eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs

Nurlan Kerimov[1,2,†], James D. Hayhurst[2,3,†], Jonathan R. Manning[2,3], Peter Walter[3], Liis Kolberg[1], Kateryna Peikova[1], Marija Samoviča[1], Tony Burdett[2,3], Simon Jupp[2,3], Helen Parkinson[2,3], Irene Papatheodorou[2,3], Daniel R. Zerbino[2,3,*], Kaur Alasoo[1,2,*]

[1]Institute of Computer Science, University of Tartu, Tartu, 51009, Estonia
[2]Open Targets, South Building, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[3]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[†]These authors contributed equally to this work.
*These authors jointly supervised this work.
Correspondence should be addressed to D.R.Z (zerbino@ebi.ac.uk) or K.A. (kaur.alasoo@ut.ee)

Abstract

An increasing number of gene expression quantitative trait locus (QTL) studies have made summary statistics publicly available, which can be used to gain insight into human complex traits by downstream analyses such as fine-mapping and colocalisation. However, differences between these datasets in their variants tested, allele codings, and in the transcriptional features quantified are a barrier to their widespread use. Here, we present the eQTL Catalogue, a resource which contains quality controlled, uniformly re-computed QTLs from 19 eQTL publications. In addition to gene expression QTLs, we have also identified QTLs at the level of exon expression, transcript usage, and promoter, splice junction and 3' end usage. Our summary statistics can be downloaded by FTP or accessed via a REST API and are also accessible via the Open Targets Genetics Portal. We demonstrate how the eQTL Catalogue and GWAS Catalog APIs can be used to perform colocalisation analysis between GWAS and QTL results without downloading and reformatting summary statistics. New datasets will continuously be added to the eQTL Catalogue, enabling systematic interpretation of human GWAS associations across a large number of cell types and tissues. The eQTL Catalogue is available at https://www.ebi.ac.uk/eqtl/.

Introduction

Gene expression and splicing QTLs are a powerful tool to link disease-associated genetic variants to putative target genes. Despite efforts by large-scale consortia such as GTEx [1] and eQTLGen [2] to provide comprehensive eQTL annotations for a large number of human tissues, most eQTL datasets are still scattered across individual publications. Multiple databases have been developed that collect eQTL summary statistics [3–9]; however, these efforts have relied on the heterogeneous set of summary statistics calculated by the original authors.

Relying on publicly available eQTL summary statistics has several limitations. First, many downstream use cases such as fine-mapping [10,11] and colocalisation [12,13] require full summary statistics from the region of interest, but some studies have only released either eQTL lead variants or variants below a certain p-value threshold. Second, studies often test a different subset of variants in the *cis* region of each gene, meaning that variants tested in one study might be missing from another study. Third, even though the eQTL effect direction relative to a GWAS signal is critical for interpreting disease associations, information about the effect allele is often either missing or ambiguous. Finally, even though both splicing [1,14] and other transcript-level QTLs [15] contribute to complex traits, these analyses have not been performed on many earlier RNA-seq-based eQTL datasets. Where splicing or transcript-level QTL summary statistics have been released, these are still difficult to compare between studies due to large differences in analysis strategy and the types of transcript-level changes captured by different methods [15].

To overcome these limitations, we have reprocessed the raw data from 19 eQTL studies. We have applied uniform data analysis and quality control procedures to all of these datasets. In addition to gene expression QTLs, we have identified QTLs at the level of exons, transcripts and transcriptional events covering alternative promoters, splicing events and transcript 3' ends. This allowed us to detect novel QTLs in existing datasets that would have otherwise remained hidden. Our full summary statistics are available on the eQTL Catalogue FTP server and via a REST API. As an example, we use the eQTL Catalogue and GWAS Catalog APIs to identify a transcript usage QTL in stimulated macrophages at the *CD40* locus which colocalises with a rheumatoid arthritis GWAS signal. Access to individual-level data will enable us to recompute QTL summary statistics as improved RNA-seq analysis methods become available.

Results

Data analysis workflow

To uniformly process a large number of eQTL studies, we designed a modular and robust data analysis workflow (Figure 1). We first downloaded the raw gene expression and genotype data and converted the data to common input formats (VCF for genotypes and fastq for RNA-seq). We performed extensive quality control of genotypes (see Methods) and imputed them to the 1000 Genomes Phase 3 [16] reference panel. For RNA sequencing data, we started with the nf-

core [17] RNA-seq pipeline written in the Nextflow [18] framework and modified it to support the quantification of four different molecular phenotypes: gene expression, exon expression [19], transcript usage, and promoter, splicing and 3' end usage events defined by txrevise [15] (Supplementary Figure 3). Using the same quantification workflow ensured that molecular phenotype identifiers (genes, transcripts, exons and events) were consistent between individual studies. Furthermore, we harmonised sample metadata between studies and mapped all biological samples (cell types and tissues) to a common set of 24 distinct ontology terms from UBERON [20], Cell Ontology [21] and Experimental Factor Ontology [22]. This will allow users to easily find if the same cell types or tissues has been profiled in multiple studies (Table 1). The normalised molecular phenotype matrices and imputed genotypes were fed into our QTL mapping workflow that was also developed using the Nextflow framework. The full association summary statistics have been made publicly available via the eQTL Catalogue FTP site as well as the REST API. All our data analysis workflows have been released under a permissive licence (see Software Availability).

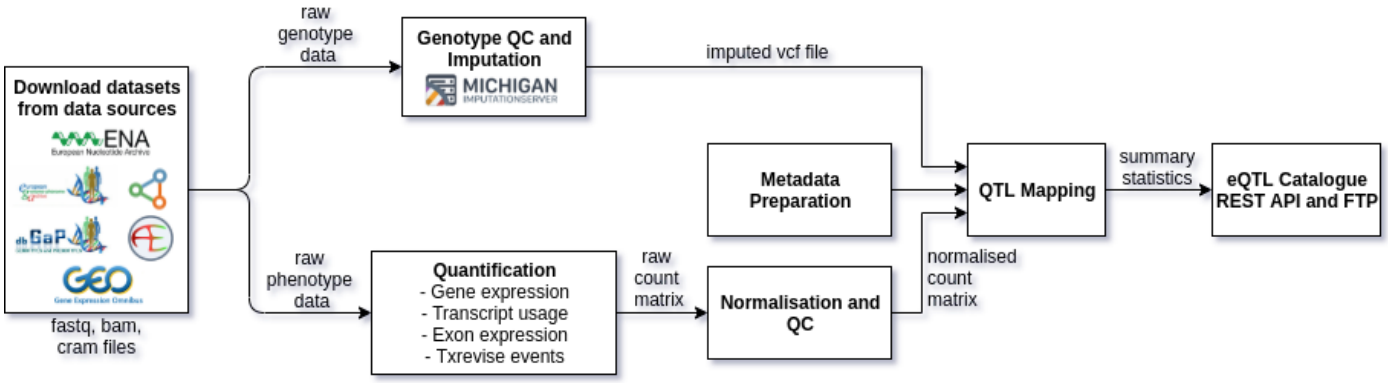


Figure 1. High-level representation of the uniform eQTL mapping process. The txrevise [15] events capture alternative promoters, splicing events and transcript 3' ends. Schematic illustration of the different quantification methods is provided in Supplementary Figure 3.

Datasets included in the eQTL Catalogue

We downloaded raw gene expression and genotype data from 14 RNA-seq and 5 microarray studies from various repositories. This included 8,115 RNA-seq samples and 4,631 microarray samples from 4,685 unique donors (Table 1), covering 24 cell types or tissues (Table 1) and 13 stimulated conditions (Supplementary Material 1) (called 'biological contexts'). Even though these samples were profiled in different laboratories using a wide range of RNA-seq protocols (Supplementary Tables 2 and 3) and sequencing depth (Supplementary Figure 2), they predominantly clustered by cell type or tissue of origin in multidimensional scaling analysis (MDS) (Figure 2A). Projecting the genotype data of the donors to 1000 Genomes Phase 3 [16] reference panel, we found that although 88% of the donors were of European origin, the datasets also included 487 (~10%) donors from African populations and a small number of samples from other populations (Table 2, Supplementary Table 1).

Table 1. Overview of the cell types and tissues included in the eQTL Catalogue. For each cell type or tissue, the table highlights the studies that profiled it as well as the total sample size across studies. Cell types and tissues were mapped to common ontology terms (Supplementary Material 1). *Some cell types (monocytes, macrophages, CD4+ and CD8+ T cells) were profiled in multiple conditions, described in Supplementary Material 1. DLPFC - dorsolateral prefrontal cortex, iPSC - induced pluripotent stem cell, LCL - lymphoblastoid cell line.

Cell type or tissue	RNA-seq studies	RNA-seq sample size	Microarray studies	Microarray sample size	Total
monocyte*	BLUEPRINT [23], Quach_2016 [24], Schmiedel_2018 [25]	1251	CEDAR [26], Fairfax_2014 [27]	1657	2908
DLPFC	BrainSeq [28], ROSMAP [29]	1055	-	0	1055
LCL	GENCORD [30], GEUVADIS [31], TwinsUK [32]	1053	-	0	1053
CD4+ T cell*	BLUEPRINT [23], Schmiedel_2018 [25]	344	CEDAR [26], Kasela_2017 [33]	570	914
macrophage*	Alasoo_2018 [34], Nedelec_2016 [35]	829	-	0	829
CD8+ T cell*	Schmiedel_2018 [25]	177	CEDAR [26], Kasela_2017 [33]	546	723
blood	Lepik_2017 [36], TwinsUK [32]	666	-	0	666
B cell	Schmiedel_2018 [25]	91	CEDAR [26], Fairfax_2012 [37]	543	634
neutrophil	BLUEPRINT [23]	196	CEDAR [26], Naranbhai_2015 [38]	373	569
adipose	TwinsUK [32]	381	-	0	381
skin	TwinsUK [32]	370	-	0	370
iPSC	HipSci [39]	322	-	0	322
fibroblast	GENCORD [30]	186	-	0	186
T cell	GENCORD [30]	184	-	0	184
Th17 cell	Schmiedel_2018 [25]	177	-	0	177
pancreatic islet	van_de_Bunt_2015 [40]	117	-	0	117
sensory neuron	Schwartzentruber_2018 [41]	98	-	0	98
CD16+ monocyte	Schmiedel_2018 [25]	90	-	0	90
NK cell	Schmiedel_2018 [25]	90	-	0	90
Tfh cell	Schmiedel_2018 [25]	89	-	0	89
Th2 cell	Schmiedel_2018 [25]	89	-	0	89

Treg memory	Schmiedel_2018 [25]	89	-	0	89
Treg naive	Schmiedel_2018 [25]	89	-	0	89
Th1 cell	Schmiedel_2018 [25]	82	-	0	82
Total:		**8115**	**-**	**3689**	**11804**

Our quality control of the gene expression and genotype datasets included removing outlier samples from the gene expression datasets, ascertaining the genetic sex of the samples using the expression of sex-specific genes, detecting genotype concordance between RNA-seq and genotype samples, and detecting cross-contamination between samples within a study using both sex-specific gene expression as well as genotype data (see Methods, Supplementary Figure 4). We excluded a total of 2,418 samples during the quality control procedure (Supplementary Table 3).

