of all regions of the nuclear genome is expected, duplicate reads are often removed. However, in RNA-sequencing the read coverage of different genes varies substantially, and highly expressed genes are expected to have reads with the same coordinates simply due to saturation of the sequence space, in which case duplicate reads are not due to technical PCR artifacts.

To analyze whether duplicate reads in our data are driven by high mRNA expression levels – i.e. biology – or by PCR artifacts, we calculated exon quantifications of 5 samples with and without duplicate reads (Fig. S6). The results show a high overall correlation of read counts with and without duplicates; however, the quantifications of highly expressed exons are most affected by duplicate removal in a consistent logarithmic manner, indicating that duplicate reads in these exons are due to saturation of the read space. PCR artifacts would be expected to occur also in lower expressed exons, but we observe very few cases of sudden drops in expression levels by duplicate removal.

In miRNA quantifications, removing duplicates in this manner is not possible since miRNA genes are so small that all quantifications are essentially derived from reads with identical mapping coordinates. This is a general caveat of all small RNA sequencing and to our knowledge, no correction method has been suggested for this. However, miRNA fold-changes estimated by miRNA-seq and

Agilent arrays (which do not use amplification) correlate very well (e.g. [37]). There is evidence that some miRNAs are more easily amplified than others[38], which is an important factor to consider in analyses of relative expression levels of different miRNAs. However, the reported amplification biases are highly reproducible, so they are likely to cancel out in our analysis comparing patterns between individuals.

Altogether, in eQTL and coexpression network analyses, amplification biases are likely to increase noise from some additional technical variation rather than create false positive signals. The results in Fig. S6 indicate that removing duplicates in mRNA sequencing would lead to considerable downsampling of highly expressed genes, and more importantly for our analysis, create a saturation threshold where variation between individuals can no longer be detected. Thus, we did not remove duplicate reads from any of our analysis.

5. mRNA quantifications

The gene annotation used in this project was Gencode v12[39]. Duplicate reads were included in all mRNA quantifications.

5.1. Exons

Exon quantifications were calculated for protein-coding and linc-RNA transcripts. All overlapping exons of a gene were merged into meta-exons with identifier of type ENSG000001.1_exon.start.pos_exon.end.pos. Read counts over these elements were calculated without using information of read pairing, except for excluding reads where the pairs map to two different genes. We counted a read in an exon if either its start or end coordinate overlapped an exon. For split reads, we counted the exon overlap of each split fragment, and added counts per read as 1/(number of overlapping exons per gene).

5.2. Transcripts, genes, and splicing (Fig. S13)

Quantifications of transcripts and splice junctions by the Flux Capacitor approach[40] are based on the annotation-mapped genomic mappings considering transcript structures of the Gencode transcriptome annotation: mappings of read pairs that were completely included within the annotated exon boundaries and paired in the expected orientation have been taken into account. Reads belonging to single transcripts were predicted by deconvolution according to observations of paired reads mapping across all exonic segments of a locus (see Fig S13 of descriptive statistics of transcript quantifications). Gene quantificatons were calculated as the sum of all transcript RPKMs per gene. Transcript ratios were calculated as the proportion of each transcript quantification (in RPKM) of the sum of all transcripts per gene. Annotated splice junctions were quantified using split read information, counting the number of reads supporting a given junction. Exon inclusion levels were calculated as the Percentage Splice In (PSI)[41,42], defined as the ratio between reads that support inclusion of an exon over the total inclusion plus exclusion reads.

5.3. Transcribed repeats

We quantified transcription of repeat elements using the following approach: First, we extracted all repetitive elements from UCSC's repeat masker table, and excluded all elements that overlapped UCSC or Gencode genes by at least one nucleotide. This left us with 2.5M regions, in which we then counted the number of overlapping RNA-seq reads in each region for each sample. Reads that were partially overlapping are only counted for the part that is overlapping. Since we observed that rRNA elements had strong differences between laboratories, we excluded them from further analysis.

6. small RNA (sRNA) data processing

6.1. Improved miRNA gene annotations

Our annotation builds on miRBase version 18[32] but with important improvements. In the cases where only one miRNA strand was annotated, the position and sequence of the other strand was estimated using RNA structure prediction[43]. Furthermore, for the mature and hairpin miRNAs which overlapped SNP or indel variants that were polymorphic in our genotype data, sequences carrying the nonreference alleles were generated and used for downstream analyses together with the reference sequences. This is important for avoiding allelic mapping bias that can easily occur for short sequences.

6.2. sRNA read data processing

Small RNA reads with homo-polymer and low PHRED scores were removed. Ligation adapters were clipped using the AdRec.jar program from the seqBuster suite[44] with the following options: java -jar AdRec.jar 1 8 0.3. A custom search subsequently clipped shorter adapters: if there were no matches to the first 8 nts, then matches to the first 7 nts of the adapter were searched in the last 7 nts of the read, then matches of the first 6 to the last 6 positions and so on. Reads that had no matches were retained, but not clipped. Last, reads shorter than 18 nts were discarded.

6.3. sRNA mapping and quantification

For tracing the reads to their genomic source for quality control purposes, reads were mapped to the hg19 genome concatenated with unassembled parts of the human genome and genomes of known human viral pathogens (available upon demand) with this command line: bowtie -f -v 1 -a --best --strata. sRNA reads were assigned to annotations based on the genome mappings. Annotations used were from Gencode v8[39] supplemented with rRNA and LINE and Alu transposon annotations from RepBase[45] and snoRNA[46] and miRNA[32] annotations. Annotations were first resolved so that each nucleotide on each strand had exactly one annotation. In case of nucleotides with more than one annotation, conflicts were resolved using a confidence-based floating hierarchy (as in [47]):

mitochondrion > virus > miRNA > snoRNA > rRNA > tRNA > snRNA > misc_RNA > lincRNA > processed_transcript > pseudogenes > protein_coding > LINE > Alu > intron_coding > intron_non_coding > intergenic. Each read mapping was weighted inversely to the number of genome mappings for the read, e.g. a read mapping to two genomic locations would get an assigned weight of 0.5. Each mapping was counted towards the annotation of the nucleotide in the middle of the mapping.

miRNA quantifications for analysis were calculated as read counts using miraligner.jar from the seqBuster suite using the following options: java -jar miraligner.jar 1 3 1, and using the improved annotations as the reference. Reads that mapped equally well to two or more miRNAs are counted fully towards each miRNA.

7. RNA-seq quality control

A more detailed analysis of technical variation of this dataset can be found in 't Hoen et al. (in preparation).

