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

Nurlan Kerimov[1,2,†], James D. Hayhurst[2,3,†], Kateryna Peikova[1], Jonathan R. Manning[2,3], Peter Walter[3], Liis Kolberg[1], Marija Samoviča[1], Manoj Pandian Sakthivel[2,3], Ivan Kuzmin[1], Stephen J. Trevanion[2,3], Tony Burdett[2,3], Simon Jupp[2,3], Helen Parkinson[2,3], Irene Papatheodorou[2,3], Andrew Yates[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 (eQTL) studies have made summary statistics publicly available, which can be used to gain insight into complex human 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. Consequently, target genes for most GWAS signals have still not been identified. Here, we present the eQTL Catalogue (https://www.ebi.ac.uk/eqtl/), a resource which contains quality controlled, uniformly re-computed QTLs from 21 eQTL studies. We find that for matching cell types and tissues, the eQTL effect sizes are highly reproducible between studies, enabling the integrative analysis of these data. Although most *cis*-eQTLs were shared between most bulk tissues, the analysis of purified cell types identified a greater diversity of cell-type-specific eQTLs, a subset of which also manifested as novel disease colocalisations. Our summary statistics can be downloaded by FTP, accessed via a REST API, and visualised on the Ensembl genome browser. New datasets will continuously be added to the eQTL Catalogue, enabling the systematic interpretation of human GWAS associations across many cell types and tissues.

Introduction

Gene expression and splicing QTLs are a powerful tool to link disease-associated genetic variants to putative target genes. However, 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, target genes and relevant biological contexts for most GWAS signals have not been found yet. Systematic colocalisation efforts based on GTEx data have identified putative target genes for 47% of the GWAS loci (*3*). Still, these genetic effects mediate only 11% of disease heritability (*4*), suggesting that many regulatory effects cannot be detected in bulk tissues at a steady-state (*5*). In contrast, profiling specialised disease-relevant cell types such as induced pluripotent stem cells (*6*), peripheral immune cells (*7*), microglia (*8*, *9*) or dopaminergic neurons (*10*) often identifies additional colocalisations that are missing in GTEx. While several databases have been developed to collect eQTL summary statistics from individual studies (*11*–*17*), these efforts have relied on the heterogeneous set of files provided by the original authors. These results often contain only a small subset of significant associations or lack essential details such as effect alleles, standard errors or sample sizes, which limit the downstream colocalisation and Mendelian randomisation analyses that can be performed (*18*).

Moreover, there is considerable technical variation between studies in sample collection, RNA sequencing, genotyping and data analysis. Thus, it is currently unclear how strongly eQTL effect sizes are influenced by technical differences in sample collection, how many eQTLs are broadly shared, and what fraction are specific to a given cell or tissue type and could thus give rise to novel disease colocalisations. While analyses based on GTEx data have generally estimated high levels of eQTL sharing between most bulk tissues (*1*, *19*), smaller studies have often estimated much lower levels of sharing between purified cell types (*20*, *21*). However, these analyses are sensitive to how sharing is defined, which genes and variants are included in the analysis and which analytical approaches are used (*19*, *22*). Thus, it is impossible to directly compare the estimates of eQTL sharing between studies without re-analysing the individual-level data with uniform methods.

Recent methodological advances have made it feasible to fine map genetic associations to small credible sets of putative causal variants and distinguish between multiple independent genetic signals in the region (*23*, *24*). These fine mapping results can be directly used in colocalisation analysis (*25*). They can also help avoid the many false negative colocalisations missed by approaches that assume a single causal variant in the region of interest (*18*). However, reliable fine mapping requires precise information about in-sample linkage disequilibrium (LD) between genetic variants which is usually not available (*26*, *27*).

To overcome these limitations, we have uniformly re-processed (see Figure 1) individual-level eQTL data from 112 datasets across 21 independent studies (see Figure 2). We find that eQTL effect sizes from matched cell types or tissues are generally highly reproducible between studies. Using both eQTL sharing and matrix factorisation approaches on fine mapped eQTL signals, we find that differences in eQTL effect sizes between datasets are dominated by

biological differences between cell types and tissues rather than technical differences in sample processing. Uniformly processed summary statistics provided us with a unique opportunity to characterise eQTL diversity across 69 distinct cell types and tissues. Consistent with previous analyses by the GTEx project, we find high levels of *cis*-eQTL sharing between most bulk tissues. In contrast, we find that a much smaller proportion of eQTLs are shared between purified cell types and bulk tissues, and between different cell types. This eQTL diversity also manifests itself at the level of disease colocalisation, where we detect many novel colocalisations that are missed when analysing GTEx data alone. Finally, in addition to gene expression QTLs, we have identified QTLs at the levels of exon expression, transcript usage, and splicing, which were often absent from the original studies. Our uniformly processed QTL summary statistics and fine mapping results are available from the eQTL Catalogue FTP server and REST API and they can also be explored using the Ensembl Genome Browser (*28*) (Figure 1B).

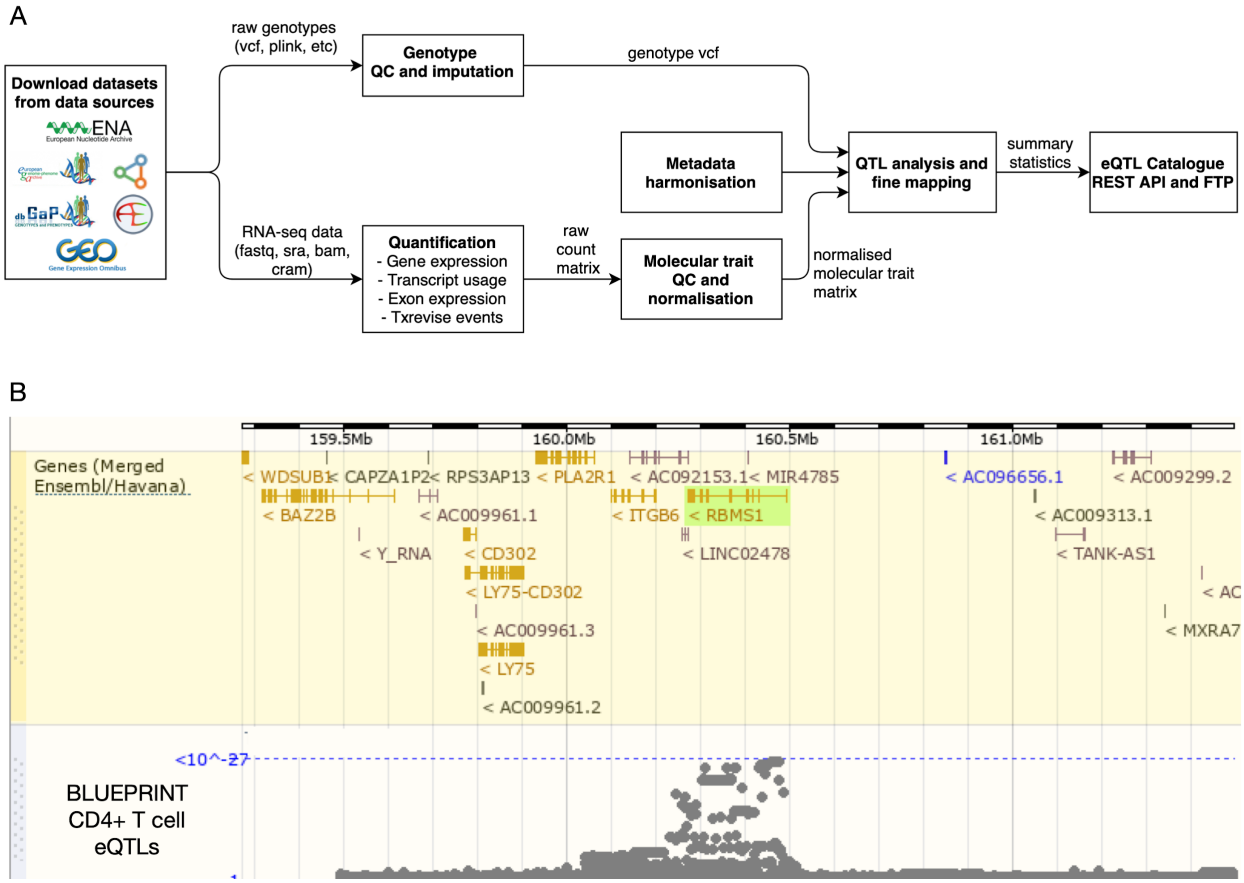


Figure 1. Overview of the eQTL Catalogue database. (**A**) A high-level representation of the uniform data harmonisation and eQTL mapping process. Supplementary Figure 1 provides a schematic illustration of the different quantification methods. (**B**) eQTL Catalogue summary results for the *RBMS1* gene in BLUEPRINT CD4+ T cells, viewed via the Ensembl Genome Browser.

Results

Studies, datasets and samples included in the eQTL Catalogue

We downloaded raw gene expression and genotype data from 16 RNA-seq and five microarray studies from various repositories. The RNA-seq data consisted of 17,210 samples spanning 95 datasets (defined as distinct cell types, tissues or contexts in which eQTL analysis was performed separately). These 95 datasets originated from 66 distinct cell types and tissues and ten stimulated conditions (Figure 2A). Similarly, the 4,631 microarray samples spanned 17 datasets from eight distinct cell types and tissues and three stimulated conditions (Figure 2B). While most cell types and tissues were profiled only by two of the largest studies (GTEx (*1*) and Schmiedel_2018 (*21*), Figure 2A), 13 cell types or tissues were captured by multiple studies, allowing us to characterise both technical and biological variability between datasets and studies. The total number of unique donors across studies was 5,714, of which 89% had predominantly European ancestries and only 9% had African or African American ancestries, with other ancestries being rare (Figure 2C, Supplementary Table 1). Thus, similarly to most GWAS studies, published eQTL studies also suffer from a lack of genetic diversity (*29*).

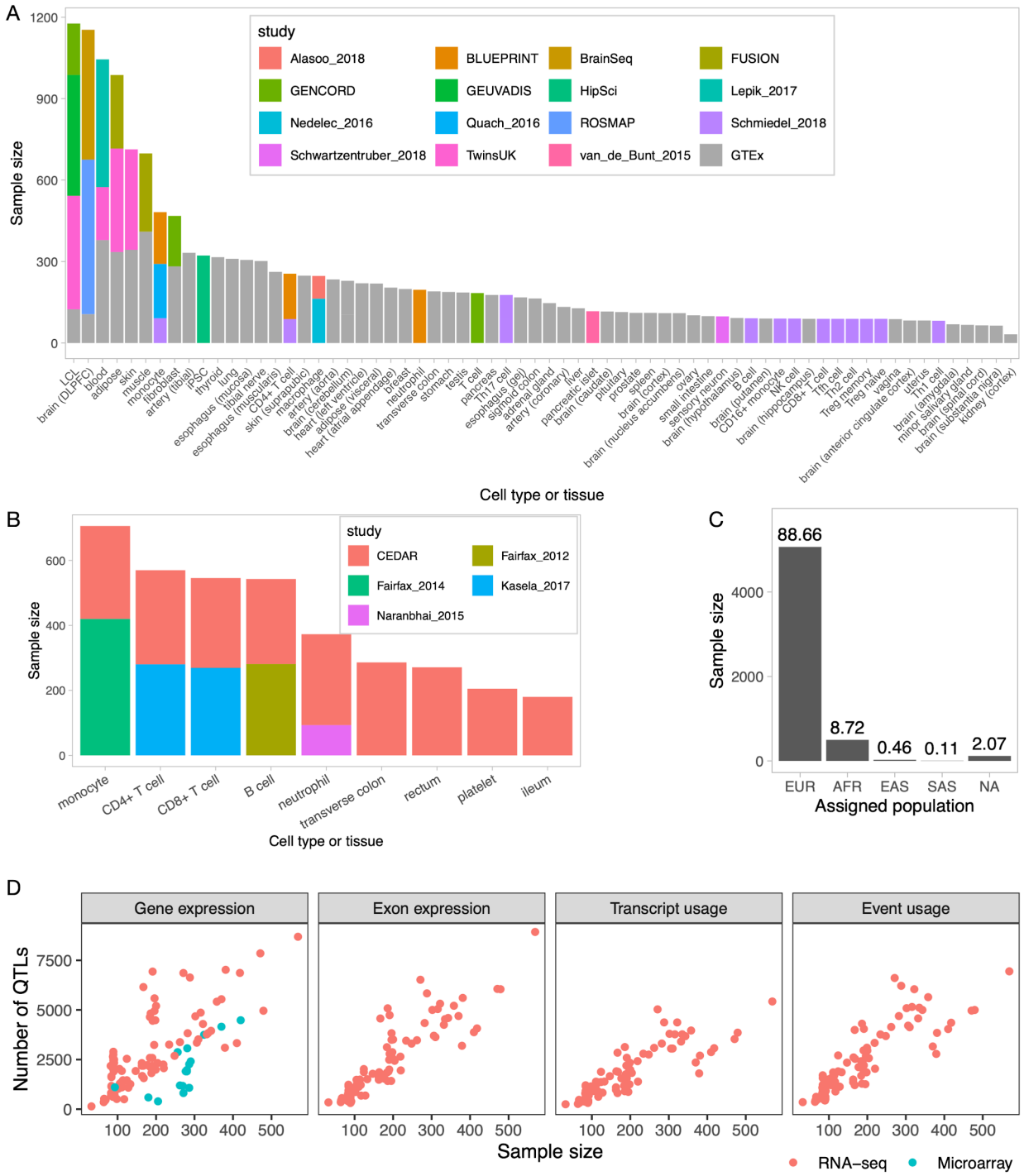


Figure 2. Overview of studies and samples included in the eQTL Catalogue. (**A**) Cumulative RNA-seq sample size for each cell type and tissue across 16 studies. Datasets from stimulated conditions have been excluded to improve readability. DLPFC - dorsolateral prefrontal cortex, iPSC - induced pluripotent stem cell, LCL - lymphoblastoid cell line. (**B**) The cumulative microarray sample size for each cell type and tissue across five studies. Datasets from stimulated conditions have been excluded to improve readability. (**C**) The 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. Superpopulation codes: EUR - European, AFR - African, EAS - East Asian, SAS - South Asian, NA - unassigned. (**D**) The relationship between the sample size of each dataset and the number of associations detected with each quantification method. The number of QTLs on the y-axis is defined as the number of genes with at least one significant QTL (FDR < 0.05).

To uniformly process a large number of eQTL studies, we designed a modular and robust data analysis workflow (Figure 1A). First, we performed extensive quality control and imputed missing genotypes using the 1000 Genomes Phase 3 reference panel (*30*) (Supplementary Table 3). For RNA-seq datasets, we performed QTL mapping for the four molecular traits described above (Figure 1A, Supplementary Figure 1). The QTL analysis was performed separately in each dataset (i.e. separately for each cell type or tissue within each study). We found the largest number of QTLs at the level of gene expression, but for all molecular traits the number of significant associations scaled approximately linearly with the sample size (Figure 2D, Supplementary Material 1). For microarray datasets, we performed the analysis only at the gene level but found the same linear trend (Figure 2D, Supplementary Material 1). Our remaining analyses focus on the RNA-seq-based eQTL datasets as they cover a more comprehensive range of cell types and tissues, and account for most of the samples in the eQTL Catalogue.

Biological and technical variability between studies and datasets

First, we assessed if the gene expression and eQTL signals were dominated by technical differences between studies (Supplementary Tables 2-3) rather than true biological differences between cell types and tissues. We visualised median transcripts per million (TPM) gene expression estimates from each dataset using multidimensional scaling (MDS). Reassuringly, we found that the datasets clustered predominantly by cell type or tissue of origin, rather than by studies or other technical factors (Figure 3A). Notably, except for brain tissues, whole blood and testis, most other bulk tissues had relatively similar gene expression profiles (Figure 3A). In contrast, datasets from purified cell types such as lymphoblastoid cell lines (LCLs), monocytes, neutrophils, induced pluripotent stem cells (iPSCs), and B and T lymphocytes had more distinct gene expression profiles (Figure 3A).

