nature genetics

Article

https://doi.org/10.1038/s41588-023-01494-9

Analysis of blood methylation quantitative trait loci in East Asians reveals ancestry-specific impacts on complex traits

Supplementary Note

I. Supplementary Protocols

Two mQTL mapping methods and comparison

FastQTLmapping. FastQTLmapping is a linear regression solver in C++ for exhaustive regression analysis between two extraordinary large matrices allowing for covariates, which is particularly useful for mQTL-like studies[1]. The input contains three text files of floating-point matrices for independent variables ***X***, dependent variables ***Y***, and covariates ***C***. Note that rows represent variables and columns represent observations for faster loading. FastQTLmapping also supports Plink binary file format. The output file records the test statistics of all regressions, optionable under a preset significance threshold. FastQTLmapping is developed on Linux using MKL (https://software.intel.com/tools/onemkl) and GSL (http://www.gnu.org/software/gsl/) library, and is run from the command line. C++ source code, an example run, and documentation of fastQTLmapping are freely available at https://github.com/Fun-Gene/fastQTLmapping.

MatrixEQTL. MatrixEQTL (R package v2.3) is widely-used QTL mapping tool which processed large data with large matrix operations with no loss of precision. Linear addictive models were calculated, adjusted for age, sex, batch, bisulfite slide number, array position, top ten genomic PCs and predicted blood cell fractions (B cells, $CD4^+$ and $CD8^+$ T cells, NK cells, monocytes and neutrophils). SNPs within 1 Mb of the CpGs were considered as cis-mQTLs. SNPs located further than 1Mb of the CpGs or in other chromosomes were considered as trans-mQTLs. The *P*-value threshold for defining mQTLs was 1.00×10^{-8}.

Both fastQTLmapping and MatrixEQTL produced the exact results under all investigated settings. The operating time of both fastQTLmapping and MatrixEQTL was largely linear to the computation size. In the single-thread setting, fastQTLmapping completed the analyses of sizes 10^9, 10^{10}, and 10^{11} in 24, 210, and 2,058 seconds, respectively. MatrixEQTL completed these analyses in 103, 714, and 6,733 seconds, respectively, meaning that fastQTLmapping was 3 - 4 times faster than MatrixEQTL. In the 32 CPU threads setting, fastQTLmapping completed these analyses of sizes 10^9, 10^{10}, and 10^{11} in 3, 14, and 100 seconds, respectively. MatrixEQTL completed these analyses in 33, 162, and 682 seconds, respectively, meaning that fastQTLmapping was 6-11 times faster than MatrixEQTL. In addition, the peak memory usage of fastQTLmapping under the computation size of 10^{11} and 32-thread (17.6 GB) was much small than that of MatrixEQTL (105.4GB). Applying fastQTLmapping in the analysis of our data (8.6M SNPs × 811K CpGs × 3,523 samples) consumed 6.79 hours and 321GB memory under 32-threads parallel.

Estimating the heritability of DNAm

Narrow-sense heritability of each CpG (h_{CpG}^2) was estimated by a self-designed pipeline. We first adjust DNAm M values for age, sex, batch, bisulfite slide number, top ten genomic PCs and predicted blood cell fractions (B cells, CD4$^+$ and CD8$^+$ T cells, NK cells, monocytes and neutrophils). In our algorithms, we used the same hybrid estimation scheme in GCTA v1.94.1[2] and utilize the factored spectral transformation (FaST)[3] to accelerate the calculation. By parameterization, we also transformed the constrained optimization problem into an unconstrained. The run time and memory footprint were linear in the cohort size, making it far efficient than GCTA.

Functional annotation and enrichment of mQTLs and mCpGs

Enrichment of mQTLs and mCpGs in different genomic regions

The SNPs were annotated by ANNOVAR v2020-06-07[4] with its pre-built RefSeq Gene annotation files (hg19 build, date: 2020.11.2). The annotations contained: Upstream (within 1500bp region upstream of transcription start site, TSS), 5´ UTR (within a 5´ untranslated region), Exon (within a coding region), Splice (within 2bp of a splicing junction), Intron (overlapping an intron), ncRNA (non-coding RNA items) and 3´ UTR (within a 3´ untranslated region).

The CpGs were annotated by referring to the manufacturer's manifest files[5] (hg19 build, date: 2020.10.21). The genomic annotations contained: Enhancer (in FANTOM5[6] project defined enhancer regions), TSS 1500 (200-1500 bases upstream of the TSS), TSS 200 (0-200 bases upstream of the transcriptional start site), 5´ UTR (within the 5´ untranslated regions), 1st Exon (the first exon), Exon Bnd (within 20 bases of an exon boundary, i.e. the start or end of an exon), Body (gene body) and 3´ UTR (within a 3´ untranslated region). The CpG island annotations included: N Shelf (upstream 2-4Kbp from CpG islands), N Shore (upstream 0-2Kbp from CpG islands), Island, S Shore (downstream 0-2Kbp from CpG islands) and S Shelf (downstream 2-4Kbp from CpG islands). Two-tailed hypergeometric tests were used to evaluate if the identified cis-, lcis- and trans-mQTL SNPs and CpGs showed enrichment for these genomic functional features. *P*-values were adjusted by Bonferroni method.

We also collected functional genomic regions from the Functional Annotation of ANimal Genomes (FAANG) Project for neutrophil cells[7], totally yielding 598,398 functional regions of CTCF binding sites, enhancers, open chromatin regions, promoters, promoter flanking regions, and TF binding sites. We examined the functional classes to which the SNPs and CpGs of mQTL pairs belonged and counted the number of annotated mQTL pairs. One-tailed hypergeometric tests were used to evaluate if the identified cis-, lcis- and trans-mQTL SNPs and CpGs showed enrichment for these genomic functional classes. *P*-values were adjusted by Bonferroni method. Then, we generated random SNP-CpG pairs at the same sample size as mQTLs and annotated those random pairs to functional regions. For each combination of two functional classes, the fold change is calculated as the ratio between the number of annotated mQTL pairs and that of annotated random SNP-CpG pairs.

Enrichment of mQTLs and mCpGs in GWAS and EWAS signals

We downloaded the curated significant GWAS results from GWAS Catalog[8] (*P*-value < 1×10^{-5}, date: 2020.12.25) including 227,262 records and 137,450 sites, and significant EWAS results from EWAS Atlas[9] (*P*-value < 1×10^{-4} or *Q*-value < 0.05, date: 2020.12.25) containing 416,331 records and 226,959 sites. We applied permutation test to examine if mQTL SNPs were enriched for the known GWAS signals. First, we randomly sampled sites from NSPT tested SNPs to keep the same number and MAF distribution (bin = 0.05) as the mQTL SNPs, which was replicated 10^4 times. Then based on the 10^4 sampled sets, we calculated the coverage ratio of GWAS signals for each set and got the null distribution. We compared the actual ratio with the null distribution and got the permutation *P*-value. A similar procedure was used to evaluate the enrichment of EWAS signals on mCpGs.

We also tested the enrichment of mQTLs and mCpGs in different Experimental Factor Ontology (EFO) parent categories using one-tailed hypergeometric test with NSPT tested SNPs and CpGs as background.

Enrichment of mQTLs and mCpGs in Gene Ontology and KEGG pathway

The Gene Ontology (GO) and KEGG pathway enrichment of genes annotated (using ANNOVAR v2020-06-07[4] as mentioned before) by clumped (limiting LD r^2<0.2) mQTLs were conducted by R package clusterProfiler v4.8.1[10] (one-tailed hypergeometric test) with genes annotated by all NSPT tested SNPs as background. The Gene Ontology (GO) and KEGG pathway enrichment of genes annotated by mCpGs with genes annotated by all NSPT tested CpGs as background were conducted by R package missMethyl v1.24.0[11] (one-tailed Wallenius' noncentral hypergeometric test) which could deal with two different bias: 1) the different number of probes per gene on the array, and 2) CpGs that annotated to multiple genes.

Enrichment of cis-mQTLs in cis-eQTLs

We downloaded the significant cis-eQTL data in whole blood (N = 670) from GTEx V8[ref[12]], then we harmonized the alleles (switch and flip if necessary) to match SNPs between the cis-eQTLs and our tested SNPs. We performed one-tailed hypergeometric test to evaluate the enrichment of cis-mQTL SNPs for cis-eQTL SNPs.

Functional analysis for the overlaps between mQTLs and eQTLs

We collected 3804 house-keeping (HK) genes from Eisenberg et al (2013)[13] and checked the proportion of those HK genes in the overlapped SNPs between mQTLs and eQTLs, denoted as A. Then, we calculated the proportion of HK genes in all genes derived from Ensembl release v104[ref[14]], denoted as B. The fold change is then defined as the ratio of A and B. In addition, we downloaded GO terms, KEGG pathways, and Reactome pathways from the Molecular Signatures Database[15]. We checked the enrichment of the overlapped mQTLs and eQTLs in the functional terms/pathways using one-tailed hypergeometric tests.

Relation between mQTL effect sizes and biological implication as well as reproducibility

We attempted to study whether the effect sizes of mQTLs were associated with stronger biological implication. The absolute values of mQTL effect sizes (calculated based on methylation M values) were divided into gradually increasing groups with an interval of 0.5 (only kept the mQTL associations with absolute effect sizes <=4.5, accounting for ~99.998% of all significant mQTL associations in NSPT). We then calculated 4 different terms in each group: 1. the proportion of mQTL SNPs which were cis-eQTLs (significant cis-eQTL data in whole blood (N = 670) from GTEx V8[ref[12]] as mentioned before); 2. the proportion of mQTL SNPs which were known GWAS signals (data from GWAS Catalog[8], *P*-value < 1×10^{-5}, date: 2020.12.25, as mentioned before); 3. the proportion of targeted CpGs which were known EWAS signals (EWAS Atlas[9], *P*-value < 1×10^{-4} or

Q-value < 0.05, date: 2020.12.25, as mentioned before); 4. the proportion of mQTLs which were replicated (FDR < 0.05 and consistent effect direction) in CAS (n=1,060, as mentioned before).