Table 2. Number of unique donors assigned to the four major superpopulations in the 1000 Genomes Phase 3 reference dataset. Detailed assignment of donors to the four superpopulations in each study is presented in Supplementary Table 1. Visual explanation of the population assignment is provided in Supplementary Figure 5. Superpopulation codes: EUR - European, AFR - African, EAS - East Asian, SAS - South Asian.

Assigned population	Sample Size	Percent
EUR	4138	0.883
AFR	487	0.104
EAS	21	0.004
SAS	5	0.001
Unassigned	34	0.007
TOTAL	**4685**	**1**

For RNA-seq datasets, we performed QTL mapping for four different molecular phenotypes described above (Figure 1, Supplementary Figure 3). The QTL analysis was performed separately in each biological context of each study. In general, we found the largest number of QTLs at the level of gene expression, but for all molecular phenotypes, the number of significant associations scaled approximately linearly with the sample size (Figure 2B, Supplementary Material 1). For microarray datasets, we performed the analysis only at the gene level, but found the same linear trend (Figure 2B, Supplementary Material 1).

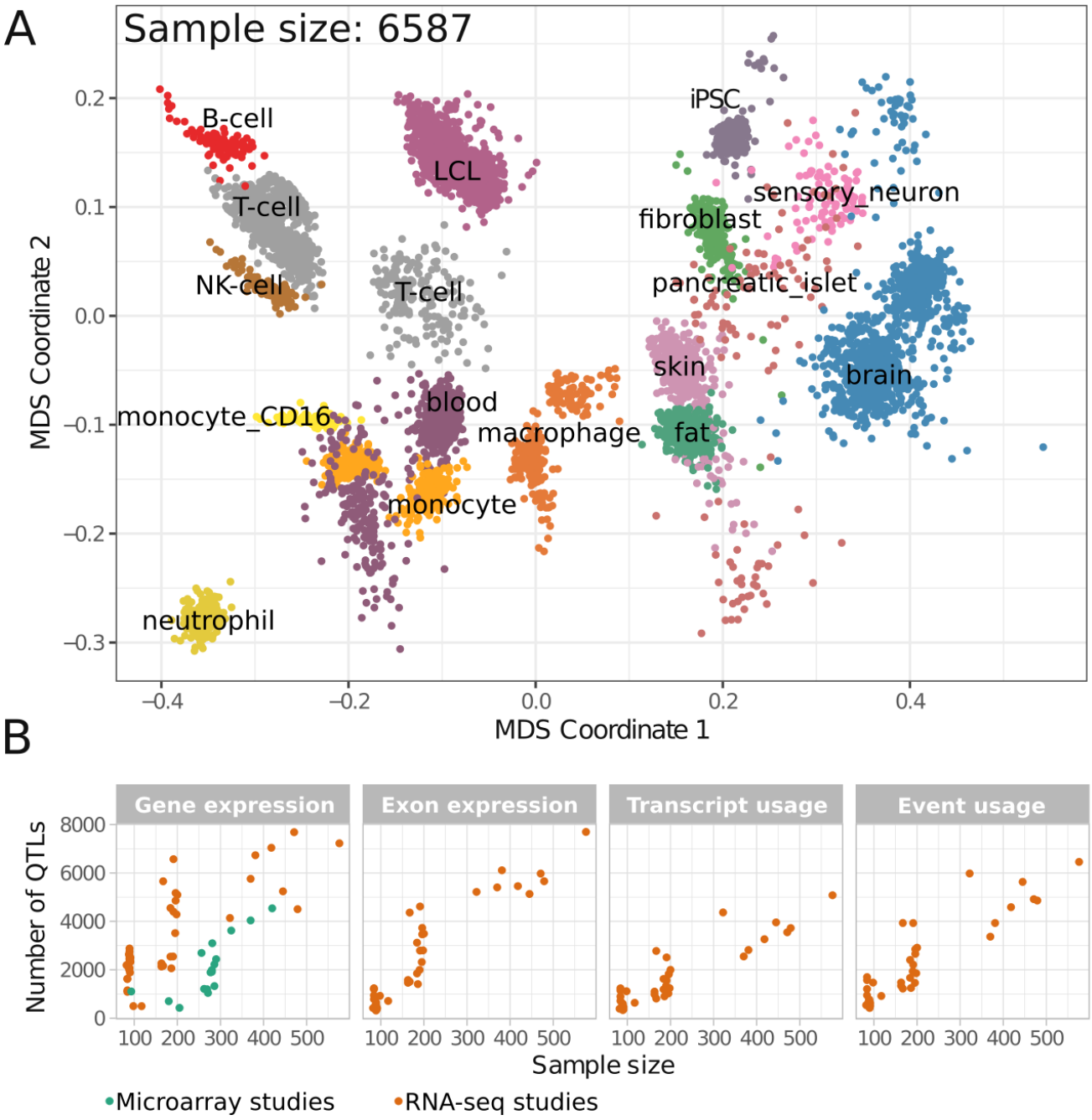


Figure 2. Overview of the datasets included in the eQTL Catalogue. **(A)** Multidimensional scaling (MDS) analysis of the RNA-seq samples. To improve clarity, only 6,587 samples from the unstimulated conditions passing quality control are included in the plot. Various T cell subsets from Table 1 have been grouped together. A similar MDS plot for the microarray samples can be found in Supplementary Figure 1. **(B)** The relationship between the sample size of each study and the number of associations detected using each quantification method. The x-axis represents sample size of a biological context in a study. The y-axis represents number of significant associations (FDR < 0.05) found in each biological context.

Example use case

To demonstrate the utility of the eQTL Catalogue and REST API for interpreting disease-associated genetic variants, we explored the *CD40* locus associated with rheumatoid arthritis (RA) [42]. We have previously demonstrated that the RA GWAS signal at this locus colocalises with a promoter usage QTL for *CD40* in macrophages stimulated with interferon-gamma [34]. To assess whether this association could be detected in other tissues and cell types, we queried the eQTL Catalogue API using the GWAS lead variant from the *CD40* RA locus (rs4239702). We found a number of molecular phenotypes strongly associated with the lead variant (nominal p-value < 10^{-4}) (Figure 3A). In particular, there was a strong association with the total expression level of *CD40* in four independent monocyte eQTL studies covering both RNA-seq and microarrays studies [23,24,26,27] (Figure 3A).

To test if these eQTLs are likely to share the same causal variant with the RA GWAS signal, we used colocalisation analysis [12]. We fetched the full association summary statistics from the *CD40* locus (GRCh38 chr20:45,980,000-46,200,000). This analysis replicated the previously reported colocalisation with *CD40* promoter usage in stimulated macrophages [15] (Figure 3B); however, the same analysis applied to monocyte-specific eQTLs strongly supported a model of distinct causal variants underlying the eQTL and GWAS association in all four studies (Figure 3C). This was consistent with the low linkage disequilibrium (LD) of $r^2 = 0.13$ between the monocyte eQTL (rs745307) and RA GWAS lead variants (rs4239702). This highlights the importance of having access to full summary statistics from the region. Although the GWAS variant was strongly associated with *CD40* expression in monocytes, this was likely due to a very strong independent eQTL signal nearby (nominal p-value < 10^{-50} in the Fairfax_2014 dataset) that was in low LD with the GWAS lead variant. It is possible that the promoter usage QTL detected in stimulated macrophages (Figure 3B) is a weak secondary eQTL in the monocyte samples, but this would still indicate that *CD40* expression in naive monocytes does not directly contribute to RA disease risk, because a much stronger eQTL in that context is not associated with the disease [43]. The complete RMarkdown document to reproduce this analysis is available from GitHub (see Software Availability).