7.1. Outlier and laboratory effect detection (Fig. 1a, S4-5, S7, S10-S11)

The read and gene count distribution of mRNA-seq data were very uniform (Fig. S4). To further estimate sample quality, we calculated Spearman rank correlation between all samples using exon counts and transcript RPKMs. From these data, we calculated the median correlation of one sample against all the other samples. 2 samples in mRNA data and 4 samples in miRNA data were excluded from analysis due to low correlation with other samples. We analyzed sample correlations for replicate samples and for the whole sample set used in analysis.

While differences between laboratories were not nonexistent in our data set, the analysis of replicate samples shows that this variation is less than variation between individuals – which is already very slight. Normalization of the quantifications further reduced laboratory effects, and in our study design we carefully randomized the samples across laboratories to make sure that e.g. population differences are not confounded by technical artefacts.

7.2. Sample swap and contamination analysis

Allele-specific expression (ASE) analysis of mRNA-seq data was used to detect sample swaps, which we did not find. Based on analysis of increased heterozygosity in ASE results and mixed expression pattern of sex chromosome specific genes, we excluded 5 samples because of possible contamination ('t Hoen et al. submitted).

7.3. miRNA data quality control (Fig. 1a, S4-5,,S7, S10-11, S15)

The total small RNA read count and the number of miRNA reads were relatively similar across samples, but the proportion of miRNA reads per sample showed large variation from close to 0 to 60%. This is likely caused by variation in the library preparation step and sequencing of a large number of non-miRNA reads

in some samples. However, the number of quantified miRNAs is very uniform, and is not correlated to the proportion of miRNA reads, and only 8 samples were excluded due having low mapping rate, coverage, or gene count. This indicates that while in some samples sequencing depth is lost on non-miRNA reads, this hardly affects our miRNA detection and quantification. Notably, correlations between miRNA samples were high, and population clustering clearly more pronounced than clustering by laboratory even before normalization.

8. Normalization of quantifications (Fig S8-11, Table S3)

All read count quantifications were corrected for variation in sequencing depth between samples by normalizing the reads to the median number of well-mapped reads (45M) for mRNA, and to the median number of miRNA reads (1.2M) for miRNA. In general, we used only elements quantified in >50% of individuals unless mentioned otherwise (Table S3).

Normalization of expression data in general and RNA-seq data to remove technical biases that can mask or mimic biological effects under study is an area of intensive study (see e.g. [48-51]). Different study designs and questions necessitate different types of normalization: for instance, in characterization of the transcriptome different genes must be compared to each other, which requires sophisticated correction of sequence context biases (e.g. [52]). However, in most applications of RNA-seq the comparisons are across individuals and not across genes (e.g. [48]), and the aim of normalization is to make quantification distributions from different samples better comparable. Differential expression analyses comparing two (or more) groups to each other is one of the most common applications of RNA-seq data, and in this analysis it is important to take the variance pattern of RNA-seq read count quantifications into account, and consider possible confounding technical differences between the groups ([53,54]). In the analysis of population variation e.g. for the purposes of eQTL analysis, the sensitivity to technical variation depends on the scope and statistical methods. While it is possible to correct out measured technical covariates such as sequencing date and GC content ([40]), it has been shown that detecting synthetic covariates from the expression data itself by PCA or related methods such as SVA[49] or PEER [55] can greatly increase the number of eQTL discoveries especially in cis[56]. Some novel methods can simultaneously correct for confounders and map eQTLs[57] (see Fig. S12 and section 11.1), which is beneficial especially in trans-eQTL analysis where one must be careful not to correct out biological variation that is being studied.

Furthermore, after adjusting the variation between samples, the data may need further transformations depending on eQTL analysis methods. Widely used linear regression is sensitive to outliers and assumes a normal distribution of the expression values of each gene across the samples. In RNA-seq data outliers are much more common than in expression array data due to the wider dynamic range, and thus transformation of the expression values to a normal distribution is necessary.

In summary, all expression quantifications are affected by technical variation that reduces power to find biological variation, and the choice of the

normalization method appropriate for the study question and the analysis method is crucial for successful and unbiased analysis of RNA-seq data.

In this study, we removed technical variation for the cis-eQTL and miRNA-mRNA correlation analysis by PEER[56], which finds synthetic covariates from quantification data that can then be regressed out. Correcting out synthetic covariates detected from expression data itself naturally implies that major genome-wide biological effects might be corrected out, which can be a problem in some study settings, such as trans-eQTL analysis. However, our analysis of cis-eQTLs, i.e. local rather than genome-wide effects, is unlikely to be affected by this, and hardly any cis-eQTL signals are lost in PEER correction. Figures S8-9 shows that PEER-normalization helps to find much more eQTLs compared to unnormalized or quantile normalized data, with minimal loss of eQTLs in data with less normalization. For the analysis of mirQTL trans-effects, we also tested a more conservative normalization by regression of only major technical covariates such as sequencing lab and GC content, but this resulted hardly any additional discoveries and altogether much fewer significant signals in the PEER-normalized data, suggesting low power due to high amount of technical variation (data not shown). In 't Hoen et al. (under review) we show that the PEER factors strongly correlate with various technical covariates that were measured from the data, such as mean GC content, which further indicates that the normalization is indeed removing technical variation from the data. Importantly, we also correct out much of the population differences – i.e. a biological source of variation – as this is desirable for our eQTL analysis where population differences would be a potential confounding factor.

Altogether, we concluded that PEER normalization was suitable for our data set and analysis, efficiently removing technical variation and leading to no or minimal loss of biological signals. However, as this data set is used in diverse analyses in subsequent studies with different questions or even different eQTL mapping methods (see Section 11.1), we encourage careful consideration of whether another normalization method should be applied on the raw quantification data.

The PEER normalization was done for the total sample set as follows: First, for each type of quantifications, we estimated the best number of covariates (K) to correct: PEER was ran for a subset of the data (chr20, or chr20-22) using K=0,1,3,5,7,10,13,15,20, the resulting corrected quantifications were transformed to standard normal distribution, and cis-eQTL analysis was performed for each K. The number of genes with an eQTL ($p<10e-8$ and $p<10e-6$) was calculated, since eQTL discovery is a good indicator of power to find biological effects. These results can be seen in Figure S8.

Based on this analysis, we chose K=10 as the number of covariates to correct for, except for transcribed repeats where we did not use PEER correction. To normalize the final data sets, we ran PEER for 20 000 quantification units (e.g. exons or genes), adding the mean to the model. Covariates from this analysis were regressed out from all the quantifications, and the mean was added to the residuals. Correlation of samples after this normalization showed less remaining laboratory effects for mRNA data across the samples and in replicates (Fig. S10, S11). In eQTL analysis and miRNA-mRNA correlation analysis, these quantifications were further transformed to standard

normal distribution, in order to avoid false positive associations due to any outliers in the data (see section 11.1 and Fig. S9).

