Hypothesis-free identification of modulators of genetic risk factors

**Supplementary Figures**

Figure S1. Number of independent eQTLs per gene, exon, exon ratio and polyA ratio

Figure S2. Heatmap of expression of covariate genes per module in the BLUEPRINT data

Figure S3. Expression of module eQTL genes in BLUEPRINT data and their replication in previously reported cell-type-specific datasets

Figure S4. Interaction plots for various diseases

Figure S5. Combining gene- and exon-level eQTLs can help to explain gene expression regulation

Figure S6. Genotype concordance and heterozygosity rate per sample for each biobank

Figure S7. Biotypes of *cis-*regulated genes

Figure S8. Overlap of different levels of eQTLs

Figure S9. Distance of the eSNP relative to TSS

Figure S10. Distance from eSNP to the exon middle for exon ratio and exon-level eQTLs

Figure S11. Disease categories for trait-associated eQTLs

Figure S12. Distribution of exon ratio QTLs around splice donor or acceptor sites

Figure S13. Classification of splicing events affected by exon ratio QTLs

Figure S14. An example of an exon ratio QTL overlapping with a hit in the GWAS catalogue

Figure S15. Distance of polyA ratio QTLs to annotated polyA sites

Figure S16. An example of a polyA ratio QTL overlapping with a hit in the GWAS catalogue

Figure S17. Plot of an eQTL having opposite allelic directions in BBMRI and Geuvadis

Figure S18. Independent effects in GEUVADIS dataset and in a random subset of BIOS of the same size

Figure S19. Histogram of RNA blood sampling age distribution per biobank

Figure S20. Plots of picard metrics results and the outlier samples removed based on these metrics

Figure S21. Accuracy of cell count prediction method

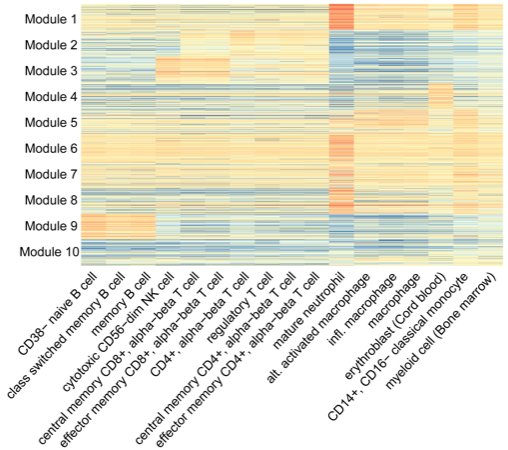
# Figure S1. Number of independent eQTLs per gene, exon, exon ratio and polyA ratio

These bar plots show the number of unique eQTL features having primary, secondary, tertiary and further effects for gene-level eQTLs (A), exon-level eQTLs (B), exon ratio QTLs (C) and polyA ratio QTLs (D). Panels (E), (F), (G) show the number of unique genes for exon-level, exon-ratio and polyA-ratio QTL mapping. In red the number of genes *cis*-regulated by a GWAS hit or its close proxy (r2 ≥ 0.8) is shown.

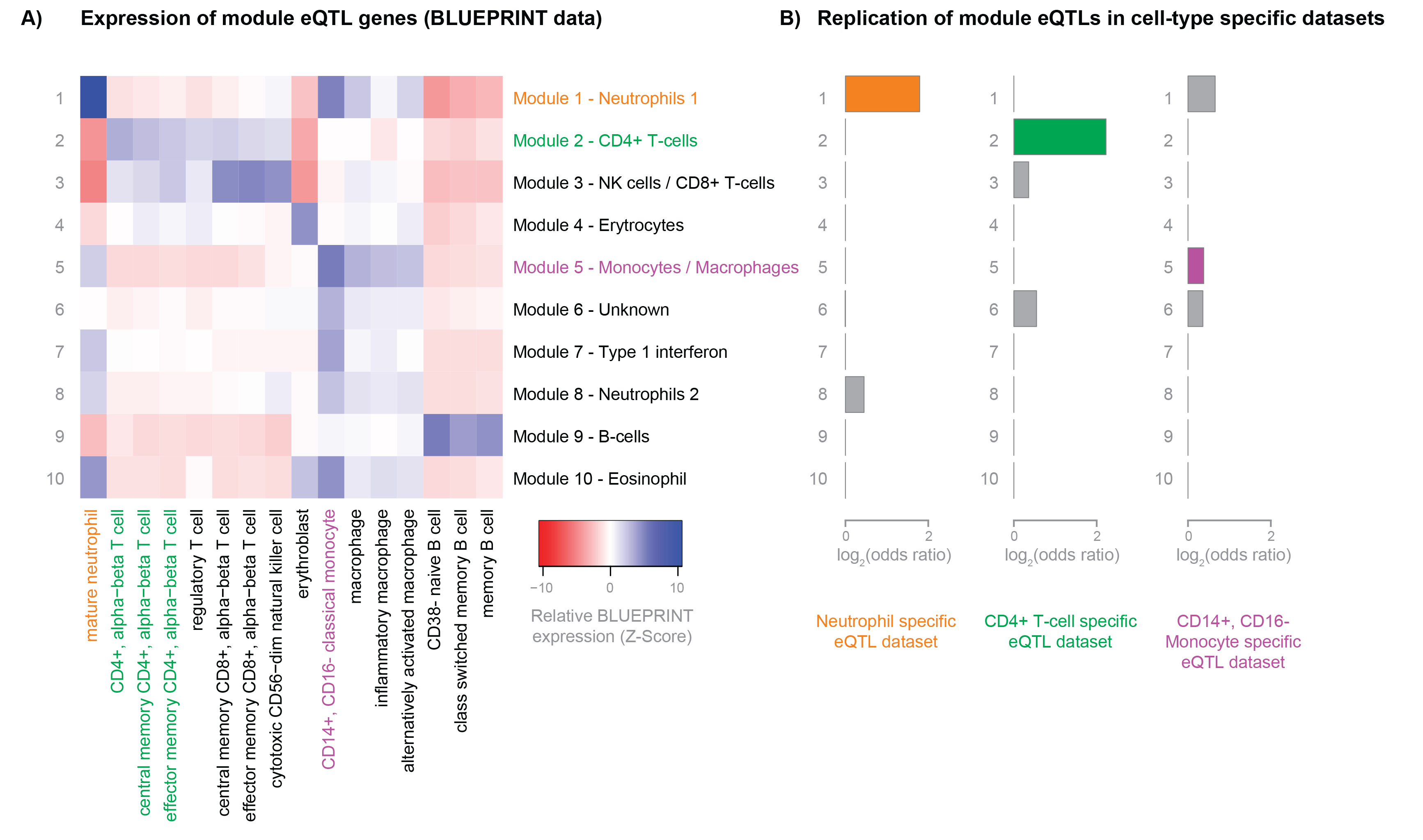
A description...