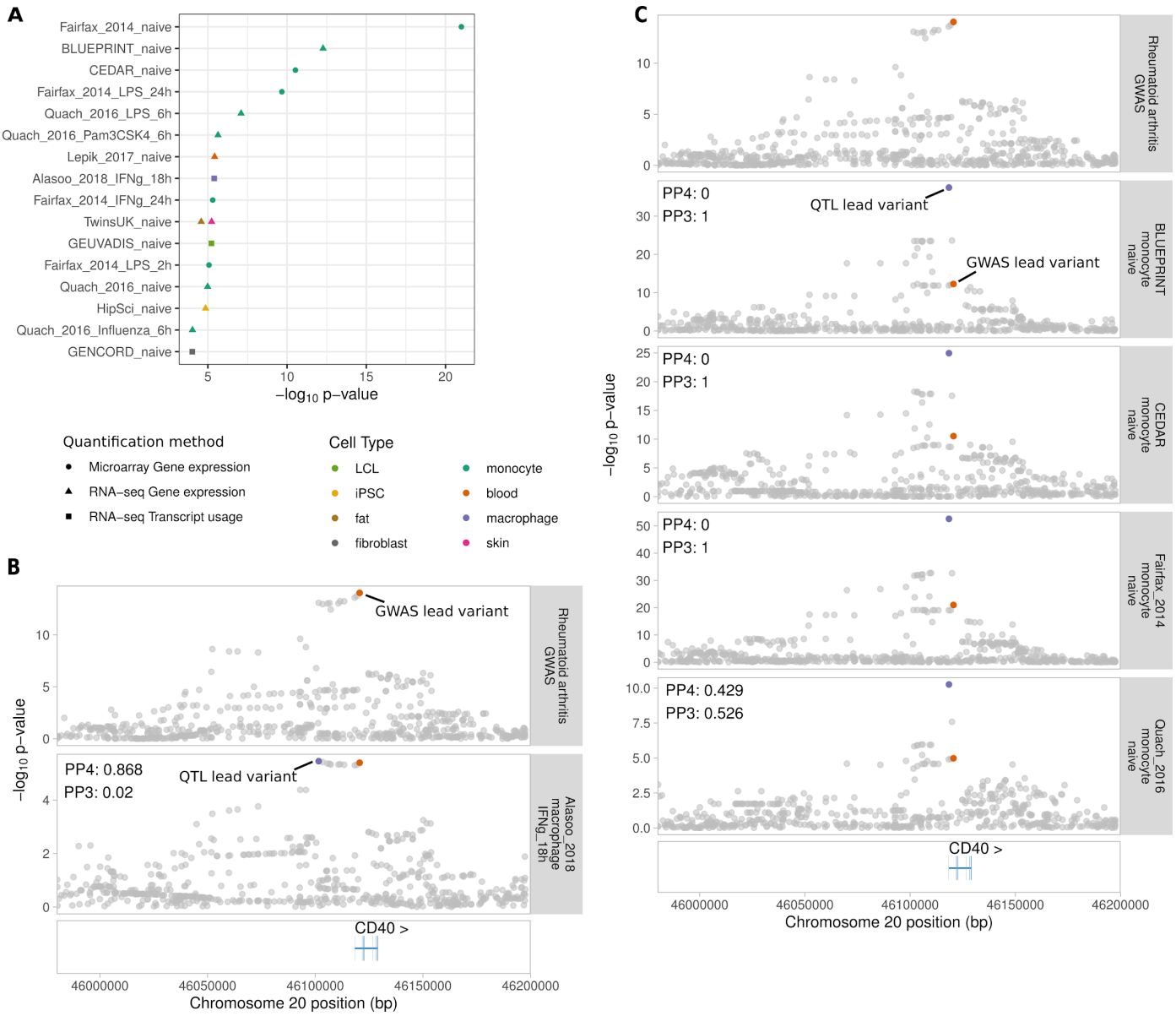


Figure 3. Example colocalisation with a transcript/gene QTLs and disease using eQTL Catalogue and GWAS Catalog APIs. **(A)** Strong gene expression and transcript usage QTL associations with rheumatoid arthritis (RA) GWAS lead variant (rs4239702) from the CD40 locus. The significant eQTLs (nominal p-value < 10^{-4}) are from three quantification methods and eight distinct cell types or tissues. **(B)** Colocalisation analysis between RA GWAS and *CD40* transcript usage QTL in macrophages stimulated with interferon-gamma from the Alasoo_2018 study. The GWAS lead variant is shown in orange and the QTL lead variant is shown in purple. **(C)** Colocalisation analysis between RA GWAS and *CD40* eQTL in naive monocytes from two RNA-seq (BLUEPRINT, Quach_2016) and two microarray studies (CEDAR, Fairfax_2014). PP3, posterior probability of a model with two distinct causal variants; PP4, posterior probability of a model with one common causal variant.

Comparison with existing databases

The largest collection of eQTLs is currently hosted by the QTLbase [5] database. Although QTLbase contains some splicing QTLs, this is limited to the summary statistics provided by the original study authors. Secondly, although QTLbase has harmonized variant identifiers and effect sizes across studies, these are not accessible programmatically and the downloadable files only contain p-values of nominally significant associations ($p < 0.05$) without the effect sizes. Thus, QTLbase summary statistics are not suitable for fine-mapping and colocalisation applications. Both QTLizer [4], PhenoScanner [7] provide programmatic access to their summary statistics, but the QTLizer summary statistics have not been harmonized and PhenoScanner contains data from only ten studies. Finally, both FUMA [6] and ImmuneRegulation [8] provide access to some eQTL summary statistics via their web interface, but the full data cannot be downloaded for local computational analyses.

All eQTL Catalogue summary statistics are available under the Creative Commons Attribution 4.0 International License, enabling third parties to build their own tools and services on top of the released summary statistics and the REST API. To avoid downloading large text files, slices of the summary statistics can be accessed using tabix [44] (see Data Availability).

Discussion

The eQTL Catalogue provides a resource of uniformly processed human gene-level and transcript-level QTL summary statistics, with the aim of supporting biomedical genetic research. This resource will be progressively expanded to all accessible human datasets. We are currently analysing raw data from GTEx v8 [1], the CommonMind Consortium [45] and the FUSION study [46]. We are also setting up data access agreements for additional datasets on an ongoing basis.

We have paid particular attention to making the summary statistics as usable as possible. By mapping cell types and tissues to common ontology terms, we make it easy to discover which studies contain the tissues and cell types of interest for the users. This will also enable summary-level meta-analysis [2] across studies containing the same cell types and tissues. We have imputed most genotype datasets to the same reference panel and reference genome version, ensuring that similar set of genetic variants is present in most studies. Finally, we use a consistent set of molecular phenotype identifiers (genes, exons, transcripts, events) across all datasets, ensuring that genetic effects can directly be compared across datasets. Our summary statistics have already been used to interpret GWAS associations for Alzheimer's disease [47].

We welcome feedback on ways to improve our methods. In the next release planned for June 2020, we plan to include LeafCutter [48] splice junction usage QTLs as the fifth molecular phenotype quantified from RNA-seq data. We are also exploring ways to systematically fine-map [10,11] the QTL signals to identify multiple independent associations for each gene and make the credible sets of causal variants publicly available. This can help to further characterise loci with multiple independent signals, such as the *CD40* locus described above (Figure 3).

Finally, we are exploring approaches to handle related samples and population stratification by using either linear mixed models or performing eQTL analysis in each population separately. These modifications would not be possible without access to individual-level genotype and RNA-seq data.

We are always looking for additional datasets to be included in the eQTL Catalogue. Unfortunately, we were unable to obtain access to all of the datasets that we would have liked to include in the analysis due to consent limitations or restrictions on sharing individual-level genetic data (Supplementary Table 4). These limitations could be overcome in the future by federated data analysis approaches, where the eQTL analysis is performed at remote sites using our analysis workflows, and only summary statistics are shared with the eQTL Catalogue. To this end, we will continue to improve the usability and portability of our data analysis workflows and will make them available via community efforts such as the nf-core [17] repository. Researchers interested in contributing their datasets to the eQTL Catalogue should contact us at eqtlcatalogue@ebi.ac.uk.

Methods

Data access and informed consent

Gene expression and genotype data from two studies (GEUVADIS and CEDAR) were available for download without restrictions from ArrayExpress. For all other datasets, we applied for access via the relevant Data Access Committees. The database accessions and contact details of the individual Data Access Committees can be found on the eQTL Catalogue website (http://www.ebi.ac.uk/eqtl/Datasets/). In our applications, we explained the project and our intent to publicly share the association summary statistics. Although this was acceptable for the 19 studies currently included in the eQTL Catalogue, some of our data access requests were rejected either because informed consent obtained from the study participants did not allow the sharing of genotype data with other researchers or the data were restricted for research into specific diseases (Supplementary Table 4). Ethical approval for the project was obtained from the Research Ethics Committee of the University of Tartu (approval 287/T-14).

Genotype data

Pre-imputation quality control. We aligned the strands of the genotyped variants to 1000 Genomes Phase 3 reference panel using Genotype Harmonizer [49]. We excluded genetic variants with Hardy-Weinberg p-value < 10^{-6}, missingness > 0.05 and minor allele frequency < 0.01 from further analysis. We also excluded samples with more than 5% of their genotypes missing.

Genotype imputation and QC. We imputed the genotypes to the 1000 Genomes Phase 3 reference panel [16] using a local installation of the Michigan Imputation Server v1.0.4 [50]. After imputation, we converted the coordinates of genetic variants from GRCh37 reference

genome to GRCh38 using CrossMap v0.2.8 [51]. We used bcftools v1.9.0 to exclude variants with minor allele frequency (MAF) < 0.01 and imputation quality score R2 < 0.4 from downstream analysis.

Assigning individuals to reference populations. We used PLINK [52] v1.9.0 to perform LD pruning of the genetic variants and LDAK [53] to project new samples to the principal components of the 1000 Genomes Phase 3 reference panel [16]. To assign each genotyped sample to one of four superpopulations, we calculated the Euclidean distance in the principal component space from the genotyped individual to all individuals in the reference dataset. Distance from a sample to a reference superpopulation cluster is defined as a mean of distances from the sample to each reference sample from the superpopulation cluster. We explored distances between samples and reference superpopulation cluster using different number of PCs and found that using 3 PCs worked best for inferring superpopulation of a sample. Then, we assigned each sample to a superpopulation if the distance to the closest superpopulation cluster was at least 1.7 times smaller than to the second closest one (Supplementary Figure 5). We used this relatively relaxed threshold, because our aim was to get an approximate estimate of the number of individuals belonging to each superpopulation. Performing a population-specific eQTL analysis would probably require a much more stringent assignment of individuals to populations.

Microarray data

Data normalisation. All five microarray datasets currently included in the eQTL Catalogue (CEDAR, Fairfax_2012, Fairfax_2014, Naranbhai_2015, Kasela_2017) used the same Illumina HumanHT-12 v4 gene expression microarray. The database accessions for the raw data can be found on the eQTL Catalogue website (http://www.ebi.ac.uk/eqtl/Datasets/). Batch effects, where applicable, were adjusted for with the function removeBatchEffect from the limma v.3.40.6 R package [54]. The batch adjusted $\log_2$ intensity values were quantile normalized using the lumiN function from the lumi v.2.36.0 R package [55]. Only the intensities of 30,353 protein-coding probes were used. The raw intensity values for the five microarray datasets have been deposited to Zenodo (doi: https://doi.org/10.5281/zenodo.3565554).

