# Hypothesis-free identification of modulators of genetic risk factors

# Online Methods

# Cohort description

The four cohorts used in our BIOS (Biobank-based Integrative Omics Study) study are briefly described below. The age range of the individuals differed for the different biobanks (Figure S19). The number of samples per cohort used in our study can be found in Table S1.

## CODAM

The Cohort on Diabetes and Atherosclerosis Maastricht1 (CODAM) consists of a selection of 547 subjects from a larger population-based cohort.2 Inclusion of subjects into CODAM was based on a moderately increased risk of developing cardiometabolic diseases such as type 2 diabetes and/or cardiovascular disease. Subjects were included if they were of Caucasian descent, over 40 years of age and additionally met at least one of the following criteria: increased BMI (>25), a positive family history of type 2 diabetes, a history of gestational diabetes and/or glycosuria, or use of anti-hypertensive medication.

## LLD

The LifeLines-DEEP (LLD) cohort 3 is a sub-cohort of the LifeLines cohort.4 LifeLines is a multi-disciplinary prospective population-based cohort study examining the health and health-related behaviors of 167,729 individuals living in the northern parts of the Netherlands using a unique three-generation design. It employs a broad range of investigative procedures assessing the biomedical, socio-demographic, behavioral, physical and psychological factors contributing to health and disease in the general population, with a special focus on multi-morbidity and complex genetics. A subset of 1,500 LifeLines participants also take part in LLD3. For these participants, additional molecular data is generated, allowing for a more thorough investigation of the association between genetic and phenotypic variation.

## LLS

The aim of the Leiden Longevity Study5 (LLS) is to identify genetic factors influencing longevity and examine their interaction with the environment in order to develop interventions to increase health at older ages. To this end, long-lived siblings of European descent were recruited together with their offspring and their offspring’s partners, on the condition that at least two long-lived siblings were alive at the time of ascertainment. For men the age criteria was 89 or older, for women age 91 or over. These criteria led to the ascertainment of 944 long-lived siblings from 421 families, together with 1,671 of their offspring and 744 partners.

## RS

The Rotterdam Study6 is a single-center, prospective population-based cohort study conducted in Rotterdam, the Netherlands. Subjects were included in different phases from the start of the study in 1998, with a total of 14,926 men and women aged 45 and over included as of late 2008. The main objective of the Rotterdam Study is to investigate the prevalence and incidence of and risk factors for chronic diseases to contribute to a better prevention and treatment of such diseases in the elderly.

# RNA data preparation and sequencing

Total RNA from whole blood was deprived of globin using Ambions GLOBINclear kit and subsequently processed for sequencing using Illumina’s Truseq version 2 library preparation kit. Paired-end sequencing of 2x50bp was performed using Illumina’s Hiseq2000, pooling samples at 10 per lane, and aiming for >15M read pairs per sample. Finally, read sets per sample were generated using CASAVA, retaining only reads passing Illumina’s Chastity Filter for further processing.

## RNA-seq data processing

All data processing was performed on the Dutch Life Science Grid. All files and their md5 checksums are linked to sample identifiers and metadata in a centralized meta-database running Apache CouchDB (v1.4.0). The quality control (QC), alignment and quantification pipeline ran on the execution nodes of the Dutch Life Science Grid. This network of clusters has access to the SRM (storage resource manager). Employing the same database as a token-pool server, dozens of samples were analyzed simultaneously across the different university computer clusters in the Netherlands.

## Preprocessing

The quality of the raw reads was checked using FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/]. The adaptors identified by FastQC (v0.10.1) were clipped using cutadapt (v1.1) applying default settings (min overlap 3, min length). Sickle (v1.200) [https://github.com/najoshi/sickle] was used to trim low quality ends of the reads (min length 25, min quality 20).

## Alignment

Read alignment was performed using STAR 2.3.0e 7. To avoid reference mapping bias, all GoNL SNPs with MAF > 0.01 in the reference genome were masked with N 8. Read pairs with at most 8 mismatches, mapping to at most 5 positions, were used.

## Alignment statistics

Mapping statistics from the BAM files were acquired through Samtools flagstat (v0.1.19-44428cd). The 5’ and 3’ coverage bias, duplication rate and insert sizes were assessed using Picard tools (v1.86).