Next, we performed the same similarity analysis on eQTL effect sizes. To overcome the high uncertainty associated with effect size estimates, especially in datasets with small sample sizes, we used the recently developed multiple adaptive shrinkage (mash) model (*19*). Mash improves eQTL effect size estimates by sharing information both across datasets as well as individual eQTLs. We limited our analysis to 54,733 fine mapped eQTLs (see Methods) and defined two eQTLs to be shared between a pair datasets if they had the same sign and their effect sizes did not differ more than two-fold. We calculated pairwise eQTL sharing estimates for all 95 RNA-seq datasets (including 48 tissues from GTEx v7) and projected those onto two dimensions using MDS. Reassuringly, we found that if the same cell type or tissue was profiled in multiple

studies, then their eQTL effect sizes often showed a high degree of concordance (Figure 3B, Supplementary Figures 2-3). For example, LCLs from TwinsUK, GENCORD and GEUVADIS clustered together with LCLs from GTEx (Figure 3B) and exhibited median sharing of ~80% (Figure 3C, Supplementary Figures 2-3). The same was also true for the brain (GTEx, ROSMAP and BrainSeq studies), whole blood (GTEx, TwinsUK and Lepik_2017 studies), muscle (GTEx and FUSION), skin (GTEx and TwinsUK) and adipose tissues (GTEx, TwinsUK and FUSION), which all had median intra-tissue sharing of ~70% (Figure 3C). Moreover, the two-dimensional MDS plot of pairwise eQTL similarity (Figure 3B) was broadly similar to the pairwise gene expression similarity plot presented above (Figure 3A), suggesting that high gene expression similarity and a high degree of eQTL sharing both reflect similarity in the underlying regulatory state of cells.

Finally, we focussed on the patterns of sharing between different cell types and tissues. We found that 46-80% (median 62%) of the eQTLs were shared between most pairs of bulk tissues (Figure 3C). The exception to this pattern were the brain tissues and whole blood that formed separate clusters in the MDS analysis (Figure 3B) and shared a median of 45% and 35% of the eQTLs with other tissues, respectively (Figure 3C). In contrast, purified immune cell types (LCLs, neutrophils, monocytes, macrophages and lymphocytes) formed distinct clusters on the MDS plot (Figure 3B) and had much lower eQTL sharing both with whole blood as well as other bulk tissues (Figure 3C). Thus, although our results reconfirm the generally high level of *cis*-eQTL sharing between bulk tissues, they also reveal a much greater *cis*-eQTL diversity between purified cell types and especially immune cells. Importantly, this diversity is missed when analysing highly tissue-focused eQTL studies such as GTEx.

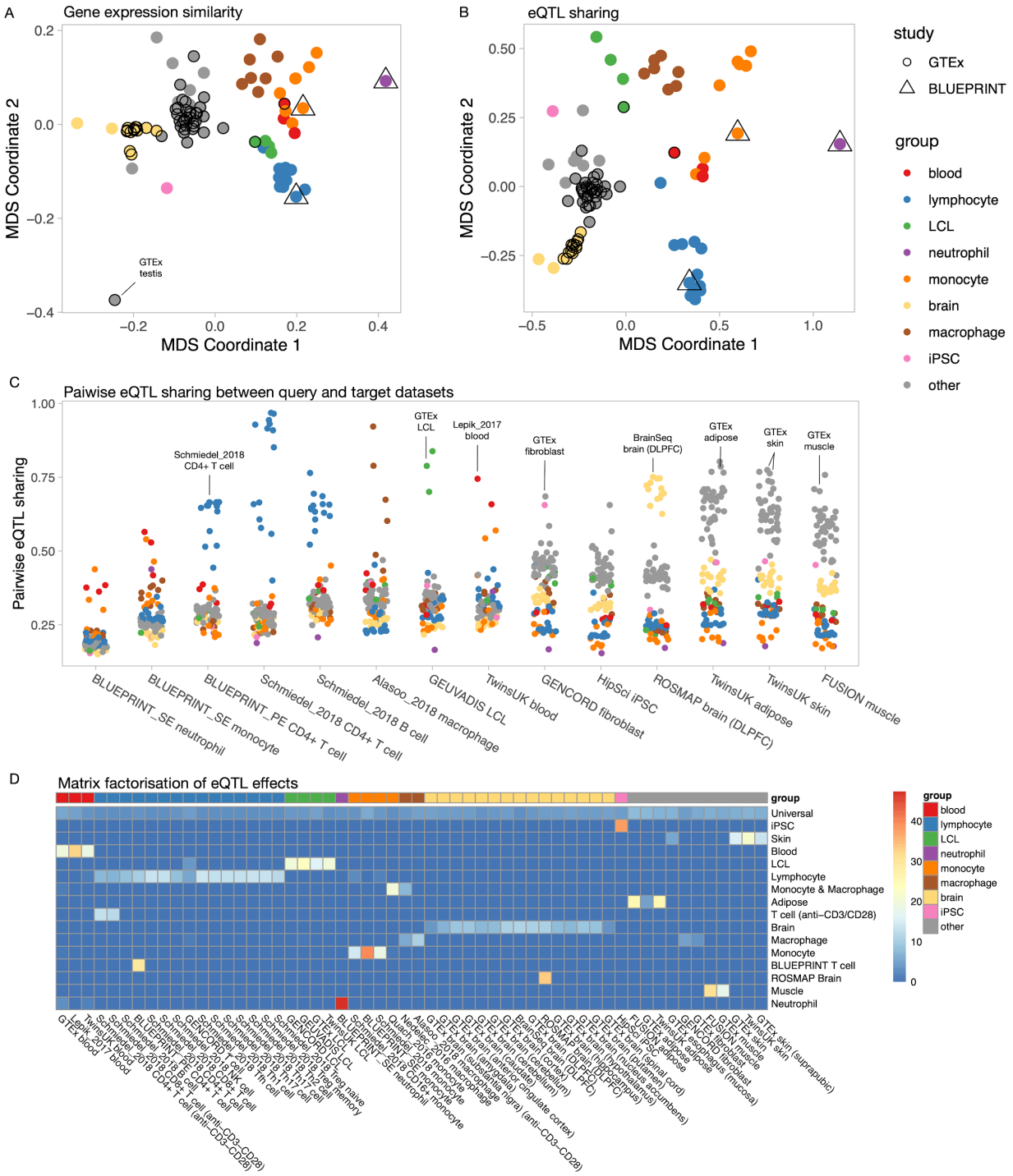


Figure 3. Gene expression similarity between datasets predicts eQTL similarity. (**A**) Multidimensional scaling (MDS) analysis of median gene expression across datasets. The pairwise similarity between datasets was calculated using Pearson's correlation. Datasets from GTEx and BLUEPRINT studies have been highlighted to demonstrate that they cluster with other matching cell types and tissues. (**B**) MDS analysis of eQTL sharing across datasets. Pairwise eQTL sharing between datasets was estimated using the Mash model (*19*). The complete matrix is presented in Supplementary Figure 2. (**C**) Visualisation of eQTL sharing

estimates between selected representative tissues (x-axis) and all other cell types and tissues in the eQTL Catalogue. The individual points have been coloured according to the major cell type and tissue groups from panel A. (**D**) Matrix factorisation of the eQTL effect sizes across all eQTL Catalogue datasets. Only datasets with non-zero loading on one or more cell-type- and tissue-specific factors (excluding the universal factor) are shown.

Matrix factorisation identifies cell-type- and tissue-specific latent factors shared across datasets

To better understand the eQTL sharing patterns between cell types and tissues, we turned to a recently developed semi-nonnegative sparse matrix factorisation (sn-spMF) model that can directly identify latent factors from eQTL summary statistics (*31*). When applied to the fine mapped eQTL Catalogue summary statistics, sn-spMF detected 16 independent factors (Figure 3D). The largest universal factor was broadly shared between all datasets and accounted for ~37.5% of the independent fine mapped eQTLs (Supplementary Figure 4). The remaining 15 factors captured cell-type- and tissue-specific effects (Figure 3D). Overall, matrix factorisation identified many of the same patterns detected in the pairwise eQTL sharing analysis (Figure 2B). For example, lymphocytes, LCLs, iPSC, monocytes, macrophages, neutrophils, stimulated T cells as well as brain and blood tissues all had their individual factors. Notably, these cell-type- and tissue-specific factors were shared across multiple studies (Figure 3D).

Although most eQTLs were highly shared between bulk tissues (Figure 3B-C), our factor analysis still detected independent factors capturing eQTLs that were specific to muscle, skin and adipose tissues from the FUSION (*32*), GTEx (*1*) and TwinsUK (*33*) studies. Brain, blood, adipose, muscle and skin tissues had larger sample sizes than other bulk tissues and purified cell types (Figure 2A), allowing us to obtain more accurate eQTL effect size estimates. Thus, we expect to detect additional tissue-specific factors as the sample sizes of the respective tissues increase (*31*). Finally, only two of the 16 factors were specific to a single dataset (BLUEPRINT CD4+ T cells and ROSMAP brain samples), suggesting that although batch effects between datasets exist, they are not a major factor confounding our analysis.

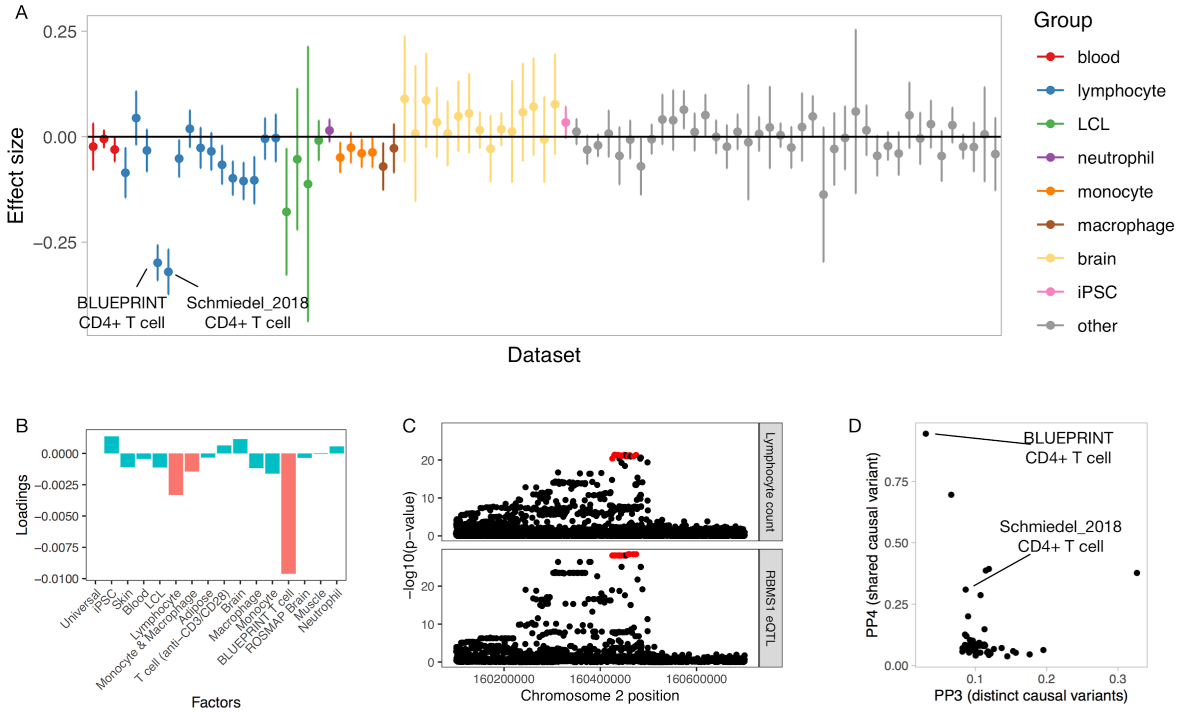


Figure 4. CD4+ T cell-specific eQTL at the *RBMS1* locus colocalises with a GWAS hit for lymphocyte count. (**A**) Effect sizes and 95% confidence intervals for the *RBMS1* eQTL across all eQTL Catalogue datasets (naive conditions only). (**B**) Factor loadings for the *RBMS1* lead variant (rs6753933) from the sn-spMF model. **(C**) Regional association plot for lymphocyte count (top panel) and *RBMS1* eQTL in the BLUEPRINT CD4+ T cells. The fine mapped eQTL credible set is highlighted in red. (**D**) Colocalisation posterior probabilities of a shared causal variant (PP4) and two distinct causal variants (PP3) for the fine mapped *RBMS1* lead variant (rs6753933) across all eQTL Catalogue datasets.

A major advantage of the matrix factorisation is that it allows us to focus on a small number of biologically meaningful factors shared between one or more datasets rather than comparing the eQTL effect sizes in 95 individual datasets. This level of summarisation is going to be increasingly important as the number of datasets included in the eQTL Catalogue increases. For example, a *cis*-eQTL for *RBMS1* had large effects in both BLUEPRINT and Schmiedel_2018 CD4+ T cell datasets and smaller significant effects in multiple other T cell subsets from Schmiedel *et al*. (Figure 4A). Consequently, the two factors with the largest loadings for this eQTL were the BLUEPRINT CD4+ T cell factor and the general lymphocyte factor (Figure 4B). The *RBMS1* eQTL also colocalised with a GWAS signal for lymphocyte count (*34*) in BLUEPRINT CD4+ T cells (PP4 = 0.94) (Figure 4C), illustrating how a lymphocyte-specific eQTL might contribute to the regulation of lymphocyte count in whole blood. Notably, we did not detect this colocalisation in any of the 49 GTEx tissues.

eQTL Catalogue finds novel colocalisations missed in GTEx

Our eQTL sharing analysis demonstrated that the eQTL Catalogue contains many additional eQTLs not present in GTEx. To quantify how these novel eQTLs might improve the interpretation of complex trait and disease associations, we performed colocalisation between GWAS summary statistics for 14 traits and either the eQTL Catalogue datasets or all GTEx v8 tissues. To ensure that each independent GWAS locus was counted only once, we first partitioned GWAS summary statistics into approximately independent LD blocks (*35*). Overall, we detected at least one colocalising eQTL (PP4 > 0.8) for 4,429 independent loci across 14 traits, 373 (8.4%) of which were only detected in one of the eQTL Catalogue datasets and not captured by GTEx v8 (max PP4 < 0.8). The fraction of novel colocalising loci varied from 5% for height to 14% for lupus (Supplementary Figure 5), suggesting that a substantial fraction of trait colocalisations might be missed if the analysis is only restricted to GTEx.

However, we often detected many novel colocalisations even in those eQTL Catalogue datasets that were already captured by GTEx (e.g. blood, skin, muscle, adipose and brain tissues, Figure 2A). These additional colocalisations could be either due to thresholding effects (just below or above the PP4 > 0.8 threshold), increased sample sizes in the eQTL Catalogue, and biological and population differences between datasets or other technical factors. For example, we found that the number of novel colocalisations detected for height GWAS increased linearly with the eQTL sample size with no particular dataset standing out (Figure 5A). In contrast, for some trait and eQTL dataset pairs, we detected considerably more colocalisations than we would have expected at the given sample size. For example, we observed six novel colocalisations with lymphocyte count in BLUEPRINT CD4+ T cells (including the *RBMS1* example in Figure 4C), which was three times more than in any other dataset of comparable sample size (Figure 5B).