Detecting sample mixups. We used Genotype harmonizer [49] v1.4.20 to convert the imputed genotypes into TRITYPER format. We used MixupMapper [56] v1.4.7 to detect sample swaps between gene expression and genotype data. We detected 155 sample swaps in the CEDAR dataset, most of which affected the neutrophil samples. We also detected one sample swap in the Naranbhai_2015 dataset.

RNA-seq data

Pre-processing. For each study, we downloaded the raw RNA-seq data from one of the six databases (European Genome-phenome Archive (EGA), European Nucleotide Archive (ENA), Array Express, Gene Expression Omnibus (GEO), Database of Genotypes and Phenotypes (dbGaP), Synapse). If the data were already in fastq format then we proceeded directly to

quantification. If the raw data were shared in BAM or CRAM format, we used samtools v1.6 [57] to first collate paired-end reads with `samtools collate` and then used `samtools fastq` command with '`-F 2816 -c 6`' flags to convert the CRAM or BAM files to fastq. Since samples from GEO and dbGaP were stored in SRA format, we used the `fastq-dump` command with '`--split-files --gzip --skip-technical --readids --dumpbase --clip`' flags to convert those to fastq. The pre-proccessing scripts are available from the rnaseq quantification pipeline GitHub repository (https://github.com/eQTL-Catalogue/rnaseq).

Quantification. We quantified transcription at four different levels: (1) gene expression, (2) exon expression, (3) transcript usage and (4) transcriptional event usage. Quantification was performed with a Nextflow-based [18] pipeline that we developed by adding new quantification methods to nf-core rnaseq pipeline [17]. Before quantification, we used Trim Galore v0.5.0 to remove sequencing adapters from the fastq files.

For gene expression quantification, we used HISAT2 v2.1.0 [58] to align reads to the GRCh38 reference genome (Homo_sapiens.GRCh38.dna.primary_assembly.fa file downloaded from Ensembl). We counted the number of reads overlapping the genes in the GENCODE V30 [59] reference transcriptome annotations with featureCounts v1.6.4 [60]. To quantify exon expression, we first created exon annotation file (GFF) using GENCODE V30 reference transcriptome annotations and `dexseq_prepare_annotation.py` script from the DEXSeq [19] package. We then used the aligned RNA-seq BAM files from the gene expression quantification and featureCounts with flags '`-p -t exonic_part -s ${direction} -f -O`' to count the number of reads overlapping each exon.

We quantified transcript and event expression with Salmon v0.13.1 [61]. For transcript quantification, we used GENCODE V30 (GRCh38.p12) reference transcript sequences (fasta) file to build Salmon index. For transcriptional event usage, we downloaded pre-computed txrevise [15] alternative promoter, splicing and alternative 3' end annotations corresponding to Ensembl version 96 from Zenodo (https://doi.org/10.5281/zenodo.3232932) in GFF format. We then used gffread to generate fasta sequences from the event annotations and built Salmon indexes for each event set as we did for transcript usage. Finally, we quantified transcript and event expression using `salmon quant` with '`--seqBias --useVBOpt --gcBias --libType`' flags. All expression matrices were merged using csvtk v0.17.0. The pipeline is publicly available at https://github.com/eQTL-Catalogue/rnaseq. Our reference transcriptome annotations are available from Zenodo (https://doi.org/10.5281/zenodo.3366280).

Detecting outliers from gene expression data. We performed the quality control measures using only gene expression counts matrix. In all downstream analyses, we only included 35,367 protein coding and non-coding RNA genes belonging to one of the following Ensembl gene types: lincRNA, protein_coding, IG_C_gene, IG_D_gene, IG_J_gene, IG_V_gene, TR_C_gene, TR_D_gene, TR_J_gene, TR_V_gene, 3prime_overlapping_ncrna, known_ncrna, processed_transcript, antisense, sense_intronic, sense_overlapping. For PCA and MDS analyses, we first filtered out invalid gene types (23,458) and genes in sex chromosomes (1,247), TPM normalised [62] the gene counts, filtered out genes having median normalised expression value less than 1 and $\log_2$ transformed the matrix. We performed principal

component analysis with `prcomp` R stats package (center = true, scale = true). For multidimensional scaling (MDS) analysis, we used the isoMDS method from MASS R package with k=2 dimensions. As a distance metric for isoMDS we used 1 - Pearson correlation as recommended previously [63]. We plotted these two-dimensional scatter plots to visually identify outliers (Supplementary Figure 4A-B).

Sex-specific gene expression analysis. Previous studies have successfully used the expression of *XIST* and Y chromosome genes to ascertain genetic sex of RNA samples [64]**.** In our analysis, we extracted all protein coding genes from the Y chromosome and *XIST* gene (ENSG00000229807) expression values and TPM normalised them. Then, we calculated mean value of expressions of Y chromosome genes. Finally, we plotted $\log_2$ scatter plot of *XIST* gene expression (X axis) against the mean expression of Y chromosome genes (Y axis) (Supplementary Figure 4C). In addition to detecting samples with incorrectly labeled genetic sex, this analysis also allowed us to identify cross-contamination between samples (*XIST* and Y chromosome genes expressed simultanously, Supplementary Figure 4C).

Concordance between genotype data and RNA-seq samples. We used the Match Bam to VCF (MBV) method from QTLTools [65] which directly compares the sample genotypes in VCF to an aligned RNA-seq BAM file. MBV is a good method to detect sample swaps, genotypes from the same donor and cross-contaminated genotypes in VCF. In some cases, such cross-contamination was confirmed by the both sex-specific gene expression and MBV analyses (Supplementary Figure 4D).

Normalisation. We filtered out samples which failed the QC step. We normalised the gene and exon-level read counts using the conditional quantile normalisation (cqn) R package v1.30.0 [66]. We downloaded the gene GC content estimates from Ensembl biomaRt and calculated the exon-level GC content using bedtools v2.19.0 [67]. We also excluded lowly expressed genes, where 95 per cent of the samples within a biological context had TPM normalised expression less than 1. To calculate transcript and transcriptional event usage values, we obtained the TPM normalised transcript (event) expression estimates from Salmon and divided those by the total expression of all transcripts (events) from the same gene (event group). Subsequently, we used the inverse normal transformation to standardise the transcript and event usage estimates. Normalisation scripts together with containerised software is publicly available at https://github.com/eQTL-Catalogue/qtl_norm_qc.

Metadata harmonisation

We mapped all RNA-seq and microarray samples to a minimal metadata model. This included consistent sample identifiers, information about the cell type or tissue of origin, biological context (e.g. stimulation), genetic sex, experiment type (RNA-seq or microarray) and properties of the RNA-seq protocol (paired-end vs single-end; stranded vs unstranded; poly(A) selection vs total RNA). To ensure that cell type and tissue names were consistent between studies and to facilitate easier integration of additional studies, we used Zooma (https://www.ebi.ac.uk/spot/zooma/) to map cell types and tissues to controlled vocabulary of ontology terms from Uber-anatomy ontology (Uberon) [20], Cell Ontology [21] or Experimental

Factor Ontology (EFO) [22]. We opted to use *ad-hoc* controlled vocabulary to represent biological contexts as those often included terms and combinations of terms that were missing from ontologies.

Association testing

We developed a Nextflow based pipeline which takes normalised phenotype expression matrix, genotype VCF file and metadata files and produces association summary statistics for all molecular phenotypes. We performed association testing separately in each biological context (also known as 'qtl group') and used a +/- 1 megabase *cis* window centered around the start of each gene. First, we excluded molecular phenotypes that had less than 5 genetic variants in their *cis* window, as these were likely to reside in regions with poor genotyping coverage. We also excluded molecular phenotypes with zero variance across all samples and calculated phenotype principal components using `prcomp` R stats package (center = true, scale = true). We calculated genotype principal components using plink2 v1.90b3.35. We used the first six genotype and phenotype principal components as covariates in QTL mapping. For association testing, we used QTLtools v1.1 [68] nominal and permutation passes in *cis*. For nominal pass, we used the '`--window 1000000 --nominal 1`' flags to find all associations in 1 Mb *cis* window. For permutation pass, we used '`--window 1000000 --permute 1000 --grp-best`' flags in order to calculate empirical p-values based on 1000 permutations. The '`--grp-best`' option ensured that the permutations were performed across all phenotypes within the same 'group' (e.g. multiple probes per gene in microarray data or multiple transcripts or exons per gene in the exon-level and transcript-level analysis) and the empirical p-value was calculated at the group level.

Colocalisation

We used the GWAS Catalog [69] API (https://www.ebi.ac.uk/gwas/docs/api) to download the rheumatoid arthritis [42] GWAS summary statistics (accession GCST002318) from the *CD40* locus (GRCh38 coordinates: chr20:45,980,000-46,200,000). We downloaded the eQTL summary statistics from the eQTL Catalogue API and performed colocalisation using the coloc R package [12] with default prior probabilities.

Software availability

Data analysis pipelines:

- RNA-seq quantification: https://github.com/eQTL-Catalogue/rnaseq
- Normalisation and QC: https://github.com/eQTL-Catalogue/qtl_norm_qc
- Genotype QC: https://github.com/eQTL-Catalogue/genotype_qc
- Association testing: https://github.com/eQTL-Catalogue/qtlmap

Example use cases:

- Colocalisation in R using GWAS Catalog and eQTL Catalogue APIs: https://github.com/eQTL-Catalogue/eQTL-Catalogue-resources/blob/master/scripts/eQTL_API_usecase.Rmd
- Python example for querying the HDF5 files: https://github.com/eQTL-Catalogue/eQTL-SumStats/blob/master/querying_hdf5_basics.ipynb

Data availability

The full association summary statistics in HDF5 and TSV format can be downloaded from the eQTL Catalogue website (https://www.ebi.ac.uk/eqtl/Data_access/). Slices of the TSV files can be accessed using tabix. All of the summary statistics are also available via the REST API (https://www.ebi.ac.uk/eqtl/api-docs/). Database accessions for the raw gene expression and genotype datasets are listed on the eQTL Catalogue website (https://www.ebi.ac.uk/eqtl/Datasets/). Our summary statistics have also been integrated to the Open Targets Genetic Portal (https://genetics.opentargets.org/) and gene expression matrices will be made available via the EMBL-EBI Expression Atlas [70]

Author contributions

NK and KA developed the data analysis and quality control workflows and performed quality control of the data. NK processed the RNA-seq datasets and performed the QTL analysis. JH developed and implemented the eQTL Catalogue API. JM processed the gene expression data for the Expression Atlas. LK performed microarray gene expression data normalisation and quality control. KP and MS developed the initial version of the population assignment workflow. TB, SJ, IP, DZ and KA supervised the work. NK and KA wrote the manuscript with input from all authors.