9. mRNA variation in populations

9.1. Quantitative versus qualitative variation (Fig 1b, S14)

We estimated the contribution of alternative splicing and gene expression on the total transcript abundance variation using approaches in Gonzales-Porta et al. 2012[58]. Briefly, for each gene, the samples are represented in the R^T space using transcript expression levels (T=number of expressed transcripts for this gene), from which we can calculate the total variability (Vt) in this space. Projecting the samples in a model of constant splicing ratios gives us an estimate of expression level variation (Vls). The ratio Vls/Vt estimate the contribution of gene expression in the transcript abundance variability, where Vls/Vt ≈ 1 implies that only gene expression contributes to transcript variability, and Vls/Vt ≈ 0 implies that only transcript usage variation contributes to transcription variability of the gene. In this analysis, we used only protein-coding genes expressed in at least 20 individuals per population with at least two expressed (RPKM >=0.01) transcripts, after verifying that our results were generally robust to differences in total gene expression levels.

We further extended this model to between-population variation. Representing the samples in the space of the transcript expression, between-group variation was computed removing the within-group variation from the total variation. Then all the samples were projected on a line, which represents the model of constant splicing ratios. The between-group variation of these projected points was computed, and the estimator of gene expression level variation between populations is the ratio of between-group variation of the projected points over between- group variation of the original points. A value close to one means that the projection did not remove variation, so gene expression is the one mainly contributing to between-population variation.

9.2. Differentially transcribed genes (Fig. 1c, S14)

In addition to the genome-wide quantitative analysis of transcriptome variation outlined above, we also wanted to identify genes with significantly different expression levels and/or differential transcript structure between populations. Only protein coding genes were used in this analysis.

We performed gene differential expression (DE) analysis using tweeDEseq[59], a method that uses a Poisson-Tweedie family of distributions and is well suited to compare groups with more than 15 samples. 16,583 genes with more than 5 counts per million in at least 1 sample were analyzed in pairwise population comparisons. Genes with FDR < 0.05 and log2 fold change greater than 2 were considered significant.

In order to identify genes with differential transcript structure between populations, we used two independent methods. First, we calculated the ratio of each transcript quantification of the total expression level of the gene, and compared the distributions of these transcript ratios between populations by

Wilcoxon-Mann-Whitney rank sum test to identify transcripts with significantly different relative abundances between population pairs, with p-values of the individual comparisons were adjusted using the Benjamini–Hochberg FDR method. Genes with differential transcript usage were then obtained by mapping those transcripts to the associated gene ids (see Fig. 1c for results). As a second method we used DEXSeq [60] that measures differences in exon usage rather than whole transcripts – thus verifying that any bias in transcript quantifications is not affecting our analysis (see Fig S15 for results).

10. miRNA effects on the transcriptome

10.1. miRNA family and target definition

In the analysis of association of miRNA-mRNA quantifications we used 449 samples with both miRNA and mRNA expression data. For defining miRNA-targets we used the TargetScan version 5.2 predictions.[31] Specifically, we downloaded the seed families of all known miRNAs conserved in vertebrates or mammals, and the corresponding conserved target sites (http://www.targetscan.org/). The target sites were lifted from REFSEQ annotations by mapping the 3'UTR sequences to the hg19 genome and intersecting the coordinates with our merged exon annotations (see mRNA Quantifications). The validity of the lift was confirmed at the sequence level by matching the seed sites of targets with the reverse complement of the miRNA seeds. For quantifying miRNA seed expression, we summed up read counts for all miRNAs with the same TargetScan seed sequences. E.g. the expression of the miR-141/200a seed was found by summing the read counts from hsa-miR-141-3p and hsa-miR-200a-3p. For mRNA expression data, we used the count data for the exon containing the predicted miRNA binding site. Both microRNA and mRNA expression data were corrected for hidden confounding factors with PEER and the resulting residuals were transformed to standard normal (see Normalization of quatifications). The final analysis included 100 microRNA-families and 126,698 exons.

10.2. Integrated analysis of mina and mRNA expression (Fig. 1d, Table S4)

The integrated analysis is based on the globaltest [61] and is further described in (Iterson et al., Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions, submitted). Previously, it was shown that a global test-based integration model is robust and sensitive to identify sets of genes whose expression is affected by copy number[62]. In this context, the globaltest was used for testing of the association of a group of genes – the predicted targets – with a microRNA expression profile. It is specifically designed for the situation of more samples then genes (p>>n). Furthermore, the test overcomes the large multiple testing problem that arises when each target is tested individually for association with a microRNA expression profile. P-values for a set of target mRNAs sharing a predicted miRNA seed sequence were obtained by 100,000 permutations of the sample labels and corrected for

multiple testing using Holm's procedure. Within each set of predicted mRNA targets, P-values for individual associations between expression of predicted mRNA targets and miRNA expression levels were corrected by the Bonferroni multiple testing procedure.

A useful interpretation of the global test is as a sum of squared covariances between a set of predictors $X_{n\times p}$, and responses, $y_{n\times 1}$ (see section 5 of [61]). Consider the sample covariance, $r_{y,x}$ between a miRNA expression profile $y_{n\times 1}$ and a single target $x_{n\times 1}$ given by:

$$r_{y,x} = \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y}_n)(x_k - \bar{x}_n) = \frac{(\mathbf{x}-\bar{\mathbf{x}}_n)^T(\mathbf{y}-\bar{\mathbf{y}}_n)}{n-1},$$

where bar y_n and bar x_n denote the sample means of miRNA and mRNA expression profiles, bar y_n and bar x_n are vectorized versions (note that $r_{y,x} = r_{x,y}$). For multiple mRNA profiles $X_{n\times 1}$ the p×1 vector of the sample covariances, $r_{y,X}$ can be expressed as:

$$\mathbf{r}_{y,\mathbf{X}} = \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y}_n)(X_{kj} - \bar{X}_j) = \frac{(\mathbf{X}-\bar{\mathbf{X}})^T(\mathbf{y}-\bar{\mathbf{y}})}{n-1}.$$

Note that this expression is valid even when the number of targets exceeds the number of samples p > n, and again $r^T_{y,X} = r_{X,y}$. Now the global test test-statistics,

$$\frac{(\mathbf{y}-\bar{\mathbf{y}}_n)^T\mathbf{X}\mathbf{X}^T(\mathbf{y}-\bar{\mathbf{y}}_n)}{(\mathbf{y}-\bar{\mathbf{y}}_n)^T(\mathbf{y}-\bar{\mathbf{y}}_n)} \propto \mathbf{r}^T_{y,\mathbf{X}}\mathbf{r}_{y,\mathbf{X}}$$

is proportional to the squared sample covariance.