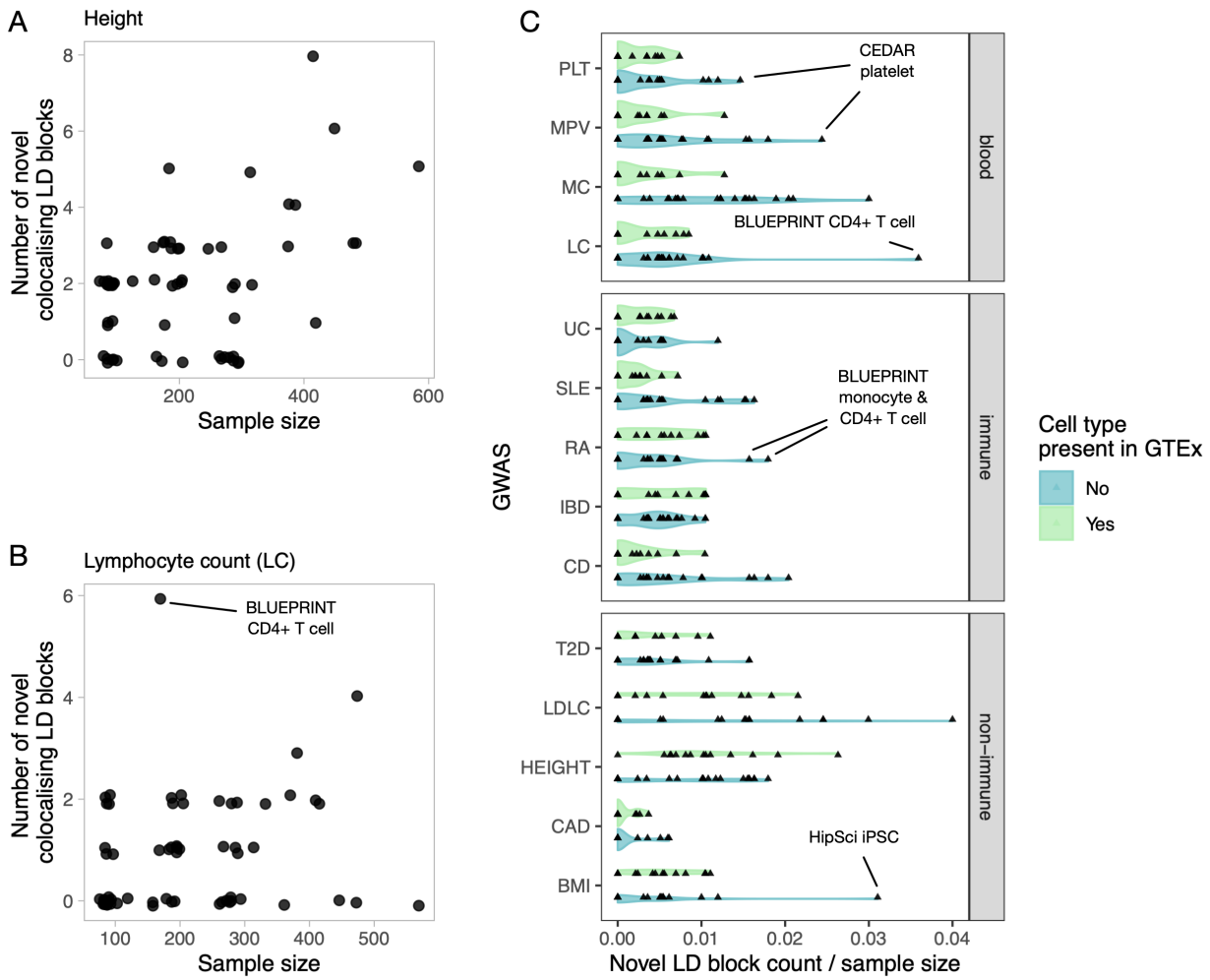


Figure 5. Additional GWAS colocalisations detected in the eQTL Catalogue relative to GTEx v8. (**A**) The number of novel height GWAS loci that colocalise with eQTLs in each cell type or tissues as a function of eQTL dataset size. (**B**) The number of novel lymphocyte count GWAS loci that colocalise with eQTLs in each cell type or tissues as a function of eQTL dataset size. (**C**) The number of novel colocalising loci detected for the 14 GWAS traits in each cell type and tissue from eQTL Catalogue divided by the eQTL sample sizes. The eQTL Catalogue cell types and tissues were grouped according to whether they were present in GTEx (blood, LCL, adipose, muscle, skin, brain) or not (T cells, B cells, monocytes, macrophages, neutrophils and iPSCs). GWAS traits: PLT - platelet count, MPV - mean platelet volume, MC - monocyte count, LC - lymphocyte count, UC - ulcerative colitis, SLE - systemic lupus erythematosus, RA - rheumatoid arthritis, IBD - inflammatory bowel disease, CD - Crohn's disease, T2D - type 2 diabetes, height, CAD - coronary artery disease, BMI - body mass index, LDLC - LDL cholesterol.

To assess if some eQTL datasets were particularly relevant for specific GWAS traits, we assigned each dataset a ‘novelty score’ by dividing the number of novel colocalisations detected in that dataset by its sample size. For each GWAS trait, we then asked if the novelty scores were higher for datasets from cell types and tissues missing in GTEx compared to the datasets

that were already well captured by GTEx. While there was considerable overlap between the two distributions (Figure 5C), we detected several trait-dataset pairs where the number of novel colocalisations observed was higher than expected for a given sample size. For example, the largest number of novel colocalisations for platelet count (PLT) and mean platelet volume (MPV) was detected in the CEDAR (*36*) platelet dataset (Figure 5C). Similarly, we observed most novel colocalisations for monocyte and lymphocyte count in BLUEPRINT monocyte and CD4+ T cell datasets, respectively. These results suggest that many novel colocalisations detected in the eQTL Catalogue relative to GTEx cannot be explained by sampling or technical variation alone and are likely to reflect cell-type-specific genetic effects.

A subset of colocalisations manifest at the transcript level

Multiple studies have demonstrated that some colocalisations between QTLs and complex traits only manifest at the level of RNA splicing and transcript usage (*37*, *38*). To quantify this in the eQTL Catalogue, we performed colocalisation analysis between the 14 complex traits mentioned above and all QTLs detected with the three transcript-level quantification methods (Supplementary Figure 1). We found that 586/3394 (17.2%) colocalisations in independent LD blocks were only detected using one of the three transcript-level traits and not by traditional eQTLs in any of the 95 RNA-seq datasets (Figure 6A). However, this is likely to be underestimated because transcript and gene-level QTLs could be colocalising with independent GWAS signals within the same LD block (*1*). Furthermore, our gene expression quantification was based on the total read count, which can also capture larger splicing changes, especially as the number of datasets and sample sizes increase.

To illustrate this, we looked at the colocalisation between LDL cholesterol and an exon expression QTL for *HMGCR*. The gene product of *HMGCR* is a known target for statins, and the link between exon 13 inclusion and circulating LDL cholesterol levels has been reported previously (*38*, *39*). Our analysis detected colocalisation (PP4 > 0.8) between the expression of exon 13 of the *HMGCR* gene and LDL cholesterol in 51/95 datasets. We saw the strongest association in the HipSci (*6*) induced pluripotent stem cell dataset, where we were able to fine map the exon QTL to a single causal variant (rs3846662, posterior probability = 1) (Figure 6B). The same colocalisation was also detected by transcript usage in 18/95 datasets and by txrevise in 29/95 datasets. Although the colocalisation was also seen at the level of gene expression in the FUSION (*32*) muscle dataset (PP4 = 0.99, Supplementary Figure 6), the 95% credible set contained a total of 46 variants. Furthermore, the standardised effect size of the fine mapped variant on exon expression (Figure 6C) was considerably larger than on gene expression (Figure 6D) in all datasets (Figure 6C-D). Thus, even though some transcript-level QTLs can manifest as standard eQTLs in large datasets, having access to summary statistics from different quantification methods can inform on the identity and functional impact of the causal variant as well as provide stronger genetic instruments for future Mendelian randomisation applications.

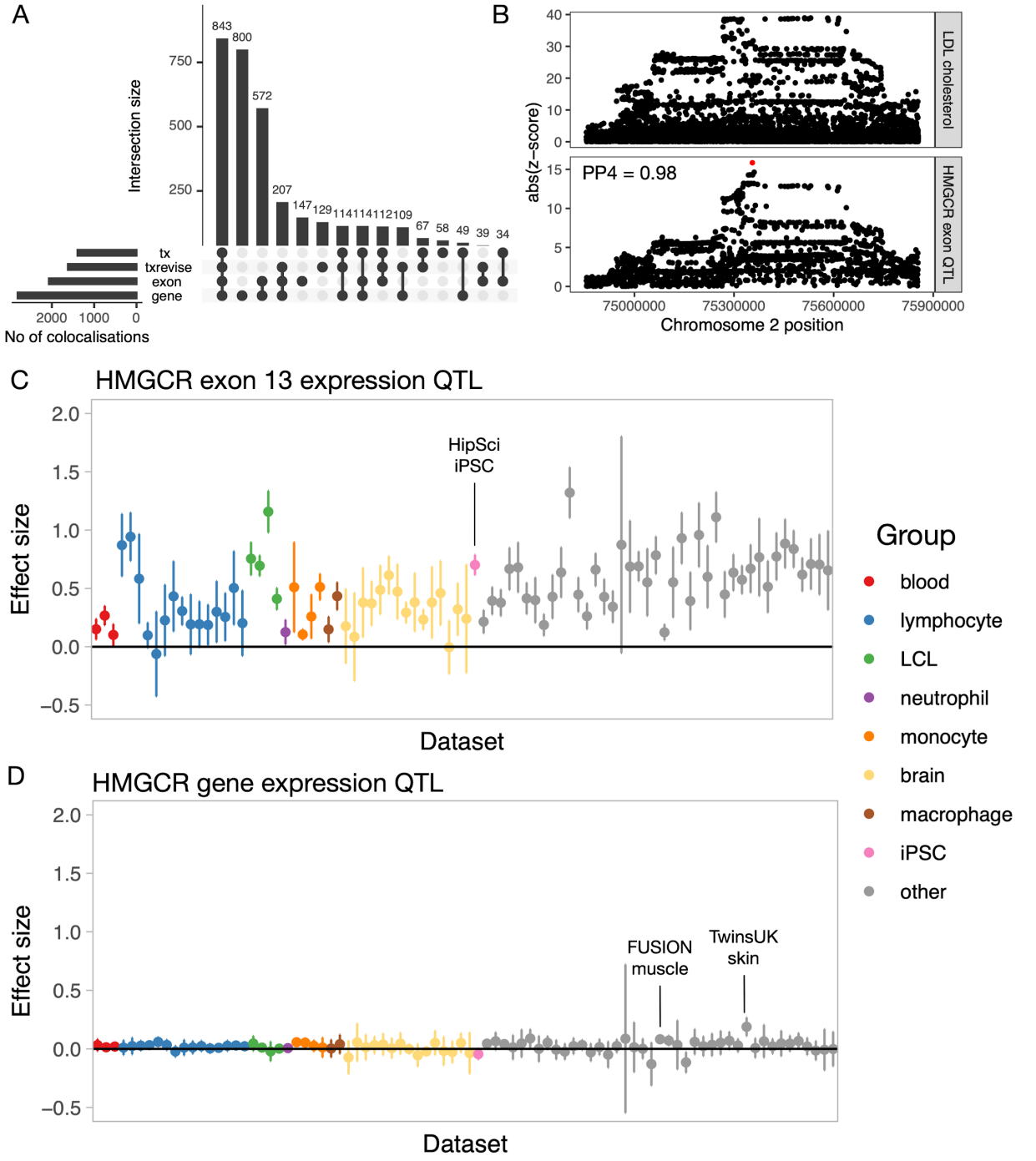


Figure 6. Colocalisation between transcript-level QTLs and complex traits. (**A**) Complex trait colocalisations (independent LD blocks) stratified by the quantification methods that they were detected with. In addition to gene-level eQTLs, we also used three transcript-level quantification methods (exon expression (exon), transcript usage (tx), and promoter, splicing and 3' end usage events (txrevise)). (**B**) Regional association plot for LDL cholesterol (top panel) and *HMGCR* exon 13 QTL in the HipSci iPSC dataset. SuSiE fine mapped the exon QTL to a single intronic variant (rs3846662, represented by the red dot) which was missing from the GWAS summary statistics. (**C**) Exon 13 expression QTL effect sizes and 95% confidence intervals for

the fine mapped causal variant (rs3846662) across eQTL Catalogue datasets. (**D**) Gene expression QTL effect sizes and 95% confidence intervals for the fine mapped causal variant (rs3846662) across eQTL Catalogue datasets.

Discussion

We believe that the main value of the eQTL Catalogue lies in the uniformly processed gene-level and transcript-level QTL summary statistics and statistical fine mapping results. We have thus sought to make the data as easy to use as possible. By mapping cell and tissue types to standard ontology terms, we make it easy to discover which studies contain the tissues and cell types of interest to the users. We have further re-imputed genotypes using the 1000 Genomes Phase 3 reference panel for all studies using genotyping microarrays, ensuring that the same set of genetic variants is present in most studies. We have used a consistent set of molecular trait identifiers (genes, exons, transcripts, events) across all datasets, ensuring that genetic effects can directly be compared across datasets (e.g. Figures 4A and 6C-D). Finally, we have released credible sets from statistical fine mapping analysis, which can help to further characterise loci with multiple independent signals and paves the way for fine-mapping-based colocalisation approaches (*25*). We will progressively expand the resource to all accessible human datasets.

The relationship between gene expression similarity and eQTL sharing has been noticed before. For example, two studies conducted in stimulated monocytes and macrophages found that the number of differentially expressed genes between cell states correlates with the number of state-specific eQTLs (*38*, *40*). This correlation raises an exciting prospect that once a *sufficient* sample size has been reached in a given cell type or tissue, the discovery of novel eQTL can be maximised by focussing on cell types and cell states with low gene expression similarity to existing eQTL datasets. Of course, the definition of what is sufficient depends on the downstream use case of interest. While many cell-type- and tissue-specific *cis*-eQTLs can be detected with a sample size of a few hundred individuals (Figure 3D), other applications such as expression-mediated heritability analysis (*4*), Mendelian randomisation (*18*) and *trans*-eQTL analysis (*2*) benefit from much larger sample sizes.

A limitation of our automated RNA-seq processing and eQTL mapping workflow is that we have not tailored our analyses to specific studies. For example, although the TwinsUK (*33*) and HipSci (*6*) studies collected samples from multiple related individuals, we used only a subset of samples (TwinsUK: 1,364 of 2,505 total, HipSci: 322 of 513 total) from unrelated individuals to avoid pseudoreplication when using linear regression. Similarly, for the six studies containing individuals from non-European and admixed populations (Supplementary Table 1), we jointly analysed all samples with six genotype principal components as covariates. However, stratified analyses (*41*) or approaches taking into account local ancestry (*42*, *43*) might be more appropriate in this specific setting. Access to individual-level data will enable us to revisit these decisions as new analytical approaches and computational workflows become available.

To ensure that the eQTL Catalogue is a comprehensive resource that encompasses tissue and human population diversity, we encourage researchers to contribute their eQTL datasets (contact eqtlcatalogue@ebi.ac.uk). Unfortunately, we have been unable to include some existing datasets due to consent limitations or restrictions on sharing individual-level genetic data. 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 (*44*) repository.

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 (*45*). 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/Studies/). In our applications, we explained the project and our intent to share the association summary statistics publicly. 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 the 1000 Genomes Phase 3 reference panel using Genotype Harmonizer (*46*). 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 quality control. We pre-phased and imputed the genotypes to the 1000 Genomes Phase 3 reference panel (*30*) using Eagle v2.4.1 (*47*) and Minimac4 (*48*). After imputation, we converted the coordinates of genetic variants from the GRCh37 reference genome to the GRCh38 using CrossMap v0.4.1 (*49*). 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. The genotype imputation and quality control steps are implemented in eQTL-Catalogue/genimpute (v20.11.1) workflow available from GitHub (see URLs).

Assigning individuals to reference populations. We used PLINK (*50*) v1.9.0 with '--indep-pairwise 50000 200 0.05' to perform LD pruning of the genetic variants and LDAK (*51*) to project new samples to the principal components (PCs) of the 1000 Genomes Phase 3 reference panel (*30*). To assign each genotyped sample to one of four superpopulations, we calculated the Euclidean distance in the PC 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 clusters using different numbers of PCs and found that using 3 PCs worked best for inferring the 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 7). 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. The population assignment steps are implemented in the eQTL-Catalogue/qcnorm (v20.12.1) workflow available from GitHub (see URLs).

Microarray data

Data normalisation. All five microarray studies currently included in the eQTL Catalogue (CEDAR (*36*), Fairfax_2012 (*52*), Fairfax_2014 (*53*), Kasela_2017 (*54*), Naranbhai_2015 (*55*)) 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/Studies/). Batch effects, where applicable, were adjusted for with the function removeBatchEffect from the limma v.3.40.6 R package (*56*). The batch adjusted $\log_2$ intensity values were quantile normalized using the lumiN function from the lumi v.2.36.0 R package (*57*). 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 (*46*) v1.4.20 to convert the imputed genotypes into TRITYPER format. We used MixupMapper (*58*) 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

Studies. eQTL Catalogue contains RNA-seq data from the following 16 studies: ROSMAP (*59*), BrainSeq (*60*), TwinsUK (*33*), FUSION (*32*), BLUEPRINT (*20*, *61*), Quach_2016 (*62*), Schmiedel_2018 (*21*), GENCORD (*63*), GEUVADIS (*64*), Alasoo_2018 (*65*), Nedelec_2016 (*66*), Lepik_2017 (*67*), HipSci (*6*), van_de_Bunt_2015 (*68*), Schwartzentruber_2018 (*69*), GTEx v7 (*1*).

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 the `samtools collate` command (*70*) to collate paired-end reads 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-processing scripts are available from the eQTL-Catalogue/rnaseq GitHub repository (see URLs).

Quantification. We quantified transcription at four different levels: (1) gene expression, (2) exon expression, (3) transcript usage and (4) transcriptional event usage (Supplementary Figure 1). Quantification was performed using a custom Nextflow (*71*) workflow that we developed by adding new quantification methods to nf-core/rnaseq pipeline (*44*). 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 (*72*) 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 (*73*) reference transcriptome annotations with featureCounts v1.6.4 (*74*). To quantify exon expression, we first created an exon annotation file (GFF) using GENCODE V30 reference transcriptome annotations and `dexseq_prepare_annotation.py` script from the DEXSeq (*75*) 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 (*76*). For transcript quantification, we used the GENCODE V30 (GRCh38.p12) reference transcript sequences (fasta) file to build the Salmon index. For transcriptional event usage, we downloaded pre-computed txrevise (*38*) 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 (*77*) to generate fasta sequences from the event annotations and built Salmon indices 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. All of these quantification methods are implemented in the eQTL-Catalogue/rnaseq workflow available from GitHub (see URLs). Our reference transcriptome annotations are available from Zenodo (https://doi.org/10.5281/zenodo.3366280).