Acknowledgements

The RNA-seq quantification and QTL analyses were performed at the High Performance Computing Center, University of Tartu. We thank Eleri Vako from the Grant Office of the University of Tartu, and Holly Foster and Paris Litterick from Open Targets for assistance in setting up data access agreements. We thank Jeremy Schwartzentruber for his helpful comments on the manuscript; Daniel Gaffney for guidance in setting up this project.

Funding

NK, JH and JM were supported by a grant from Open Targets (OTAR2-046). TB, SJ, IP, HP and DZ were supported on core EMBL funds. KA was supported by the European Regional Development Fund and the programme Mobilitas Pluss (MOBJD67). KA also received funding from the European Union's Horizon 2020 research and innovation programme (grant number

825775) and Estonian Research Council (grants IUT34-4 and PSG415). LK was supported by the Estonian Research Council grant PSG59. KA, NK and LK were also supported by Estonian Centre of Excellence in ICT Research (EXCITE) funded by the European Regional Development Fund.

Funding for datasets in the eQTL Catalogue

BLUEPRINT. This study makes use of data generated by the Blueprint Consortium. A full list of the investigators who contributed to the generation of the data is available from www.blueprint-epigenome.eu. Funding for the project was provided by the European Union's Seventh Framework Programme (FP7/2007- 2013) under grant agreement no 282510 - BLUEPRINT.
Fairfax_2012, Fairfax_2014 and Naranbhai_2015. Funding for the project was provided by the Wellcome Trust under awards Grants 088891 [B.P.F.], 074318 [J.C.K.] and 075491/Z/04 to the core facilities at the Wellcome Trust Centre for Human Genetics, the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) (281824 to J.C.K.), the Medical Research Council (98082, J.C.K.) and the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre.
TwinsUK. TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London.
BrainSeq. This research was supported by the Intramural Research Program of the NIMH (NCT00001260, 900142).
Schmiedel_2018. This work was funded by the William K. Bowes Jr Foundation (P.V.) and NIH grants R24AI108564 (P.V., B.P., A.R., M.K.), S10RR027366 (BD FACSAria II), and S10OD016262 (Illumina HiSeq 2500).
ROSMAP. Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P304G10161, R014G15819, R014G17917, R01AG30!46, R014G36836, U014G32984, U014G46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute.
GENCORD. Emmanouil T Dermitzakis was supported by grants from the European Research Council (260927), Swiss National Science Foundation (31003A_130342, CRSI33_130326) Louis-Jeantet Foundation, and the Blueprint Consortium. Stylianos E Antonarakis was supported by grants from the European Research Council (249968), Swiss National Science Foundation (144082), and the Blueprint Consortium.
van_de_Bunt_2015. MvdB is supported by a Novo Nordisk postdoctoral fellowship run in partnership with the University of Oxford. ALG is a Wellcome Trust Senior Research Fellow in Basic Biomedical Science (095010/Z/10/Z). MIM is a Wellcome Trust Senior Investigator (WT098381) and a National Institute of Health Research Senior Investigator. PEM holds the Canada Research Chair in Islet Biology. This work was supported in part in Oxford, UK, by grants from the Medical Research Council (MRC; MR/L020149/1) and National Institutes of Health (NIH; R01 MH090941), and in Edmonton, Canada, by operating grants to PEM from the

Canadian Institutes of Health Research (CIHR; MOP244739) and the ADI/Johnson & Johnson Diabetes Research Fund. Human islet isolations at the Alberta Diabetes Institute IsletCore were funded by the Alberta Diabetes Foundation and the University of Alberta. The National Institute for Health Research, Oxford Biomedical Research Centre funded islet provision at the Oxford Human Islet Isolation facility. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary Materials

Supplementary Table 1. Samples assigned to the 1000 Genomes Phase 3 reference populations in each study. Note that three studies based on HipSci samples (HipSci, Alasoo_2018, Schwartzentruber_2018) and two studies based on Estonian Biobank samples (Kasela_2017, Lepik_2017) share a subset of donors by design. Furthermore, Fairfax_2012 and Naranbhai_2015 studies have been excluded, because donors in these two studies are subset of donors in Fairfax_2014. Thus, the total number of donors (n = 4,917) in this table exceeds the number of unique donors (n = 4685) presented in Table 2. Superpopulation codes: EUR - European, AFR - African, SAS - South Asian, EAS - East Asian.

Study	Sample size	EUR	AFR	SAS	EAS	Unassigned	Proportion
Alasoo_2018	84	84	0	0	0	0	0.017
BLUEPRINT	197	197	0	0	0	0	0.04
BrainSeq	479	232	226	1	0	20	0.097
CEDAR	322	322	0	0	0	0	0.065
Fairfax_2014	423	421	0	0	0	2	0.086
GENCORD	196	194	0	0	0	2	0.04
GEUVADIS	445	358	87	0	0	0	0.091
HipSci	322	320	0	1	0	1	0.065
Kasela_2017	295	295	0	0	0	0	0.06
Lepik_2017	471	471	0	0	0	0	0.096
Nedelec_2016	168	96	64	0	0	8	0.034
Quach_2016	200	100	100	0	0	0	0.041
ROSMAP	576	576	0	0	0	0	0.117
Schmiedel_2018	91	53	4	3	20	11	0.019
Schwartzentruber_2018	98	98	0	0	0	0	0.02
TwinsUK	433	432	0	0	0	1	0.088
van_de_Bunt_2015	117	117	0	0	0	0	0.024
Total	**4917**	**4366**	**481**	**5**	**20**	**45**	**1**

Supplementary Table 2. Overview of the the transcriptomic samples included in the eQTL Catalogue. The samples have been classified according to RNA-seq type (single-end vs paired-end), strandedness (unstranded vs stranded), read length (50bp, 75bp, 100bp), assay type (microarray vs RNA-seq) and genotype data type (imputed vs not imputed).

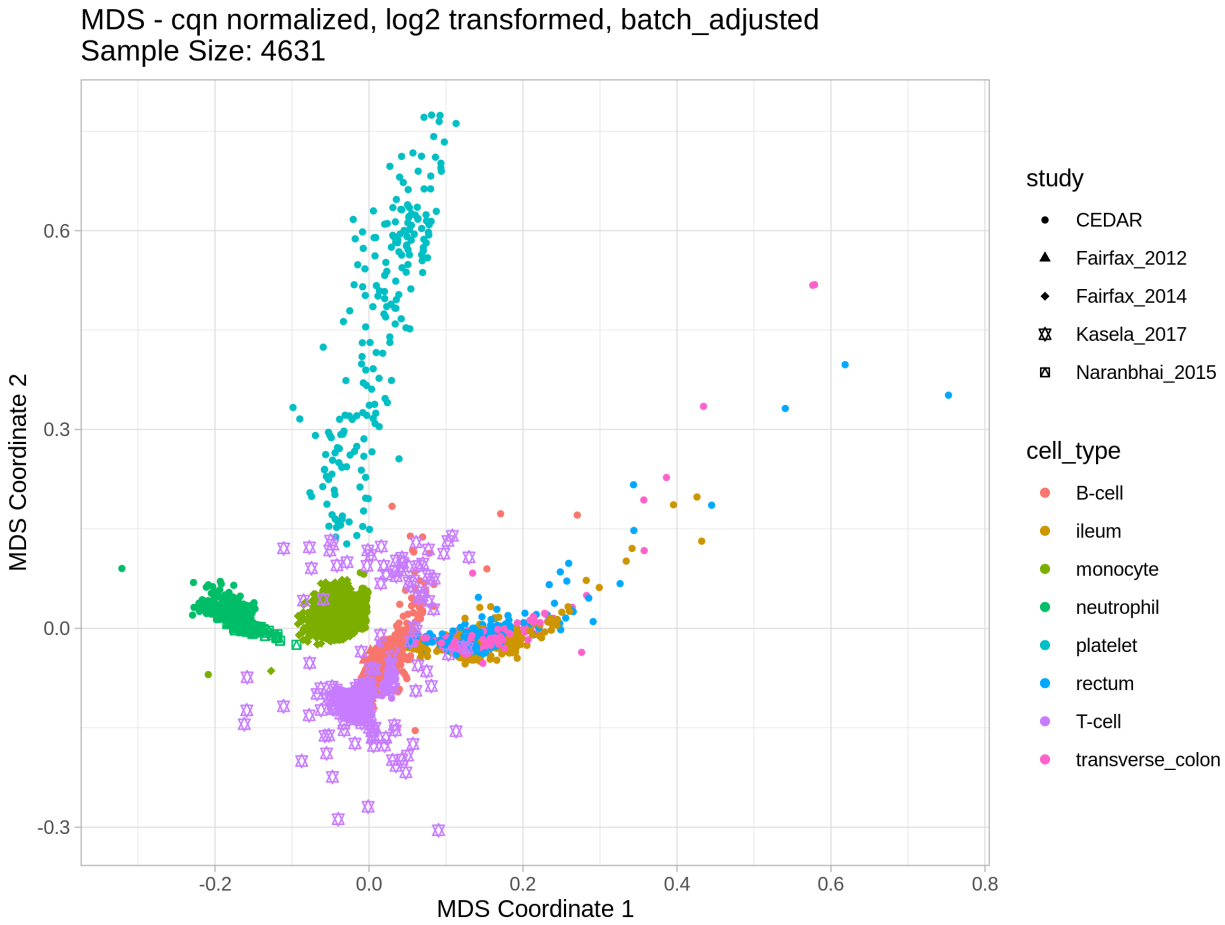
Group	Sample size	Number of studies	Proportion of studies
Single-end	3180	4	0.267
Paired-end	4935	11	0.733
Unstranded	3831	5	0.333
Stranded	4284	10	0.667
100bp	3188	7	0.467
50bp	3726	4	0.267
75bp	1201	4	0.267
microarray	4631	5	0.25
RNA-seq	8115	15	0.75
Not imputed	2834	5	0.25
Imputed	9912	15	0.75

Supplementary Table 3. Overview of the studies included in the eQTL Catalogue. *TwinsUK and HipSci studies contain related individuals by design. These were excluded in the quality control step to enable eQTL analysis with a linear model.