EA-specific mQTLs

EA-specific mQTLs overlapping with signals reported in GWAS Catalog

Significant genetic associations for known disease/trait-related sites were downloaded from GWAS Catalog[8] (*P*-value < 1×10^{-5}, date: 2020.12.25, as mentioned before). A total of 118,705 associations reported in GWAS Catalog (*P*-value < 5×10^{-8}), referring to 82,392 signals and 3,200 traits, were included in this analysis. We calculated the overlap of EA-specific mQTLs with genetic associations, signals and traits in GWAS Catalog. Parallelly, we also calculated the overlap of shared mQTLs in East Asian (NSPT) and European (GoDMC) genetic associations, signals and traits in GWAS Catalog. The proportion of the two overlapping sets was calculated to demonstrate the contribution of EA-specific mQTLs to genetic studies in GWAS Catalog.

EA-specific mQTLs overlapping with GWAS signals in 107 shared traits in BBJ and UKBB

Firstly, we selected the shared traits reported in the BioBank Japan Project (BBJ) and the UK Biobank (UKBB). In total, GWAS summary statistics (*P*-value < 5×10^{-8}) from 107 shared traits from BBJ and UKBB were included in this analysis, that is 357,783 associations and 187,454 signals in BBJ, and 1,550,927 associations and 755,229 signals in UKBB. Secondly, similarly to GWAS Catalog, we calculated the proportion of EA-specific and shared mQTLs that overlapped with genetic associations, signals and traits in BBJ and UKBB respectively. Thirdly, we split genetic associations and signals in BBJ and UKBB into three groups as BBJ-specific, UKBB-specific and shared, which indicates the genetic associations or signals were detected in BBJ or UKBB only, or both of them. Then we calculated the proportion of EA-specific mQTLs overlapped with associations and signals in three groups. Additionally, the distributions of minor allele frequencies of signals in three groups were demonstrated based on NSPT and European data from the 1000 Genomes Phase 3. Fourth, these analyses were repeated in subsets with shared genetic variants presented in BBJ and UKBB, that is 6,566,666 shared SNPs, referring to 315,944 associations and 168,085 signals in BBJ, and 1,109,373 associations and 521,373 signals in UKBB.

EA-specific mQTLs help prioritization of EA-specific variant for body height

ELF1 variants have been shown to have a large effect on adult height in East Asian populations[16,17] with rs7335629 being an EA-specific signal in the latest human stature study[17]. This SNP and one of its associated CpG (cg21067652) are predicted to be in co-opening regions with binding sites from the same TFs. Two-sample Mendelian randomization was used to investigate the causality of the colocalized CpGs on ELF1 expression and body height. CpGs colocalized with chr13q14.11 were taking as exposures and body height (in BBJ) as outcome. Instrumental SNPs were selected from the whole-genome mQTLs (including cis-, lcis- and trans-) (*P*-value $< 1\times10^{-10}$) for each CpG. The analysis was performed by using R package TwoSampleMR v0.4.26[18-20]. Specifically, we kept the results of matched SNPs (harmonized alleles where needed) between the NSPT mQTL results and the body height GWAS results. We then clumped the mQTL results to remove the SNPs that were in high LD with others and obtained the independent instrumental SNPs. LD was calculated using EAS data from the 1000 Genomes Phase 3. Clumping was performed among SNPs with LD r^2 above the specified threshold 0.01 within 10Kbp, and only the SNP with the lowest *P*-value was retained. We then performed sensitivity analyses to test for heterogeneity and horizontal pleiotropy. We also used R package MRPRESSO to exclude outliers. Finally, the results of the MR egger method were used to interpret the causality from exposures to outcomes.

Comparison of trans-mQTL colocalization signals with cis-mQTL colocalization signals

We grouped cis-mQTLs and trans-mQTLs into colocalized and non-colocalized mQTLs. Enrichment analysis was then performed using two-tailed Fisher's Exact test. We grouped trans-mQTL and cis-mQTL colocalization signals into EA-specific and EAS-EUR common colocalization signals, and then used two-tailed Fisher's Exact test for enrichment analysis.

Enrichment of functional elements and chromatin states. We collected segmented functional regions of GM12878 cell line in the blood tissue from UCSC, generated by the ENCODE analysis working group[21]. This provided us with a total of 1,573,252 annotated regions, classified into CTCF binding sites, enhancers, open chromatin regions, promoters, promoter flanking regions, and TF binding sites[22]. We checked the enrichment of cis- and trans- EA-specific mQTLs that colocalized with GWAS in each type of functional regions using hypergeometric tests with all mQTLs as background, and compared their enrichment performance. Then, we further separated the colocalized mQTLs into EA-specific ones and shared ones, and compared the functional enrichment of EA-specific colocalized mQTLs with those shared with European. To dig deeper into the mQTL enrichment in different chromatin states and draw a more robust conclusion, we also downloaded the annotations of chromatin states that trained by chromHMM[23] on data from high quality epigenomes in blood tissue, published by Roadmap Epigenomics[24]. There are 15 different chromatin states involved in the annotation, including Quies (Quiescent/Low), ReprPCWk (Weak Repressed), BivFlnk (Flanking Bivalent), EnhBiv (Bivalent Enhancer), ReprPC (Repressed PolyComb), Het (Heterochromatin), TssA (Active TSS), TssAFlnk (Flanking Active TSS), TxWk (Weak transcription), Enh (Enhancers), Tx (Strong transcription), EnhG (Genic enhancers), TxFlnk (Transcript at gene 5´ and 3´), TssBiv (Bivalent/Poised TSS), and ZNF/Rpts (ZNF genes & repeats). Similarly, we compared the enrichment of cis- and trans- EA-specific colocalized mQTLs, as well as EA-specific colocalized mQTLs versus shared colocalized mQTLs, in different chromatin states. All enrichment significance and fold changes are calculated using the two-tailed Fisher's Exact tests.

Trans-colocalization signal at chr21q22.2

Enrichment of the 233 CpGs for TF motifs. To assess whether the 233 CpGs were enriched for any TF motifs, we first downloaded human TF motif data (949 motifs in total) from JASPAR 2022 database[25]. We then tested whether the flanking regions (±100bp) of the 233 CpGs enriched for any TF motifs compared with the background CpGs (all 378,342 matched CpGs between NSPT and GoDMC) by R package PWMEnrich v4.30.0[26] which using a lognormal threshold-free approach (comparing the average affinity of the interested sequences to the average affinity of length-matched sequences from the background). We got significant enrichment results at a Bonferroni-corrected threshold, i.e., P-value < $0.05/949=5.3\times10^{-5}$.

Enrichment of the 233 CpGs in the blood cell ChIP-seq binding signals of TFs. We used the online tool 'Enrichment Analysis' on the ChIP-Atlas website[27,28] (https://chip-atlas.org) to evaluate whether the 233 CpGs (flanking regions, ±100bp) enriched in TF ChIP-seq binding signals (peak-caller MACS2 Q-value<1×10^{-5}) compared with the background CpGs (all 378,342 matched CpGs between NSPT and GoDMC) in blood cells. The P-values were calculated using the two-tailed Fisher's Exact test.

Two-sample Mendelian Randomization (MR) test. We performed two-sample MR analysis with taking the 233 CpGs as exposures and the seven traits (basophil count, eosinophil count, monocyte count, white blood cell count, urticaria, pericarditis and asthma, GWASs from BBJ) as outcomes. Instrumental SNPs were selected from the whole-genome mQTLs (including cis-, lcis- and trans-) (P-value < 1×10^{-10}) for each CpG. The analysis was performed by using R package TwoSampleMR v0.5.6 and with reference to previous studies[18-20]. For each trait, we kept the results of matched SNPs (harmonized alleles if needed) between the NSPT mQTL results and the GWAS results. Then, we clumped the mQTL results to remove the SNPs that were in high LD with others in NSPT and obtained the independent instrumental SNPs. Clumping was performed among the SNPs that had LD

r^2>0.2 within 1000Kbp, and only the SNP with the lowest *P*-value was retained. We used the IVW method to assess causality. For the results showing heterogeneity (*P*-value < 0.05), we re-evaluated the causality using the multiplicative random effects-inverse variance weighted method. The MR *P*-values were corrected by the Benjamini-Hochberg method to control the false discovery rate for each trait. We also performed the reverse MR analyses (the seven traits as exposures and the 233 CpGs as outcomes) using the similar strategy. For CpGs with only two instruments, we obtained significant MR results by 2 criteria: 1) MR FDR < 0.05; 2) reverse MR was not significant (*P*-value > 0.05). For CpGs with more than two instruments, we obtained significant MR results by 4 criteria: 1) MR FDR < 0.05; 2) horizontal pleiotropy *P*-value > 0.05; 3) leave-one-out analysis *P*-value < 0.05; 4) reverse MR was not significant (*P*-value > 0.05).

PheWAS. We used a publicly available database (https://gwas.mrcieu.ac.uk/phewas) to investigate the association of rs80109907 with different traits. The database contains GWAS results from Japanese, UK and Finnish biobanks, the GWAS catalog and GWAS results from other studies.