10.3. Trans-eQTL effects of cis-mirQTLs (Fig. S16)

Variants that associate to miRNA expression levels (mirQTLs) can potentially be trans-eQTLs for the target genes of these miRNAs. This effect was sought using the European data set, using QTL mapping methods outlined in section 11. The hypothesis was that a mirQTL variant should have a stronger trans-eQTL effect on the predicted targets of the miRNA than on non-targets – note that testing this for targets observed in this data would build a circular argument. This analysis is likely to be conservative since it relies on the accuracy of target predictions.

We selected all miRNA-target exon pairs based on the TargetScan predictions (see above). From these, we selected only those exons that were included in eQTL analysis (expressed in >90% samples), and only the 60 miRNAs that had a mirQTL. This left us with 11 miRNAs to test. For the best-associating variant of each of these mirQTLs, we collected trans association p-values (>5MB from the site) with exons that have a target site of the miRNA affected by the mirQTL (6242 variant-exon pairs in total, 125-1003 exons per mirQTL), and as a null with exons not in genes that have a target site of the miRNA with an mirQTL (9491 variant-exon pairs in total, 134-187 exons per eQTL). We compared these p-value distributions for negative and positive associations separately, i.e. where the cis-mirQTL allele increasing the miRNA expression has negative or positive correlation to the exon.

11. Transcriptome QTL analysis

11.1. Transcriptome QTL mapping (Table 1, Fig. S12)

The details of sample sets, data filtering and normalization are discussed above. Briefly, we did transcriptome QTL mapping separately for European (n=373) and Yoruba (n=89) populations. We used genetic variants with MAF>5% in either EUR or YRI <1MB from transcription start site, with covariates of imputation status (0|1), PCs 1-3 for Europeans and PCs 1-2 for Yoruba. For the different quantitative phenotypes, we used quantification units (e.g. exons) with quantification >0 in >90% of all the individuals unless mentioned otherwise (Table S3). Only autosomes were analyzed. The quantifications were normalized as described in Section 8 by PEER correction and transformation to standard normal distribution, which is essential to avoid false positive due to outliers that can be common especially with RNA-seq data (Fig. S9).

QTLs were mapped using a linear model implemented in Matrix eQTL[63], and FDR was estimated by permutations as follows: For exon eQTLs, we permuted the quantifications of each exon 2000 times, keeping the best p-value per exon from each round. From these data, we adjusted the FDR to 5% with Benjamini-Hochberg according to the most stringent exon of each gene, having a separate p-value threshold for each gene. For miRNAs and RNA editing sites, we ran 8000 permutations for each quantification unit, and calculated a p-value for each of them. For transcript ratio QTLs, we permuted ratios of all transcripts of randomly selected 1000 genes 3000 times and calculated a genome-wide p-value limit based on the median of the most stringent transcript per gene. For gene and repeat eQTLs, we permuted randomly selected 1000 genes and 500 repeats, and used their median as a genome-wide p-value limit.

To verify that the eQTLs discovered with linear regression analysis are robust, we ran eQTL analysis with three additional methods for chr7 gene quantifications of European samples. The tested methods were: 1) ANOVA, i.e. another straightforward statistical model comparable to linear regression, implemented in Matrix eQTL. The test was ran on the exact same genotype, expression and covariate data as linear regression. 2) The TReC[64] model that is specifically designed for RNA-seq data and models the variance in quantifications derived from counts. The test was ran on gene read counts, with the total number of mapped reads as an additional covariate. Running permutations to estimate FDR was not computationally feasible, and thus we analyzed the p-value distribution without a significance cutoff. 3) eQTL mapping method PANAMA[57] that captures gene expression heterogeneity by fitting a joint random effect covariance matrix rather than calculating residual expression levels explicitly (as we do with PEER, see section 8). The maximum number of hidden factors for training was set to 40 and the fitting process was carried out on all 13,703 genes. To satisfy the Gaussian assumptions underlying PANAMA, we employed an Anscombe transformation to approximately variance stabilize the gene expression profiles prior to fitting. The final covariance matrix determined by PANAMA was used in a mixed model association analysis to test for *cis*-eQTLs in 645 genes in chromosome 7. The covariates population ID and sequencing center ID were included as covariates in the analysis. Significance

levels of association within the *cis* region of each gene were determined by comparison with an empirical null distribution estimated from 1,000 permutations.

The results from these analyses (Fig S12) show an overall good concordance between the methods. Linear regression and ANOVA give nearly identical results, and also TReC and PANAMA replicate our linear regression –based eQTLs well. PANAMA finds also additional associations, likely due to more efficient control of confounders, but we chose to use a more estblished linear regression in the main analysis of this study. Altogether, cis-eQTL analysis appears robust to the choice of the statistical method.

11.2. Transcript ratio QTL effects (Fig. 2b)

For the transcript ratio QTLs (trQTLs), we sought to characterize the QTL effect on transcript usage. For each trQTL gene, we identified the transcript with highest association and the transcript with most negatively correlated quantifications to this. Given the annotation of these two transcripts, AStalavista[65,66] was used to classify the events for each trQTL.

11.3. Independence of QTLs (Fig. S17-18)

To estimate independent QTL signals for the same gene, we used an approach where the linear regression QTL analysis is reran using a previous association signal as a covariate – in cases where a second variant is not the same and not linked to the first one, an association signal for the gene should remain.

We applied this to estimate the number of exons with independent eQTLs from the best association for all the exons of the gene. Additionally, for 279 genes that had both a significant transcript ratio (trQTL) and a gene eQTL, we reran the eQTL analysis with the best trQTL variant as a covariate to estimate whether the trQTL signal is driving the eQTL association as well.

11.4. Null variant distribution

To compare QTL variants to a null distribution of similar variants but without regulatory association, we sampled genetic variants in cis-regions of genes expressed in our data set based on the QTL variant distributions of distance from the gene (taking upstream and downstream distance into account) and minor allele frequency. We also tried matching for the coding/noncoding status of the variants, but did not use this in the final analysis since it did not appear to have a major impact in the results.

11.5. Functional overlap of eQTLs (Fig. 2a, S20-24)

We analyzed the overlap of eQTL variants as well as the matched null variants in different functional categories according to our functional annotation of the variants, as described in section 3.1. Furthermore, we linked our eQTL findings to two earlier data sets of functional genomics data:

First, we analyzed DNase1 sensitivity QTLs (dsQTLs) [67] from their "long" list of dsQTLs, comparing the intersection of these variants with our best eQTLs

per gene (based on p-value) and with the null set of variants matched to eQTL properties (see above).