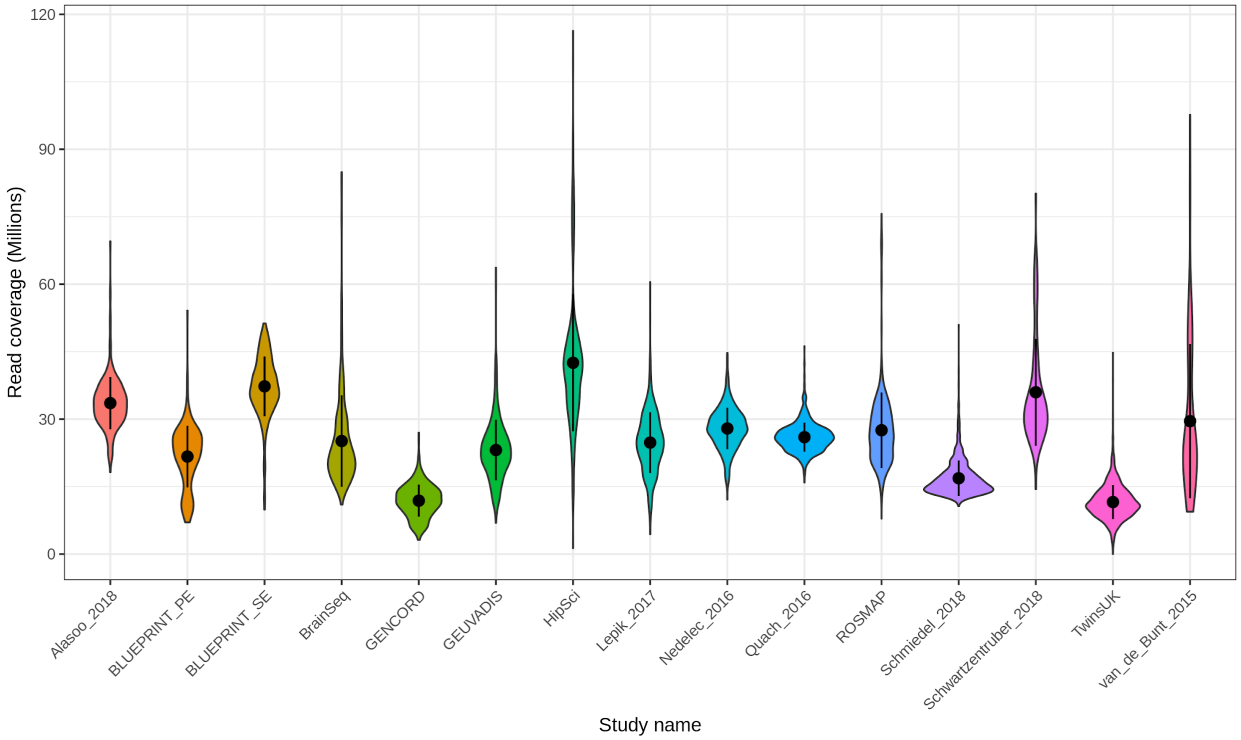
Study	Dataset Type	Imputed	Paired-end	Stranded	Read length	Pre-QC sample size	Post-QC sample size
Alasoo_2018	RNA-seq	YES	YES	YES	75bp	336	336
BLUEPRINT_PE	RNA-seq	NO	YES	YES	100bp	221	167
BLUEPRINT_SE	RNA-seq	NO	NO	YES	100bp	387	387
BrainSeq	RNA-seq	YES	YES	YES	100bp	495	479
GENCORD	RNA-seq	YES	YES	NO	50bp	567	560
GEUVADIS	RNA-seq	NO	YES	NO	75bp	462	445
HipSci	RNA-seq	YES	YES	YES	75bp	513	322
Lepik_2017	RNA-seq	NO	YES	YES	50bp	508	471
Nedelec_2016	RNA-seq	YES	NO	NO	100bp	503	493
Quach_2016	RNA-seq	YES	NO	NO	100bp	970	969
ROSMAP	RNA-seq	YES	YES	YES	100bp	581	576
Schmiedel_2018	RNA-seq	YES	NO	YES	50bp	1544	1331
Schwartzentruber_2018	RNA-seq	YES	YES	YES	75bp	130	98
TwinsUK	RNA-seq	NO	YES	NO	50bp	2505	1364
van_de_Bunt_2015	RNA-seq	YES	YES	YES	100bp	118	117
CEDAR	microarray	YES	NA	NA	NA	2967	2337
Fairfax_2012	microarray	YES	NA	NA	NA	296	281
Fairfax_2014	microarray	YES	NA	NA	NA	1384	1371
Kasela_2017	microarray	YES	NA	NA	NA	576	549
Naranbhai_2015	microarray	YES	NA	NA	NA	101	93
				RNA-seq samples		**9840**	**8115**
				Microarray samples		**5324**	**4631**
				Total samples		**15164**	**12746**

Supplementary Table 4. List of rejected data access applications. While some of the datasets were restricted for research into specific diseases (marked with red), others were rejected due to restrictions on sharing individual-level genotype data (marked with yellow). We hope to include some of the datasets with genotype data sharing restrictions in a future version of the eQTL Catalogue by employing federated analysis approaches which do not require the transfer of genotype data.

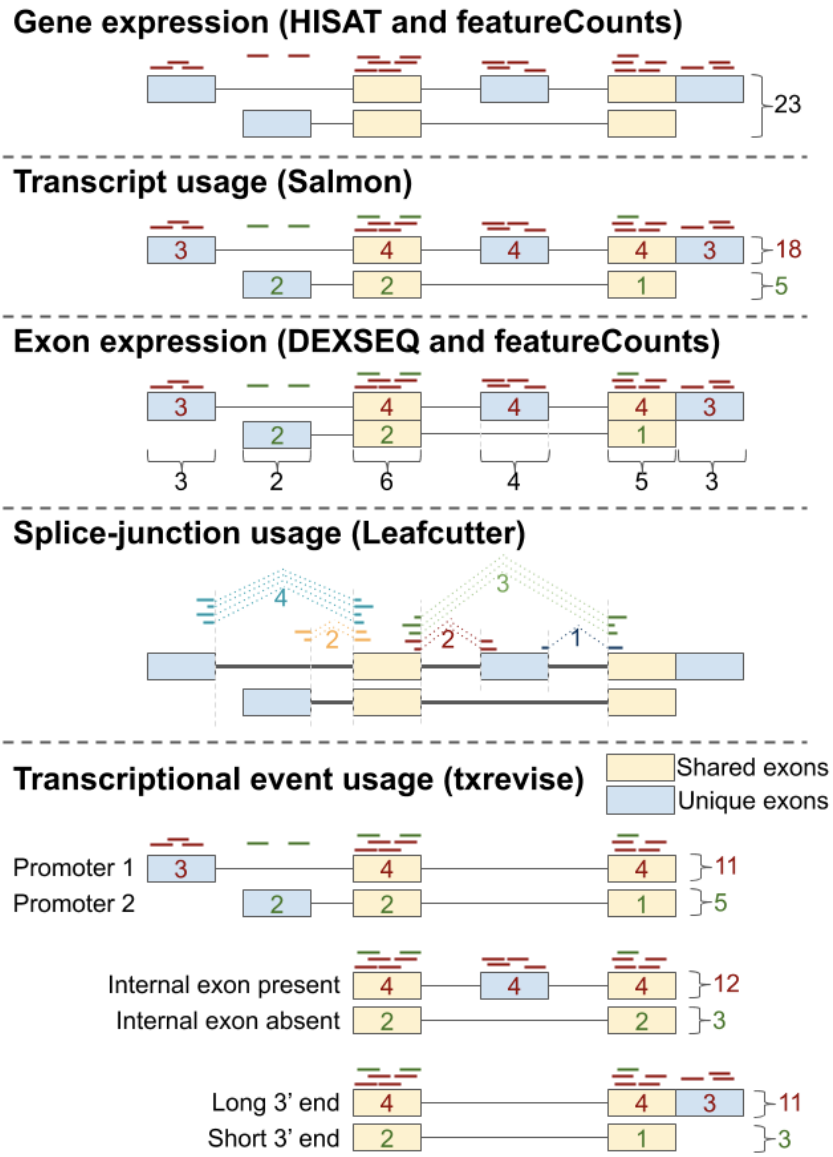
Study	Cell types or tissues	Reason for rejection
Raj_2014 [71]	monocyte, T cell	"Use of data is limited to research studying genetic variation in human immune system function."
Ye_2018 [72]	dendritic cell	"Use of data is limited to research studying genetic variation in human immune system function."
Gate_Cheng_2018 [73]	T cell	"Use of data is limited to research studying genetic variation in human immune system function."
Battle_2014 [74]	blood (n = 922)	"Study is not related to depression."
Gillies_2018 [75]	kidney (n = 187)	"No consent in place to share individual-level genotype data."
Fadista_2014 [76]	pancreatic islet (n = 89)	“Genotype data cannot be shared.”
Ishigaki_2017 [77]	T cell, monocyte, NK cell, B cell (n = 100)	“Genotype data cannot be shared.”
Qiu_2018 [78]	kidney (n = 151)	“No consent was obtained to share individual-level genotype data.”
BIOS [79]	Blood (n = 2,116)	“Genotype data can only be analysed on a centralised cloud service in the Netherlands.”