## Expression quantification

We estimated expression on the gene, exon, exon ratio and polyA ratio levels using Ensembl v.71 annotation (which corresponds to GENCODE v.16).

Overlapping exons (on either of the two strands) were merged into meta-exons and expression was quantified for the whole meta-exon. For that, custom scripts were developed which use coverage per base from coverageBed and intersectBed from the Bedtools suite (v2.17.0) 9 and R (v2.15.1). This resulted in base counts per exon or meta-exon.

Gene expression, as base count per gene, was calculated as the sum of expression values of all exons of each gene (excluding meta-exons). Overlapping gene parts are counted separately from the unique gene parts throughout this manuscript.

Expression of exons relative to their gene (exon ratio) was calculated by dividing the exon base counts by the summed base counts for all exons of the same gene. Meta-exons overlapping with multiple genes were discarded.

Overlapping 3’ UTRs for the same gene, as annotated in Ensembl, were merged by gene. A collection of polyadenylation sites was retrieved from PolyA\_DB and the annotated 3’ ends of transcripts from Ensembl. These polyadenylation sites were used to split the merged 3’ UTRs into bins. To avoid small bins that tend to give noisy ratios, we applied some filtering on the polyA sites. PolyA sites located no more than 10bp from the start or from the end of the 3'UTR were discarded. Additionally, sites that are no more than 10 bp apart were merged (if their number was even, the first site downstream was used).

For all genes with at least two bins (corresponding to at least two potential sites of polyadenylation), we calculated the ratio of base counts between every two neighboring bins (polyA ratios).

# Genotype data

## Data generation

Genotype data was generated for each cohort individually. Details on the methods used can be found in the individual papers (CODAM: van Dam et al.2; LLD: Tigchelaar et al.10; LLS: Deelen et al.11; RS: Hofman et al.6).

## Imputation and QC

The genotype data were harmonized to the Genome of the Netherlands12 (GoNL) using Genotype Hamonizer13 and subsequently imputed per cohort with Impute214 using the GoNL reference panel15 (v5). Quality control was also performed per cohort. We removed SNPs with an imputation info-score below 0.5, a HWE *P*-value smaller than 10-4, a call rate below 95% or a minor allele frequency smaller than 0.05. In total, 9,333,740 SNPs passed QC in at least one dataset.

# Quality Control

To identify low quality samples we applied several quality metrics and used a combination of them to decide whether to exclude a sample from further analyses.

## Read counts

For each sample, the total number of mapped reads was used as a quality measure. Those samples for which these counts were less than 70% were flagged and excluded from the analysis.

## Exon and gene expression correlation

For each pair of samples, Spearman correlation of their expression was calculated on gene and exon levels. From these values, the median Spearman correlation for each sample was calculated (D-statistic). Samples with D-statistics lower than 0.85 were excluded from the analysis.

## Genotype concordance

To further check the quality of the data, we compared the imputed genotypes with genotypes called from RNA-seq data. Genotype concordance could be low in cases of bad quality RNA-seq or imputed genotype data or in cases of sample mix-ups.

To call genotypes from RNA-seq data, we used the combination of samtools mpileup 16 (with the following parameters: -A -B -Q 0 -s -d10000000, calling only GoNL SNPs with MAF > 0.01) and SNVMix2 17. Only the genotypes with posterior probabilities higher than 0.8 were included. We determined genotype concordance per sample as genotype correlation of high-confidence SNPs (the SNPs which had a mean genotype correlation across all samples that was not less than 0.9). Outlier samples, for which the genotype concordance was less than 0.9, were flagged and excluded from the analysis.

## Heterozygosity rate

To detect possible contamination of RNA-seq samples, we calculated the heterozygosity rate of the genotypes called from RNA-seq data. Heterozygosity rate was calculated as a ratio of the number of heterozygous SNPs and the number of all SNPs, using only high-confidence SNPs present in both imputed genotypes and genotypes called from RNA-seq data (i.e. the same SNPs as described in the previous section). Samples with RNA-seq heterozygosity rate > 0.52 were excluded from the analysis.

## Mix-up mapping

Previously we showed that sample mix-ups occur frequently in genetical genomics datasets, introducing noise in subsequent analyses 18. We checked the data for mix-ups using this published method and flagged possible mixed samples.

In the LLS dataset, a group of samples were detected to have been swapped. Six of these samples were swapped back and verified again using the same QC measures.