GO enrichment. We performed GO pathway enrichment analysis on ERG, the 62 TFs that significantly enriched in the 233 CpG regions colocalized at chr21q22.2, and 195 genes annotated in the regions around the 233 CpGs. Enrichment analyses were performed using Metascape v3.5[29].

PPI network. We performed a protein-protein interaction network analysis using STRING V11.5[30] for the ERG and 62 TFs significantly enriched in the 233 CpG regions. We also performed protein-protein interaction network analysis on the ERG, the 62 TFs significantly enriched in the 233 CpG regions, and the 195 genes annotated in the regions around the 233 CpGs. The PPI network was then downloaded and visualized using Cytoscape v3.9[31].

Functional enrichment of mQTLs

Enrichment of mQTLs in OpenCausal

We downloaded the list of opening-causal SNPs in blood tissue from Li et al (2020)[32], where the SNPs are predicted to be sensitive to the chromatin accessibility of the regulatory elements. We considered the total number of SNPs as N, the number of opening-causal SNPs as M, the number of mQTLs as R, and the overlap between mQTLs and those opening-causal SNPs as *k*. Then, the enrichment significance *P*-value is then calculated by two-tailed Fisher's Exact test.

Enrichment of mQTLs in 3D chromatin contacts

PCHi-C data of primary blood cells including total B cells (tB), naive B cells (nB), monocytes (Mon), neutrophil (Neut), total CD4$^+$ T cells (tCD4), naive CD4$^+$ T cells (nCD4), total CD8$^+$ T cells (tCD8), naive CD8$^+$ T cells (nCD8) are collected from Javierre et al (2016)[33], where PCHi-C interactions were filtered with a threshold of CHiCAGO scores > 5[ref[34]]. The loops and TADs were directly downloaded from the website given by this study (https://osf.io/u8tzp). The H3K27ac HiChIP data for naive T cells were obtained from Mumbach et al (2017)[35]. The JuiceBox output was processed by HiCCUPS implemented in the Juicer Tools (version 0.7.5) with default parameter settings to obtain the HiChIP loops[36].

We counted a mQTL pair as validated if its SNP and CpG can respectively locate within the two anchors of a 3D interacting loop derived from PCHi-C or HiChIP data. We built two groups of randomly selected SNP-CpG pairs as controls to evaluate the enrichment of mQTLs in chromatin contacts. To take the MAF of SNP and the variation of CpG methylation level into consideration, we only involve SNPs and CpGs that appears in mQTLs to generate control groups. The first control group was generated by randomly sampling SNP-CpG pairs from all SNP-CpG combinations, regarded as genomic background. The other control group was sampled SNP-CpG pairs with the same distance distribution as mQTL pairs, named distance-matched SNP-CpG group. We compared the number of validated mQTL pairs with that of genomic background and distance-matched SNP-CpG pairs. The enrichment significance and fold changes are calculated with the one-tailed hypergeometric test.

Comparison of distance distributions of mQTLs and that of 3D loops

We checked the distance distribution of mQTL pairs and that of 3D chromatin loops and found that our mQTLs tend to capture interactions whose SNP and CpG are quite close to each other: 39.26% of mQTL interactions occur within sub-50Kbp distances, typically a region of low sensitivity for 3C-based techniques (2.9% Hi-C loops, 13.2% PCHi-C loops, 2.7% HiChIP loops). This suggests that many functional interactions may be below the resolution of conventional 3C-based techniques, which also partly explained why a proportion of mQTLs are not colocalized with chromatin loops.

Calculation of co-opening and co-activity of mQTL pairs

The peaks of DNase-seq data for B cells and T cells and the H3K27ac ChIP-seq data of neutrophil cells were all downloaded from ENCODE[37]. We derived the chromatin opening degree of SNPs and CpGs by mapping them to the DNase-seq peaks and assigned the opening score of the overlapped peak to SNPs or CpGs. Similarly, the chromatin activity of SNPs and CpGs is derived from ChIP-seq data. We define a SNP-CpG pair as co-opening/co-active if both the SNP and the CpG are overlapped with the peaks. Then, we generated two control groups by sampling SNP-CpG pairs from genomic background or with the same distance distribution as mQTL pairs. We compared the number of co-opening/co-active mQTL pairs with that of genomic background and distance-matched SNP-CpG pairs. The enrichment significance and fold changes are calculated with the one-tailed hypergeometric test.

Evaluating the sensitivity of mQTLs to chromatin accessibility

We quantified the influence of mQTL SNPs to the regulatory elements where they locate, using a tool named OpenCausal[32]. OpenCausal first predicts the chromatin accessibility score of a given region using the TF expression and genomic sequence information as input, where the sequence information can be derived from the reference genome, or alternatively, it can be formed based on the SNP information. Then, OpenCausal uses the change of chromatin accessibility scores before and after SNP mutation to measure the influence of a variant on the regulatory element. In this study, we collected the processed fragments per kilobase million (FPKMs) of RNA-seq data for blood tissue from GTEx project. Using the TF expression and mQTL information as input, OpenCausal calculated an opening-causal score for each mQTL. We consider the mQTLs with non-zero scores as the ones that are sensitive to the chromatin accessibility.

A TF is more likely to be near a trans-mQTL

The list of human transcription factors (TFs) was downloaded from Lambert et al (2018)[38], which included 1,639 manually examined TF genes. We checked if it's more likely to find a TF near (<1Mbp) a trans-mQTL than a random SNP from the NSPT tested SNPs using two-tailed Fisher's Exact test.

Motif and ChIP-seq enrichment of trans-mCpGs

To assess if trans-mQTLs can be explained by differential transcription factor binding, we first filtered trans-mQTLs where the SNP is associated in-trans with 100 or more CpGs. This was done to ensure adequate power to detect an enrichment of TF-binding motifs or ChIP-seq peaks. Next, SNPs driving such trans-mQTLs were filtered further by the requirement that the SNP be a GWAS hit for a disease/trait, as recorded by the GWAS Catalog database[8] and PhenoScanner[39,40]. We then filtered SNPs further by the requirement that a transcription factor (TF) be in-cis (<1Mbp) with the SNP. The TF list was downloaded from The Human Transcription Factors database[38] which contains 1,639 human TFs (and their motifs). Next, we examined if the trans-mQTL CpGs associated with a given SNP were enriched for either the corresponding TF-binding motif or TF ChIP-seq signal. ChIP-seq data were downloaded from the ChIP-seq atlas[28]. For each relevant TF, the ChIP-seq atlas provides the binding intensity score to potential target genes defined at 3 different distances from the TSS (±1Kbp, ±5Kbp and ±10Kbp). Trans-mQTL CpGs were assigned corresponding binding intensity scores if they co-localized to within ±1Kbp, ±5Kbp and ±10Kbp of the target gene's TSS. To obtain a *P*-value of enrichment we performed a one-tailed Wilcoxon rank sum test comparing these binding intensity values to those of 5,000 randomly selected CpGs. Separate to this, we also performed enrichment for TF-binding motifs using Position Weight Matrices (PWMs) from MotifDB using R package PWMEnrich. In more detail, we scanned 200bp sequences (±100bp from the trans-mQTL CpGs) and statistical enrichment was assessed against background sequences randomly selected from a pre-compiled threshold-free lognormal background sequence distribution of human promoters. From the TF ChIP-seq and binding-motif analysis we thus obtained a union list of trans-mQTLs with the GWAS trait SNP linked in-cis to a TF and with the corresponding trans-mQTL CpGs enriched for TF-binding sites (as determined either by ChIP-seq or motif-enrichment). For traits from the GWAS Catalog, we obtained a total of 2,182 unique SNP-Trait-TF validated trans-mQTL networks. For traits from Phenoscanner, we got 125,133 unique SNP-Trait-TF validated trans-mQTL networks.

Cell-lineage specific mQTL mapping using CellDMC

Validation of cell-lineage specific mQTL in BLUEPRINT data

We validated the cell-lineage specific mQTLs inferred in previous step against the BLUEPRINT data[41], which contains matched genotype and DNAm profiles for purified neutrophils, monocytes and CD4$^+$ T cells from approximately 200 samples. Specifically, we first downloaded the BLUEPRINT mQTL summary statistics for the three primary blood cell subtypes (CD4$^+$ T cells, monocytes and neutrophils), which we note only includes the summary statistics of significant mQTLs defined at a threshold of $6.3\lozenge10^{-5}$ and only for the corresponding index SNPs.

For the validation analysis, we focused on the subset of mQTLs derived at 2 cell-lineage resolution and which displayed clear specificity (lymphoid lineage and myeloid lineage), defined as those with

P-values less than 10^{-8} in one cell lineage but larger than 0.9 in the other cell lineage. With this stringent definition, there were 58,105 lymphocyte-specific mQTLs and 407,069 myeloid-specific mQTLs. Then, we matched these mQTLs with the index mQTLs in $CD4^+$ T cells, monocytes and neutrophils as reported in BLUEPRINT, obtaining an overlap of 22 index mQTLs for $CD4^+$ T cells, 221 for monocytes and 278 for neutrophils. Statistical significance of the validation was assessed at the level of (i) overlap number using a binomial test against a null distribution derived from the expected overlap as obtained from the opposite lineage, as well as (ii) in terms of the consistency of directionality, as determined from a linear regression of the corresponding signed $-\log_{10}(P\text{-values})$.

