Hypothesis-free identification of modulators of genetic risk factors

# Supplementary Text

We sequenced globin-depleted whole peripheral blood of 2,176 healthy adults from four different Dutch cohorts. On average 96% of reads were mapped to the genome, 89% of those were mapped to exons (Table S17). To filter out low quality samples, we performed extensive quality control.

# Quality Control

Our quality control was based on the following metrics: total number of mapped reads, exon and gene expression correlation for each pair of samples, concordance of genotypes called from RNA-seq data with imputed genotypes and heterozygosity rate.

## QC results

RNA sequencing was performed on 2,176 individuals, and 44 samples did not pass our quality control criteria. We excluded three samples having less than 5 million sequenced read pairs, two additional samples having less than 70% of reads mapped, one sample for which the correlation of exon counts to all other samples was lower (0.86) than for other samples (median correlation for all samples passing QC was 0.93). A total of 46 samples failed our genotype concordance and contamination criteria (Figure S6). A possible cause for genotyping discordances are sample mixups. Using MixupMapper 1, we identified a total of 12 sample swaps (6 pairs of samples), which were also identified in the genotype concordance analysis. Three of these sample swaps could be traced back in the RNA-seq sample preparation process and were swapped back, after which they showed high genotype concordance. Overall, 40 samples were removed due to genotype discordance or contamination.

Of the remaining 2,116 samples, 29 had less than 15 million read pairs, these were resequenced and fastq files of both runs were merged. This resulted in 2116 samples with an average 47.4 million reads that were used in all subsequent analyses (Table S1).

## eQTL mapping

Apart from eQTLs for protein-coding genes, we found eQTLs for various types of non-coding genes (Figure S7). Of the 23,060 *cis*-regulated genes, 74% were coding and 26% were non-coding. We saw an enrichment for coding *cis*-regulated genes (63% of all tested protein-coding genes and 18% of tested non-coding genes were *cis*-regulated, Fisher exact p-value < 2.2 x 10-6).

Between 35 and 40% of features tested (genes, exons, exon ratios, polyA ratios) were found to be *cis*-regulated (Table S3). The eQTLs for the different features partially overlap (Figure S8). As expected, there is a large overlap between gene-level and exon-level eQTLs, because gene quantifications are derived from the sum of the exon quantifications for the gene. The exon-level eQTLs are a combination of effects on total gene expression and the inclusion rate of that exon due to alternative transcription, splicing or polyadenylation events. Exon-level eQTLs that are also exon ratio QTLs are likely caused by alternative mRNA processing, whereas the exon-level eQTLs that are not exon ratio QTLs are more likely to affect total gene expression levels (Fisher exact p-value < 2.2 x 10-16, OR = 1.8). The polyA ratio QTLs are partially overlapping with the exon ratio eQTLs because SNPs influencing the length of the 3’-UTR will affect the expression quantification of the exon containing the 3’-UTR.

## mRNA expression and processing affected by multiple independent SNPs

Next, we analyzed whether multiple SNPs can influence expression or mRNA processing independently from each other by doing stepwise regression of the effects of the top SNPs. There were fewer exon ratios and polyA ratios with multiple independent effects, which is in line with the larger heterogeneity expected in features influencing total gene expression (Figure 1a, Figure S1). The number of observed independent effects is higher in highly expressed genes (Wilcoxon p-value < 2 x 10-16, comparison of gene expression levels between genes with multiple and single effects). Primary eQTLs are mostly located close to Transcription Start Site (TSS), whereas secondary eQTLs have a more diffuse distribution in the gene (Figure S9). Primary eQTLs are therefore more likely to affect basic transcription regulation at the promoter, whereas secondary eQTLs are more likely to affect mRNA processing. We observed that exon-ratio QTLs tend to have a more local effect than exon-level eQTLs (Figure S10).

## Disease- and trait-associated eQTLs

We found that trait- or disease-associated SNPs (or their close proxies) affect many non-coding genes (Figure S7b). eQTL effects were not solely confined to SNPs associated with hematological or immune-related traits, but also apply to other trait categories (Figure S11).

## Co-localization of eQTL SNPs and GWAS loci.

It has been argued that co-localization between eQTL SNPs and GWAS SNPs often occurs by chance and that therefore these eQTLs are not drivers of complex diseases. To test this explicitly, we ascertained whether SNPs for an immune-related disease (e.g. inflammatory bowel disease, IBD) more often are in strong LD with a blood *cis*-eQTL than expected by chance.