Detecting outliers from gene expression data. The quality of the RNA-seq samples was assessed using the 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 on the sex chromosomes (1,247), TPM normalised (*78*) 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 the prcomp R package (center = true, scale = true). For multidimensional scaling (MDS) analysis, we used the isoMDS method from the MASS R package with k=2 dimensions. As a distance metric for isoMDS, we used 1 - Pearson's correlation as recommended previously (*79*). We plotted these two-dimensional scatter plots to visually identify outliers (Supplementary Figure 8A-B).

Sex-specific gene expression analysis. Previous studies have successfully used the expression of *XIST* and Y chromosome genes to ascertain the genetic sex of RNA samples (*80*)**.** In our analysis, we extracted all protein-coding genes from the Y chromosome, and the *XIST* gene (ENSG00000229807) expression values and TPM normalised them. Then, we calculated the mean expression level of the genes on the Y chromosome. Finally, we plotted the log_2 of *XIST* expression level (X-axis) against the mean expression level of the genes on the Y chromosome (Y-axis) (Supplementary Figure 8C). In addition to detecting samples with incorrectly labelled genetic sex, this analysis also allowed us to identify cross-contamination between samples (*XIST* and Y chromosome genes expressed simultaneously, Supplementary Figure 8C).

Concordance between genotype data and RNA-seq samples. We used the Match Bam to VCF (MBV) method from QTLtools (*81*) which directly compares the sample genotypes in VCF format to an aligned RNA-seq BAM file. MBV can detect sample swaps, multiple samples from the same donor, and cross-contamination between RNA-seq samples. In some cases, such cross-contamination was confirmed by both the sex-specific gene expression and MBV analyses (Supplementary Figure 8D).

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 (*82*) with gene or exon GC nucleotide content as a covariate. We downloaded the gene GC content estimates from Ensembl biomaRt and calculated the exon-level GC content using bedtools v2.19.0 (*83*). We also excluded lowly expressed genes, where 95 per cent of the samples within a dataset 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. We then divided those transcript (event) expression estimates 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 are publicly available at https://github.com/eQTL-Catalogue/qcnorm.

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 and tissue types to a controlled vocabulary of ontology terms from Uber-anatomy ontology (Uberon) (*84*), Cell Ontology (*85*) or Experimental Factor Ontology (EFO) (*86*). We opted to use an *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 performed association testing separately in each dataset and used a +/- 1 megabase *cis* window centred around the start of each gene. First, we excluded molecular traits with less than five genetic variants in their *cis* window, as these were likely to reside in regions with low genotyping coverage. We also excluded molecular traits with zero variance across all samples and calculated phenotype principal components using the 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. We calculated nominal eQTL summary statistics using the GTEx v6p version of the FastQTL (*87*) software (https://github.com/francois-a/fastqtl) that also estimates standard errors of the effect sizes. We used the '`--window 1000000 --nominal 1`' flags to find all associations in 1 Mb *cis* window. For permutation analysis, we used QTLtools v1.1 (*88*) with '`--window 1000000 --permute 1000 --grp-best`' flags to calculate empirical p-values based on 1000 permutations. The '`--grp-best`' option ensured that the permutations were performed across all molecular traits 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. The steps described above are implemented in the eQTL-Catalogue/qtlmap v20.07.2 Nextflow workflow available from GitHub (see URLs).

Statistical fine mapping

We performed QTL fine mapping using the Sum of Single Effects Model (SuSiE) (*23*) implemented in the susieR v0.9.0 R package. We converted the genotypes from VCF format to a tabix-indexed dosage matrix with bcftools v1.10.2. We imported the genotype dosage matrix into R using the Rsamtools v1.34.0 R package. We used the same normalised molecular trait matrix used for QTL mapping and further applied a rank-based inverse normal transformation to each molecular trait to ensure that they were normally distributed. We regressed out the first six phenotype and genotype PCs separately from the phenotype and genotype matrices. We performed fine mapping with the following parameters: L = 10, estimate_residual_variance = TRUE, estimate_prior_variance = TRUE, scaled_prior_variance = 0.1, compute_univariate_zscore = TRUE, min_abs_corr = 0. Finally, we extracted the 95% credible sets and the 95% posterior inclusion probabilities for each variant belonging to the credible set. The steps described above are implemented in the eQTL-Catalogue/susie-workflow v20.08.3 Nextflow workflow available from GitHub (see URLs).

Quantifying eQTL sharing between tissues, cell types and conditions

Identifying independent signals based on fine mapping. We extracted independent signals from the variants included in fine-mapped credible sets. At first, we selected credible sets with less than 50 variants in size and with a univariate z-score of at least 3. For every gene, we built connected components of credible sets to represent independent signals. From every connected component, we picked the lead variant – the variant with the smallest p-value across all eQTL datasets. As a result, 54,733 eQTLs remained.

Calculating Spearman's correlation. We aggregated the eQTL data into a matrix of effect sizes, where each row represents a lead variant and each column an eQTL dataset. We noticed that this matrix contained many missing values. While most of the missing values were caused by the gene not being expressed in a particular cell type or tissue, some of the missing values were also caused by low allele frequency or low imputation quality score. Thus, we substituted all missing values with 0. We then calculated pairwise Spearman's correlation between the columns of the matrix to estimate the eQTL similarity between datasets.

Running Mash. As an alternative to Spearman's correlation, we used the multiple adaptive shrinkage (Mash) (*19*) model to estimate the pairwise sharing of eQTLs between datasets. Betas and standard errors of lead effects were input to the Mash model as Bhat and Shat. We set missing eQTL effect sizes to 0 and standard errors to 1. The model was fitted with alpha = 1 (exchangeable effects model). To find candidate covariance matrices, we discovered strong effects that are significant in at least one dataset with the `get_significant_results` method. Then we performed PCA on identified strong effects to obtain covariance matrices with `cov_pca` function and applied extreme deconvolution to them with `cov_ed`. Resulting matrices were set as a candidate covariance matrices into the model fitting. We estimated pairwise eQTL sharing between datasets with `get_pairwise_sharing` method by magnitude (factor of 0.5) and sign of posterior effect estimates.

Factor analysis

We performed factor analysis using the semi-nonnegative sparse matrix factorisation (sn-spMF) model (*31*). We included the 54,733 independent gene-variant pairs detected using statistical fine mapping (see above). The input files contained effect sizes and standard errors as reciprocal of weights of lead effects. The missing values made up 27% of the input effect size matrix. If the effect size estimate was missing in a given cell type or tissue then the effect size and weight were set to zero. To find hyper-parameter K (initial number of factors), and regularization parameters alpha and lambda, we performed a two-level grid search. In the first level, K was set to 20, 30, 40, 50, lambda and alpha were set in a range of 800 to 1800 with optimisation number of iterations = 10. In the second level, we fine-tuned the parameters by narrowing the search space to those values that lead to higher sparsity of the loading and factor matrices in the first level. At the second level, we ran the parameter optimization for 50 iterations. We picked the final matrix with a very high cophenetic coefficient (0.99) and 16 factors.

Colocalisation

We performed colocalisation analysis on QTLs in the eQTL Catalogue against GWAS summary statistics from 14 studies downloaded from the IEU OpenGWAS database in VCF format (*89*, *90*). Our analysis included summary statistics for inflammatory bowel disease (IBD) and its two subtypes (Crohn's disease (CD) and ulcerative colitis (UC)) (*91*); rheumatoid arthritis (RA) (*92*), systemic lupus erythematosus (SLE) (*93*), type 2 diabetes (T2D) (*94*), coronary artery disease (CAD) (*95*), LDL cholesterol (*96*), four blood cell type traits (lymphocyte count (LC), monocyte count (MC), platelet count (PLT), mean platelet volume (MPV)) (*34*) and two anthropometric traits (height, body mass index (BMI)) from the UK Biobank (*96*). The variant coordinates of the GWAS summary statistics were lifted to the GRCh38 reference genome using CrossMap (*49*). Allele frequencies of variants in five of the GWAS (IBD, CD, UC, RA, SLE) were extracted from the 1000 Genomes Phase 3 reference panel (*30*). For all eQTL and GWAS dataset pairs, we performed colocalisation in a ± 200,000 window around each of the 54,733 fine mapped eQTL credible set lead variants (see fine mapping above). This ensured that colocalisation was also performed separately for multiple independent eQTLs of the same gene and colocalisation results were obtained in datasets in which no significant eQTL was detected for a particular gene. However, since we did not use masking or conditional analysis, many secondary eQTL colocalisations could still have been missed (*18*, *97*). Since transcript usage, exon expression and txrevise contained many more redundant phenotypes (e.g. multiple exons of the same gene), we limited colocalisation analysis for those molecular traits to the significant lead QTL variants in each dataset only (FDR < 0.01), using the same ± 200,000 *cis* window as above. We used version 3.1 of the coloc R package (*98*). All analysis steps are implemented in the eQTL-Catalogue/colocalisation (v20.11.1) workflow (see URLs).

Quantification of novel colocalisations at the transcript level. We only included QTL and complex trait pairs with strong evidence of colocalisations (PP4 > 0.8) in our analysis. Inspired by the study by Barbeira *et al.* (*3*), we summarised colocalisations at the level of approximately independent LD blocks (*35*). Positions of approximately independent LD blocks were obtained from Berisa and Pickrell (*35*) and converted to GRCh38 coordinates using CrossMap (*49*). If the colocalisation *cis* window overlapped two or more LD blocks, then the colocalising QTL was assigned to the LD block where the QTL lead variant was located. The number of LD blocks for which we detected at least one colocalising QTL with each quantification method was visualised using the upsetR R package (*99*).

Comparative analysis with GTEx V8. Current version of the eQTL Catalogue (release 3) contains two versions of the GTEx summary statistics: uniformly processed summary statistics from GTEx v7 and the official GTEx v8 summary statistics downloaded from Google Cloud (gs://gtex-resources/GTEx_Analysis_v8_QTLs/GTEx_Analysis_v8_eQTL_all_associations). Since the sample size of GTEx v8 is approximately two times larger than GTEx v7, we decided to use the official GTEx v8 summary statistics in our comparative colocalisation analysis. This ensured that we were as conservative as possible when identifying novel colocaliations. For each GWAS trait, we summarised colocalisation signals at the level of independent LD blocks and defined an LD block to harbour a novel colocalisation signal if there was no colocalisation

detected within that LD block in any of the GTEx v8 tissues. We further excluded datasets with small sample sizes (n < 150) due to their low power to detect colocalisations.

URLs

Data analysis workflows:

- RNA-seq quantification: https://github.com/eQTL-Catalogue/rnaseq
- Normalisation and QC: https://github.com/eQTL-Catalogue/qcnorm
- Genotype imputation: https://github.com/eQTL-Catalogue/genimpute
- Association testing: https://github.com/eQTL-Catalogue/qtlmap
- Statistical fine mapping: https://github.com/eQTL-Catalogue/susie-workflow
- Colocalisation: https://github.com/eQTL-Catalogue/colocalisation

Example use cases:

- Accessing eQTL Catalogue summary statistics with tabix: https://github.com/eQTL-Catalogue/eQTL-Catalogue-resources/blob/master/tutorials/tabix_use_case.md
- Python example for querying the HDF5 files: https://github.com/eQTL-Catalogue/eQTL-SumStats/blob/master/querying_hdf5_basics.ipynb

Data availability

All eQTL Catalogue summary statistics are available under the Creative Commons Attribution 4.0 International License. The full association summary statistics and fine mapped credible sets 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 (*100*) and seqminer (*101*). All of the summary statistics are also available via the REST API (https://www.ebi.ac.uk/eqtl/api-docs/). Fine mapped credible sets can be browsed using our interactive web interface (https://elixir.ut.ee/eqtl/). Database accessions for the raw gene expression and genotype datasets are listed on the eQTL Catalogue website (https://www.ebi.ac.uk/eqtl/Studies/). Our summary statistics have also been integrated into third party services such as the Open Targets Genetics Portal (*102*) and FUMA (*13*). The gene expression matrices will be made available via the EMBL-EBI Expression Atlas (*103*).

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.

KP performed eQTL similarity and matrix factorisation analyses. MPS connected the Ensembl display to the eQTL Catalogue. IK created the interactive credible set browser. TB, SJ, HP, AY, ST, 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 Pihlapuu 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, Emily Steed, Silva Kasela and Urmo Võsa for their helpful comments on the manuscript; Masahiro Kanai, Jacob Ulirsch and Hilary Finucane for feedback on the fine mapping workflow; Daniel Gaffney for guidance in setting up this project.

Funding

NK, JH, MS, ST and JM were supported by a grant from Open Targets (OTAR2-046). TB, SJ, IP, HP, AY and DZ were supported by the European Molecular Biology Laboratory. 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). KP and NK were supported by the Estonian Research Council grant PSG415. LK was supported by the Estonian Research Council grant PSG59. KA, NK, KP, IK and LK were also supported by Estonian Centre of Excellence in ICT Research (EXCITE) funded by the European Regional Development Fund. IK was supported by A Distributed Infrastructure for Life-Science Information ELIXIR, European Regional Development Fund project 2014-2020.4.01.16-0271.

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.
FUSION. Support for the FUSION Tissue Biopsy Study dataset was contributed by the NHGRI intramural projects ZIAHG000024 and Z1BHG000196, NIDDK grants DK062370, DK072193, and DK099240, NHGRI grant HG003079, American Diabetes Association Pathway to Stop Diabetes Grant 1-14-INI-07, and grants from the Academy of Finland.

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 a subset of donors in Fairfax_2014. Thus, the total number of donors (n = 5,714) in this table slightly exceeds the number of unique donors. Superpopulation codes: EUR - European, AFR - African, SAS - South Asian, EAS - East Asian.

Study	Donors	EUR	AFR	SAS	EAS	Unassigned
Alasoo_2018	84	84	0	0	0	0
BLUEPRINT	197	197	0	0	0	0
BrainSeq	479	231	195	1	0	52
CEDAR	322	322	0	0	0	0
Fairfax_2014	423	421	0	0	0	2
FUSION	297	297	0	0	0	0
GENCORD	196	192	0	0	0	4
GEUVADIS	445	358	87	0	0	0
GTEx	507	421	61	1	6	18
HipSci	322	318	0	1	0	3
Kasela_2017	295	295	0	0	0	0
Lepik_2017	471	471	0	0	0	0
Nedelec_2016	168	96	52	0	0	20
Quach_2016	200	100	100	0	0	0
ROSMAP	576	576	0	0	0	0
Schmiedel_2018	91	48	3	3	20	17
Schwartzentruber_2018	98	98	0	0	0	0
TwinsUK	433	432	0	0	0	1
van_de_Bunt_2015	117	117	0	0	0	0
Total	**5714**	**5066**	**498**	**6**	**26**	**118**

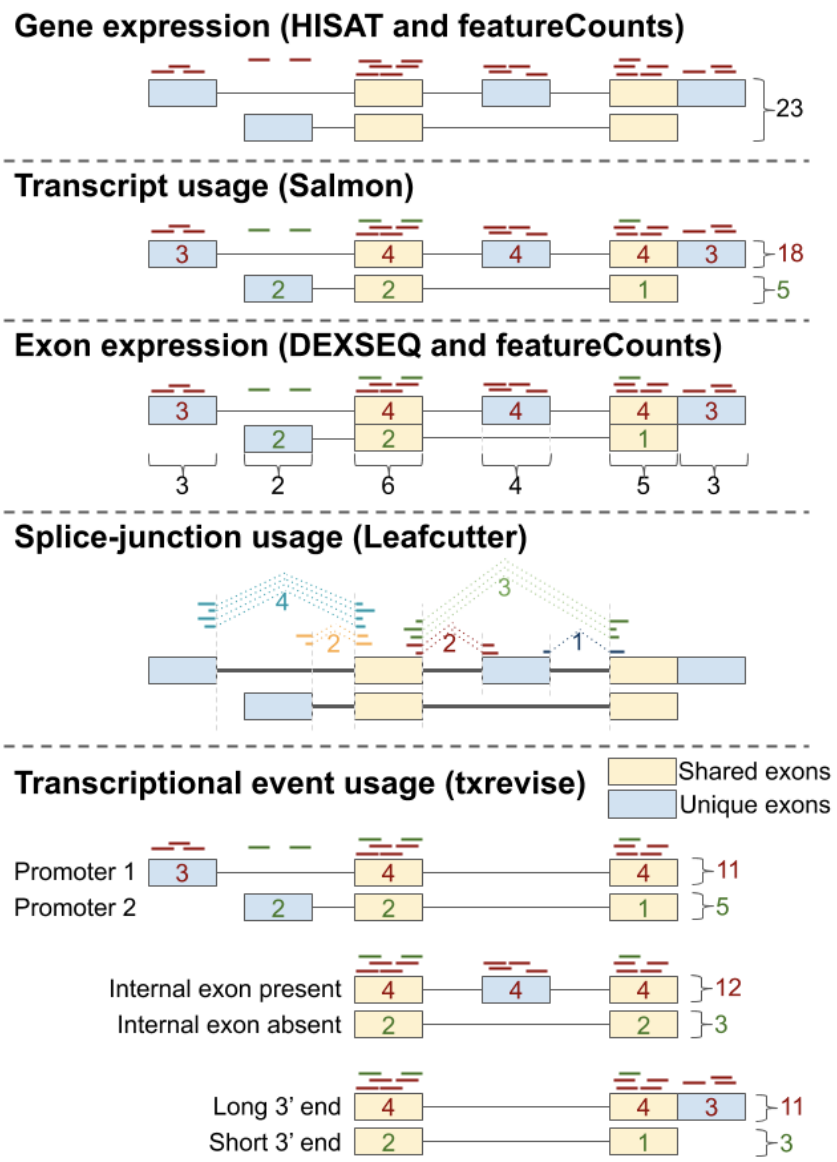
Supplementary Table 2. Overview of the transcriptomic samples included in the eQTL Catalogue. The samples have been classified according to RNA-seq type (single-end vs paired-end), strandedness (unstranded vs stranded), read length (50bp, 75bp, 100bp, 250bp), assay type (microarray vs RNA-seq) and genotype data type (whole-genome sequencing (WGS) vs genotyping array). Only genotyping array samples have been re-imputed by us.