# Figure S2. Heatmap of expression of covariate genes per module in the BLUEPRINT data

When clustering the module eQTL genes subset using BLUEPRINT RNAseq data on several different blood cell types, different modules show cell-type-specificity. For example, Module I eQTL genes are highly expressed in mature neutrophils and Module 9 genes are mainly expressed in B-cells.



# Figure S3. Expression of module eQTL genes in BLUEPRINT data and their replication in previously reported cell-type-specific datasets

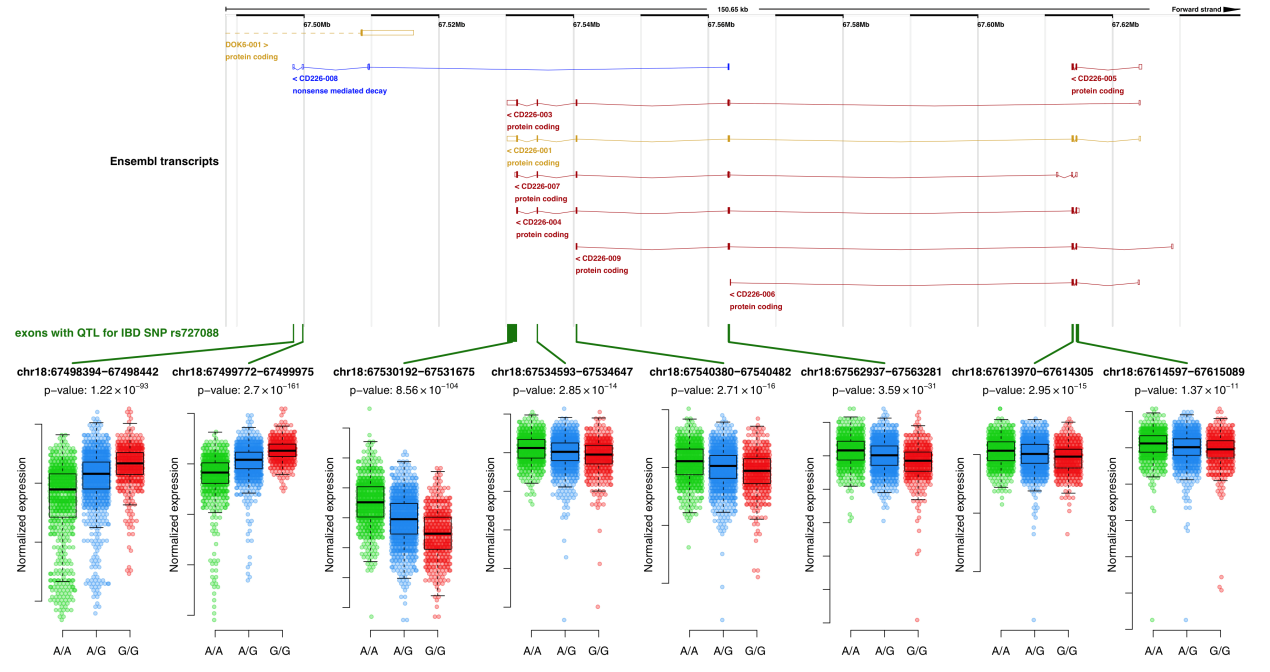
(A) For each module we selected the genes with an eQTL affected by this module. We looked up the expression of these genes within the relevant BLUEPRINT cell types and used a t-test to compare the expression within each cell type to the other cell types. The z-score of each test is shown on the heatmap. (B) We show the overlap with cell-type-specific eQTLs reported previously in cell sorted datasets (neutrophil-specific eQTLs: Naranbhai *et al.*, Nat. Commun*.* 2015; T-cell and monocyte-specific eQTLs: Raj *et al*., Science, 2014). eQTLs significantly interacting with the 10 modules and their proxies with r2 > 0.8 in the GoNL dataset were used to calculate the overlap. Fisher exact test was performed to compare the replication of each module to other modules. log2 transformed odds ratios of enrichments are plotted (significant effects in color, non-significant effects in grey). We saw that neutrophil-specific eQTLs reported in Naranbhai et al. indeed show a strong enrichment among module 1 eQTLs (Fisher exact p-value = 4 x 10-22); CD4+ T-cell eQTLs replicate better in module 2 eQTLs, as compared to eQTL in other modules (p-value = 5.57 x 10-12).Replication in monocytes did not yield significant results for monocyte / macrophage module 5.