Enrichment of cell-lineage mQTLs in lineage-DMCs

We performed enrichment analysis of four types of mQTLs (shared, myeloid-specific, lymphoid-specific, none) among lymphoid and myeloid differentially methylated cytosines (DMCs). The DMCs were identified using the EPIC DNAm data from Salas et al (2022)[42], by comparing the DNAm values of myeloid and lymphoid cells with the empirical Bayes limma method. DMCs were called at a significance level of FDR < 0.05. For this analysis, mQTLs passing the stringent 10^{-14} threshold were used, and these were divided into the four types mentioned above. Shared mQTLs were called in both myeloid and lymphoid lineages. By “none” we mean mQTLs that were not significant in both myeloid and lymphoid compartments. Odds Ratios and *P*-values of enrichment were derived from a one-tailed Fisher’s Exact tests.

eFORGE analysis of myeloid and lymphoid mQTLs

To test whether the cell-lineage specific mQTLs capture the tissue specific regulatory component at CpG level, we conducted a functional enrichment analysis of the cell-lineage specific mQTLs derived at 2 cell-lineage resolution using eFORGE2.0[43,44]. Given a set of CpGs, eFORGE2.0 assesses if these CpGs map to cell-type-specific DHS regions, which are naturally enriched for functional regulatory elements such as enhancers. eFORGE2.0 makes use of a large database of DHS signals as derived from the Epigenome Roadmap, BLUEPRINT and ENCODE. eFORGE2.0 analysis was performed for the highly specific mQTLs defined earlier and due to eFORGE2 server restrictions, we used as input the top 1000 lymphocyte and myeloid specific mQTLs.

Identification of mQTL hotspots

Enrichment of OpenCausal variants in trans-mQTL hotspots

For each trans-mQTL hotspots, there were trans-mQTLs with significant OpenCausal scores. We examined if the trans-mQTLs in each trans-mQTL hotspot were enriched with significant OpenCausal scores compared with randomly selected genomic regions in the same length. For each trans-mQTL hotspot, we first generated 1,000 random numbers according to the chromosome length. Then we generate a region of 2Mbp with each random number as the center. Next, we calculated the proportion of SNPs with significant OpenCausal scores in these regions, and take their distribution as a null distribution. At last, we checked the posterior probability of the real value on this distribution.

Enrichment of TAD, HChIP in trans-mQTL hotspots

The signatures of trans-mQTL hotspots include trans-mQTLs located in TAD, HChIP promoter-promoter and promoter-other interaction regions. To test if these signatures were enriched in trans-mQTL hotspots, we compared these signatures with randomly selected SNPs in the whole genome. We applied one-tailed hypergeometric tests to evaluate the enrichment of these signatures in trans-mQTL hotspots.

Enrichment of trans-mCpGs associated with trans-mQTL hotspots in biological process

We attempted to analyze the biological effects of trans-mCpGs (n = 6,195) associated with the index mQTLs in the 16 mQTL hotspots. We conducted the biological process enrichment on them with all the trans-mCpGs (n = 26,415) as the background. The enrichment was conducted by R package missMethyl v1.24.0[11] as mentioned before. There were 58 significant terms with FDR < 0.05. We further simplified the 58 terms by clustering on the semantic similarity. This was performed by R package simplifyEnrichment v1.0.0[45] with the default parameters.

Downstream effects of trans-mQTLs on diseases/traits

Enrichment of genes annotated by trans-mCpGs (rs4666078) in diseases/traits. The trans-mCpGs associated with the index mQTL rs4666078 of the hotspot H2 annotated to 151 genes. We tested if these genes were enriched in human diseases/traits based on the DisGeNET knowledgebase[46] (date: 2021.6.9), i.e., one of the largest publicly available collections of genes and variants associated to human diseases. The enrichment analyses over the diseases/traits in DisGeNET were conducted by R package disgenet2r v0.99.2[46] with all human genes (according to the NCBI) as background. One-tailed Fisher's Exact tests were performed, and *P*-values were corrected by the Benjamini-Hochberg method to control FDR < 0.05.

Enrichment of trans-mCpGs (rs4666078) in tissue eosinophilia associated CpGs. We downloaded the summary statistics about differentially methylated positions (DMPs) associated with nasal tissue eosinophilia which was closely related to blood eosinophil count[47] (as there was no EWAS of blood eosinophil count to date) from a published study of Korean population (n = 147, including 57 eosinophilic nasal polyp tissues (case group), 72 noneosinophilic nasal polyp tissues (control group) and 18 normal nasal tissues)[48]. The DMPs were defined as: analysis of variance (ANOVA) FDR< 0.05 (the multiple testing problem of ANOVA analyses was adjusted by Benjamini-Hochberg method) and Tukey's honest significant difference *P*-value < 0.05. They identified DMPs with significant differences among all groups or only between the case group and control group. They only kept 24,114 DMPs with a mean methylation level (β value) difference between case and control at least 1%. Based on the genome-wide significant CpGs associated with tissue eosinophilia, we evaluated the enrichment of the 232 trans-mCpGs that were associated with rs4666078 compared with background (all NSPT tested CpGs) using one-tailed hypergeometric test.

We also analyzed if these trans-mCpGs were more significantly associated with tissue eosinophilia compared with background CpGs using one-tailed Kolmogorov-Smirnov test.

Two-sample MR test to identify potential causality between CpGs and eosinophil count. We downloaded the GWAS summary statistics of blood eosinophil count from a recent GWAS in East Asians (n = 86,890)[49]. We conducted Two-sample MR analysis with taking the 232 trans-mCpGs (associated with rs4666078) as exposures and blood eosinophil count as outcome. Instrument SNPs were selected from the whole-genome mQTLs (including cis-, lcis- and trans-) (*P*-value < 1×10^{-10}) for each trans-mCpG. The analysis was performed by using R package TwoSampleMR v0.5.6 and referring to previous studies[18-20] . In detail, we kept the results of matched SNPs (harmonized alleles if needed) between NSPT mQTL results and eosinophil count GWAS results. Then, we clumped the mQTL results to remove the SNPs which were in high LD with others and got the independent instrument SNPs. LD was calculated using EAS data from the 1000 Genomes Phase 3. Clumping amongst those SNPs that had LD r^2 above the specified threshold 0.2 within 1000Kbp, and only the SNP with the lowest P-value was retained. We excluded the mCpGs which had less than three instruments, and used the inverse variance weighted (IVW) method on the left mCpGs (n = 136) to perform MR. The MR *P*-values were corrected by Benjamini-Hochberg method to control false discovery rate. Next, we did sensitivity analyses to test heterogeneity and horizontal pleiotropy. We also did the reverse MR analyses (blood eosinophil count as exposure and CpGs as outcomes) using the similar strategy. We got significant MR results by 5 criteria: 1) MR FDR < 0.05; 2) heterogeneity *P*-value > 0.05; 3) horizontal pleiotropy *P*-value > 0.05; 4) leave-one-out analysis *P*-value < 0.05; 5) reverse MR was not significant (FDR > 0.05).

BMI EWAS. We obtained a list of 364 BMI-associated validated CpGs from 4 EWAS BMI studies[50-53]. To test if these high-confidence BMI-associated CpGs were correlating with BMI in our NSPT cohort, we performed linear regressions of DNAm against BMI with age, sex, batch, array position and array slide as covariates. We derived the signed $\log_{10}$(P-values) for a total of 302 BMI-associated CpGs with representation in our EPIC array, and compared them with the corresponding values reported in the BMI EWAS studies.

Two-sample MR analysis to identify potential causality between CpGs and obesity. To identify causality of *NFKB1* trans-mQTLs and BMI, we conducted Mendelian Randomization analysis using R package TwoSampleMR. We focused on a list containing 554 BMI-associated CpGs in *NFKB1* trans-mQTL list and 364 BMI-associated validated CpGs from 4 EWAS BMI studies as candidate CpGs in MR model. Firstly, to estimate the causal effects, we run causal test in Two-sample MR model with candidate CpGs as exposures and BMI as outcome. We filtered the NSPT mQTLs list with candidate CpGs at the threshold P-value < 1×10^{-8} as exposure input and took the BMI GWAS summary statistics of instrument SNPs as outcome input. SNPs with LD score larger than 0.2 within 1000Kbp were removed by clumping, retaining only SNP with the lowest P-value. We used the inverse variance weighted (IVW) method to obtain an estimate of the causal effect when performing MR analysis on the candidate CpGs. The MR P-values were adjusted by Benjamini-Hochberg method. We then did sensitivity analyses including testing heterogeneity and horizontal pleiotropy. Secondly, we conducted the inverse test with BMI as exposures and CpGs as outcome. Here, we used the GWAS summary statistics of 32 BMI associated SNPs from a large GWAS study[54] as exposure input and a list of mQTL statistics of candidate CpGs at the threshold P-value < 1×10^{-8} as outcome input. Similar strategy as casual test was conducted in the inverse consequential test including LD clumping, IVW MR analysis and sensitivity test. After excluding CpGs that did not pass the sensitivity test, we obtained 170 causal CpGs of BMI in *NFKB1* trans-mQTL list, 58 consequential CpGs and 3 causal CpGs in BMI EWAS CpG list.

BMI and SNP interaction on CpGs. To interpret the relation between BMI and *NFKB1*-mediated trans-mQTLs, we carried out interaction analysis of BMI and trans-mQTLs on trans-mCpGs via linear regression in R.

$$mCpG = \alpha + \beta_{BMI}BMI + \beta_{SNP}G + \beta_{interaction}G * BMI \qquad (2)$$

where, $\beta_{interaction}$ is the effect of interaction between mQTL genotype G and BMI whilst α, β_{SNP} and β_{BMI} are intercept, marginal effect of mQTL genotype G and BMI respectively.