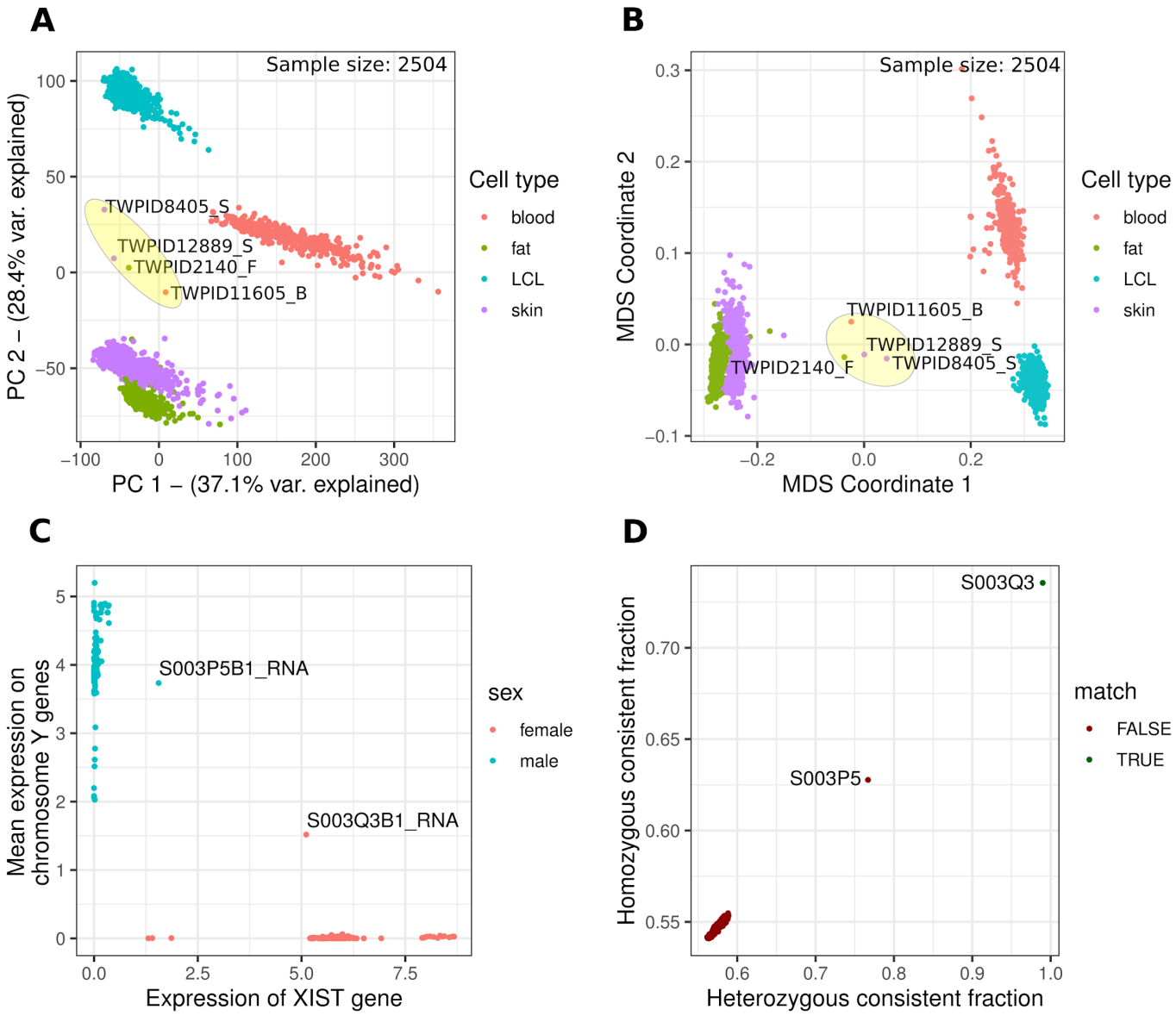
Supplementary Figure 1. Multidimensional scaling plot of the 4631 microarray samples that passed quality control (QC) from the five microarray studies included in the eQTL Catalogue (CEDAR, Fairfax_2012, Fairfax_2014, Kasela_2017, Naranbhai_2015). Similar MDS plots for individual studies can be found in the QC reports available from the eQTL Catalogue website (http://www.ebi.ac.uk/eqtl/Datasets/).



Supplementary Figure 2. Distribution of sample RNA sequencing read depth in each study. Samples from the BLUEPRINT [23] study have been split into paired-end (PE) and single-end (SE), because they were sequenced in two different laboratories using different RNA-seq protocols.

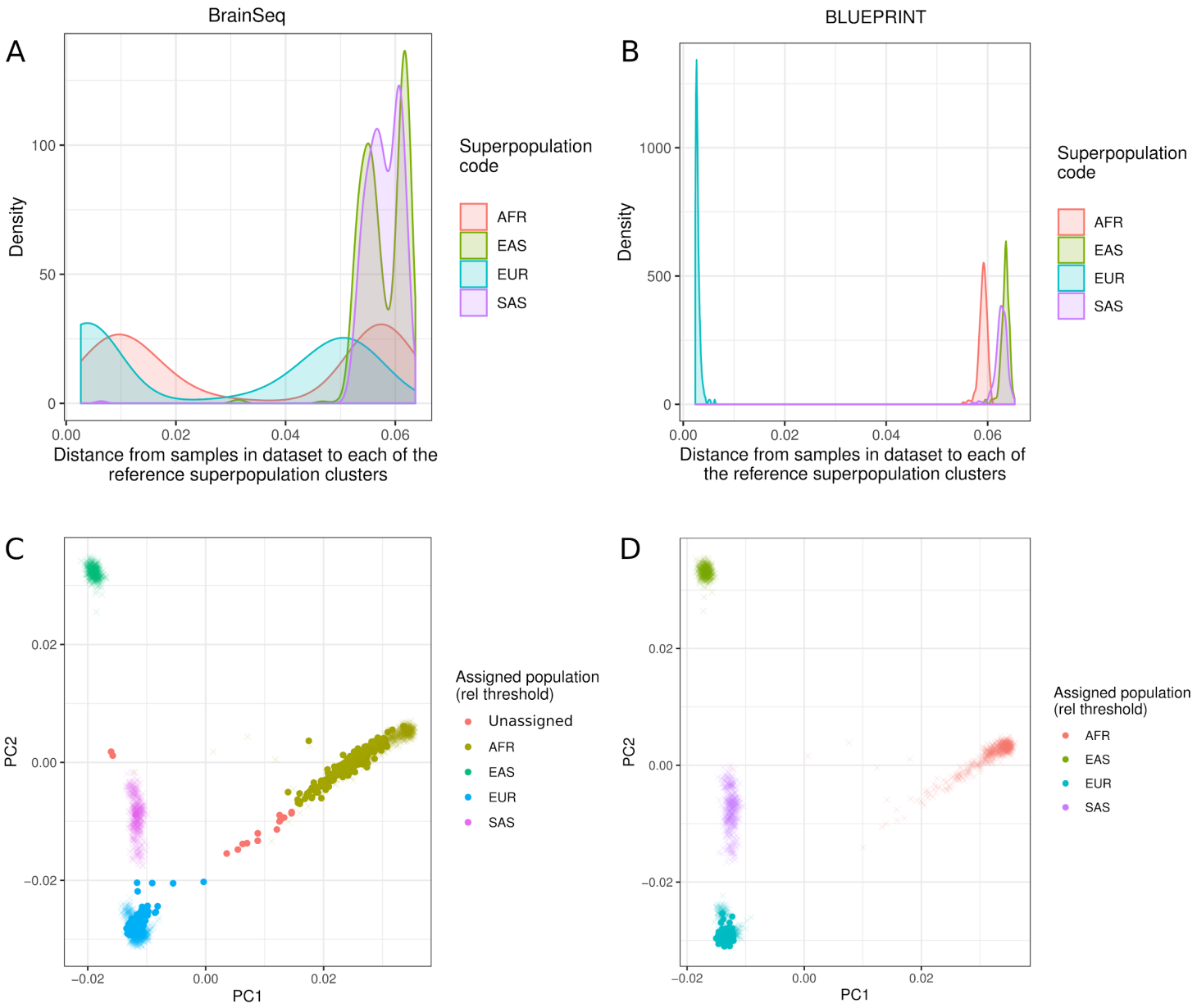


Supplementary Figure 3. Quantification methods of molecular phenotypes in the eQTL Catalogue. Symbolic representation of 23 read fragments assigned to 1 gene (aligned with HISAT2 [58], quantified with featureCounts [60]) consisting of 2 transcripts (quantified with Salmon [61]) and 6 exonic parts (annotated with DEXSeq [19], quantified with featureCounts). The gene also has 5 distinct introns which are identified and quantified by Leafcutter [48]. Transcriptional event usage is quantified with txrevise [15]. Txrevise uses shared exons as a scaffold to identify independent transcriptional events corresponding to alternative promoters, internal exons and 3' ends. Leafcutter splice junction QTLs will be included in a future version of the eQTL Catalogue.



Supplementary Figure 4. Overview of the Quality Control (QC) measures applied to all of the datasets in the eQTL Catalogue. QC reports for individual datasets can be found on the eQTL Catalogue website (https://www.ebi.ac.uk/eqtl/Datasets/). **(A)** Principal component analysis of the TwinsUK dataset. **(B)** Multidimensional scaling analysis of the TwinsUK dataset. Four outlier samples (highlighted in yellow) from the PCA and MDS analysis were excluded from QTL mapping. (**C)** Sex-specific gene expression analysis. Expression of the female-specific *XIST* gene is plotted against the mean expression the protein coding genes on the Y chromosome. Samples from two donors (S003P5 (male) and S003Q3 (female)) expressed both *XIST* and genes from the Y chromosome, indicating potential cross-contamination with RNA from a sample of different genetic sex. **(D)** Genetic similarity of S003Q3B1 RNA sample to all of the genotyped donors in the BLUEPRINT VCF file as calculated by the QTLtools mbv command [65]. As expected, the genotypes of the S003Q3B1 RNA sample are most similar to the genotype data from the same donor and most other donors are equally dis-similar, forming a

separate cluster at the bottom left corner. However, the S003Q3B1 RNA sample also displays higher-than-expected genetic similarity with genotype data from the S003P5 donor. Together with evidence presented on panel C, this suggests that cross-contamination has occurred between the S003Q3B1 and S003P5B1 RNA samples. As a result, we decided to remove these two samples from downstream analysis.