# Figure S4. Interaction plots for various diseases

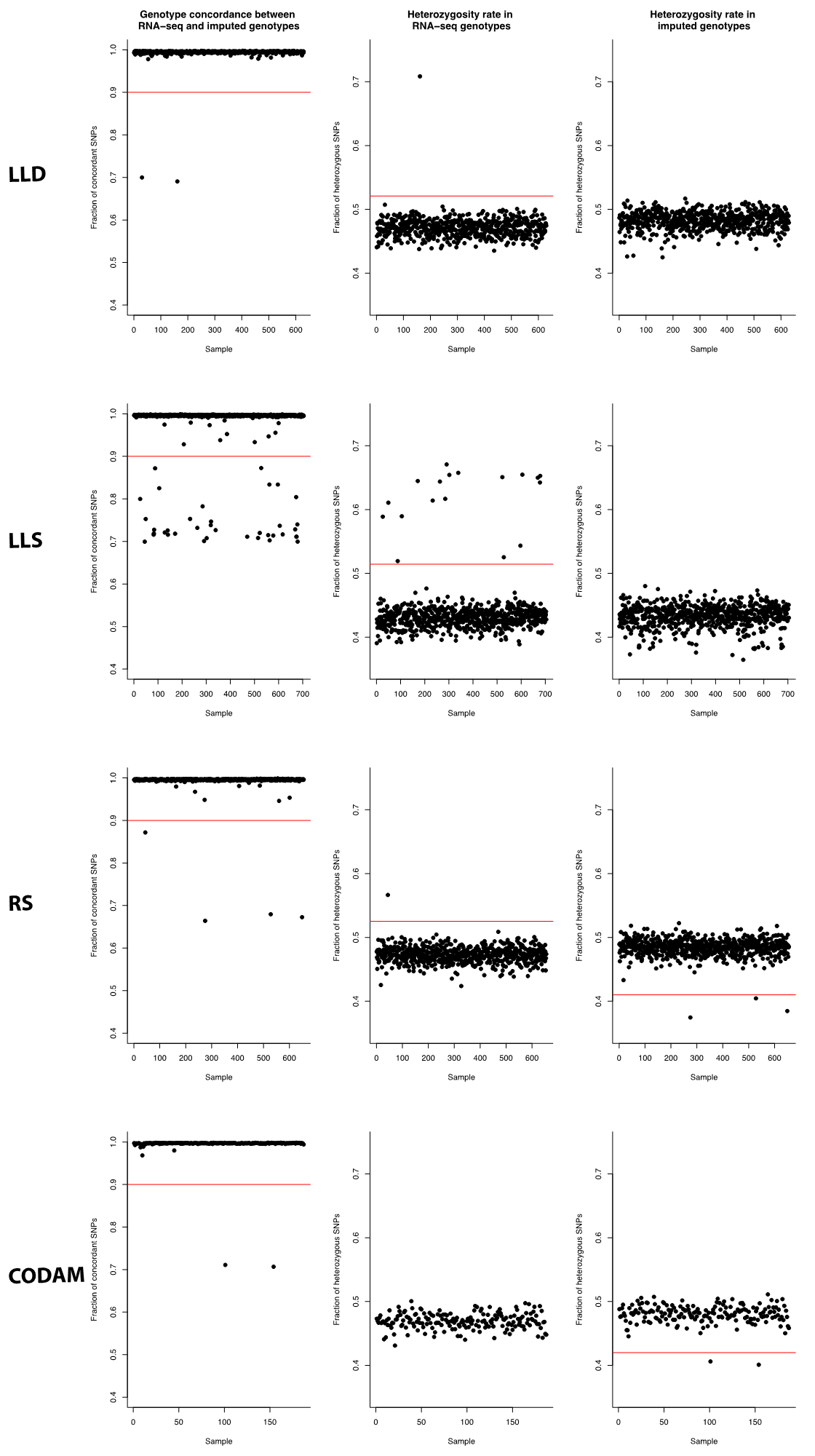
*Provided separately. For legend, please consult legend to main Figure 3.*

# Figure S5. Combining gene- and exon-level eQTLs can help to explain gene expression regulation

The *CD226* locus is shown, and exon-level eQTLs for SNP rs727088 are shown. Exons unique to the NMD transcripts show eQTL effects, whereas other exons do not.