# QTL mapping

We used our previously described pipeline 19 to perform eQTL mapping. We mapped QTLs using a Spearman's rank correlation on imputed genotype dosages in each cohort and then ran a meta-analysis combining the results by weighted z-score method. To control the false discovery rate (FDR) at 0.05, we created a null distribution by permuting sample labels of the expression data, repeating this process 10 times.

## Expression data normalization

Expression data on the gene and exon level were first normalized using Trimmed Mean of M-values (TMM)20. Expression values were then log2 transformed, probe and sample means centered to zero and their standard deviation was scaled to 1. To correct for batch effects, principal component analysis (PCA) was run on the sample Spearman correlation matrix and the first 25 PCs were removed19. We observed that removing these PCs resulted in the detection of the highest number of eQTLs. To ascertain that none of these 25 PCs are under genetic control, we ran separate QTL mapping on each principal component and ensured that there were no SNPs associated with them.

Exon ratio and polyA ratio expression data were not normalized, as ratios are not dependent on the library size and we used non-parametric statistics.

## *cis-*QTL mapping

For running cis-QTL mapping we tested genes (or exons, exon ratios and polyA ratios) and SNPs located within 250kb from the gene (or exon) center. Only SNPs with minor allele frequency (MAF) > 0.05, call rate (CR) > 0.95 and Hardy-Weinberg equilibrium p-value > 0.001 were included.

## Set of background SNP for functional enrichment analyses.

For assessment of functional enrichment of eSNPs on each QTL level, we created a background list of SNPs which we compared to the real set. For each eQTL SNP, we selected the variants within a 50,000 bp window, with a MAF differing not more than 0.05 point from the eQTL SNP, and a linkage disequilibrium r2 < 0.5. From the variants that match these criteria, we selected the variant that is physically closest to the eQTL SNP as a background SNP.

## Replication of *cis*-eQTLs

For replication of our eQTL results, we used two dataset. The first dataset was Geuvadis RNA-seq data of lymphoblastoid cell cines (LCLs)21. For replication, we took raw RNA-seq reads of 373 European samples and processed them using the same alignment and quantification pipeline as we used on the BIOS data. For eQTL mapping, we regressed out the first 20 PCs from the expression data (due to the smaller sample size of Geuvadis dataset). We ran eQTL mapping on gene, exon, exon ratio and polyA ratio level. To replicate BIOS eQTLs in Geuvadis, we took all significant eQTLs (top SNP per gene) from BIOS and ran eQTL mapping in Geuvadis, testing only these eQTLs. We then checked how many eQTLs out of all those tested were replicated and for how many of the replicated eQTLs the allelic direction was opposite. We did the same in the other direction, testing how many of the Geuvadis eQTLs were replicated in the BIOS data.

The second dataset was a meta-analysis of 5,311 microarray peripheral blood samples published by Westra et al.19. As raw data were not available for these data, we used all significant eQTLs (FDR < 0.05) identified in the meta-analysis, mapped the microarray probes to gene and exons using Ensembl v.71 gene annotation, and then tested these SNP-gene and SNP-exon combinations in the BIOS data.

## GWAS annotation

For annotating the eQTLs with known disease/trait associations, we used a set of 6,321 SNPs derived from the NHGRI GWAS catalog and a set of reported ImmunoChip associations, each with reported P < 5 x 10-8 (Table S18).

# Interaction analysis

The interaction analysis was performed using the following linear model:

where is the eQTL gene expression, is the eSNP genotype, is the proxy gene, is the interaction term between the proxy gene and the genotype, is the intercept, and , and are regression coefficients.

As a linear model is parametric, and thus more sensitive to outliers and non-normal distributions than our non-parametric eQTL model, we performed stricter quality control. We found that several metrics introduced outliers in our data that confounded the linear regression analyses. These were percentage of coding bases, median 3’ bias, percentage of uniquely mapped reads and percentage of mRNA bases (Figure S20). Based on these metrics we removed 75 samples and used the remaining 2,041 samples in the interaction analyses. We confined the interaction analysis to genes with at least one mapped read in all samples; this criterion was used for both the proxy genes and the eQTL genes. As a result, we tested 29,750 genes as potential proxies and 17,291 eQTL effects.