Group	N samples	N studies	List of studies
Single-end	3180	4	BLUEPRINT, Nedelec_2016, Quach_2016, Schmiedel_2018
Paired-end	14023	13	Alasoo_2018, BLUEPRINT, BrainSeq, GTEx, FUSION, GENCORD, GEUVADIS, HipSci, Lepik_2017, ROSMAP, Schwartzentruber_2018, TwinsUK, van_de_Bunt_2015
Unstranded	12367	6	GTEx, GENCORD, GEUVADIS, Nedelec_2016, Quach_2016, TwinsUK
Stranded	4836	10	Alasoo_2018, BLUEPRINT, BrainSeq, FUSION, HipSci, Lepik_2017, ROSMAP, Schmiedel_2018, Schwartzentruber_2018, van_de_Bunt_2015
100bp	3751	8	BLUEPRINT, BrainSeq, GTEx, FUSION, Nedelec_2016, Quach_2016, ROSMAP, van_de_Bunt_2015
250bp	4	1	GTEx
50bp	3726	4	GENCORD, Lepik_2017, Schmiedel_2018, TwinsUK
75bp	9722	5	Alasoo_2018, GTEx, GEUVADIS, HipSci, Schwartzentruber_2018
microarray	4631	5	CEDAR, Fairfax_2014, Kasela_2017, Naranbhai_2015, Fairfax_2012
RNA-seq	17203	16	Alasoo_2018, BLUEPRINT, BrainSeq, GTEx, FUSION, GENCORD, GEUVADIS, HipSci, Lepik_2017, Nedelec_2016, Quach_2016, ROSMAP, Schmiedel_2018, Schwartzentruber_2018, TwinsUK, van_de_Bunt_2015
WGS	10006	4	BLUEPRINT, GTEx, GEUVADIS, Lepik_2017
Genotyping array (imputed)	11828	17	Alasoo_2018, BrainSeq, FUSION, GENCORD, HipSci, Nedelec_2016, Quach_2016, ROSMAP, Schmiedel_2018, Schwartzentruber_2018, TwinsUK, van_de_Bunt_2015, CEDAR, Fairfax_2014, Kasela_2017, Naranbhai_2015, Fairfax_2012

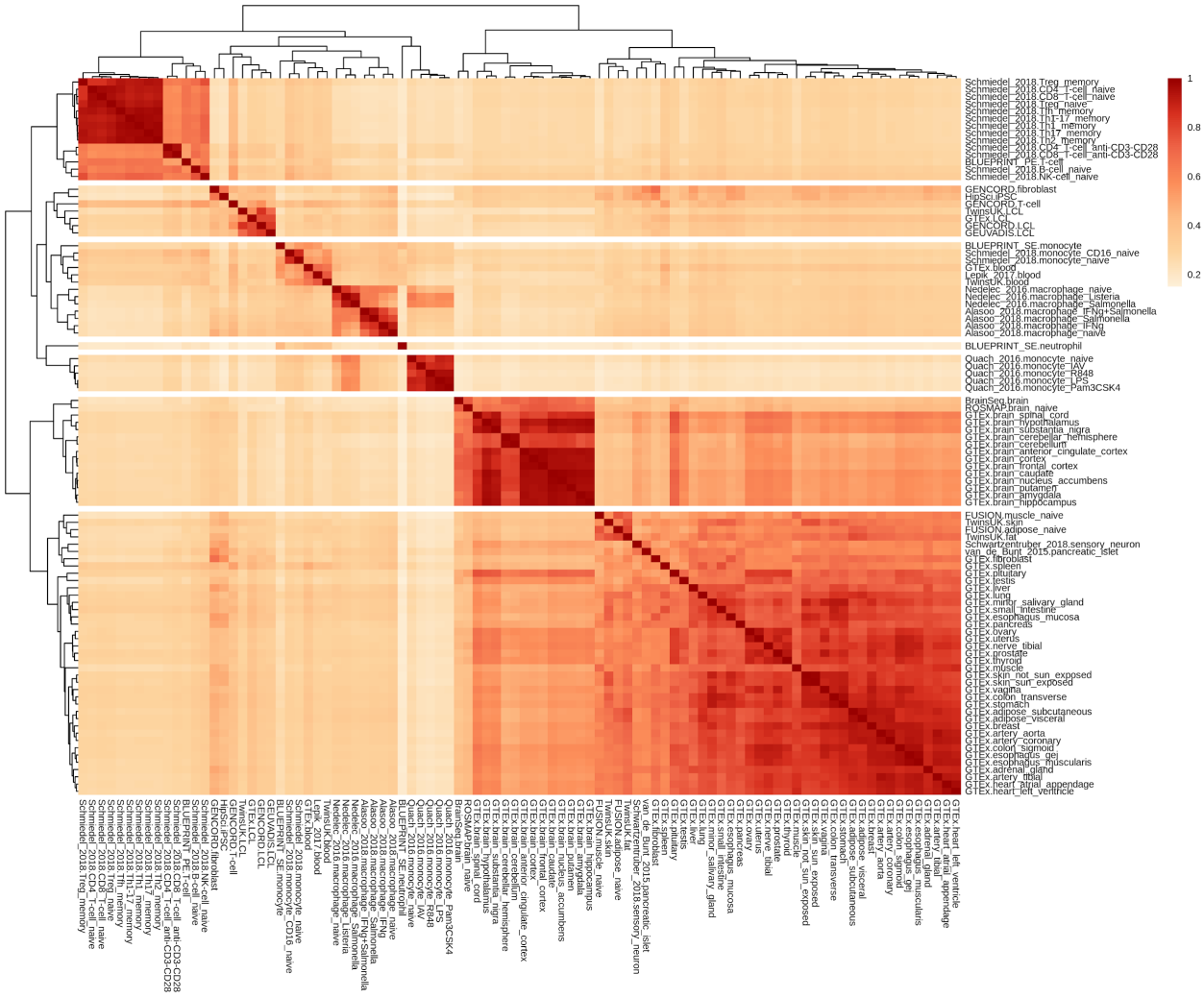
Supplementary Table 3. Overview of the studies included in the eQTL Catalogue. For four studies based on whole genome sequencing (BLUEPRINT, GTEx, GEUVADIS and Lepik_2017), we relied on final genotype files provided by the authors. All of the other genotypes were imputed using the 1000 Genomes Phase 3 reference panel (see Methods).
[1]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.
[2]Small fraction GTEx samples have RNA-seq read lengths of 100 and 250 bp.

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
FUSION	RNA-seq	YES	YES	YES	100bp	575	559
GENCORD	RNA-seq	YES	YES	NO	50bp	567	560
GEUVADIS	RNA-seq	NO	YES	NO	75bp	462	445
GTEx_v7	RNA-seq	NO	YES	NO	75bp[2]	8879	8536
HipSci[1]	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[1]	RNA-seq	YES	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		**19294**	**17210**

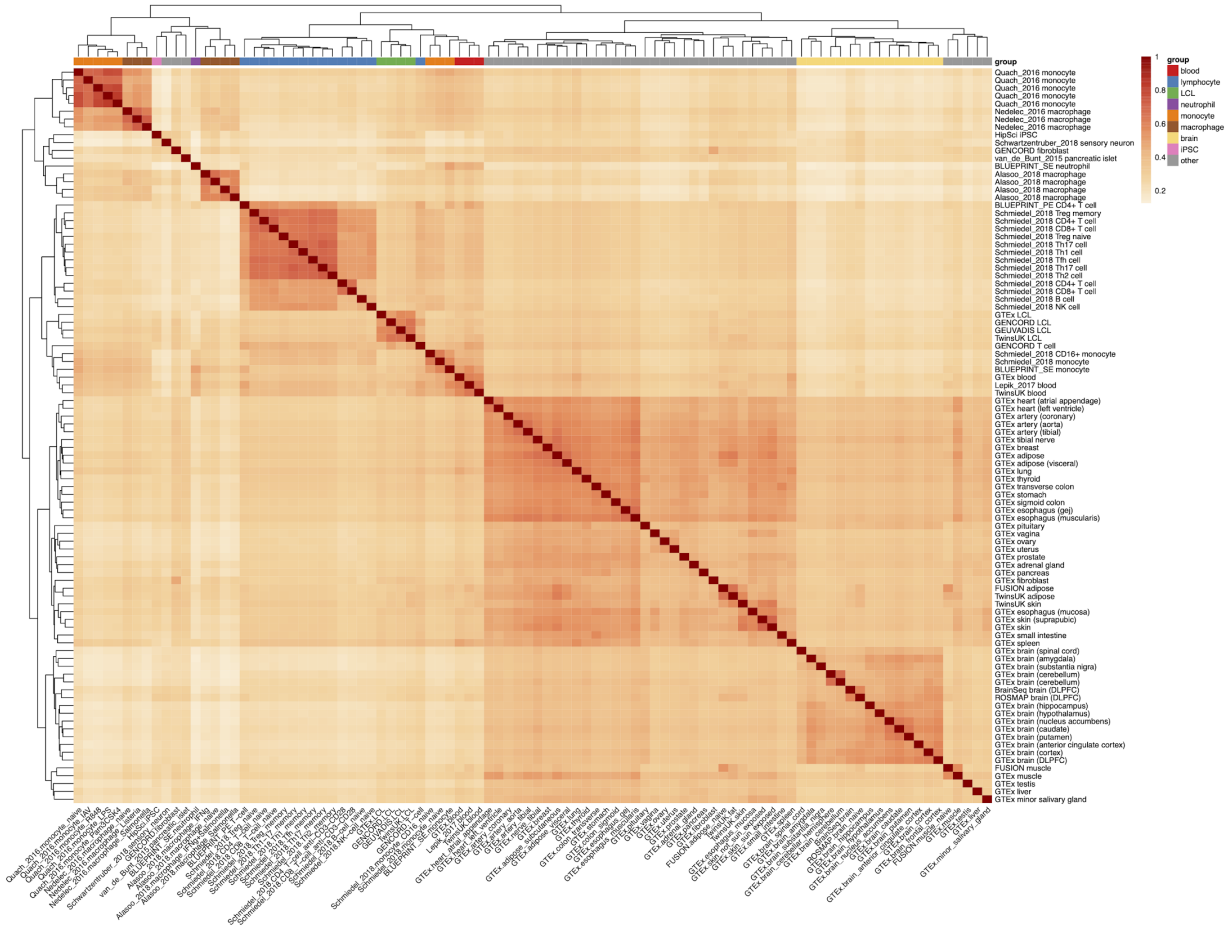
	Microarray samples	**5324**	**4631**
	Total samples	**24618**	**21841**



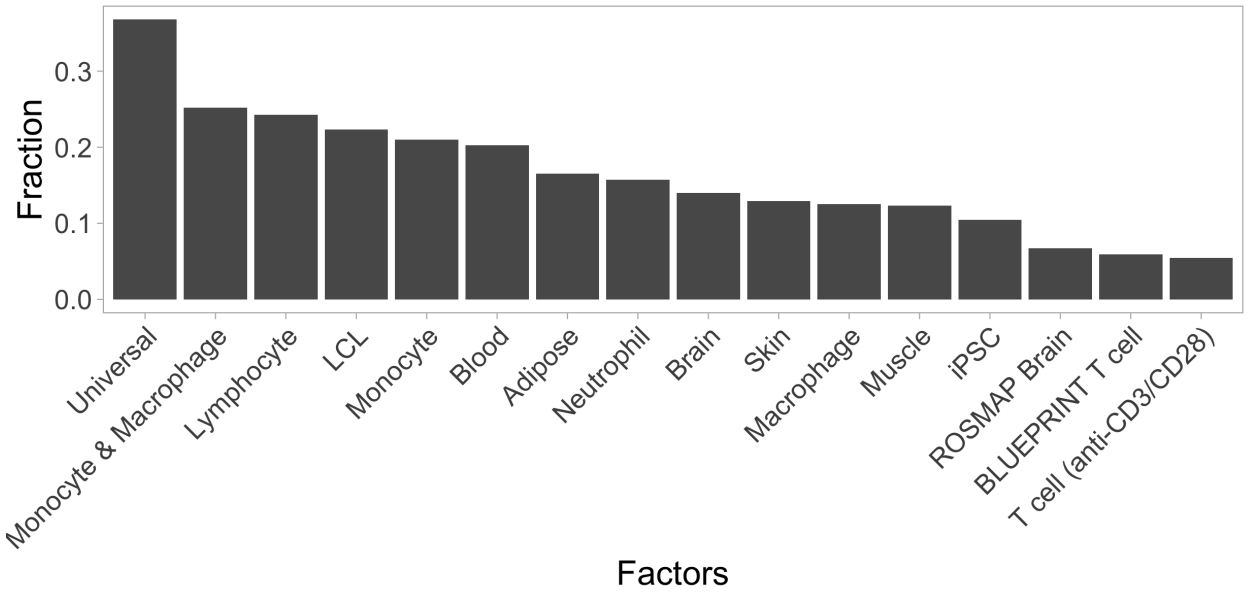
Supplementary Figure 1. Quantification methods of molecular traits in the eQTL Catalogue. Symbolic representation of 23 read fragments assigned to 1 gene (aligned with HISAT2 (*72*), quantified with featureCounts (*74*)) consisting of 2 transcripts (quantified with Salmon (*76*)) and 6 exonic parts (annotated with DEXSeq (*75*), quantified with featureCounts). The gene also has 5 distinct introns which are identified and quantified by Leafcutter (*104*). Transcriptional event usage is quantified with txrevise (*38*). 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.



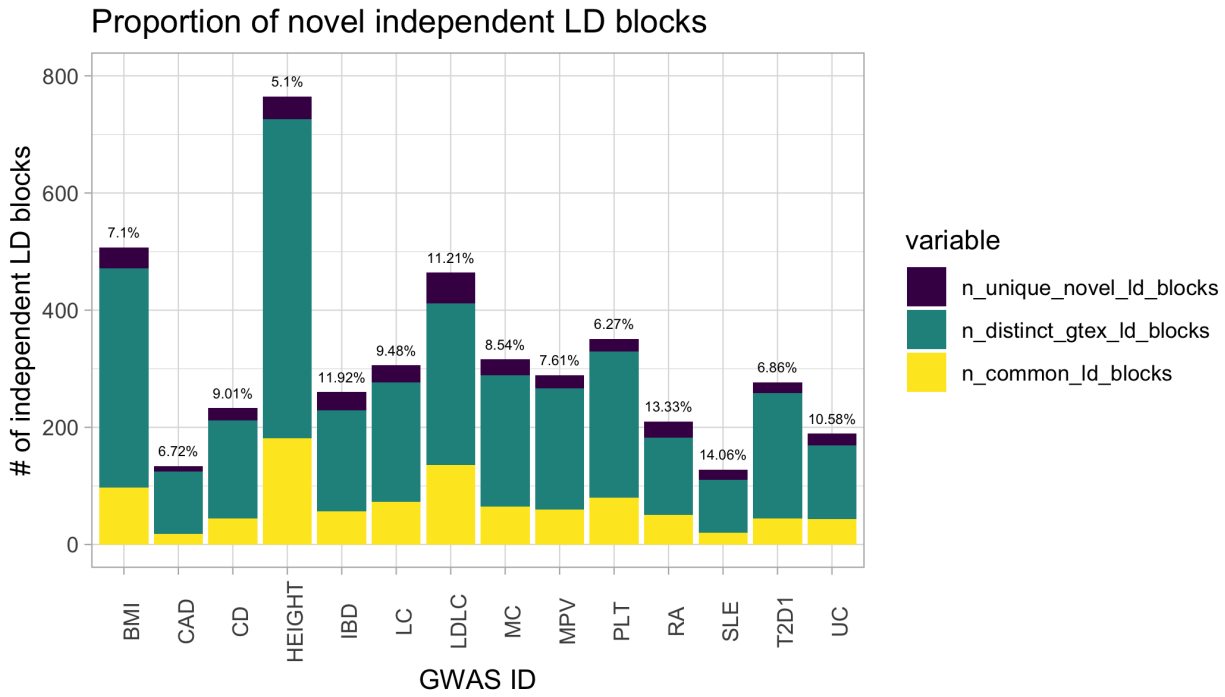
Supplementary Figure 2. Pairwise eQTL sharing between 95 datasets estimated with the Mash model. We used 54,733 independent gene variant pairs from the fine mapping analysis (see Methods) and used the Mash model to estimate eQTL sharing between all pairs of the 95 datasets measured with RNA-seq.