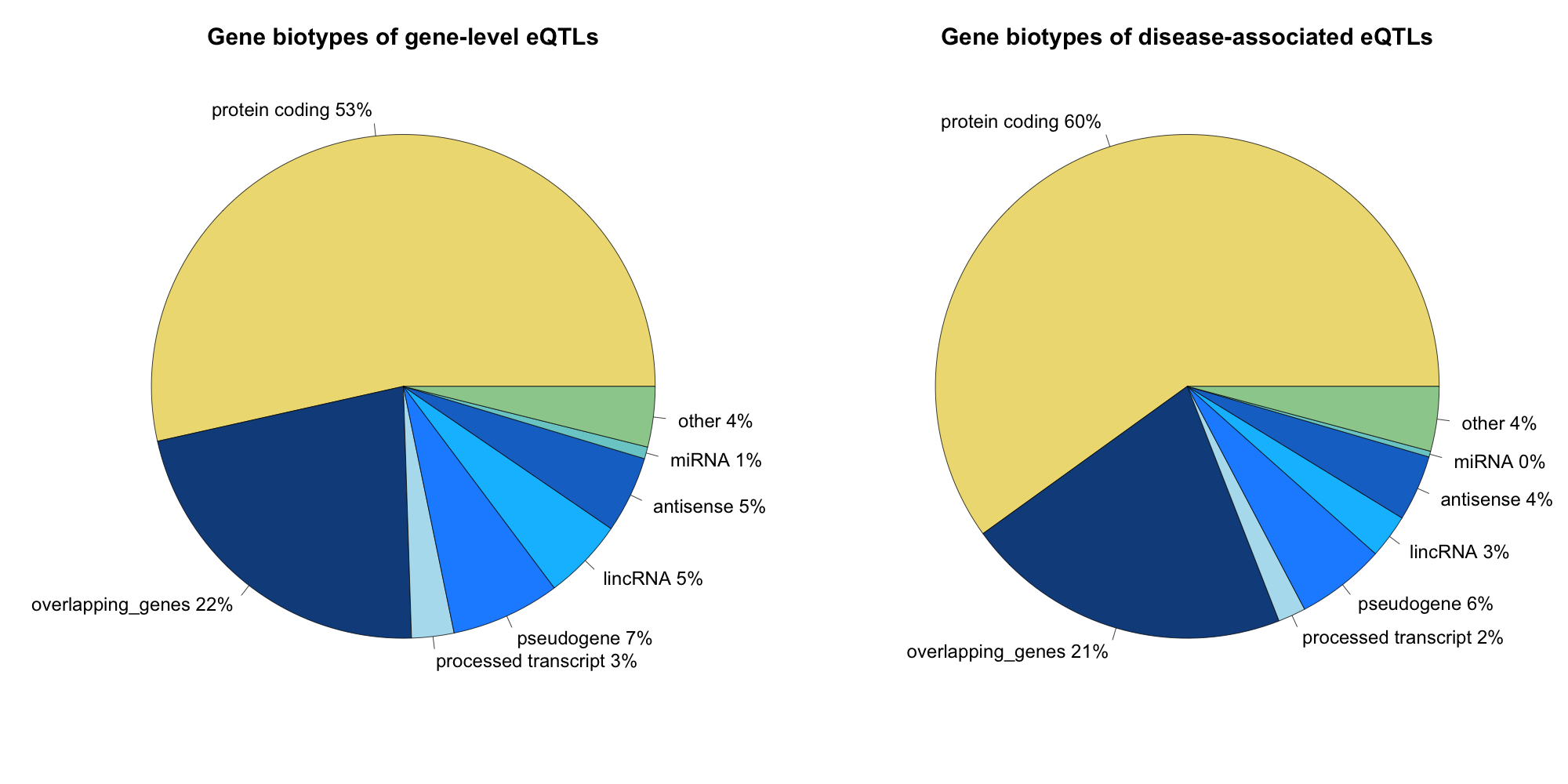


# Figure S6. Genotype concordance and heterozygosity rate per sample for each biobank

Genotype concordance was calculated per sample as the correlation of imputed microarray genotypes with genotypes called from RNA-seq data. The heterozygosity rate was estimated for both imputed microarray genotypes and RNA-seq genotypes as a ratio of heterozygous samples to all samples. In red we show the empirical threshold used to identify outliers.

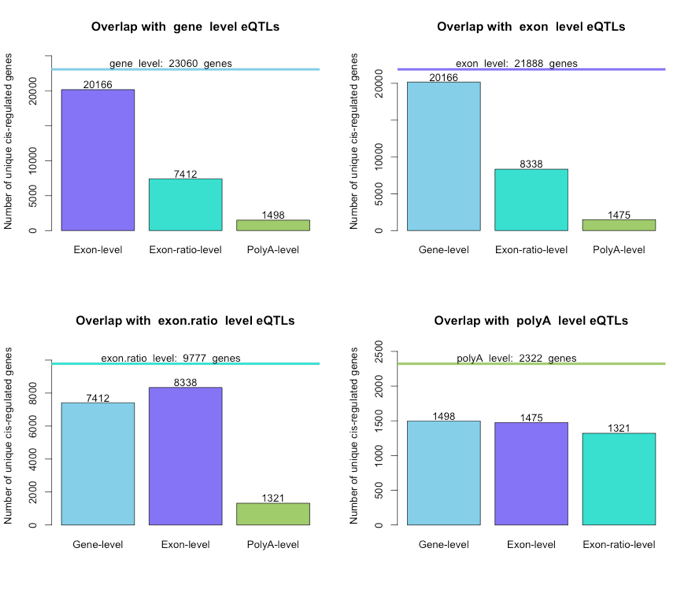
# Figure S7. Biotypes of *cis-*regulated genes

(A) The pie chart shows the biotypes of all *cis*-regulated genes (detected in gene-level eQTL mapping) according to Ensembl v.71 annotation. “Meta-exons” represent overlapping genes. (B) Gene biotypes were also assigned to genes that we found to be regulated by a GWAS SNP (see methods for description) or by its close proxy (r2 ≥ 0.8).