Sensitivity analysis

Influence of quality control of DNAm by 3*s.d.. We applied a two-step analysis strategy for mQTL mapping, including the excluding of outlying methylation values (outside the range of mean ± 3s.d.) in the 2nd step. We compared the mQTLs that generated in step 1 and step 2. Most of the mQTLs in two steps showed associations in the same direction (mQTLs in the same direction > 99.996%).

Influence of adjusting or not of DNAm PCs on mQTLs. To evaluate the influence of adjusting DNAm PCs on mQTLs, we also carried out mQTL calculation based on regressions of DNA methylation M values on SNPs along with bisulfite slide number, batch, age, sex, predicted blood cell fractions, top ten genomic principal components (PC). We then compared these mQTLs with that were adjusted for 2 DNAm PCs, where we found that these two sets of mQTLs repeated each other perfectly. The association directions of the two sets of mQTLs were completely the same, except for slight differences in P-values. This indicates that adjusting two DNAm PCs or not would not change mQTLs discovered in our study.

Sensitivity analysis of impute.knn. We carried out a simulation to assess the impact of sample size on the performance of impute.knn. We randomly selected 100 CpGs from the CpG data in NSPT. And then we generated samples with proportion of missing values (20% and 40%) at varied sample sizes (4, 18, 700, 1400, 2100, 2800 and 3500). We then applied impute.knn to each of these data to impute missing values. We applied Pearson correlation and mean absolute deviation (MAD) to evaluate the performance of impute.knn under varied simulation scenario. As expected, this indicates that performance of impute.knn is much better in the large sample size setting.

Validation of cell-lineage mQTLs. We applied a flexible statistical method, named mashr[55], for estimating and testing the sharing effect in cell-lineage mQTLs. The sharing of effect is 0.98 for myeloid and whole blood, 0.81 for lymphoid and blood, and 0.79 for myeloid and lymphoid (**Supplementary Fig. 16a**), yielding results consistent with previous studies[41,56,57]. Similar results were obtained when the mQTLs were split into cis- and trans-ones (**Supplementary Fig. 16b&15c**). Looking at the *FOSL2* and *NFKB1* hotspots, the sharing of effect for lymphoid and myeloid are different as it is 0.70 for *NFKB1* hotspot and 0.41 for *FOSL2* (**Supplementary Fig. 16d&e**), suggesting that the *FOSL2* hotspot may tend to have a cell-specific regulatory pattern.

II. Supplementary References

1. Gao, X. *et al.* FastQTLmapping: an ultra-fast package for mQTL-like analysis. (bioRxiv, 2021).
2. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
3. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat Methods* **8**, 833-5 (2011).
4. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
5. Zhou, W., Laird, P.W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic acids research* **45**, e22 (2017).
6. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome biology* **16**, 22 (2015).
7. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-8 (2010).
8. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-d1012 (2019).
9. Li, M. *et al.* EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res* **47**, D983-d988 (2019).
10. Yu, G., Wang, L.G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* **16**, 284-7 (2012).
11. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics (Oxford, England)* **32**, 286-288 (2016).
12. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).
13. Eisenberg, E. & Levanon, E.Y. Human housekeeping genes, revisited. *Trends in Genetics* **29**, 569-574 (2013).
14. Howe, K.L. *et al.* Ensembl 2021. *Nucleic Acids Res* **49**, D884-d891 (2021).
15. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
16. Akiyama, M. *et al.* Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat Commun* **10**, 4393 (2019).
17. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704-712 (2022).
18. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**(2018).
19. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS genetics* **13**, e1007081 (2017).
20. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nature genetics* **52**, 1122-1131 (2020).
21. Consortium., E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
22. Hoffman, M.M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**, 827-41 (2013).
23. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods* **9**, 215-216 (2012).

24. Bernstein, B.E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nature biotechnology* **28**, 1045-1048 (2010).

25. Castro-Mondragon, J.A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **50**, D165-d173 (2022).

26. Stojnic, R. & Diez, D. PWMEnrich: PWM enrichment analysis. (R package version 4.30.0, 2021).

27. Zou, Z., Ohta, T., Miura, F. & Oki, S. ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Res* **50**, W175-82 (2022).

28. Oki, S. *et al.* ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO reports* **19**(2018).

29. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, 1523 (2019).

30. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**, D638-d646 (2023).

31. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).

32. Li, W., Duren, Z., Jiang, R. & Wong, W.H. A method for scoring the cell type-specific impacts of noncoding variants in personal genomes. *Proc Natl Acad Sci U S A* **117**, 21364-21372 (2020).

33. Javierre, B.M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19 (2016).

34. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol* **17**, 127 (2016).

35. Mumbach, M.R. *et al.* Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet* **49**, 1602-1612 (2017).

36. Durand, N.C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-8 (2016).

37. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

38. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665 (2018).

39. Kamat, M.A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**, 4851-4853 (2019).

40. Staley, J.R. *et al.* PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207-3209 (2016).

41. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).

42. Salas, L.A. *et al.* Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. *Nat Commun* **13**, 761 (2022).

43. Breeze, C.E. *et al.* eFORGE: A Tool for Identifying Cell Type-Specific Signal in Epigenomic Data. *Cell Rep* **17**, 2137-2150 (2016).

44. Breeze, C.E. *et al.* eFORGE v2.0: updated analysis of cell type-specific signal in epigenomic data. *Bioinformatics* **35**, 4767-4769 (2019).

45. Gu, Z. & Hübschmann, D. Simplify enrichment: A bioconductor package for clustering and visualizing functional enrichment results. *Genomics, Proteomics & Bioinformatics* (2022).

46. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research* **48**, D845-D855 (2020).

47. Wang, K. *et al.* Concordant systemic and local eosinophilia relates to poorer disease control in patients with nasal polyps. *The World Allergy Organization journal* **12**, 100052 (2019).

48. Kim, K.W. *et al.* Integrated genetic and epigenetic analyses uncover MSI2 association with allergic inflammation. *The Journal of allergy and clinical immunology* **147**, 1453-1463 (2021).

49. Chen, M.H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182**, 1198-1213.e14 (2020).

50. Dick, K.J. *et al.* DNA methylation and body-mass index: a genome-wide analysis. *Lancet* **383**, 1990-8 (2014).

51. Mendelson, M.M. *et al.* Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. *PLoS Med* **14**, e1002215 (2017).

52. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81-86 (2017).

53. Demerath, E.W. *et al.* Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Hum Mol Genet* **24**, 4464-79 (2015).

54. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* **42**, 937-48 (2010).

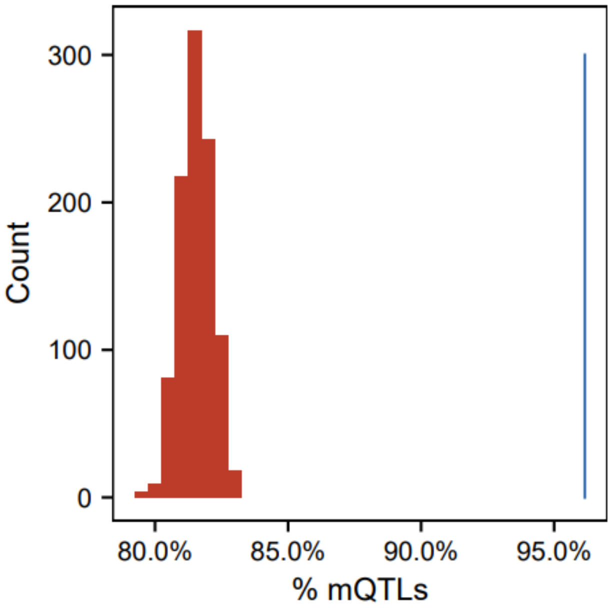
55. Urbut, S.M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet* **51**, 187-195 (2019).

56. Oliva, M. *et al.* DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat Genet* **55**, 112-122 (2023).

57. Hawe, J.S. *et al.* Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. *Nat Genet* **54**, 18-29 (2022).

58. Marcus, J.H. & Novembre, J. Visualizing the geography of genetic variants. *Bioinformatics* **33**, 594-595 (2017).

III. Supplementary Figures



Supplementary Fig.1 Enrichment of our mQTLs in blood-cell traits GWAS signals
Blood-cell traits GWAS signals were downloaded from the most recent and to-date the largest GWAS study (746,667 trans-ethnic individuals)[49]. We calculated the proportion of mQTLs (blue line in the plot) in the matched SNPs between the GWAS signals and NSPT tested SNPs. We randomly sampled the same number of SNPs for 1000 times and kept the distribution of allele frequencies in each sampled set the same as the matched SNPs. Then we calculated the proportion of mQTLs in each sampled set as the null distribution (red bars in the plot).