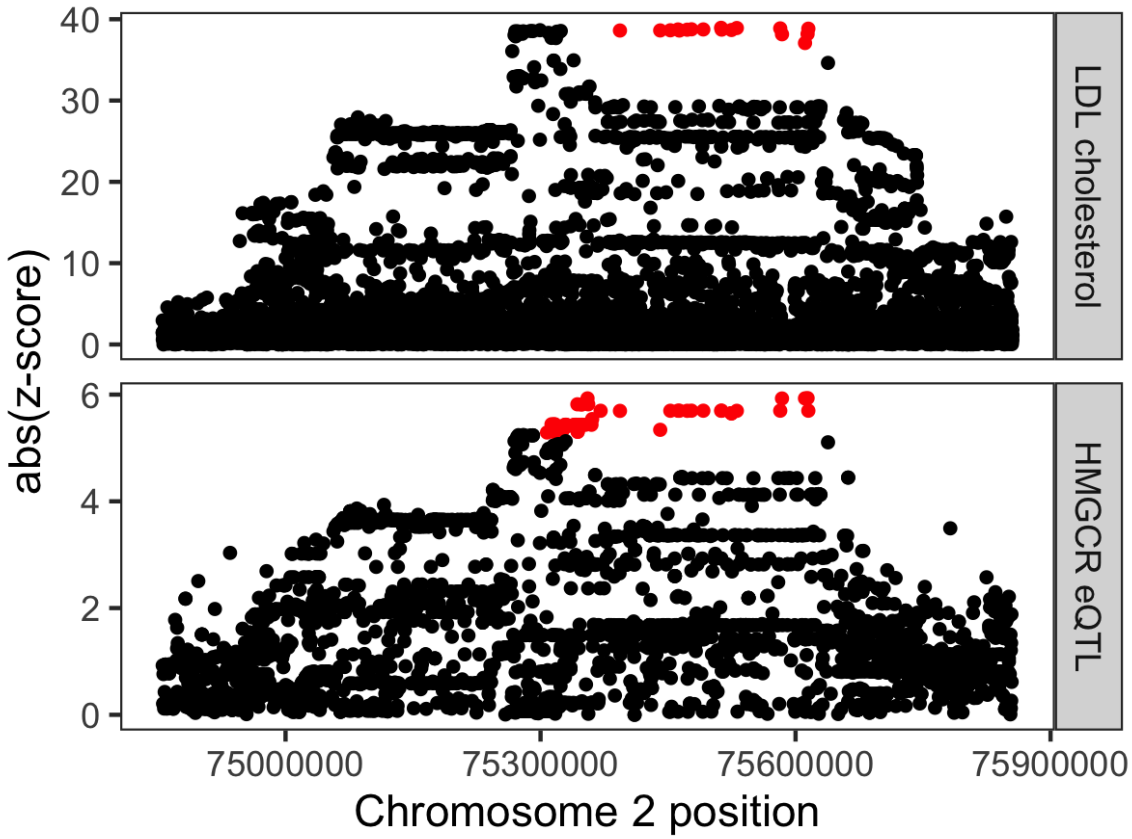
Supplementary Figure 3. Pairwise eQTL similarity between 95 datasets estimated with Spearman correlation. We used 54,733 independent gene variant pairs from the fine mapping analysis (see Methods) and used the Spearman correlation of eQTL effect sizes to estimate eQTL sharing between all pairs of the 95 datasets measured with RNA-seq.



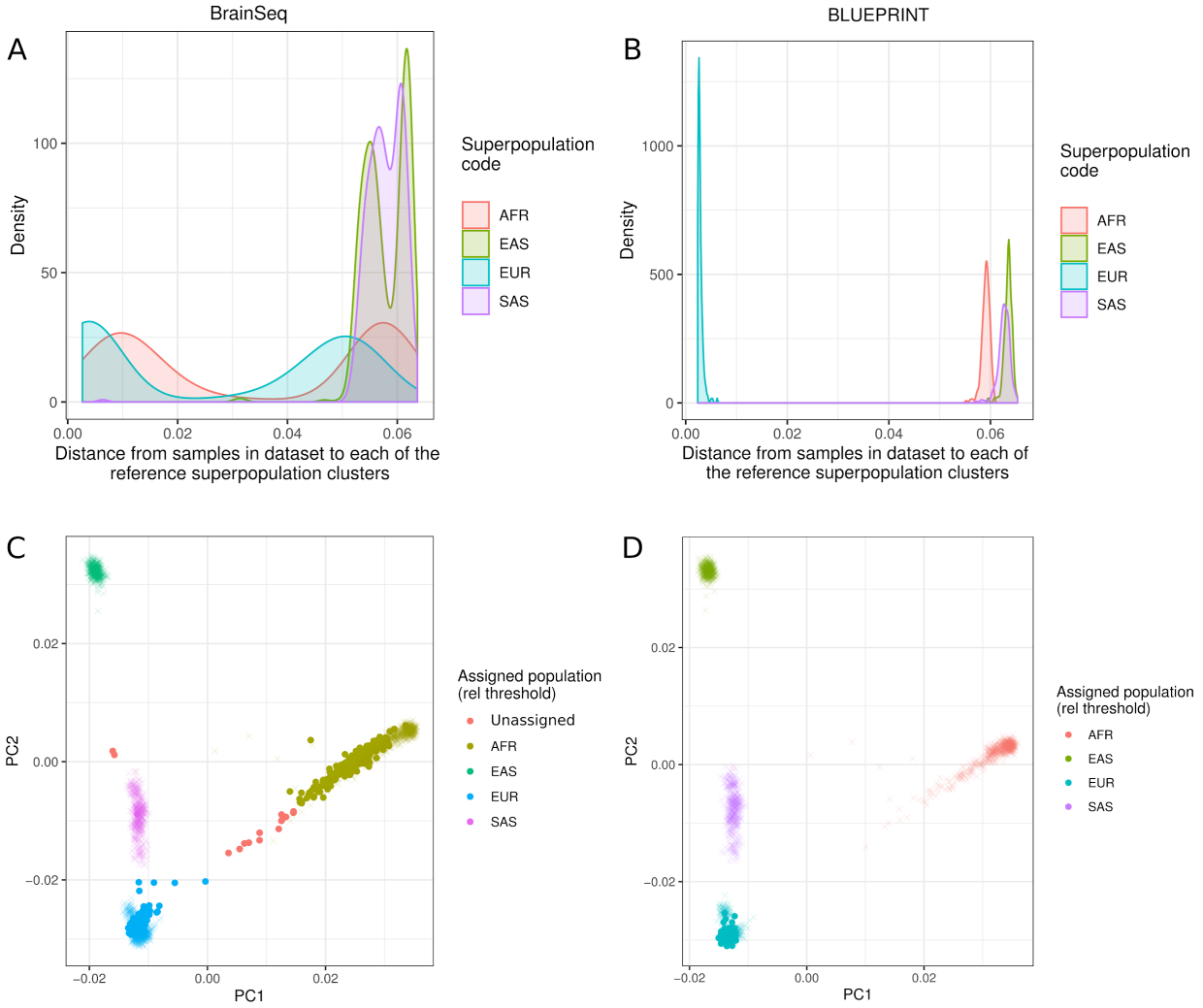
Supplementary Figure 4. The fraction of fine mapped eQTLs assigned to each of the 16 factors detected by the sn-spMF method.



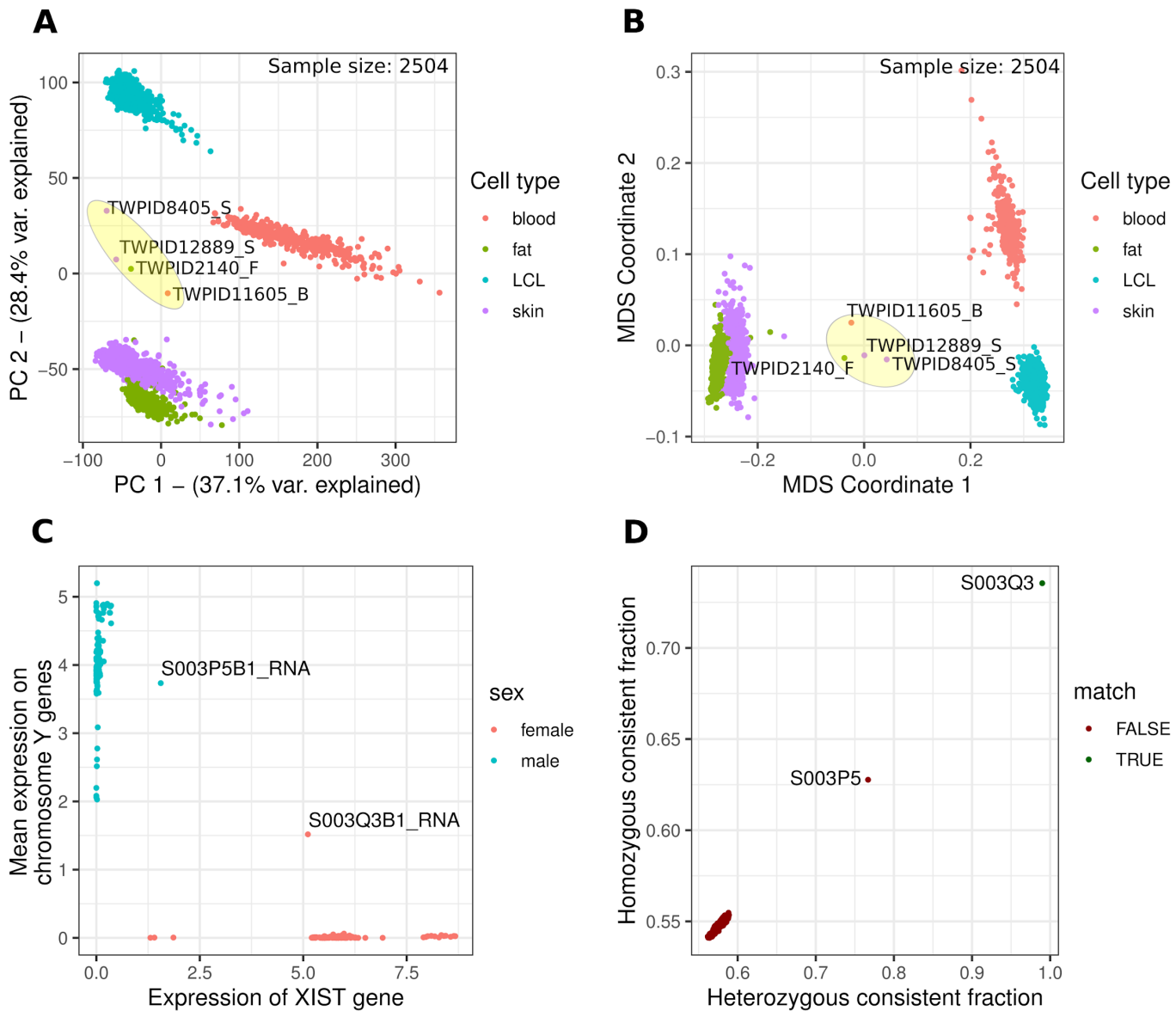
Supplementary Figure 5. The number of shared and novel colocalisations detected for the 14 traits and diseases.



Supplementary Figure 6. Regional association plot for LDL cholesterol (top panel) and *HMGCR* eQTL in the FUSION muscle dataset (bottom panel). The eQTL signal was fine mapped to 46 variants represented by red dots on both panels.



Supplementary Figure 7. Assigning genotyped samples to the four 1000 Genomes superpopulations. **(A)** Density plot of distances between each sample in BrainSeq (*60*) dataset and each superpopulation cluster in the 1000 Genomes Phase 3 reference dataset (*30*). First three principal components of the genotype data are used to calculate distances. The 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 (*20*) dataset and each superpopulation cluster in the 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.



Supplementary Figure 8. 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/Studies/). **(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 of 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 the opposite 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 (*81*). As expected, the genotypes of the S003Q3B1 RNA sample are most similar to the genotype data from the same donor (S003Q3) and most other donors are equally dis-similar,

forming a separate cluster in 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 the evidence presented in 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.

References

1. The GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. **369**, 1318–1330 (2020).

2. U. Võsa, A. Claringbould, H.-J. Westra, M. J. Bonder, P. Deelen, B. Zeng, H. Kirsten, A. Saha, R. Kreuzhuber, S. Kasela, N. Pervjakova, I. Alvaes, M.-J. Fave, M. Agbessi, M. Christiansen, R. Jansen, I. Seppälä, L. Tong, A. Teumer, K. Schramm, G. Hemani, J. Verlouw, H. Yaghootkar, R. Sönmez, A. A. Andrew, V. Kukushkina, A. Kalnapenkis, S. Rüeger, E. Porcu, J. Kronberg-Guzman, J. Kettunen, J. Powell, B. Lee, F. Zhang, W. Arindrarto, F. Beutner, BIOS Consortium, H. Brugge, i2QTL Consortium, J. Dmitrieva, M. Elansary, B. P. Fairfax, M. Georges, B. T. Heijmans, M. Kähönen, Y. Kim, J. C. Knight, P. Kovacs, K. Krohn, S. Li, M. Loeffler, U. M. Marigorta, H. Mei, Y. Momozawa, M. Müller-Nurasyid, M. Nauck, M. Nivard, B. Penninx, J. Pritchard, O. Raitakari, O. Rotzschke, E. P. Slagboom, C. D. A. Stehouwer, M. Stumvoll, P. Sullivan, P. A. 't, J. Thiery, A. Tönjes, J. van Dongen, M. van Iterson, J. Veldink, U. Völker, C. Wijmenga, M. Swertz, A. Andiappan, G. W. Montgomery, S. Ripatti, M. Perola, Z. Kutalik, E. Dermitzakis, S. Bergmann, T. Frayling, J. van Meurs, H. Prokisch, H. Ahsan, B. Pierce, T. Lehtimäki, D. Boomsma, B. M. Psaty, S. A. Gharib, P. Awadalla, L. Milani, W. H. Ouwehand, K. Downes, O. Stegle, A. Battle, J. Yang, P. M. Visscher, M. Scholz, G. Gibson, T. Esko, L. Franke, Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* (2018), p. 447367.

3. A. N. Barbeira, R. Bonazzola, E. R. Gamazon, Y. Liang, Y. Park, S. Kim-Hellmuth, G. Wang, Z. Jiang, D. Zhou, F. Hormozdiari, B. Liu, A. Rao, A. R. Hamel, M. D. Pividori, F. Aguet, GTEx GWAS Working Group, L. Bastarache, D. M. Jordan, M. Verbanck, R. Do, GTEx Consortium, M. Stephens, K. Ardlie, M. McCarthy, S. B. Montgomery, A. V. Segrè, C. D. Brown, T. Lappalainen, X. Wen, H. K. Im, Exploiting the GTEx resources to decipher the mechanisms at GWAS loci (2020), p. 814350.

4. D. W. Yao, L. J. O'Connor, A. L. Price, A. Gusev, Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).

5. B. D. Umans, A. Battle, Y. Gilad, Where Are the Disease-Associated eQTLs? *Trends Genet.* (2020), doi:10.1016/j.tig.2020.08.009.

6. H. Kilpinen, A. Goncalves, A. Leha, V. Afzal, K. Alasoo, S. Ashford, S. Bala, D. Bensaddek, F. P. Casale, O. J. Culley, P. Danecek, A. Faulconbridge, P. W. Harrison, A. Kathuria, D. McCarthy, S. A. McCarthy, R. Meleckyte, Y. Memari, N. Moens, F. Soares, A. Mann, I. Streeter, C. A. Agu, A. Alderton, R. Nelson, S. Harper, M. Patel, A. White, S. R. Patel, L. Clarke, R. Halai, C. M. Kirton, A. Kolb-Kokocinski, P. Beales, E. Birney, D. Danovi, A. I. Lamond, W. H. Ouwehand, L. Vallier, F. M. Watt, R. Durbin, O. Stegle, D. J. Gaffney, Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature*. **546**, 370–375 (2017).