The normalization for the expression levels of the eQTL genes requires different normalization then the expression of the proxy genes. The gene expression data of the eQTL genes was corrected using covariates for the source biobank, the first 25 PCs, gender, median 3’ bias, median 5’ bias, GC content and the percentage of intronic bases. In order to detect biologically meaningful interaction effects, we also regressed out the interaction effects for gender, median 3’ bias, median 5’ bias, GC content and the percentage of intronic bases. The expression data used in the interaction term was processed in a similar manner, with the exception that we did not correct for the principal components, as this would have removed correlations to cell types, and we did not correct for interactions with the technical covariates.

We excluded interactions where the eQTL SNP showed a significant eQTL effect on the tested proxy gene, as we wanted to exclude the cases in which the gene giving the interaction effect was in the same locus as the tested eQTL gene.

We then performed an iterative interaction analysis by regressing the top covariate in a stepwise manner. After the first round of interaction analysis, we identified the covariate having the highest chi2sum (, where e is an eQTL out of the set of all eQTLs (E) and is the squared interaction z-score of the current covariate with the eQTL e) over all interaction z-scores. We regressed out this covariate from covariate and gene expression data and repeated the interaction analysis. This procedure was repeated 10 times. For each top covariate, we identified a set of covariates (module) with a similar interaction pattern by taking the top 100 covariates having the highest chi2sum difference between the current interaction analysis step and the previous step (effectively identifying co-expressed genes). These covariates are mostly highly coexpressed with the top covariate in the module (Figure 2c).

To determine the significance level of interactions, we permuted genotype sample labels and ran the interaction analysis. This enabled us to determine which eQTLs significantly interact with the top covariate of the module with a FDR < 0.05.

We ran the interaction analysis on exon and exon ratio levels in a similar manner as for the gene level.

## Interaction module functions

To find the prevalent cell type for each module, we used several sources of information. Some of the BIOS biobanks had cells counts available, making it possible to correlate the top 100 covariates of each module with cell type percentages.

As an additional source of evidence, we used expression profiles for isolated populations of 17 of the major cell types in blood generated by the BLUEPRINT consortium 22.

To determine the putative function of each module, we performed pathway enrichment analysis using GeneNetwork 23,24 on the top 100 covariates in the module and on all eQTL genes having a significant interaction with the top covariate of the module.

To gain more insight into the function of the modules we identified, we overlapped the interaction results with those of several papers which studied stimulated cells and response QTLs (reQTLs): a study of PBMCs infected with rhinovirus 25 and a study of monocytes infected by LPS (at two time points: after 2 and 24 hours) and IFN 26. To investigate whether our interaction modules represent anti-viral or anti-bacterial response, we checked for enrichment of differentially expressed genes reported for each stimulation (with -1 < log FC < 1) within the top 100 covariates from each interaction module by performing a one-tailed Fisher’s exact test to determine the significance. We also checked whether the reported reQTLs showed significantly stronger interaction with the top covariate of each module by performing a Wilcoxon rank-sum test on interaction z-scores.

We also checked whether we see an enrichment of binding of particular transcription factors (TFs) using ChIP-seq data from ENCODE 27. First, we determined which TFs overlapped with the eQTL SNP or a variant in very strong LD (r2 ≥ 0.99). Then, using a Fisher exact test, we determined if we found any enrichment in overlap between the genes assigned to a module and the genes not significantly assigned to this module.

## Using interaction modules to better understand disease mechanisms

We extracted the genes regulated by any type of top QTL variant in strong LD (r2 ≥ 0.8) with the top GWAS hits. Co-expression was calculated in our data for these genes and Cytoscape 3.2.1 28 was used to create network plots. Assignment to specific clusters was performed using the R implementation of Affinity Propagation 29,30. Cell-type-specific expression levels were based on the RNA-seq generated by the BLUEPRINT consortium 22 and plotted using gplots. We performed gene function enrichment analysis using GeneNetwork 23.