Second, we analyzed allele-specific binding in CTCF. Raw ChIP-seq data for CTCF binding in two parent-offspring trios from the 1000 Genomes pilot project[68] was obtained from McDaniell *et al*.[69]. All subsequent CTCF and ChIP-seq related methods are described in detail in Kilpinen *et al.* submitted. Briefly, 36 bp single-end reads were mapped against the hg19 build of the human reference genome using BWA[70], and reads for biological replicates were merged after mapping. For peak calling, final mapped reads (MAPQ>=10) from the six trio individuals (NA12878, NA12891, NA12892, NA19238, NA19239, and NA19240) were pooled, and peaks called from this metasample using HOMER[71], excluding duplicate reads. Called peaks were extended to the expected fragment length of 200 bp. Allele-specific (AS) analysis of CTCF binding was based on binomial testing of allelic ratios over heterozygous SNP sites of each individual, similarly to ASE analysis described in section 12.1. We required both alleles to be observed in the data, a minimum coverage of 10 reads per site, included only SNP sites located within CTCF metasample peaks, and filtered SNPs with unreliable mapping as in section 12.1. We also applied two simulation-based filtering steps to exclude individual SNP with low complexity library artifacts (Waszak et al. submitted). From these data, we extracted the sites that were our top eQTLs, or part of the matched null variants, and compared the signal of allele-specific binding.

11.6. Causal regulatory variant estimation (Fig. S23)

We estimated the probability of the best associating EUR and YRI eQTL variants being the causal regulatory variants by comparing the annotation enrichment of all loci to enrichment in those that are very likely to be causal. Specifically, we first calculated the annotation enrichment $c_{all,y}$ of the best eQTLs relative to the matched null across all eQTL loci, separately for each annotation class *y*. Then, we defined a subset of eQTLs where the best eQTL is likely to be causal: we binned eQTLs according to –log10 p-value difference between the first and the second variant (Δp), hypothesizing that for very large Δp, the first variant can be safely declared as the causal variant. To determine the Δp threshold where this point is reached, we calculated the annotation enrichment between the 1st and the 2nd variant b_y for eQTLs in each Δp bin. In both EUR and YRI, b_y saturates at Δp = 1.5, similarly in all annotation categories; thus, we reasoned that eQTLs with Δp>1.5 can be used to estimate the amount of annotation enrichment $c_{causal,y}$ for the eQTLs where the best variant is causal. Finally, from these data, we calculated the proportion of all eQTL loci where the 1st variant is causal as $p_y = (c_{all,y} - 1) / (c_{causal,y} - 1)$.

11.7. GWAS overlap of eQTLs (Fig. S26-27, 2d, Table S5)

We first estimated a simple overlap with exon eQTL variants and 6473 published GWAS SNPs [72] that were part of the 1000 Genomes Phase 1 data set. As a null, we collected 14 000 variants matched to the minor allele frequency spectrum of the GWAS variants. To estimate whether the GWAS overlap is particularly pronounced in the top eQTLs which would be expected if the causal variant is the

same, for each GWAS variant (and the null set) that overlapped significant eQTLs, we calculated the highest eQTL rank.

The large number of significant eQTL variants and GWAS variants gives a large overlap even under the null, and with genome sequencing data we are testing a very different set of variants than the GWAS studies. This makes it challenging to identify those GWAS SNPs that are truly driven by an eQTL signal. To this end, we used a published dataset of 1213 GWAS SNPs that have been statistically shown by the RTC method to be likely to tag the same causal variant as an eQTL signal[73,74]. From these data, we extracted 91 GWAS-eQTL SNPs where both the original GWAS variant and the eQTL variant were found in our data, and the recombination interval containing the original eQTL and the GWAS SNP contains a significant eQTL in our EUR data that is the strongest association for that gene. For these GWAS variants, we report the top eQTL variants in our study as putative causal GWAS variants (Table S5).

12. Allele-specific analysis

12.1. Allele-Specific Expression (ASE) (Fig. 3, S2, S28-29, S31-33)

Allele-specific expression analysis was based on binomial testing of each allelic ratio of heterozygous sites within each individual. First, we excluded sites that are susceptible to allelic mapping bias: 1) sites with 50bp mapability < 1 based on the UCSC mapability track, implying that the 50bp flanking region of the site is non-unique in the genome, and 2) simulated RNA-seq reads overlapping the site show >5% difference in the mapping of reads that carry the reference or non-reference allele (Fig S29; see also section 4.1). In all the analyses, we only used uniquely mapping reads (mapping quality >150), NM>=6, and sites with base quality >10.

Next, we calculated the expected reference allele ratio for each individual by summing up reads across all sites separately for each SNP allele combination after down-sampling reads of sites in the top 25th coverage percentile in order to avoid the highest covered sites having a disproportionally large effect on the ratios. These expected REF/TOTAL ratios correct for any remaining genome-wide mapping bias as well as GC bias in each individual (Fig. S28).

Finally, for all the sites covered by >=8 reads in each individual, we calculated a binomial test of the REF/NONREF allele counts, using the expected ratio described above. Except for the NMD analysis (see below), we used only sites with >=16 reads, and sites where both alleles are observed in RNA-sequencing data in order to verify that the genotype is a true heterozygote (Fig. S28).

In many analyses, differing coverage between sites creates noise due to difference in power to call ASE. To correct for this, in many analyses we used only sites with >=30 reads (Fig. S28), and sampled all sites to exactly 30 reads. In a further analysis of ASE differences between individuals, we calculated allelic expression distances between all sample pairs as the median of absolute REF/TOTAL ratio differences of all the shared heterozygous sites between individuals after sampling the reads to 30.

12.2. Allele-Specific Transcript Structure (ASTS) (Fig. 3, S2, S30, S33)

Allele-specific transcript structure (ASTS) is a novel sister method of ASE, and aims at detecting differences in transcripts between the two haplotypes of an individual. As in ASE, we look at reads overlapping heterozygous coding sites, and the allele of this site in the RNAseq data tells the haplotype origin of each read fragment. The distribution of the reads to exons is then quantified.

For every sample, we first retrieved all heterozygous sites that are covered by >= 20 RNAseq reads, after mapability filter as in ASE analysis. Using the pysam package (http://code.google.com/p/pysam/), we scanned the bam file to extract all the reads and their mates that overlap the site, separated them to reads with REF or ALT allele, and printed out a pseudo-sam file that contains information of which SNP each read overlaps, and if it carries the REF or NONREF allele.