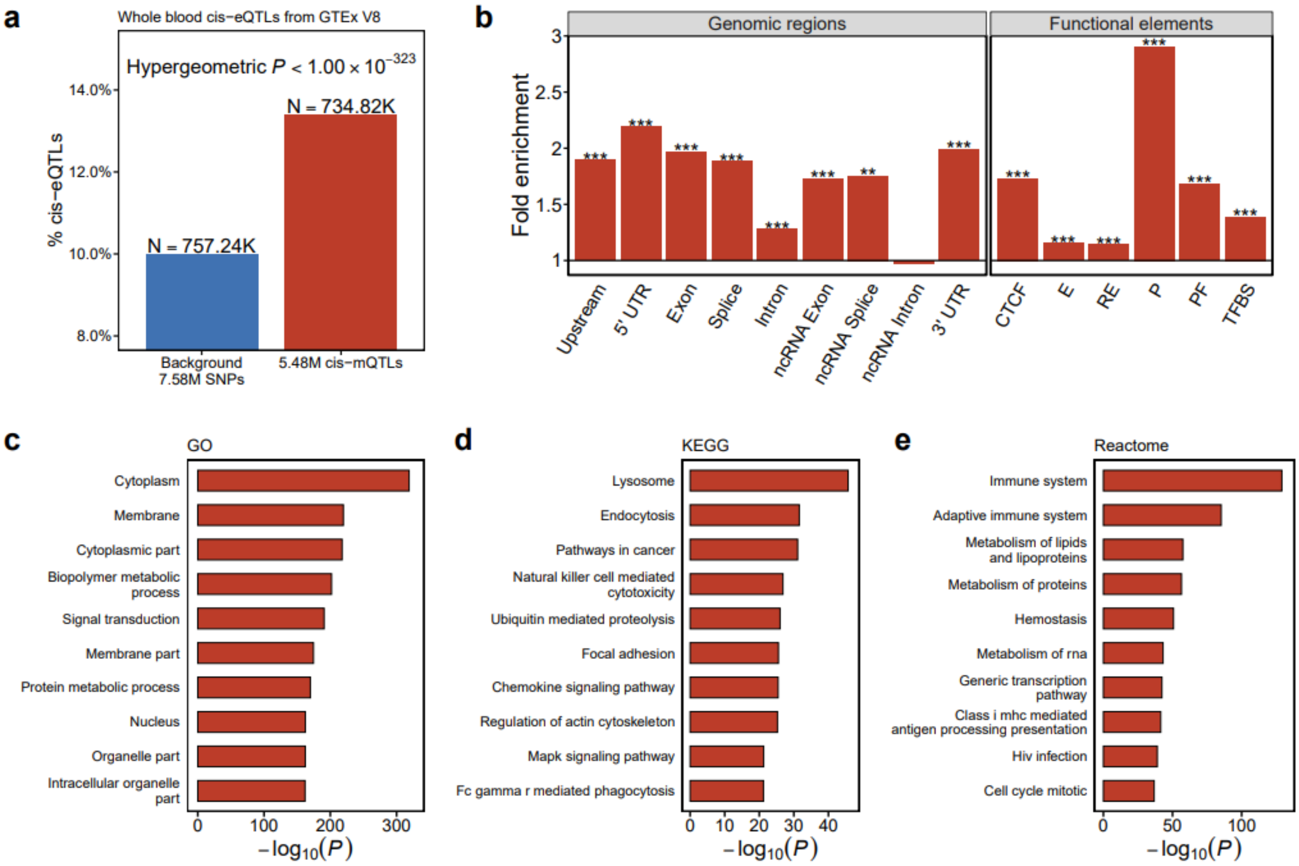


Fig. S2 cis-mQTLs are enriched for cis-eQTLs (whole blood, GTEx v8)

a, Bars show the proportion of cis-eQTL SNPs in background (blue) and cis-mQTLs (red). The *P*-value of the cis-mQTLs enrichment for cis-eQTLs (one-tailed hypergeometric test) is labeled on the top of the bars. **b,** Enrichment of cis-mQTLs that are also cis-eQTLs in different genomic regions (left) and functional elements (right). The y-axis indicates the fold enrichment compared with all cis-mQTLs. Significance of one-tailed hypergeometric test is denoted by different symbols on each bar, i.e., *, the *P* values adjusted by Bonferroni method < 0.05; **, < 0.01; ***, < 0.001. **c-e,** Enrichment of the overlaps between cis-mQTLs and cis-eQTLs in GO terms (**c**), KEGG pathways (**d**), and Reactome pathways (**e**). The x-axis indicates the -$\log_{10}(P)$ from one-tailed hypergeometric test.

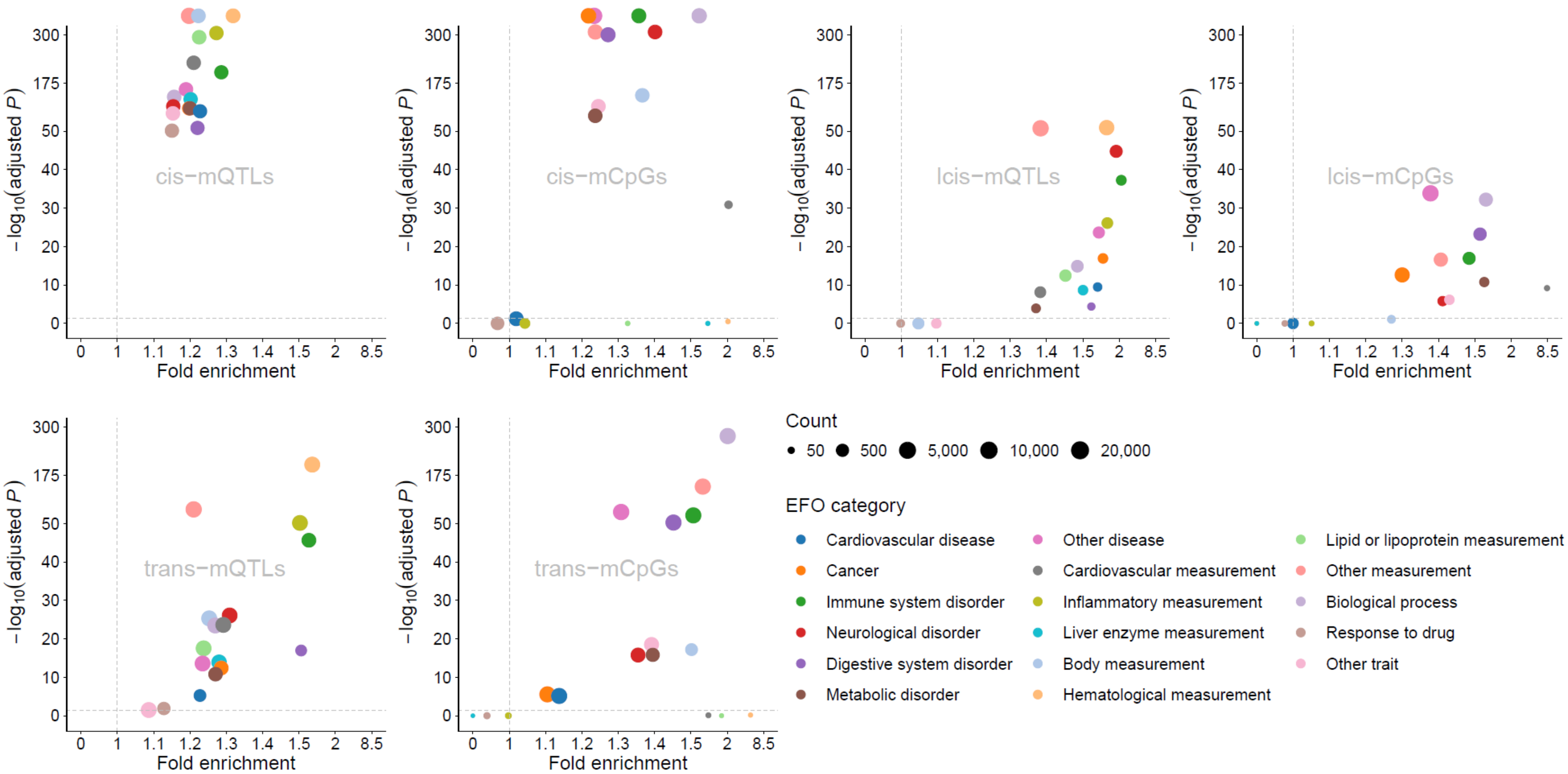


Fig. S3 Enrichment of mQTLs and mCpGs in different EFO parent categories

These scatter plots showing the enrichment of mQTLs (left plots) and mCpGs (right plots) in different Experimental Factor Ontology (EFO) parent categories using one-tailed hypergeometric test based on data from GWAS Catalog[8] and EWAS Atlas[9]. The *P*-values of enrichment are corrected by Bonferroni method.