## Cell-type-specific eQTL mapping

The cell-type-specific eQTL were identified using the same method we used for the gene-based interaction analyses. However, here we used the cell-type percentages instead of the expression of other genes. As not all cohorts measured cell counts, we estimated for these cohorts. RNA-seq data and cell count measurements in 628 samples from the LL and 650 samples from the LLS cohorts were used to build prediction models for cell counts using a proposed method based on expression of genes showing cell-type-specific expression (Di Tommaso et al, manuscript in preparation) for neutrophils, lymphocytes, monocytes, eosinophils and basophils. Subsequently, these models were applied to RNA-seq data of 185 samples from the CODAM cohort and 14 samples from the LLS cohort for prediction of cell counts of the five cell types. In addition, the prediction models were applied to estimate cell counts for neutrophils, eosinophils and basophils, using RNA-seq data from 652 samples from the RS cohort in which the cell count measurements of lymphocytes and monocytes were available.

In order to evaluate the performance of cell count prediction, RNA-seq from 1,278 samples of two cohorts (LLD and LLS) and cell counts data were used. We randomly split the dataset into training and prediction sets. Based on the model built by the training set, the cell counts of prediction set were estimated. The predicted cell counts were then compared with the measured data using both Spearman and Pearson correlation coefficients. To demonstrate the performance of the proposed method, this process was repeated 100 times (cross-validation) using data from the same cohort and between different cohorts (Figure S21).

# BLUEPRINT tissue-specific expression data analysis

BLUEPRINT data was downloaded from their ftp site (ftp://[ftp.ebi.ac.uk/pub/databases/blueprint](http://ftp.ebi.ac.uk/pub/databases/blueprint)). All venous blood-, myeloid cell- and erythroblast-derived RNAseq data was downloaded. Read counts were obtained according to the gene quantification performed by the Center for Genomic Regulation. Subsequently, TMM normalization 20 was performed. The averaged normalized log-counts per million per cell type were used for drawing the heatmaps. For each module, we extracted the corresponding genes based on their ENSEMBL gene identifier (for 'meta-exons' we used the first Ensembl id and three noncoding RNAs could not be extracted from the BLUEPRINT data). Furthermore, the R-package pheatmap (1.0.7) was used to generate the heatmaps.

# References

1. van Greevenbroek, M. M. J. *et al.* The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). *Eur. J. Clin. Invest.* **41,** 372–379 (2011).

2. Van Dam, R. M., Boer, J. M. a, Feskens, E. J. M. & Seidell, J. C. Parental history off diabetes modifies the association between abdominal adiposity and hyperglycemia. *Diabetes Care* **24,** 1454–1459 (2001).

3. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open*  **5 ,** (2015).

4. Scholtens, S. *et al.* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* **44,** 1172–80 (2015).

5. Schoenmaker, M. *et al.* Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *Eur. J. Hum. Genet.* **14,** 79–84 (2006).

6. Hofman, A. *et al.* The rotterdam study: 2014 objectives and design update. *Eur. J. Epidemiol.* **28,** 889–926 (2013).

7. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 1–7 (2012). doi:10.1093/bioinformatics/bts635

8. Liu, Z. *et al.* Comparing computational methods for identification of allele-specific expression based on next generation sequencing data. *Genet. Epidemiol.* **38,** 591–8 (2014).

9. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–2 (2010).

10. Tigchelaar, E. F. *et al.* *An introduction to LifeLines DEEP: study design and baseline characteristics*. 0–21 (2014). doi:10.1101/009217

11. Deelen, J. *et al.* Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum. Mol. Genet.* **23,** 4420–32 (2014).

12. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **advance on,** (2014).

13. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7,** 901 (2014).

14. 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,** (2009).

15. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in European samples using the ‘Genome of The Netherlands’. *Eur J Hum Genet* **22,** 1321–1326 (2014).

16. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma.*  **25 ,** 2078–2079 (2009).

17. Goya, R. *et al.* SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26,** 730–6 (2010).

18. 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).

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

20. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11,** R25 (2010).

21. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–11 (2013).

22. Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol.* **30,** 224–226 (2012).

23. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47,** 115–125 (2015).

24. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6,** 5890 (2015).

25. Çalışkan, M., Baker, S. W., Gilad, Y. & Ober, C. Host genetic variation influences gene expression response to rhinovirus infection. *PLoS Genet.* **11,** e1005111 (2015).

26. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343,** 1246949 (2014).

27. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22,** 1813–31 (2012).

28. Cline, M. S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2,** 2366–82 (2007).

29. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315,** 972–6 (2007).

30. Bodenhofer, U., Kothmeier, A. & Hochreiter, S. APCluster: an R package for affinity propagation clustering. *Bioinformatics* **27,** 2463–4 (2011).