7. Z. Mu, W. Wei, B. J. Fair, J. Miao, P. Zhu, Y. Li, Impact of cell-type and context-dependent regulatory variants on human immune traits. *bioRxiv* (2020), p. 2020.07.20.212753.

8. A. Young, N. Kumasaka, F. Calvert, T. R. Hammond, A. J. Knights, N. Panousis, J. Schwartzentruber, J. Liu, K. Kundu, M. Segel, N. Murphy, C. E. McMurran, H. Bulstrode, J. Correia, K. P. Budohoski, A. Joannides, M. R. Guilfoyle, R. Trivedi, R. Kirollos, R. Morris,

M. R. Garnett, H. Fernandes, I. Timofeev, I. Jalloh, K. Holland, R. Mannion, R. Mair, C. Watts, S. J. Price, P. J. Kirkpatrick, T. Santarius, N. Soranzo, B. Stevens, P. J. Hutchinson, R. J. M. Franklin, D. J. Gaffney, A map of transcriptional heterogeneity and regulatory variation in human microglia. *bioRxiv* (2019), p. 2019.12.20.874099.

9. K. de Paiva Lopes, G. J. L. Snijders, J. Humphrey, A. Allan, M. Sneeboer, E. Navarro, B. M. Schilder, R. A. Vialle, M. Parks, R. Missall, W. van Zuiden, F. Gigase, R. Kübler, A. B. van Berlekom, C. Böttcher, J. Priller, R. S. Kahn, L. D. de Witte, T. Raj, Atlas of genetic effects in human microglia transcriptome across brain regions, aging and disease pathologies. *Cold Spring Harbor Laboratory* (2020), p. 2020.10.27.356113.

10. J. Jerber, D. D. Seaton, A. S. E. Cuomo, N. Kumasaka, J. Haldane, J. Steer, M. Patel, D. Pearce, M. Andersson, M. J. Bonder, E. Mountjoy, M. Ghoussaini, M. A. Lancaster, HipSci Consortium, J. C. Marioni, F. T. Merkle, O. Stegle, D. J. Gaffney, Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *bioRxiv* (2020), p. 2020.05.21.103820.

11. K. Xia, A. A. Shabalin, S. Huang, V. Madar, Y.-H. Zhou, W. Wang, F. Zou, W. Sun, P. F. Sullivan, F. A. Wright, seeQTL: a searchable database for human eQTLs. *Bioinformatics*. **28**, 451–452 (2012).

12. Z. Zheng, D. Huang, J. Wang, K. Zhao, Y. Zhou, Z. Guo, S. Zhai, H. Xu, H. Cui, H. Yao, Z. Wang, X. Yi, S. Zhang, P. C. Sham, M. J. Li, QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. *Nucleic Acids Res.* **48**, D983–D991 (2020).

13. K. Watanabe, E. Taskesen, A. van Bochoven, D. Posthuma, Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

14. M. A. Kamat, J. A. Blackshaw, R. Young, P. Surendran, S. Burgess, J. Danesh, A. S. Butterworth, J. R. Staley, PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics*. **35**, 4851–4853 (2019).

15. S. Kalayci, M. E. Selvan, I. Ramos, C. Cotsapas, R. R. Montgomery, G. Poland, B. Pulendran, J. Tsang, R. J. Klein, Z. H. Gumus, ImmuneRegulation: A web-based tool for identifying human immune regulatory elements. *bioRxiv* (2018), p. 468124.

16. C.-H. Yu, L. R. Pal, J. Moult, Consensus Genome-Wide Expression Quantitative Trait Loci and Their Relationship with Human Complex Trait Disease. *OMICS*. **20**, 400–414 (2016).

17. M. Munz, I. Wohlers, E. Simon, T. Reinberger, H. Busch, A. S. Schaefer, J. Erdmann, Qtlizer: comprehensive QTL annotation of GWAS results. *Sci. Rep.* **10**, 20417 (2020).

18. J. Zheng, V. Haberland, D. Baird, V. Walker, P. C. Haycock, M. R. Hurle, A. Gutteridge, P. Erola, Y. Liu, S. Luo, J. Robinson, T. G. Richardson, J. R. Staley, B. Elsworth, S. Burgess, B. B. Sun, J. Danesh, H. Runz, J. C. Maranville, H. M. Martin, J. Yarmolinsky, C. Laurin, M. V. Holmes, J. Z. Liu, K. Estrada, R. Santos, L. McCarthy, D. Waterworth, M. R. Nelson, G. D. Smith, A. S. Butterworth, G. Hemani, R. A. Scott, T. R. Gaunt, Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).

19. S. M. Urbut, G. Wang, P. Carbonetto, M. Stephens, Flexible statistical methods for

estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).

20. L. Chen, B. Ge, F. P. Casale, L. Vasquez, T. Kwan, D. Garrido-Martín, S. Watt, Y. Yan, K. Kundu, S. Ecker, A. Datta, D. Richardson, F. Burden, D. Mead, A. L. Mann, J. M. Fernandez, S. Rowlston, S. P. Wilder, S. Farrow, X. Shao, J. J. Lambourne, A. Redensek, C. A. Albers, V. Amstislavskiy, S. Ashford, K. Berentsen, L. Bomba, G. Bourque, D. Bujold, S. Busche, M. Caron, S.-H. Chen, W. Cheung, O. Delaneau, E. T. Dermitzakis, H. Elding, I. Colgiu, F. O. Bagger, P. Flicek, E. Habibi, V. Iotchkova, E. Janssen-Megens, B. Kim, H. Lehrach, E. Lowy, A. Mandoli, F. Matarese, M. T. Maurano, J. A. Morris, V. Pancaldi, F. Pourfarzad, K. Rehnstrom, A. Rendon, T. Risch, N. Sharifi, M.-M. Simon, M. Sultan, A. Valencia, K. Walter, S.-Y. Wang, M. Frontini, S. E. Antonarakis, L. Clarke, M.-L. Yaspo, S. Beck, R. Guigo, D. Rico, J. H. A. Martens, W. H. Ouwehand, T. W. Kuijpers, D. S. Paul, H. G. Stunnenberg, O. Stegle, K. Downes, T. Pastinen, N. Soranzo, Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*. **167**, 1398–1414.e24 (2016).

21. B. J. Schmiedel, D. Singh, A. Madrigal, A. G. Valdovino-Gonzalez, B. M. White, J. Zapardiel-Gonzalo, B. Ha, G. Altay, J. A. Greenbaum, G. McVicker, G. Seumois, A. Rao, M. Kronenberg, B. Peters, P. Vijayanand, Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell*. **175**, 1701–1715.e16 (2018).

22. T. Flutre, X. Wen, J. Pritchard, M. Stephens, A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* **9**, e1003486 (2013).

23. G. Wang, A. Sarkar, P. Carbonetto, M. Stephens, A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **25**, 1 (2020).

24. C. Benner, C. C. A. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti, M. Pirinen, FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*. **32**, 1493–1501 (2016).

25. Q. S. Wang, D. R. Kelley, J. Ulirsch, M. Kanai, S. Sadhuka, R. Cui, C. Albors, N. Cheng, Y. Okada, The Biobank Japan Project, F. Aguet, K. G. Ardlie, D. G. MacArthur, H. K. Finucane, Leveraging supervised learning for functionally-informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. *Cold Spring Harbor Laboratory* (2020), p. 2020.10.20.347294.

26. C. Benner, A. S. Havulinna, M.-R. Järvelin, V. Salomaa, S. Ripatti, M. Pirinen, Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. *Am. J. Hum. Genet.* **0** (2017), doi:10.1016/j.ajhg.2017.08.012.

27. O. Weissbrod, F. Hormozdiari, C. Benner, R. Cui, J. Ulirsch, S. Gazal, A. P. Schoech, B. van de Geijn, Y. Reshef, C. Márquez-Luna, L. O'Connor, M. Pirinen, H. K. Finucane, A. L. Price, Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).

28. A. D. Yates, P. Achuthan, W. Akanni, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, K. Billis, S. Boddu, J. C. Marugán, C. Cummins, C. Davidson, K. Dodiya, R. Fatima, A. Gall, C. G. Giron, L. Gil, T. Grego, L.

Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, M. Kay, I. Lavidas, T. Le, D. Lemos, J. G. Martinez, T. Maurel, M. McDowall, A. McMahon, S. Mohanan, B. Moore, M. Nuhn, D. N. Oheh, A. Parker, A. Parton, M. Patricio, M. P. Sakthivel, A. I. Abdul Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, M. Sycheva, M. Szuba, K. Taylor, A. Thormann, G. Threadgold, A. Vullo, B. Walts, A. Winterbottom, A. Zadissa, M. Chakiachvili, B. Flint, A. Frankish, S. E. Hunt, G. IIsley, M. Kostadima, N. Langridge, J. E. Loveland, F. J. Martin, J. Morales, J. M. Mudge, M. Muffato, E. Perry, M. Ruffier, S. J. Trevanion, F. Cunningham, K. L. Howe, D. R. Zerbino, P. Flicek, Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).

29. M. C. Mills, C. Rahal, The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat. Genet.* **52**, 242–243 (2020).

30. P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, 1000 Genomes Project Consortium, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, J. O. Korbel, An integrated map of structural variation in 2,504 human genomes. *Nature*. **526**, 75–81 (2015).

31. Y. He, S. B. Chhetri, M. Arvanitis, K. Srinivasan, F. Aguet, K. G. Ardlie, A. N. Barbeira, R. Bonazzola, H. K. Im, C. D. Brown, A. Battle, GTEx Consortium, sn-spMF: matrix factorization informs tissue-specific genetic regulation of gene expression. *Genome Biol.* **21**, 235 (2020).

32. D. L. Taylor, A. U. Jackson, N. Narisu, G. Hemani, M. R. Erdos, P. S. Chines, A. Swift, J. Idol, J. P. Didion, R. P. Welch, L. Kinnunen, J. Saramies, T. A. Lakka, M. Laakso, J. Tuomilehto, S. C. J. Parker, H. A. Koistinen, G. Davey Smith, M. Boehnke, L. J. Scott, E. Birney, F. S. Collins, Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 10883–10888 (2019).

33. A. Buil, A. A. Brown, T. Lappalainen, A. Viñuela, M. N. Davies, H.-F. Zheng, J. B. Richards, D. Glass, K. S. Small, R. Durbin, T. D. Spector, E. T. Dermitzakis, Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2015).

34. W. J. Astle, H. Elding, T. Jiang, D. Allen, D. Ruklisa, A. L. Mann, D. Mead, H. Bouman, F. Riveros-Mckay, M. A. Kostadima, J. J. Lambourne, S. Sivapalaratnam, K. Downes, K. Kundu, L. Bomba, K. Berentsen, J. R. Bradley, L. C. Daugherty, O. Delaneau, K. Freson, S. F. Garner, L. Grassi, J. Guerrero, M. Haimel, E. M. Janssen-Megens, A. Kaan, M. Kamat, B. Kim, A. Mandoli, J. Marchini, J. H. A. Martens, S. Meacham, K. Megy, J. O'Connell, R. Petersen, N. Sharifi, S. M. Sheard, J. R. Staley, S. Tuna, M. van der Ent, K. Walter, S.-Y. Wang, E. Wheeler, S. P. Wilder, V. Iotchkova, C. Moore, J. Sambrook, H. G. Stunnenberg, E. Di Angelantonio, S. Kaptoge, T. W. Kuijpers, E. Carrillo-de-Santa-Pau, D. Juan, D. Rico,

A. Valencia, L. Chen, B. Ge, L. Vasquez, T. Kwan, D. Garrido-Martín, S. Watt, Y. Yang, R. Guigo, S. Beck, D. S. Paul, T. Pastinen, D. Bujold, G. Bourque, M. Frontini, J. Danesh, D. J. Roberts, W. H. Ouwehand, A. S. Butterworth, N. Soranzo, The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*. **167**, 1415–1429.e19 (2016).

35. T. Berisa, J. K. Pickrell, Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*. **32**, 283–285 (2016).

36. Y. Momozawa, J. Dmitrieva, E. Théâtre, V. Deffontaine, S. Rahmouni, B. Charloteaux, F. Crins, E. Docampo, M. Elansary, A.-S. Gori, C. Lecut, R. Mariman, M. Mni, C. Oury, I. Altukhov, D. Alexeev, Y. Aulchenko, L. Amininejad, G. Bouma, F. Hoentjen, M. Löwenberg, B. Oldenburg, M. J. Pierik, A. E. Vander Meulen-de Jong, C. Janneke van der Woude, M. C. Visschedijk, International IBD Genetics Consortium, M. Lathrop, J.-P. Hugot, R. K. Weersma, M. De Vos, D. Franchimont, S. Vermeire, M. Kubo, E. Louis, M. Georges, IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat. Commun.* **9**, 2427 (2018).

37. Y. I. Li, B. van de Geijn, A. Raj, D. A. Knowles, A. A. Petti, D. Golan, Y. Gilad, J. K. Pritchard, RNA splicing is a primary link between genetic variation and disease. *Science*. **352**, 600–604 (2016).

38. K. Alasoo, J. Rodrigues, J. Danesh, D. F. Freitag, D. S. Paul, D. J. Gaffney, Genetic effects on promoter usage are highly context-specific and contribute to complex traits. *Elife*. **8** (2019), doi:10.7554/eLife.41673.

39. R. Burkhardt, E. E. Kenny, J. K. Lowe, A. Birkeland, R. Josowitz, M. Noel, J. Salit, J. B. Maller, I. Pe'er, M. J. Daly, D. Altshuler, M. Stoffel, J. M. Friedman, J. L. Breslow, Common SNPs in HMGCR in micronesians and whites associated with LDL-cholesterol levels affect alternative splicing of exon13. *Arterioscler. Thromb. Vasc. Biol.* **28**, 2078–2084 (2008).

40. S. Kim-Hellmuth, M. Bechheim, B. Pütz, P. Mohammadi, Y. Nédélec, N. Giangreco, J. Becker, V. Kaiser, N. Fricker, E. Beier, P. Boor, S. E. Castel, M. M. Nöthen, L. B. Barreiro, J. K. Pickrell, B. Müller-Myhsok, T. Lappalainen, J. Schumacher, V. Hornung, Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nat. Commun.* **8**, 266 (2017).

41. R. E. Peterson, K. Kuchenbaecker, R. K. Walters, C.-Y. Chen, A. B. Popejoy, S. Periyasamy, M. Lam, C. Iyegbe, R. J. Strawbridge, L. Brick, C. E. Carey, A. R. Martin, J. L. Meyers, J. Su, J. Chen, A. C. Edwards, A. Kalungi, N. Koen, L. Majara, E. Schwarz, J. W. Smoller, E. A. Stahl, P. F. Sullivan, E. Vassos, B. Mowry, M. L. Prieto, A. Cuellar-Barboza, T. B. Bigdeli, H. J. Edenberg, H. Huang, L. E. Duncan, Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*. **0** (2019), doi:10.1016/j.cell.2019.08.051.

42. Y. Zhong, M. A. Perera, E. R. Gamazon, On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. *Am. J. Hum. Genet.* (2019), doi:10.1016/j.ajhg.2019.04.009.

43. Y. Zhong, T. De, C. Alarcon, C. Sehwan Park, B. Lec, M. A. Perera, Discovery of novel hepatocyte eQTLs in African Americans. *PLoS Genet.* **16**, e1008662 (2020).

44. P. Ewels, A. Peltzer, S. Fillinger, J. Alneberg, H. Patel, A. Wilm, M. Garcia, P. Di Tommaso, S. Nahnsen, nf-core: Community curated bioinformatics pipelines. *bioRxiv* (2019), p. 610741.

45. A. Athar, A. Füllgrabe, N. George, H. Iqbal, L. Huerta, A. Ali, C. Snow, N. A. Fonseca, R. Petryszak, I. Papatheodorou, U. Sarkans, A. Brazma, ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.* **47**, D711–D715 (2019).

46. P. Deelen, M. J. Bonder, K. J. van der Velde, H.-J. Westra, E. Winder, D. Hendriksen, L. Franke, M. A. Swertz, Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes*. **7**, 901 (2014).

47. P.-R. Loh, P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A Reshef, H. K Finucane, S. Schoenherr, L. Forer, S. McCarthy, G. R. Abecasis, R. Durbin, A. L Price, Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

48. S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.-R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, C. Fuchsberger, Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

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

50. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. **4**, 7 (2015).

51. D. Speed, G. Hemani, M. R. Johnson, D. J. Balding, Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).

52. B. P. Fairfax, S. Makino, J. Radhakrishnan, K. Plant, S. Leslie, A. Dilthey, P. Ellis, C. Langford, F. O. Vannberg, J. C. Knight, Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).