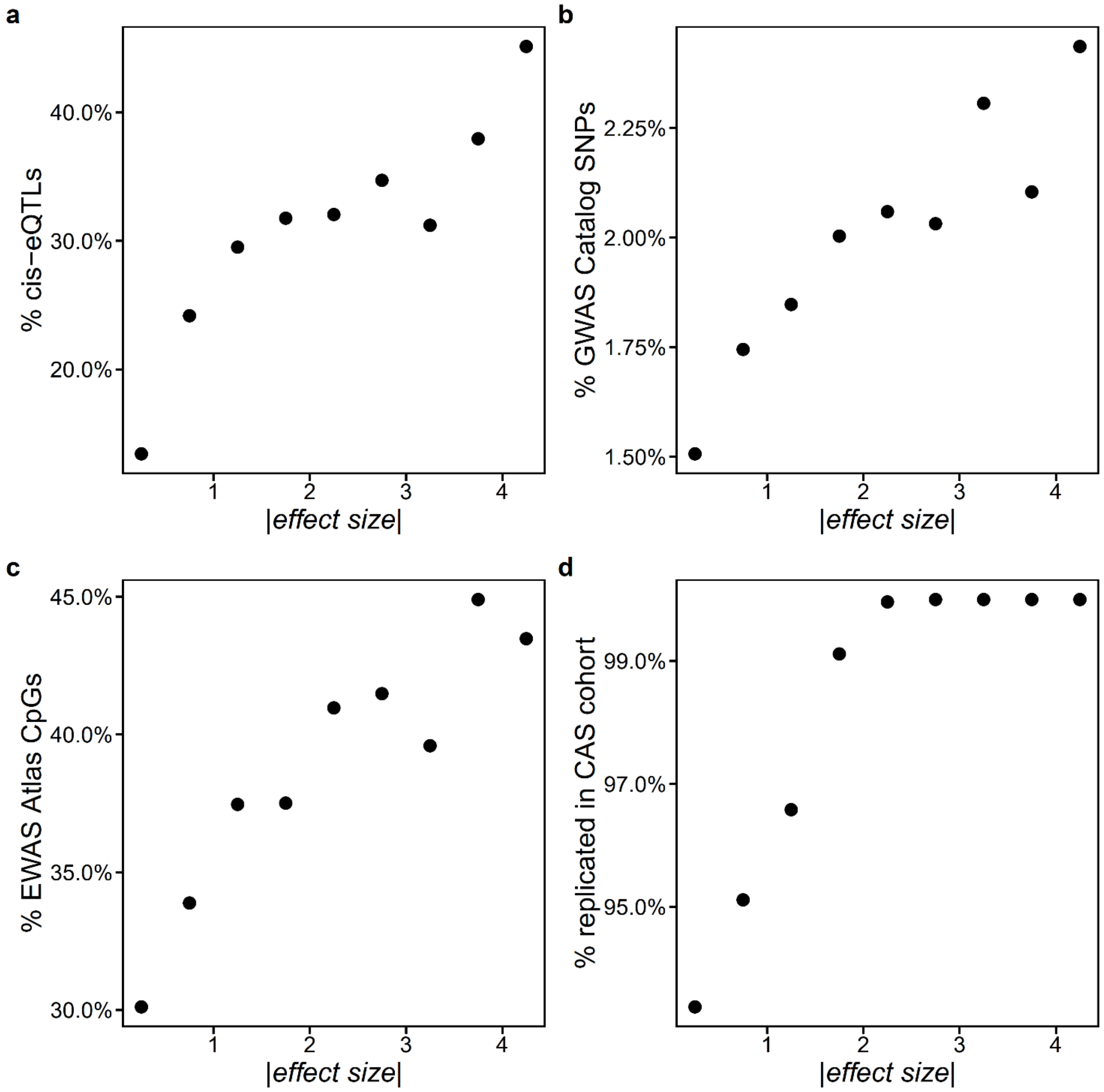


Fig. S4 The relation between effect sizes of mQTLs and their biological implications and reproducibility

a-d, Four plots showing the relationship between the absolute effect sizes (calculated based on methylation M values, only showing the mQTL associations with absolute effect sizes <=4.5 which covering ~99.998% of all significant mQTL associations in NSPT) and the proportion of mQTLs that are also cis-eQTLs **(a);** the proportion of mQTLs that are also GWAS signals (**b**); the proportion of mCpGs that are EWAS signals (**c**); the proportion of associations that are replicated in CAS (FDR < 0.05 and consistent effect direction) (**d**).

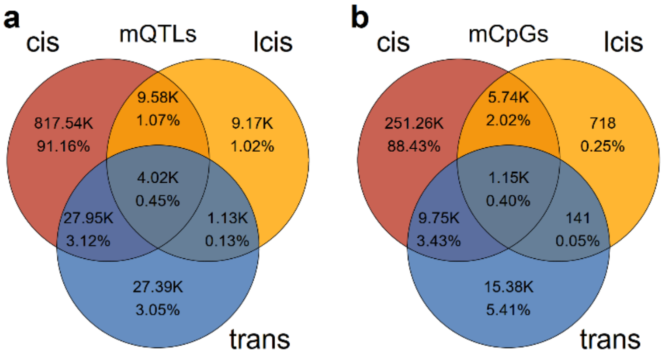


Fig. S5 Venn diagrams showing overlap of cis-, lcis- and trans-mQTLs and mCpGs

a, Venn diagrams showing overlap of clumped cis-, lcis- and trans-mQTLs; **b,** Venn diagrams showing overlap of cis-, lcis- and trans-mCpGs.

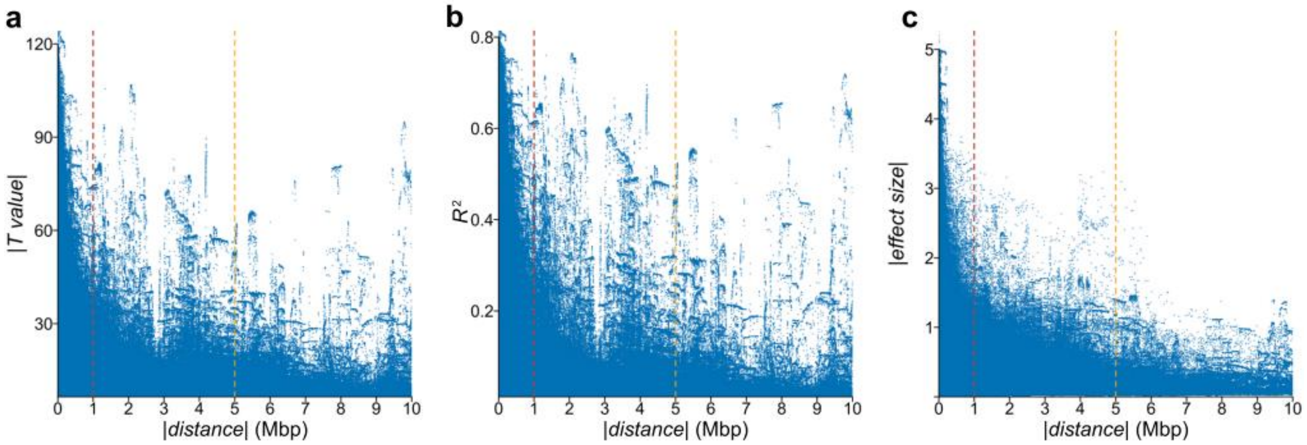


Fig. S6 mQTL effects along with distances between SNPs and CpGs

a-**c**, Three scatter plots showing the relationship between the mQTL distance (distance between a SNP and a CpG, within a local 10Mbp region, x-axis) and the mQTL significance (absolute *T* value) **(a);** the proportion of mCpG methylation variance explained by a mQTL (R^2) **(b);** the absolute mQTL effect size **(c)**. Only showing significant mQTL associations at *P*-value < 1×10^{-10} (calculated based on methylation M values).

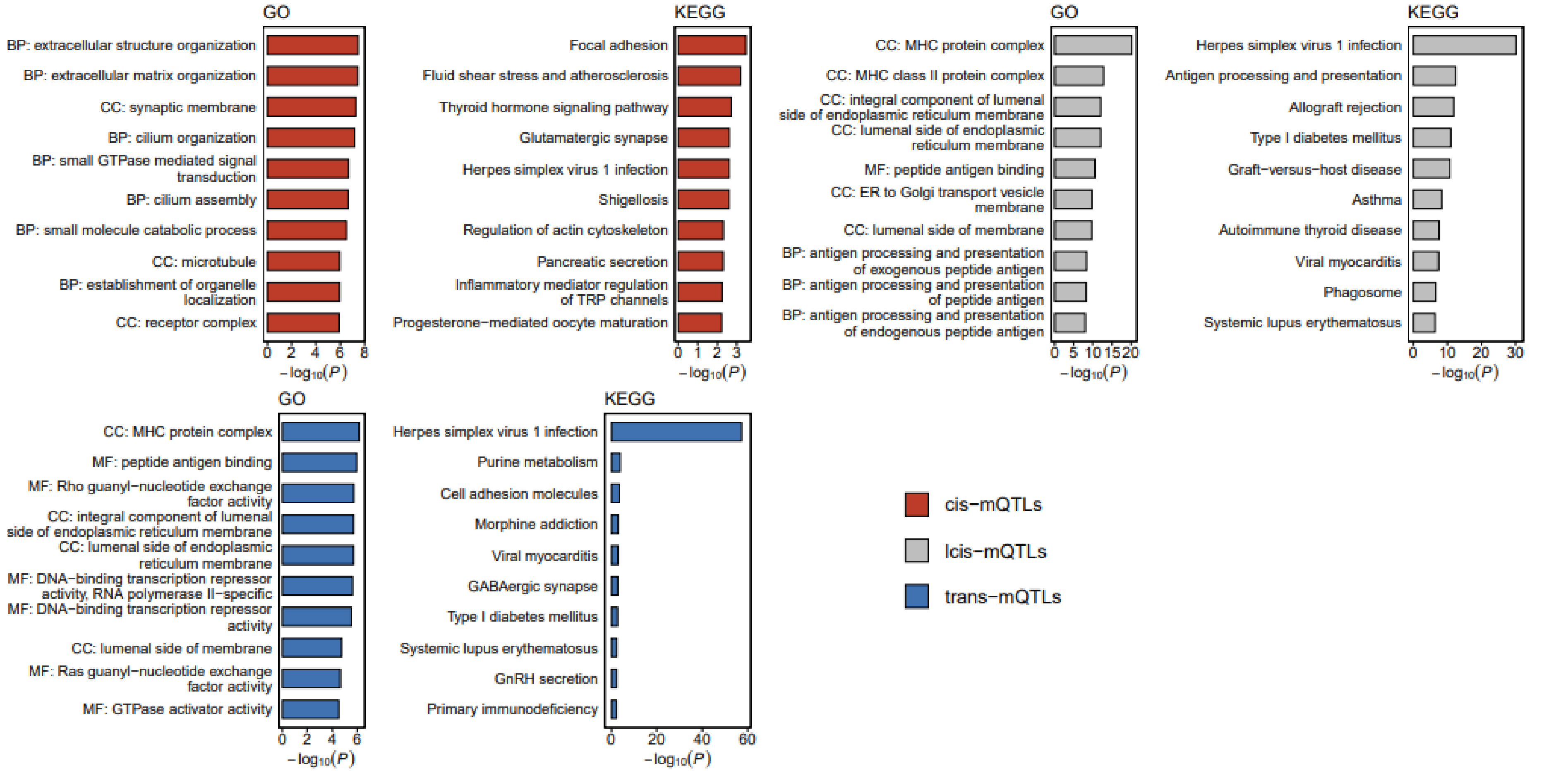


Fig. S7 Enrichment of clumped mQTL-annotated genes in GO and KEGG pathway

These barplots showing the enrichment of genes annotated by clumped mQTLs in Gene Ontology (GO, left plots) and KEGG pathway (right plots) (red, cis-; grey, lcis-; and blue, trans-). Only the top 10 significant results are displayed. The x-axis indicates the $-\log_{10}(P)$ from one-tailed hypergeometric test (**Supplementary Protocols**).