However, in order to establish this well, a realistic null-distribution needs to be established first. Previously observed *cis*-eQTL enrichments for known genetic risk factors have typically been established by using a null-distribution of SNPs that was made by matching known genetic risk factors with SNPs having a similar minor allele frequency and distance to the transcription start site of neighboring genes. This however does not exclude the possibility that other factors exist that increase the a-priori likelihood that a variant is disease-causing (e.g. SNPs within enhancer regions). Alternative approaches to yield a null-distribution have used an approach where the pruned top hits from a GWAS on a simulated, normally-distributed phenotype were used. However, the top pruned SNPs from such a permuted GWAS typically have lower minor allele frequencies, as compared to previously established genetic risk factors for disease.

To circumvent these potential issues, we used 697 height-associated SNPs as a null-distribution, because height loci are strongly enriched for skeletal development genes in chondrocytes, suggesting that these SNPs should not have profound effects on gene expression levels in blood.

We first ascertained a recent set of 166 IBD SNPs. 2,390 unique genes that are expressed in blood map within 250kb from these IBD-associated SNPs. For height 8,789 unique genes that are expressed in blood map within 250kb from previously identified height-associated SNPs. We subsequently subsampled these 8,789 positional height-candidate genes down to 2,390 unique genes, while ensuring that the 2,390 positional IBD-candidate genes and 2,390 positional-candidate height genes have a similar base-line expression distribution (both in terms of average and median expression and standard deviation, T-test p-value > 0.8 for each of these tests) while also ensuring that the positional height-candidates do not map further away from the height SNPs than IBD genes do. This sampling approach ensured that the same proportion of these 2,390 different genes had a *cis*-eQTL effect (41.5% of the IBD positional-candidate genes, compared to 39.4% of the height positional-candidate genes, Fisher’s exact p-value = 0.15). However, we observed that 102 of the 2,390 positional IBD candidate genes had a *cis*-eQTL that was in strong LD (r2 ≥ 0.8) with an IBD SNP, while we observed that only 29 of the 2,390 positional height-candidate genes had a *cis*-eQTL that was in strong LD (r2 ≥ 0.8) with a height SNP. This represents a nearly fourfold enrichment for detecting eQTL effects for IBD SNPs, as compared to height SNPs (Fisher’s exact test p-value = 1.64 x 10-10).

We performed the same analysis for all other complex traits with at least 25 reported genome-wide significant associations (Figure 1c).

## Functional annotation of QTLs

We checked whether the eQTLs identified show significant enrichment for functionally important genomic regions, such as enhancers, DNase I footprints, TF binding patterns, histone modifications etc., using the GWAVA annotation tool 2 and a set of enhancer regions predicted by the FANTOM5 consortium 3. To determine the significance of enrichment, we created a list of background SNPs (see Methods section) and compared the annotation of eQTLs with that of background SNPs. We saw significant enrichment for multiple functional annotation types (Table S4), such as region score, DNase I footprints, various histone marks and binding sites of multiple TFs. The overlap with enhancers showed a significant enrichment for general and blood-cell-type-specific enhancers but not for non-blood tissue-specific enhancers (Table S5).

Exon ratio QTLs are likely to affect the efficiency of splicing of specific exons. In accordance with this, exon ratio QTLs are enriched around the exon mid points (Figure S10). Among the 4,826 top SNPs for exon ratios QTLs (testing only the eSNPs with eQTL p-values at least 10 times smaller than the other eSNP p-values for the same exon), we identified only 15 and 7 overlapping with consensus splice donor or acceptor motifs, respectively. Many more are found in close proximity (up- or downstream) of the splice sites (Figure S12). This is expected because variants affecting the consensus motif will likely disrupt splicing completely, where variants with the nucleotides around the splice motif may alter the binding of splice factors and promote or diminish inclusion of an exon. Thus, in support of the functionality of exon ratio QTLs, they were primarily affecting alternatively spliced exons, whereas constitutive exons were depleted (Figure S13). Exon ratio QTLs affected all different types of alternative splicing events to a similar extent (Figure S13). An example of an exon ratio QTL can be found in Figure S14.