Supplementary Figure 5. Assigning genotyped samples to the four 1000 Genomes superpopulations. **(A)** Density plot of distances between each sample in BrainSeq [28] dataset and each superpopulation cluster in the 1000 Genomes Phase 3 reference dataset [16]. First three principal components of the genotype data are used to calculate distances. Majority of samples in the BrainSeq dataset are close to either European (EUR) or African (AFR) superpopulations. **(B)** Histogram of distances between each sample in the BLUEPRINT [23] dataset and each superpopulation cluster in reference dataset. All samples are close to the European (EUR) superpopulation cluster of the 1000 Genomes reference dataset. **(C)** Projection of the BrainSeq dataset to the first two principal components of the 1000 Genomes Phase 3 reference dataset. Most samples are assigned to either European or African superpopulations. Red samples are too far from all four superpopulations and thus remain unassigned. These samples are likely to represent recent admixture. **(D)** Projection of the

BLUEPRINT dataset to the first two principal components of the 1000 Genomes Phase 3 reference panel. All samples are assigned to the European superpopulation. Superpopulation codes: EUR - European, AFR - African, SAS - South Asian, EAS - East Asian.

References

1. Aguet F, Barbeira AN, Bonazzola R, Brown A, Castel SE, Jo B, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv. 2019. p. 787903. doi:10.1101/787903

2. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. bioRxiv. 2018. p. 447367. doi:10.1101/447367

3. Xia K, Shabalin AA, Huang S, Madar V, Zhou Y-H, Wang W, et al. seeQTL: a searchable database for human eQTLs. Bioinformatics. 2012;28: 451–452.

4. Munz M, Wohlers I, Simon E, Busch H, Schaefer AS, Erdmann J. Qtlizer: comprehensive QTL annotation of GWAS results. bioRxiv. 2019. p. 495903. doi:10.1101/495903

5. Zheng Z, Huang D, Wang J, Zhao K, Zhou Y, Guo Z, et al. QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. Nucleic Acids Res. 2020;48: D983–D991.

6. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8: 1826.

7. Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics. 2019;35: 4851–4853.

8. Kalayci S, Selvan ME, Ramos I, Cotsapas C, Montgomery RR, Poland G, et al. ImmuneRegulation: A web-based tool for identifying human immune regulatory elements. bioRxiv. 2018. p. 468124. doi:10.1101/468124

9. Yu C-H, Pal LR, Moult J. Consensus Genome-Wide Expression Quantitative Trait Loci and Their Relationship with Human Complex Trait Disease. OMICS. 2016;20: 400–414.

10. Wang G, Sarkar AK, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine-mapping. bioRxiv. 2018. p. 501114. doi:10.1101/501114

11. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics. 2016;32: 1493–1501.

12. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. 2014;10: e1004383.

13. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016. doi:10.1016/j.ajhg.2016.10.003

14. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a

primary link between genetic variation and disease. Science. 2016;352: 600–604.

15. Alasoo K, Rodrigues J, Danesh J, Freitag DF, Paul DS, Gaffney DJ. Genetic effects on promoter usage are highly context-specific and contribute to complex traits. Elife. 2019;8. doi:10.7554/eLife.41673

16. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526: 75–81.

17. Ewels P, Peltzer A, Fillinger S, Alneberg J, Patel H, Wilm A, et al. nf-core: Community curated bioinformatics pipelines. bioRxiv. 2019. p. 610741. doi:10.1101/610741

18. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35: 316–319.

19. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Res. 2012;22: 2008–2017.

20. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 2012;13: R5.

21. Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. J Biomed Semantics. 2016;7: 44.

22. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, et al. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 2010;26: 1112–1118.

23. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell. 2016;167: 1398–1414.e24.

24. Quach H, Rotival M, Pothlichet J, Loh Y-HE, Dannemann M, Zidane N, et al. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. Cell. 2016;167: 643–656.e17.

25. Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell. 2018;0. doi:10.1016/j.cell.2018.10.022

26. Momozawa Y, Dmitrieva J, Théâtre E, Deffontaine V, Rahmouni S, Charloteaux B, et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. Nat Commun. 2018;9: 2427.

27. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science. 2014;343: 1246949.

28. Jaffe AE, Straub RE, Shin JH, Tao R, Gao Y, Collado-Torres L, et al. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. Nat Neurosci. 2018;21: 1117–1125.

29. Ng B, White CC, Klein H-U, Sieberts SK, McCabe C, Patrick E, et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat Neurosci. 2017;20: 1418–1426.

30. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. Elife. 2013;2: e00523.

31. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501: 506–511.

32. Buil A, Brown AA, Lappalainen T, Viñuela A, Davies MN, Zheng H-F, et al. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. Nat Genet. 2015;47: 88–91.

33. Kasela S, Kisand K, Tserel L, Kaleviste E, Remm A, Fischer K, et al. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. PLoS Genet. 2017;13: e1006643.

34. Alasoo K, Rodrigues J, Mukhopadhyay S, Knights AJ, Mann AL, Kundu K, et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat Genet. 2018;50: 424–431.

35. Nédélec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, et al. Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell. 2016;167: 657–669.e21.

36. Lepik K, Annilo T, Kukuškina V, Kisand K, Kutalik Z, Peterson P, et al. C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. PLoS Comput Biol. 2017;13: e1005766.

37. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat Genet. 2012;44: 502–510.

38. Naranbhai V, Fairfax BP, Makino S, Humburg P, Wong D, Ng E, et al. Genomic modulators of gene expression in human neutrophils. Nat Commun. 2015;6: 7545.

39. Kilpinen H, Goncalves A, Leha A, Afzal V, Alasoo K, Ashford S, et al. Common genetic variation drives molecular heterogeneity in human iPSCs. Nature. 2017;546: 370–375.

40. van de Bunt M, Manning Fox JE, Dai X, Barrett A, Grey C, Li L, et al. Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors. PLoS Genet. 2015;11: e1005694.

41. Schwartzentruber J, Foskolou S, Kilpinen H, Rodrigues J, Alasoo K, Knights AJ, et al. Molecular and functional variation in iPSC-derived sensory neurons. Nat Genet. 2018;50: 54–61.

42. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis

contributes to biology and drug discovery. Nature. 2014;506: 376–381.

43. Guo H, Fortune MD, Burren OS, Schofield E, Todd JA, Wallace C. Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. Hum Mol Genet. 2015;24: 3305–3313.

44. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 2011;27: 718–719.

45. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat Neurosci. 2016;19: 1442–1453.

46. Taylor DL, Jackson AU, Narisu N, Hemani G, Erdos MR, Chines PS, et al. Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. Proc Natl Acad Sci U S A. 2019;116: 10883–10888.

47. Schwartzentruber J, Cooper S, Liu JZ, Barrio-Hernandez I, Bello E, Kumasaka N, et al. Genome-wide meta-analysis, fine-mapping, and integrative prioritization identify new Alzheimer's disease risk genes. Genetic and Genomic Medicine. medRxiv; 2020. doi:10.1101/2020.01.22.20018424

48. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, et al. Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet. 2018;50: 151–158.

49. Deelen P, Bonder MJ, van der Velde KJ, Westra H-J, Winder E, Hendriksen D, et al. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC Res Notes. 2014;7: 901.

50. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48: 1284–1287.

51. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014;30: 1006–1007.

52. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4: 7.

53. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet. 2012;91: 1011–1021.

54. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47.

55. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008;24: 1547–1548.

56. Westra H-J, Jansen RC, Fehrmann RSN, te Meerman GJ, van Heel D, Wijmenga C, et al. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics. 2011;27: 2104–2111.

57. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

58. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37: 907–915.

59. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22: 1760–1774.

60. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30: 923–930.

61. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14: 417–419.

62. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131: 281–285.

63. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. Science. 2015;348: 660–665.

64. 't Hoen PAC, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol. 2013;31: 1015–1022.

65. Fort A, Panousis NI, Garieri M, Antonarakis SE, Lappalainen T, Dermitzakis ET, et al. MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. Bioinformatics. 2017;33: 1895–1897.

66. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012;13: 204–216.

67. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842.

68. Delaneau O, Ongen H, Brown AA, Fort A, Panousis NI, Dermitzakis ET. A complete tool set for molecular QTL discovery and analysis. Nat Commun. 2017;8: 15452.

69. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47: D1005–D1012.

70. Papatheodorou I, Moreno P, Manning J, Fuentes AM-P, George N, Fexova S, et al. Expression Atlas update: from tissues to single cells. Nucleic Acids Res. 2020;48: D77–D83.

71. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, Replogle JM, et al. Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes. Science. 2014;344: 519–523.

72. Ye CJ, Chen J, Villani A-C, Gate RE, Subramaniam M, Bhangale T, et al. Genetic analysis of isoform usage in the human anti-viral response reveals influenza-specific regulation of

ERAP2 transcripts under balancing selection. Genome Res. 2018;28: 1812–1825.

73. Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, Lituiev D, et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat Genet. 2018. doi:10.1038/s41588-018-0156-2

74. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014;24: 14–24.

75. Gillies CE, Putler R, Menon R, Otto E, Yasutake K, Nair V, et al. An eQTL Landscape of Kidney Tissue in Human Nephrotic Syndrome. Am J Hum Genet. 2018;103: 232–244.

76. Fadista J, Vikman P, Laakso EO, Mollet IG, Esguerra JL, Taneera J, et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proc Natl Acad Sci U S A. 2014;111: 13924–13929.

77. Ishigaki K, Kochi Y, Suzuki A, Tsuchida Y, Tsuchiya H, Sumitomo S, et al. Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. Nat Genet. 2017;49: 1120–1125.

78. Qiu C, Huang S, Park J, Park Y, Ko Y-A, Seasock MJ, et al. Renal compartment–specific genetic variation analyses identify new pathways in chronic kidney disease. Nat Med. 2018;24: 1721–1731.

79. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017;49: 139–145.