For this file, we ran our standard exon quantification, and calculated the number of REF and ALT read overlaps in all the exons. We kept only exons with >=10 reads of each allele, and required a total of >=20 REF and NONREF reads in the remaining exons. We used Fisher test to estimate whether the read counts in exons are different for REF and NONREF reads. For each site, we calculated a quantitative measure analogous to ASE allelic ratio (maximum imbalance across all exons of a site compared to the total REF/NONREF ratio).

12.3. Mapping regulatory variants with ASE data: Method (Fig. S34)

Differential expression of the two haplotypes of an individual, or allele-specific expression, is believed to be often driven by the individual being heterozygous for a regulatory variant elsewhere in the *cis*-regulatory region (Fig. S2). Here, we developed a novel method to map regulatory variants, rSNPs (SNP used for brevity; these variants can be of other types as well) that affect allelic ratios, with significant improvements to the basic principle that we published before [75]. In these analyses the genotype of the variant (aseSNP) over which ASE is calculated is always heterozygous since otherwise ASE cannot be measured; thus all the references to the genotype in the following refer to the putative regulatory variant (prSNP).

The method is based on finding maximal concordance between the allelic ratios of an aseSNP, and the genotypes of the prSNPs in the surrounding region: for a true rSNP, we would expect heterozygote individuals to have large deviation of the null allelic ratio of 0.5, whereas homozygotes would be expected to have ratio close to 0.5. Specifically, for each aseSNP, we define r_i = ref_count$_i$ / total_count$_i$ as the reference allele ratio in individual *i* , and d_i = abs(0.5 – r_i) as the distance of the ratio from 0.5. For each individual, for each prSNP–aseSNP pair we can calculate a concordance score s_i , with $s_{i,het} = d_i^2/0.5^2$, and $s_{i,hom} = 1 - d_i^2/0.5^2$ (Fig. S34). Having phased data, we also took allelic direction into account as follows: for double heterozygotes, we calculate which prSNP allele is linked to the higher expressed aseSNP allele, and from the individuals with significant ASE ($p<0.005$), we calculated which direction is most commonly observed, and assign this as the majority direction. If there were no double heterozygotes with significant ASE, we used the direction of the individual with the lowest ASE p-

value as the majority direction. For every $s_{i,het}$ individual that has the opposite direction, we assigned $s_{i,het} = 0$, thus penalizing switches in allelic direction.

For each prSNP-aseSNP pair, we calculated score *s* as the average of s_i. We evaluated the significance of *s* by permuting the genotypes of the prSNPs and recalculating the scores. We did as many permutations as we have unique genotype combinations, here requiring at least 100 and up to 1000 permutations, thus obtaining an empirical p-value for *s*. The absolute value of s ([0,1]) is not correlated to its p-value (Fig S34).

12.4. Mapping regulatory variants with ASE data: Analysis (Fig. 3, S35-36)

We applied the prSNP method outlined above to our data of allele-specific expression. We filtered the data stringently to use only high-coverage aseSNPs sampled to 30 reads in each individual, requiring >=80 individuals with ASE data of which >1 significant ASE ($p<0.005$). Since allelic ratio is a comparison within an individual rather than a quantitative measurement in the population, this analysis is insensitive to population stratification, and we analyze all our samples together. We tested all prSNPs within 100 kb from the TSS of the gene of the aseSNP From these data, we assigned as likely rSNPs all the variants with no permuted values higher than the observed one, i.e. $p < 0.01$ to $p<0.001$ given our 100-1000 permutations. For each aseSNP separately, we calculated how many rSNPs we would expect to find by chance based on the p-values of the rSNPs and the number of tested prSNPs, and only when the observed number is greater than the expected we count the aseSNP as one with an rSNP signal. Furthermore, to analyze the properties of rSNPs, we sampled 5 rSNPs per aseSNP (or all if <5). The results for taking only a single variant were similar (data not shown) but we chose to use multiple variants per locus since our empirical p-values do not allow good ranking of rSNPs to obtain the best variant as in eQTL analysis. To compare these variants to a null set of nonsignificant SNPs, we sampled similar numbers of SNPs with $p>0.2$.

13. Loss-of-function analysis

13.1. Nonsense-mediated decay (Fig. 4, S33, S37)

To estimate the signal of nonsense-mediated decay (NMD) in premature stop variants, we quantified ASE using allelic read count data from individuals who are heterozygous for a premature stop variant, compared to other individuals where we have ASE data from the same gene as the ASE variant. We applied an EM algorithm to fit a mixture of binomial distributions where number of components, k, was set to 2, and no prior information was given for the binomial distribution parameters. The EM algorithm was run until epsilon < 1e-8; final number of iterations = 20. This was ran for all variants and for rare (minor allele counts = 1-10) premature stop variants.

13.2. Splice scores (Fig. 4, S38)

For the 1000 Genomes Phase 1 SNPs and indels that modify the splice site motif, we computed log-odd scores of variant effect in splice motifs employing the 1st order Markov Models for splice donor and acceptor sites of human U2-dependent introns from the gene prediction program GeneID[76]. The scoring has been applied to the ~478,000 splice sites currently included in the Gencode v12 reference annotation.

Splice site variants have been inferred from the 1000 relevant for the Markov model.

14. Data access

14.1. Data files (Fig. S39)

The Geuvadis data is openly accessible, and full links to all sites can be found in the project website www.geuvadis.org. Detailed documentation of the methods can be found from this paper, and in the readme files. Additional analysis results can be found in the openly accessible project wiki in geuvadiswiki.crg.es

The main accession site to the data created and analyzed by the Geuvadis RNA-sequencing project is EBI ArrayExpress, where the data is stored under three accessions: E-GEUV-1 for mRNA post-QC samples used in analyses of this paper, E-GEUV-2 for small RNA post-QC samples, and E-GEUV-3 for all the sequenced data.

Raw reads in the form of fastq files are stored in ENA under the accession ERP001942 and ERP001941, accessible also through ArrayExpress. mRNA mapped reads are stored and accessible from EBI ArrayExpress, and the bam files can be viewed in the Ensembl browser through links in ArrayExpress. Files of mapped small RNA reads are not provided due to the more complex nature of mapping to different references for different analytical purposes and the large number of multimapping reads making file sizes very large.

Genotype data that have been used in Geuvadis data analysis are available from EBI ArrayExpress site under accession E-GEUV-1. The original data created by 1000 Genomes Project are available in the 1000 Genomes web site.

Geuvadis analysis results for gene, transcript, exon and repeat quantifications and QTLs are available from EBI ArrayExpress site under accession E-GEUV-1, and corresponding data for miRNA is under accession E-GEUV-2.