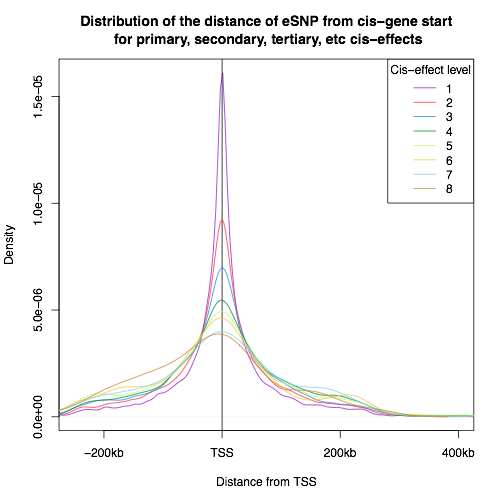
# Figure S8. Overlap of different levels of eQTLs

The figure shows the overlap between different levels of eQTLs in terms of unique SNP-gene pairs. For each level, it shows how many of the SNP-gene pairs identified are also detected on the three other levels.



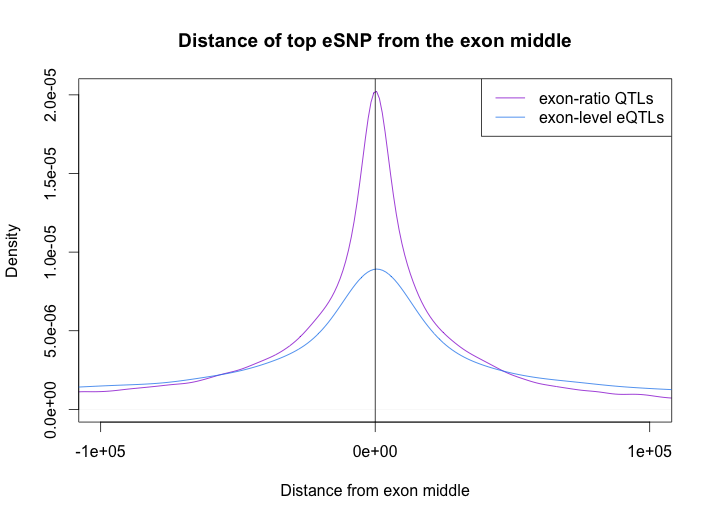
# Figure S9. Distance of the eSNP relative to TSS

The distribution of the distance from the top eSNP to the TSS site of the *cis*-regulated gene for primary, secondary, tertiary and further levels of independent effects.



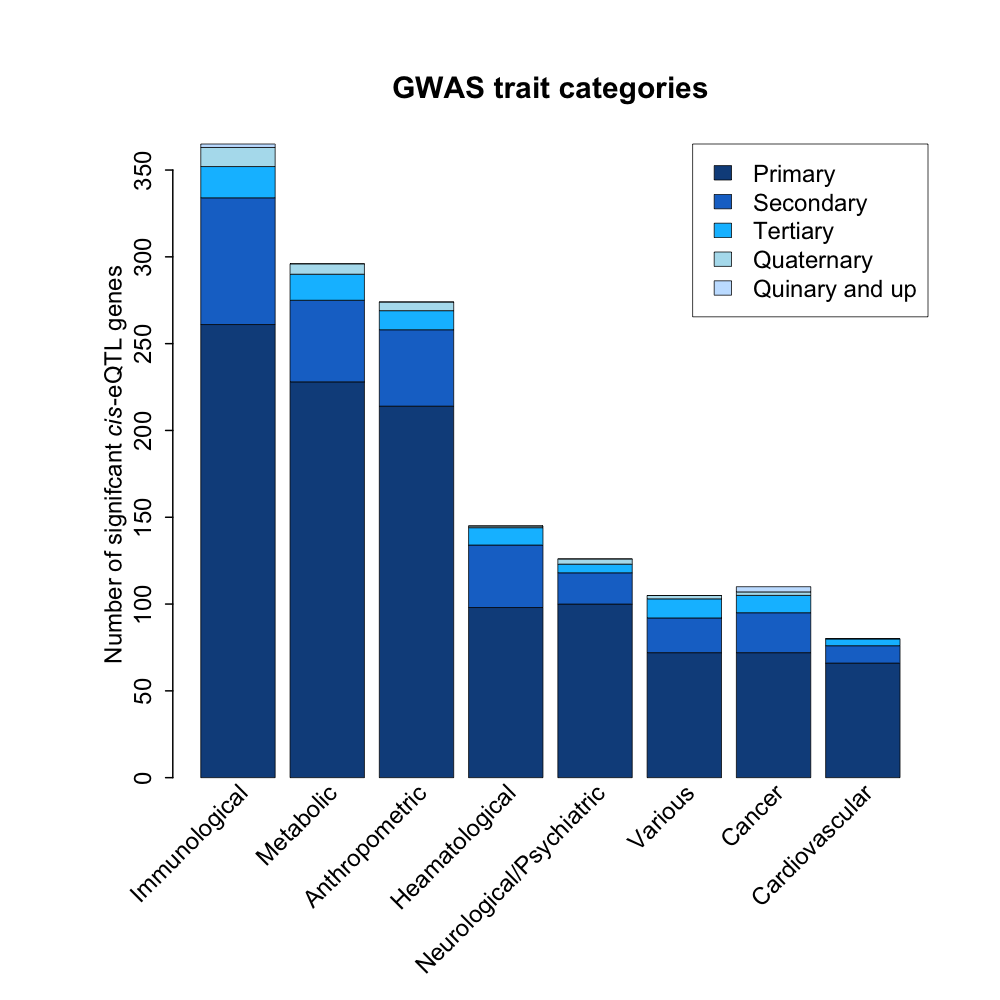
# Figure S10. Distance from eSNP to the exon middle for exon ratio and exon-level eQTLs