PolyA ratio QTLs are likely to affect the choice between alternative polyadenylation sites. As expected, they are enriched around annotated poly A sites (Figure S15). Out of 41 genes with polyA ratio QTLs detected previously with DeepSAGE technology 4, we replicated 24 in our RNA-seq data. Two were not significant at FDR of 0.05. The remaining 15 could not be assessed because our annotation sources (PolyA\_DB and Ensembl) did not contain alternative polyA sites dividing the 3’-UTR in different bins. These analyses demonstrate that polyA ratio QTLs can also be identified based on the coverage patterns in RNA-seq experiments. We did not find extensive overlap between polyA ratio QTLs and annotated polyA sites nor with canonical polyA signal motifs (AATAAA or ATTAAA), indicating that more research is needed to better understand the genetic regulation of alternative polyadenylation. An example of a polyA ratio QTL can be found in Figure S16.

# Replication studies

To compare our results with previously published studies, we performed replication of *cis-*eQTLs from two datasets: total blood microarray study published by Westra et al. 5 and the GEUVADIS RNA-seq dataset of lymphoblastoid cell lines (LCL) published by Lappalainen et al. 6.

## Gene-level eQTLs

We could replicate 84% of the *cis*-eQTLs identified by Westra et al., 10% with opposite allelic directions (Table S2).

For replication in GEUVADIS, we analyzed the GEUVADIS RNA-seq data using the same pipeline as was used for the BIOS data and performed eQTL mapping with the same settings using DNA-seq genotypes provided by 1000G. On the gene level, 34% of BIOS eQTLs were replicated in GEUVADIS data; 91% of them showing the same allelic direction. In all, 78% of GEUVADIS eQTLs were replicated in BIOS, 88% of them showing the same allelic direction. The degree of replication of exon-level eQTLs was similar (Table S2).

We investigated the genes for which SNPs show opposite allelic effects in GEUVADIS and in BIOS. An example of a strong opposite eQTL is shown on the Figure S17. The effect of this eQTL is opposite in the two datasets because it is neutrophil-specific. BIOS samples are drawn from total blood, in which neutrophils are prevalent, whereas GEUVADIS contain LCL samples. As we know that neutrophil counts are negatively correlated with lymphocyte counts, we expect that the neutrophil-specific eQTL effect may be different or even opposite in lymphoid cells. In general, we observed that the genes having opposite allelic direction more often have a cell-type-specific effect than those having the same direction (Chi2 p-value = 2.45 x 10-11, OR = 2.28).

Replication of exon ratio and polyA ratio QTLs showed similar replication rates, but a much lower percentage of opposite allelic effects, suggested that the genetic control of mRNA processing is less cell-type-dependent (Table S2).

## Multiple independent signals in GEUVADIS

We were interested to what extent secondary, tertiary and higher eQTLs could be replicated in GEUVADIS. Therefore, we performed stepwise regression of eQTLs in GEUVADIS as we had done for our own dataset. In GEUVADIS, we observed at most five significant independent eQTL effects per gene. To understand whether this lower number of independent effects compared to BIOS data was due to sample size, we randomly subsampled BIOS dataset to obtain the same number of samples as in GEUVADIS and redid the stepwise regression analysis. This resulted in the same maximum of five independent effects per gene, showing that, to a large extent, this discrepancy can be explained by the smaller number of samples investigated in GEUVADIS. However, the overall number of independent eQTLs in the BIOS subset was higher than in GEUVADIS (14,987 independent eQTLs in BIOS compared to 10,180 eQTLs in GEUVADIS) (Figure S18).

Next, we investigated whether BIOS and GEUVADIS eQTLs tagged the same causal variants by taking all independent effects from BIOS and from GEUVADIS and calculating the LD (based on GoNL genotype data) between all pairs of eSNPs per gene and taking the SNP pairs with maximum r2. On average the r2 was 0.58.

# Supplementary acknowledgements

The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists. The generation and management of GWAS genotype data for the Rotterdam Study is supported by the Netherlands Organization of Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012). This study is funded by the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) project nr. 050-060-810. We thank Pascal Arp, Mila Jhamai, Marijn Verkerk, Lizbeth Herrera and Marjolein Peters for their help in creating the GWAS database.

Work on cell count estimation was funded by NWO 863.13.011.

# References

1. Westra, H.-J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27,** 2104–11 (2011).

2. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11,** 294–6 (2014).

3. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–61 (2014).

4. Zhernakova, D. V. *et al.* DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genet.* **9,** e1003594 (2013).

5. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45,** 1238–1243 (2013).

6. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).