14.2. The Geuvadis Data Browser (Fig. S40)

For the visualisation of RNA-sequencing analysis we created the Geuvadis Data Browser (www.ebi.ac.uk/Tools/geuvadis-das). It is powered by the Genoverse browsing engine running HTML5 and Javascript and co-developed by the Ensembl and DECIPHER projects. The back-end for the browser is the EBI data sources providing the Geuvadis analysis data in real-time.

The Geuvadis RNA-sequencing analysis results consist of following tracks:

- EUR and YRI exon eQTLs and quantifications

- EUR and YRI transcript quantifications and transcript ratio QTLs,
- EUR and YRI mirQTLs and quantifications

Quantification tracks show the population minimum, average and maximum values of raw counts normalised by library size and element lengths, very similar to FPKM normalization. By clicking on the element of interest it is possible view information about each element: description, scoring information and links to other relevant data sources, for instance Ensembl Genome Browser.

QTL tracks show SNPs and indels associated with functional effects. In a similar way a click on an element of interest will provide additional information including all linked effect elements associated with eQTL along with related p-values.

Tracks at the top of the Geuvadis Data Browser provide gene and transcript element annotations. These tracks are based on Ensembl latest release of human genome GRCh37 and are given for the reference purposes. It is possible to search for genes, variants or locations, and a region selector tool allows viewing the underlying data values of the selected, which can then be saved by the user.

15. References to Supplementary Methods

28 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:btp352 [pii]
10.1093/bioinformatics/btp352 (2009).

29 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:nature11632 [pii]
10.1038/nature11632 (2012).

30 Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**, 817-825, doi:nbt.1662 [pii]
10.1038/nbt.1662 (2010).

31 Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20, doi:S0092867404012607 [pii]
10.1016/j.cell.2004.12.035 (2005).

32 Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**, D152-157, doi:gkq1027 [pii]
10.1093/nar/gkq1027 (2011).

33 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, doi:10.1371/journal.pgen.1000529 (2009).

34 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909, doi:ng1847 [pii]
10.1038/ng1847 (2006).

35 Marco-Sola, S., Sammeth, M., Guigo, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods*, doi:nmeth.2221 [pii]
10.1038/nmeth.2221 (2012).
36 Pickrell, J. K., Gaffney, D. J., Gilad, Y. & Pritchard, J. K. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics* **27**, 2144-2146, doi:btr354 [pii]
10.1093/bioinformatics/btr354 (2011).
37 Pradervand, S. *et al.* Concordance among digital gene expression, microarrays, and qPCR when measuring differential expression of microRNAs. *Biotechniques* **48**, 219-222, doi:000113367 [pii]
10.2144/000113367 (2010).
38 Linsen, S. E. *et al.* Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* **6**, 474-476, doi:nmeth0709-474 [pii]
10.1038/nmeth0709-474 (2009).
39 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:22/9/1760 [pii]
10.1101/gr.135350.111 (2012).
40 Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-777, doi:nature08903 [pii]
10.1038/nature08903 (2010).
41 Shapiro, I. M. *et al.* An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genet* **7**, e1002218, doi:10.1371/journal.pgen.1002218
PGENETICS-D-10-00244 [pii] (2011).
42 Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476, doi:nature07509 [pii]
10.1038/nature07509 (2008).
43 Buermans, H. P., Ariyurek, Y., van Ommen, G., den Dunnen, J. T. & t Hoen, P. A. New methods for next generation sequencing based microRNA expression profiling. *BMC Genomics* **11**, 716, doi:1471-2164-11-716 [pii]
10.1186/1471-2164-11-716 (2010).
44 Pantano, L., Estivill, X. & Marti, E. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res* **38**, e34, doi:gkp1127 [pii]
10.1093/nar/gkp1127 (2010).
45 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-467, doi:84979 [pii]
10.1159/000084979 (2005).
46 Lestrade, L. & Weber, M. J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res* **34**, D158-162, doi:34/suppl_1/D158 [pii]
10.1093/nar/gkj002 (2006).
47 Berninger, P., Gaidatzis, D., van Nimwegen, E. & Zavolan, M. Computational analysis of small RNA cloning data. *Methods* **44**, 13-21, doi:S1046-2023(07)00176-4 [pii]

10.1016/j.ymeth.2007.10.002 (2008).
48 Hansen, K. D., Wu, Z., Irizarry, R. A. & Leek, J. T. Sequencing technology does not eliminate biological variability. *Nat Biotechnol* **29**, 572-573, doi:nbt.1910 [pii]
10.1038/nbt.1910 (2011).
49 Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**, 733-739, doi:nrg2825 [pii]
10.1038/nrg2825 (2010).
50 Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**, 1509-1517, doi:gr.079558.108 [pii]
10.1101/gr.079558.108 (2008).
51 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628, doi:nmeth.1226 [pii]
10.1038/nmeth.1226 (2008).
52 Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38**, e131, doi:gkq224 [pii]
10.1093/nar/gkq224 (2010).
53 Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25, doi:gb-2010-11-3-r25 [pii]
10.1186/gb-2010-11-3-r25 (2010).
54 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106, doi:gb-2010-11-10-r106 [pii]
10.1186/gb-2010-11-10-r106 (2010).
55 Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500-507, doi:nprot.2011.457 [pii]
10.1038/nprot.2011.457 (2012).
56 Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* **6**, e1000770, doi:10.1371/journal.pcbi.1000770 (2010).
57 Fusi, N., Stegle, O. & Lawrence, N. D. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput Biol* **8**, e1002330, doi:10.1371/journal.pcbi.1002330
PCOMPBIOL-D-11-01209 [pii] (2012).
58 Gonzalez-Porta, M., Calvo, M., Sammeth, M. & Guigo, R. Estimation of alternative splicing variability in human populations. *Genome Res* **22**, 528-538, doi:gr.121947.111 [pii]
10.1101/gr.121947.111 (2012).
59 tweeDEseq: RNA-seq data analysis using the Poisson-Tweedie family of distributions. v.1.0.14.