The distribution of the distance from the middle of exon for the top eSNP per gene was created for exon-level eQTLs and for exon ratio QTLs. Exons belonging to multiple overlapping genes were removed. We used only the eQTLs that were not detected in gene-level eQTL mapping to concentrate on alternative splicing and the eQTLs regulating the expression of the whole gene.



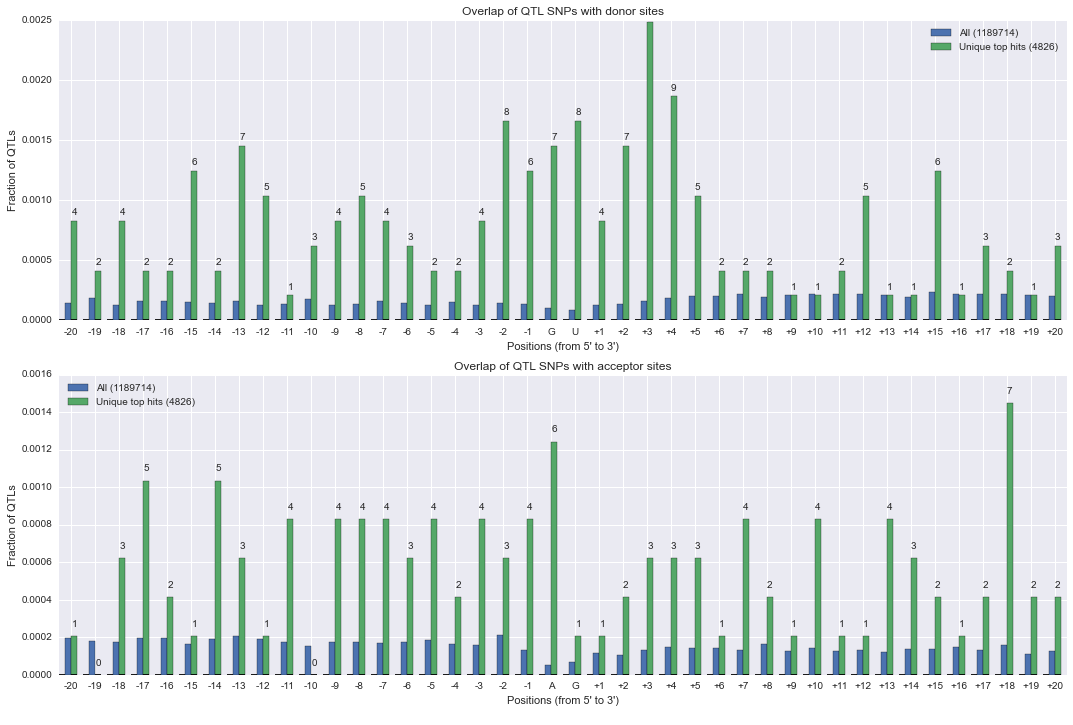
# Figure S11. Disease categories for trait-associated eQTLs

For each GWAS trait category, we show the number of *cis*-regulated genes we detect for SNPs (or proxies) associated with these traits. Different colors represent different levels of independent effects.



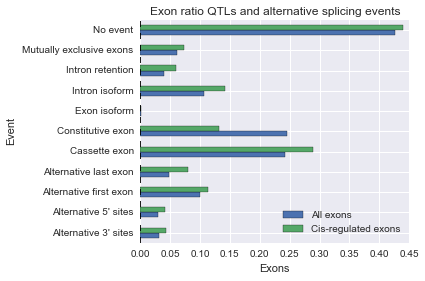
# Figure S12. Distribution of exon ratio QTLs around splice donor or acceptor sites

Histogram of the fraction and number (indicated on top of the bar) of top exon ratio QTL SNPs (green) and all exon ratio QTL SNPs (blue) around the splice donor (top panel, GU motif) and splice acceptor (bottom panel, AG motif).