53. B. P. Fairfax, P. Humburg, S. Makino, V. Naranbhai, D. Wong, E. Lau, L. Jostins, K. Plant, R. Andrews, C. McGee, J. C. Knight, Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*. **343**, 1246949 (2014).

54. S. Kasela, K. Kisand, L. Tserel, E. Kaleviste, A. Remm, K. Fischer, T. Esko, H.-J. Westra, B. P. Fairfax, S. Makino, J. C. Knight, L. Franke, A. Metspalu, P. Peterson, L. Milani, Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. *PLoS Genet.* **13**, e1006643 (2017).

55. V. Naranbhai, B. P. Fairfax, S. Makino, P. Humburg, D. Wong, E. Ng, A. V. S. Hill, J. C. Knight, Genomic modulators of gene expression in human neutrophils. *Nat. Commun.* **6**, 7545 (2015).

56. M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, G. K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids*

Res. **43**, e47 (2015).

57. P. Du, W. A. Kibbe, S. M. Lin, lumi: a pipeline for processing Illumina microarray. *Bioinformatics*. **24**, 1547–1548 (2008).

58. H.-J. Westra, R. C. Jansen, R. S. N. Fehrmann, G. J. te Meerman, D. van Heel, C. Wijmenga, L. Franke, MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics*. **27**, 2104–2111 (2011).

59. B. Ng, C. C. White, H.-U. Klein, S. K. Sieberts, C. McCabe, E. Patrick, J. Xu, L. Yu, C. Gaiteri, D. A. Bennett, S. Mostafavi, P. L. De Jager, An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* **20**, 1418–1426 (2017).

60. A. E. Jaffe, R. E. Straub, J. H. Shin, R. Tao, Y. Gao, L. Collado-Torres, T. Kam-Thong, H. S. Xi, J. Quan, Q. Chen, C. Colantuoni, W. S. Ulrich, B. J. Maher, A. Deep-Soboslay, BrainSeq Consortium, A. J. Cross, N. J. Brandon, J. T. Leek, T. M. Hyde, J. E. Kleinman, D. R. Weinberger, Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.* **21**, 1117–1125 (2018).

61. K. Kundu, A. L. Mann, M. Tardaguila, S. Watt, H. Ponstingl, L. Vasquez, N. W. Morrell, O. Stegle, T. Pastinen, S. J. Sawcer, C. A. Anderson, K. Walter, N. Soranzo, Genetic associations at regulatory phenotypes improve fine-mapping of causal variants for twelve immune-mediated diseases. *bioRxiv* (2020), p. 2020.01.15.907436.

62. H. Quach, M. Rotival, J. Pothlichet, Y.-H. E. Loh, M. Dannemann, N. Zidane, G. Laval, E. Patin, C. Harmant, M. Lopez, M. Deschamps, N. Naffakh, D. Duffy, A. Coen, G. Leroux-Roels, F. Clément, A. Boland, J.-F. Deleuze, J. Kelso, M. L. Albert, L. Quintana-Murci, Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell*. **167**, 643–656.e17 (2016).

63. M. Gutierrez-Arcelus, T. Lappalainen, S. B. Montgomery, A. Buil, H. Ongen, A. Yurovsky, J. Bryois, T. Giger, L. Romano, A. Planchon, E. Falconnet, D. Bielser, M. Gagnebin, I. Padioleau, C. Borel, A. Letourneau, P. Makrythanasis, M. Guipponi, C. Gehrig, S. E. Antonarakis, E. T. Dermitzakis, Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife*. **2**, e00523 (2013).

64. T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. C. 't Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, Geuvadis Consortium, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Häsler, A.-C. Syvänen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, E. T. Dermitzakis, Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. **501**, 506–511 (2013).

65. K. Alasoo, J. Rodrigues, S. Mukhopadhyay, A. J. Knights, A. L. Mann, K. Kundu, HIPSCI Consortium, C. Hale, G. Dougan, D. J. Gaffney, Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**,

424–431 (2018).

66. Y. Nédélec, J. Sanz, G. Baharian, Z. A. Szpiech, A. Pacis, A. Dumaine, J.-C. Grenier, A. Freiman, A. J. Sams, S. Hebert, A. Pagé Sabourin, F. Luca, R. Blekhman, R. D. Hernandez, R. Pique-Regi, J. Tung, V. Yotova, L. B. Barreiro, Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell*. **167**, 657–669.e21 (2016).

67. K. Lepik, T. Annilo, V. Kukuškina, K. Kisand, Z. Kutalik, P. Peterson, H. Peterson, eQTLGen Consortium, C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. *PLoS Comput. Biol.* **13**, e1005766 (2017).

68. M. van de Bunt, J. E. Manning Fox, X. Dai, A. Barrett, C. Grey, L. Li, A. J. Bennett, P. R. Johnson, R. V. Rajotte, K. J. Gaulton, E. T. Dermitzakis, P. E. MacDonald, M. I. McCarthy, A. L. Gloyn, 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.* **11**, e1005694 (2015).

69. J. Schwartzentruber, S. Foskolou, H. Kilpinen, J. Rodrigues, K. Alasoo, A. J. Knights, M. Patel, A. Goncalves, R. Ferreira, C. L. Benn, A. Wilbrey, M. Bictash, E. Impey, L. Cao, S. Lainez, A. J. Loucif, P. J. Whiting, A. Gutteridge, D. J. Gaffney, HIPSCI Consortium, Molecular and functional variation in iPSC-derived sensory neurons. *Nat. Genet.* **50**, 54–61 (2018).

70. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

71. P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, C. Notredame, Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).

72. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

73. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, T. J. Hubbard, GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).

74. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. **30**, 923–930 (2014).

75. S. Anders, A. Reyes, W. Huber, Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).

76. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*. **14**, 417–419 (2017).

77. G. Pertea, M. Pertea, GFF Utilities: GffRead and GffCompare. *F1000Res.* **9** (2020), doi:10.12688/f1000research.23297.2.

78. G. P. Wagner, K. Kin, V. J. Lynch, Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).

79. M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segrè, S. Djebali, A. Niarchou, Consortium, The GTEx, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, R. Guigó, The human transcriptome across tissues and individuals. *Science*. **348**, 660–665 (2015).

80. P. A. C. 't Hoen, M. R. Friedländer, J. Almlöf, M. Sammeth, I. Pulyakhina, S. Y. Anvar, J. F. J. Laros, H. P. J. Buermans, O. Karlberg, M. Brännvall, GEUVADIS Consortium, J. T. den Dunnen, G.-J. B. van Ommen, I. G. Gut, R. Guigó, X. Estivill, A.-C. Syvänen, E. T. Dermitzakis, T. Lappalainen, Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **31**, 1015–1022 (2013).

81. A. Fort, N. I. Panousis, M. Garieri, S. E. Antonarakis, T. Lappalainen, E. T. Dermitzakis, O. Delaneau, MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. *Bioinformatics*. **33**, 1895–1897 (2017).

82. K. D. Hansen, R. A. Irizarry, Z. Wu, Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*. **13**, 204–216 (2012).

83. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).

84. C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, M. A. Haendel, Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, R5 (2012).

85. A. D. Diehl, T. F. Meehan, Y. M. Bradford, M. H. Brush, W. M. Dahdul, D. S. Dougall, Y. He, D. Osumi-Sutherland, A. Ruttenberg, S. Sarntivijai, C. E. Van Slyke, N. A. Vasilevsky, M. A. Haendel, J. A. Blake, C. J. Mungall, The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics*. **7**, 44 (2016).

86. J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, H. Parkinson, Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*. **26**, 1112–1118 (2010).

87. H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, O. Delaneau, Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*. **32**, 1479–1485 (2016).

88. O. Delaneau, H. Ongen, A. A. Brown, A. Fort, N. I. Panousis, E. T. Dermitzakis, A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).

89. M. Lyon, S. J. Andrews, B. Elsworth, T. R. Gaunt, G. Hemani, E. Marcora, The variant call format provides efficient and robust storage of GWAS summary statistics (2020), p. 2020.05.29.115824.

90. B. Elsworth, M. Lyon, T. Alexander, Y. Liu, P. Matthews, J. Hallett, P. Bates, T. Palmer, V. Haberland, G. D. Smith, J. Zheng, P. Haycock, T. R. Gaunt, G. Hemani, The MRC IEU

OpenGWAS data infrastructure (2020), p. 2020.08.10.244293.

91. K. M. de Lange, L. Moutsianas, J. C. Lee, C. A. Lamb, Y. Luo, N. A. Kennedy, L. Jostins, D. L. Rice, J. Gutierrez-Achury, S.-G. Ji, G. Heap, E. R. Nimmo, C. Edwards, P. Henderson, C. Mowat, J. Sanderson, J. Satsangi, A. Simmons, D. C. Wilson, M. Tremelling, A. Hart, C. G. Mathew, W. G. Newman, M. Parkes, C. W. Lees, H. Uhlig, C. Hawkey, N. J. Prescott, T. Ahmad, J. C. Mansfield, C. A. Anderson, J. C. Barrett, Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).

92. Y. Okada, D. Wu, G. Trynka, T. Raj, C. Terao, K. Ikari, Y. Kochi, K. Ohmura, A. Suzuki, S. Yoshida, R. R. Graham, A. Manoharan, W. Ortmann, T. Bhangale, J. C. Denny, R. J. Carroll, A. E. Eyler, J. D. Greenberg, J. M. Kremer, D. A. Pappas, L. Jiang, J. Yin, L. Ye, D.-F. Su, J. Yang, G. Xie, E. Keystone, H.-J. Westra, T. Esko, A. Metspalu, X. Zhou, N. Gupta, D. Mirel, E. A. Stahl, D. Diogo, J. Cui, K. Liao, M. H. Guo, K. Myouzen, T. Kawaguchi, M. J. H. Coenen, P. L. C. M. van Riel, M. A. F. J. van de Laar, H.-J. Guchelaar, T. W. J. Huizinga, P. Dieudé, X. Mariette, S. L. Bridges Jr, A. Zhernakova, R. E. M. Toes, P. P. Tak, C. Miceli-Richard, S.-Y. Bang, H.-S. Lee, J. Martin, M. A. Gonzalez-Gay, L. Rodriguez-Rodriguez, S. Rantapää-Dahlqvist, L. Arlestig, H. K. Choi, Y. Kamatani, P. Galan, M. Lathrop, RACI consortium, GARNET consortium, S. Eyre, J. Bowes, A. Barton, N. de Vries, L. W. Moreland, L. A. Criswell, E. W. Karlson, A. Taniguchi, R. Yamada, M. Kubo, J. S. Liu, S.-C. Bae, J. Worthington, L. Padyukov, L. Klareskog, P. K. Gregersen, S. Raychaudhuri, B. E. Stranger, P. L. De Jager, L. Franke, P. M. Visscher, M. A. Brown, H. Yamanaka, T. Mimori, A. Takahashi, H. Xu, T. W. Behrens, K. A. Siminovitch, S. Momohara, F. Matsuda, K. Yamamoto, R. M. Plenge, Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*. **506**, 376–381 (2014).

93. J. Bentham, D. L. Morris, D. S. Cunninghame Graham, C. L. Pinder, P. Tombleson, T. W. Behrens, J. Martín, B. P. Fairfax, J. C. Knight, L. Chen, J. Replogle, A.-C. Syvänen, L. Rönnblom, R. R. Graham, J. E. Wither, J. D. Rioux, M. E. Alarcón-Riquelme, T. J. Vyse, Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464 (2015).

94. A. Xue, Y. Wu, Z. Zhu, F. Zhang, K. E. Kemper, Z. Zheng, L. Yengo, L. R. Lloyd-Jones, J. Sidorenko, Y. Wu, A. F. McRae, P. M. Visscher, J. Zeng, J. Yang, Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9**, 2941 (2018).

95. M. Nikpay, A. Goel, H.-H. Won, L. M. Hall, C. Willenborg, S. Kanoni, D. Saleheen, T. Kyriakou, C. P. Nelson, J. C. Hopewell, T. R. Webb, L. Zeng, A. Dehghan, M. Alver, S. M. Armasu, K. Auro, A. Bjonnes, D. I. Chasman, S. Chen, I. Ford, N. Franceschini, C. Gieger, C. Grace, S. Gustafsson, J. Huang, S.-J. Hwang, Y. K. Kim, M. E. Kleber, K. W. Lau, X. Lu, Y. Lu, L.-P. Lyytikäinen, E. Mihailov, A. C. Morrison, N. Pervjakova, L. Qu, L. M. Rose, E. Salfati, R. Saxena, M. Scholz, A. V. Smith, E. Tikkanen, A. Uitterlinden, X. Yang, W. Zhang, W. Zhao, M. de Andrade, P. S. de Vries, N. R. van Zuydam, S. S. Anand, L. Bertram, F. Beutner, G. Dedoussis, P. Frossard, D. Gauguier, A. H. Goodall, O. Gottesman, M. Haber, B.-G. Han, J. Huang, S. Jalilzadeh, T. Kessler, I. R. König, L. Lannfelt, W. Lieb, L. Lind, C. M. Lindgren, M.-L. Lokki, P. K. Magnusson, N. H. Mallick, N. Mehra, T. Meitinger, F.-U.-R. Memon, A. P. Morris, M. S. Nieminen, N. L. Pedersen, A. Peters, L. S. Rallidis, A. Rasheed, M. Samuel, S. H. Shah, J. Sinisalo, K. E. Stirrups, S. Trompet, L. Wang, K. S.

Zaman, D. Ardissino, E. Boerwinkle, I. B. Borecki, E. P. Bottinger, J. E. Buring, J. C. Chambers, R. Collins, L. A. Cupples, J. Danesh, I. Demuth, R. Elosua, S. E. Epstein, T. Esko, M. F. Feitosa, O. H. Franco, M. G. Franzosi, C. B. Granger, D. Gu, V. Gudnason, A. S. Hall, A. Hamsten, T. B. Harris, S. L. Hazen, C. Hengstenberg, A. Hofman, E. Ingelsson, C. Iribarren, J. W. Jukema, P. J. Karhunen, B.-J. Kim, J. S. Kooner, I. J. Kullo, T. Lehtimäki, R. J. F. Loos, O. Melander, A. Metspalu, W. März, C. N. Palmer, M. Perola, T. Quertermous, D. J. Rader, P. M. Ridker, S. Ripatti, R. Roberts, V. Salomaa, D. K. Sanghera, S. M. Schwartz, U. Seedorf, A. F. Stewart, D. J. Stott, J. Thiery, P. A. Zalloua, C. J. O'Donnell, M. P. Reilly, T. L. Assimes, J. R. Thompson, J. Erdmann, R. Clarke, H. Watkins, S. Kathiresan, R. McPherson, P. Deloukas, H. Schunkert, N. J. Samani, M. Farrall, A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).

96. Pan UKBB, (available at https://pan.ukbb.broadinstitute.org).

97. C. Wallace, Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet.* **16**, e1008720 (2020).

98. C. Giambartolomei, D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, V. Plagnol, Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).

99. J. R. Conway, A. Lex, N. Gehlenborg, UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*. **33**, 2938–2940 (2017).

100. H. Li, Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*. **27**, 718–719 (2011).

101. X. Zhan, D. J. Liu, SEQMINER: An R-Package to Facilitate the Functional Interpretation of Sequence-Based Associations. *Genet. Epidemiol.* **39**, 619–623 (2015).

102. M. Ghoussaini, E. Mountjoy, M. Carmona, G. Peat, E. M. Schmidt, A. Hercules, L. Fumis, A. Miranda, D. Carvalho-Silva, A. Buniello, T. Burdett, J. Hayhurst, J. Baker, J. Ferrer, A. Gonzalez-Uriarte, S. Jupp, M. A. Karim, G. Koscielny, S. Machlitt-Northen, C. Malangone, Z. M. Pendlington, P. Roncaglia, D. Suveges, D. Wright, O. Vrousgou, E. Papa, H. Parkinson, J. A. L. MacArthur, J. A. Todd, J. C. Barrett, J. Schwartzentruber, D. G. Hulcoop, D. Ochoa, E. M. McDonagh, I. Dunham, Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, D1311–D1320 (2021).

103. I. Papatheodorou, P. Moreno, J. Manning, A. M.-P. Fuentes, N. George, S. Fexova, N. A. Fonseca, A. Füllgrabe, M. Green, N. Huang, L. Huerta, H. Iqbal, M. Jianu, S. Mohammed, L. Zhao, A. F. Jarnuczak, S. Jupp, J. Marioni, K. Meyer, R. Petryszak, C. A. Prada Medina, C. Talavera-López, S. Teichmann, J. A. Vizcaino, A. Brazma, Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* **48**, D77–D83 (2020).

104. Y. I. Li, D. A. Knowles, J. Humphrey, A. N. Barbeira, S. P. Dickinson, H. K. Im, J. K. Pritchard, Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).