60 Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**, 2008-2017, doi:gr.133744.111 [pii]
10.1101/gr.133744.111 (2012).
61 Goeman, J. J., van de Geer, S. A., de Kort, F. & van Houwelingen, H. C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**, 93-99 (2004).
62 Menezes, R. X., Boetzer, M., Sieswerda, M., van Ommen, G. J. & Boer, J. M. Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics* **10**, 203, doi:1471-2105-10-203 [pii]
10.1186/1471-2105-10-203 (2009).
63 Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-1358, doi:bts163 [pii]
10.1093/bioinformatics/bts163 (2012).
64 Sun, W. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* **68**, 1-11, doi:10.1111/j.1541-0420.2011.01654.x (2012).
65 Sammeth, M., Foissac, S. & Guigo, R. A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol* **4**, e1000147, doi:10.1371/journal.pcbi.1000147 (2008).
66 Foissac, S. & Sammeth, M. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res* **35**, W297-299, doi:gkm311 [pii]
10.1093/nar/gkm311 (2007).
67 Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390-394, doi:nature10808 [pii]
10.1038/nature10808 (2012).
68 Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:nature09534 [pii]
10.1038/nature09534 (2010).
69 McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**, 235-239, doi:science.1184655 [pii]
10.1126/science.1184655 (2010).
70 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:btp324 [pii]
10.1093/bioinformatics/btp324 (2009).
71 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:S1097-2765(10)00366-7 [pii]
10.1016/j.molcel.2010.05.004 (2010).
72 Hindorff, L. A., Junkins, H. A., Hall, P. N., Mehta, J. P. & Manolio, T. A. A Catalog of Published Genome-Wide Association Studies. *www.genome.gov/gwastudies* (2010).
73 Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**, 1084-1089, doi:ng.2394 [pii]
10.1038/ng.2394 (2012).
74 Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* **6**, e1000895, doi:10.1371/journal.pgen.1000895 (2010).

75 Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E. T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet* **7**, e1002144, doi:10.1371/journal.pgen.1002144 PGENETICS-D-10-00589 [pii] (2011).

76 Blanco, E., Parra, G. & Guigo, R. Using geneid to identify genes. *Curr Protoc Bioinformatics* **Chapter 4**, Unit 4 3, doi:10.1002/0471250953.bi0403s18 (2007).

Supplementary Tables

			QC-passed			
			mRNA		miRNA	
Pop	**Full name**	**Total sequenced samples**	**Total**	**1000G Phase1**	**Total**	**1000G Phase1**
CEU	Utah residents (CEPH) with Northern and Western European ancestry	92	91	78	87	74
FIN	Finnish from Finland	95	95	89	93	87
GBR	British from England and Scotland	96	94	85	94	84
TSI	Toscani in Italia	93	93	92	89	88
YRI	Yoruba in Ibadan, Nigeria	89	89	77	89	77
TOT	**Total**	**465**	**462**	**421**	**452**	**410**

Table S1. Samples

Numbers of sequenced individuals. Replicate samples are not included in the counts.

Coding annotation (nonredundant hierarchy)	**Variants**
SPLICE_DONOR_VARIANT	4036
SPLICE_ACCEPTOR_VARIANT	2977
STOP_GAINED	6483
FRAMESHIFT_VARIANT	1186
STOP_LOST	581
INITIATOR_CODON_CHANGE	1034
INFRAME_CODON_GAIN	193
INFRAME_CODON_LOSS	531
NON_SYNONYMOUS_CODON	305959
SPLICE_REGION_VARIANT	53901
INCOMPLETE_TERMINAL_CODON_VARIANT	29
SYNONYMOUS_CODON	197584
STOP_RETAINED_VARIANT	253
CODING_SEQUENCE_VARIANT	31
COMPLEX_CHANGE_IN_TRANSCRIPT	97
MATURE_MIRNA_VARIANT	432
5_PRIME_UTR_VARIANT	101725
3_PRIME_UTR_VARIANT	381972
INTRON_VARIANT	19734371
NC_TRANSCRIPT_VARIANT	190673
Noncoding annotation (redundant)	**Variants**
MIRNA_TARGET	3324
TFMOTIF	50282
REG_FEATURE	7325520
ACTIVE_CHROM	38137117
MIRNA_MATURE	652
MIRNA_PRECURSOR	1290
NOVEL_SPLICE	431
No annotation	1027762

Table S2. Variant annotations

Numbers of variants in annotation categories.

	in >50% of samples	In QTL analysis
Genes	16433	13703
Transcripts	67603	24109
Exons	148001	122893
Splice junctions	132647	NA
Transcribed repeats	47437	43875
miRNAs	715	644

Table S3. Quantifications

Numbers of quantified transcriptome features. Gene, transcript, exon and annotated splice junction counts are from protein-coding and lincRNA genes. All eQTL counts are for autosomal genes, with a filter of quantification in >90% of samples for genes, exons, transcripts, and >50% for miRNAs and transcribed repetitive elements.

Tables S4-S5 are available as separate files:

Table S4. Associated miRNA-mRNA pairs (legend)

List of 36 significant (P<0.001, Holm) miRNA families and their associated mRNA targets (P<0.05, Bonferroni). The column descriptions are:

- Exon (exon identifier consisting of Ensembl gene id, chrom location, start and end exon containing the predicted microRNA binding site; exons are unions of all overlapping exons of the same gene)
- microRNA family: family of microRNAs with identical seed-regions
- P-value (of set): P-value indicating the strength of association of the microRNA expression profile with the set of predicted targets
- P-value (target): P-value indicating the target's individual contribution to the overall strength of association to the set
- Association: '0' indicates negative association of the microRNA expression profile with the predicted targets and '1' positive association
- Entrez Gene: Entrez gene identifier
- Gene Symbol: HGNC gene symbol

Table S5. Predicted causal GWAS variants (legend)

GWAS variants that have a signal of a shared causal variant with an eQTL (see Supplementary Methods), and the eQTL p-values of the top eQTL variants and the GWAS SNP.

	All genes			Without eQTL genes		
	Total	**Passed**	**% of total**	**Total**	**Passed**	**% of total**
Genes	2766	2674	96.7%	637	611	95.9%
aseSNPs	5479	5216	95.2%	987	925	93.7%
prSNPs	3044486	224640	7.4%	815659	37563	4.6%
prSNP-aseSNP	8677881	345750	4.0%	1351591	39421	2.9%

Table S6. Regulatory SNPs mapped using ASE data

Statistics of putative regulatory SNPs mapped based on ASE data, grouped by genes where the tested aseSNPs are located, tested aseSNPs, nonredundant prSNPs (putative regulatory SNPs), and finally all tested prSNP-aseSNP pairs. We show the statistics for all the data, and only for genes where no eQTLs were found in this study.

All rSNP-aseSNP pairs where no higher permutated scores were found are considered to be likely true rSNPs (empirical $p<0.01$ to $p<0.001$). For aseSNPs, the number here indicates the number of aseSNPs where the number of passed rSNPs is >0 and bigger than what is expected by chance, given the p-values of the passed rSNPs and the the total number of tested prSNPs for that aseSNP. For genes, we give the number of genes with any passed aseSNPs.