# Figure S13. Classification of splicing events affected by exon ratio QTLs

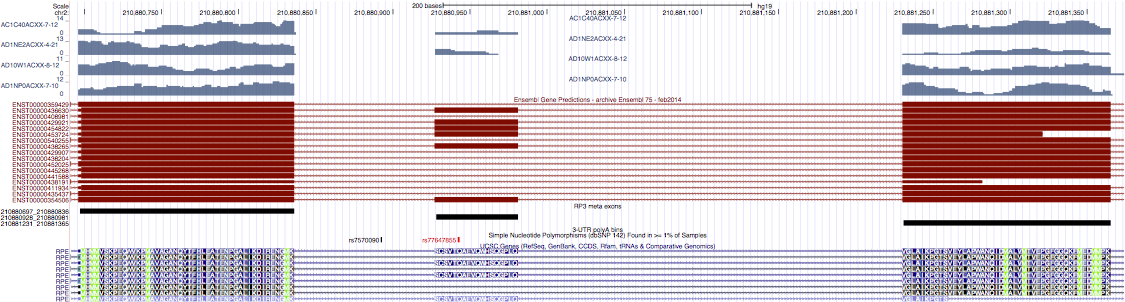
Distribution of exons affected (green) or not affected (blue) by exon ratio QTLS over diverse splicing categories, as annotated in Ensembl v71. Constitutive exons are not known to be alternatively spliced and are depleted for *cis*-regulated exons, whereas all other splicing events are enriched.



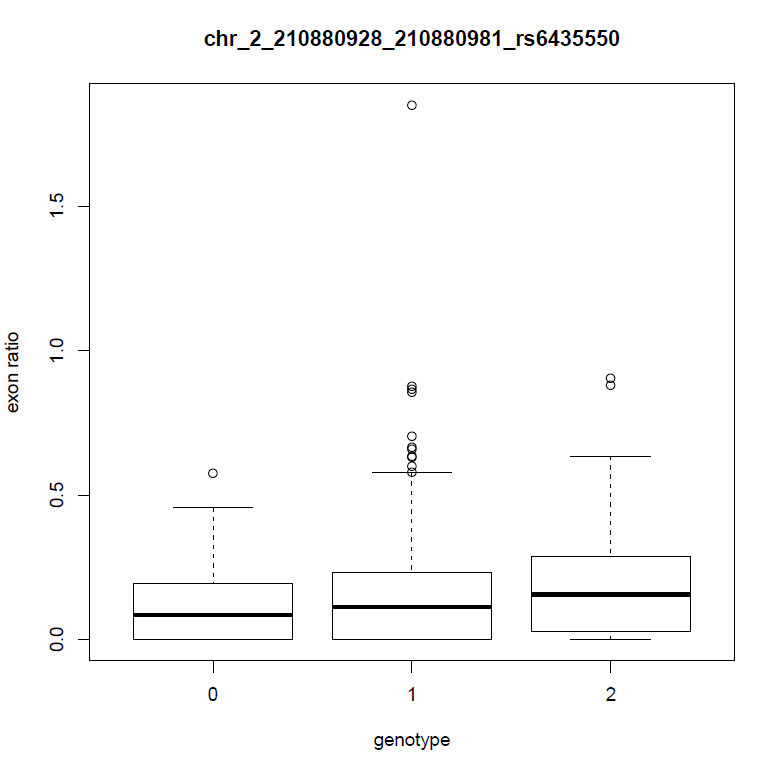
# Figure S14. An example of an exon ratio QTL overlapping with a hit in the GWAS catalogue

The top SNP (rs6435550) affects the inclusion of an in-frame 54 nucleotide exon in the *RPE* gene, adding 18 amino acids to the protein (panel A). The inclusion of the exon is dependent on the genotype (panel B). The SNP is in close LD (r2 = 0.98) with SNP rs7570090. This SNP has been associated with a yet unknown urinary metabolite from an NMR study. The inclusion or exclusion of the protein domain of 18 amino acids may affect the production of xylulose, and xylulose may therefore be the metabolite detected in the NMR study. This example shows how eQTLs for mRNA splicing may generate new hypotheses on the mechanisms underlying GWAS associations.

a.

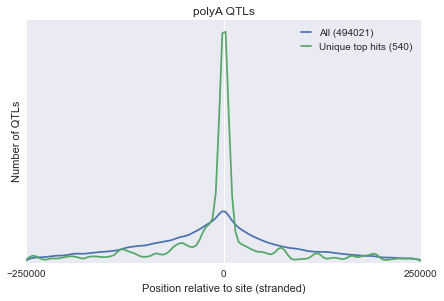


b.



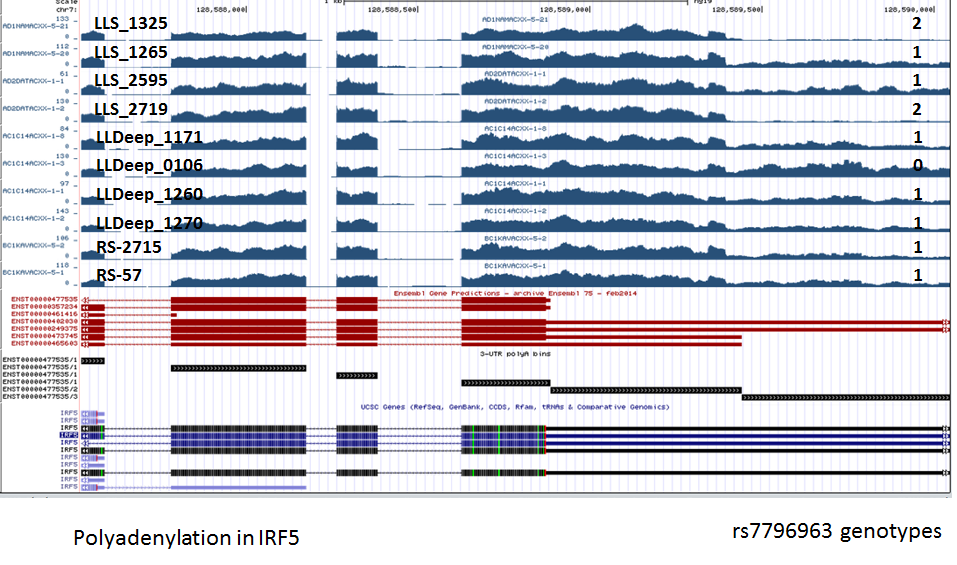
# Figure S15. Distance of polyA ratio QTLs to annotated polyA sites

Density plot of the distance of top polyA ratio QTL (green) and all polyA ratio QTL SNPs (blue) to the polyA site, for which the relative usage is affected by the SNP.



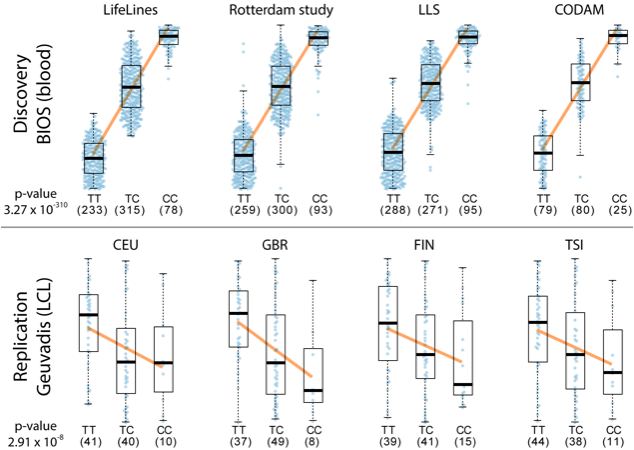
# Figure S16. An example of a polyA ratio QTL overlapping with a hit in the GWAS catalogue

PolyA sites, retrieved from polyA\_DB and the Ensembl database, were used to divide the 3’-UTRs into bins (black bars). PolyA ratios were calculated by dividing the coverage in neighboring bins. For the gene *IRF5*, we found a significant association between the rs7796963 genotype (Indicated by 0, 1, or 2 on top of the RNA-seq coverage tracks) and the ratio between the penultimate and ultimate 3’-UTR bin. This association has been found before 1,2 using targeted experiments or deepSAGE technology. We now show that polyA-QTLs can also be identified based on the coverage patterns in RNA-seq experiments. rs7796963 is in strong LD (r2 = 0.99) with rs10954213 associated with increased risk for Systemic Lupus Erythematosus.



# Figure S17. Plot of an eQTL having opposite allelic directions in BBMRI and Geuvadis

The boxplot shows normalized expression of the *PAM* gene in BIOS (upper panel) and in GEUVADIS (lower panel) samples grouped by their genotype of rs2432162.



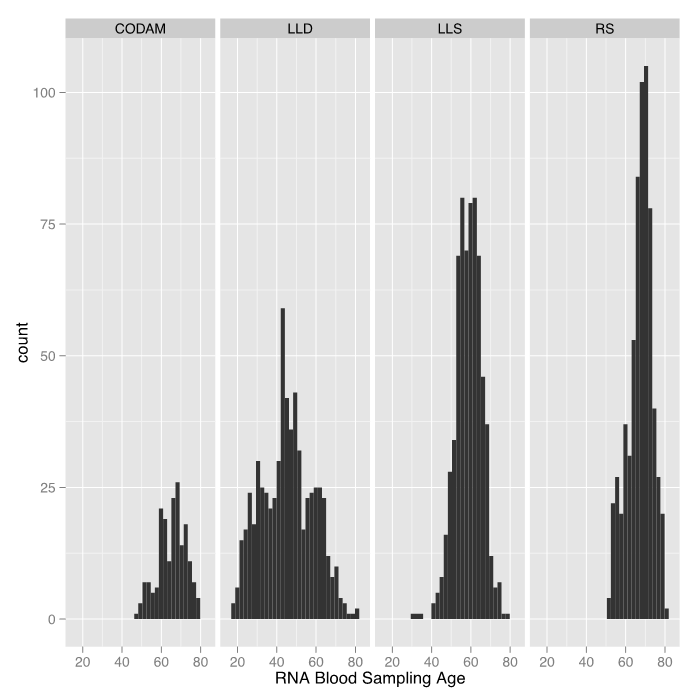
# Figure S18. Independent effects in GEUVADIS dataset and in a random subset of BIOS of the same size

The number of genes having primary, secondary, tertiary and higher level effects in gene level eQTL mapping for the GEUVADIS dataset (left) and for a random subset of BIOS samples of the same size (right).



# Figure S19. Histogram of RNA blood sampling age distribution per biobank

The histogram shows the distribution of sample age for each of the four biobanks.



# Figure S20. Plots of picard metrics results and the outlier samples removed based on these metrics

Provided separately.

To ensure normality assumptions in the linear model, used for ascertaining interactions, we excluded outlier samples. We excluded a set of samples (indicated in red) that showed Picard characteristics that deviated from the rest of the samples.

# Figure S21. Accuracy of cell count prediction method

The prediction accuracy of cell percentages in LLD and LLS cohorts (where the actual cell counts are available) is shown based on 100 times cross-validation. The correlations (Pearson and Spearman) of actual cell percentages with predicted cell percentages is shown on the boxplots for each cell type.