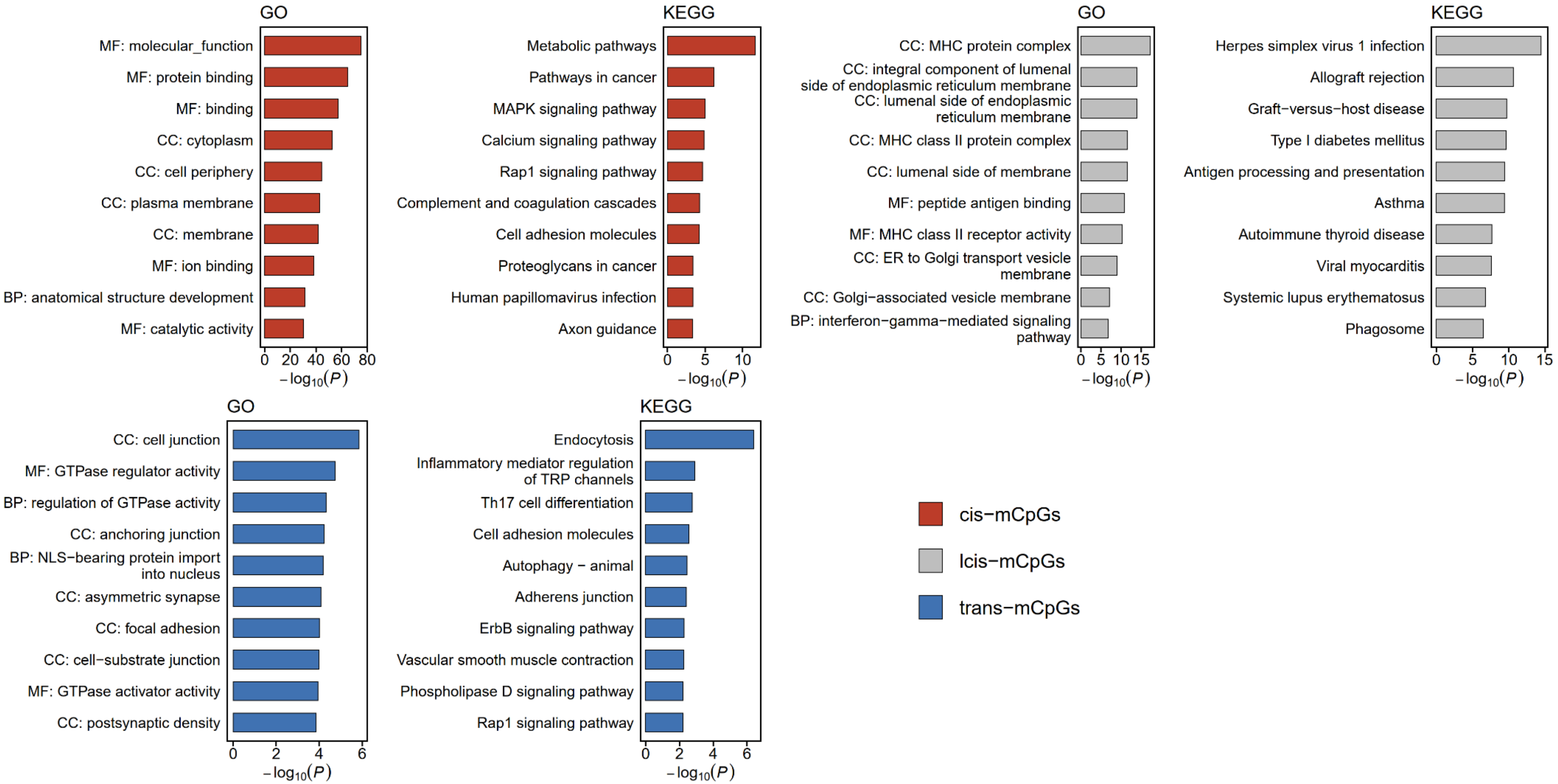


Fig. S8 Enrichment of mCpG-annotated genes in GO and KEGG pathway

These barplots showing the enrichment of genes annotated by mCpGs in Gene Ontology (GO, left plots) and KEGG pathway (right plots) (red, cis-; grey, lcis-; and blue, trans-). Only the top 10 significant results are displayed. The x-axis indicates the -$\log_{10}(P)$ from one-tailed Wallenius' noncentral hypergeometric test (**Supplementary Protocols**).

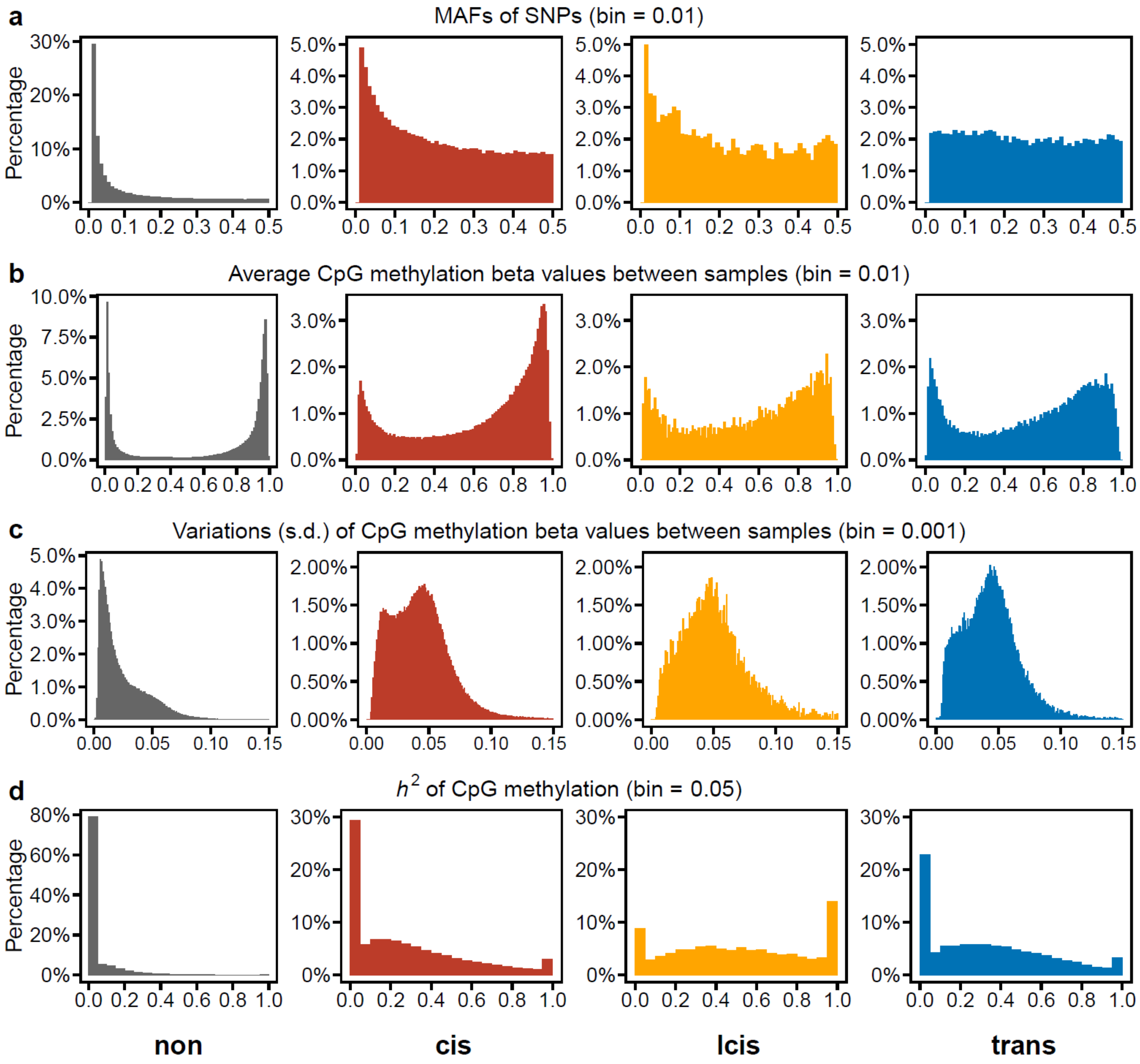


Fig. S9 Characteristics of mQTL SNPs and CpGs

a-d, these four histograms showing the differences between cis-, lcis- and trans-mQTL SNPs and CpGs. **a,** minor allele frequencies (MAFs) of SNPs; **b,** average CpG methylation beta values between samples; **c,** variations (standard deviation, s.d.) of CpG methylation beta values between samples; **d,** narrow-sense heritability of CpG methylation.

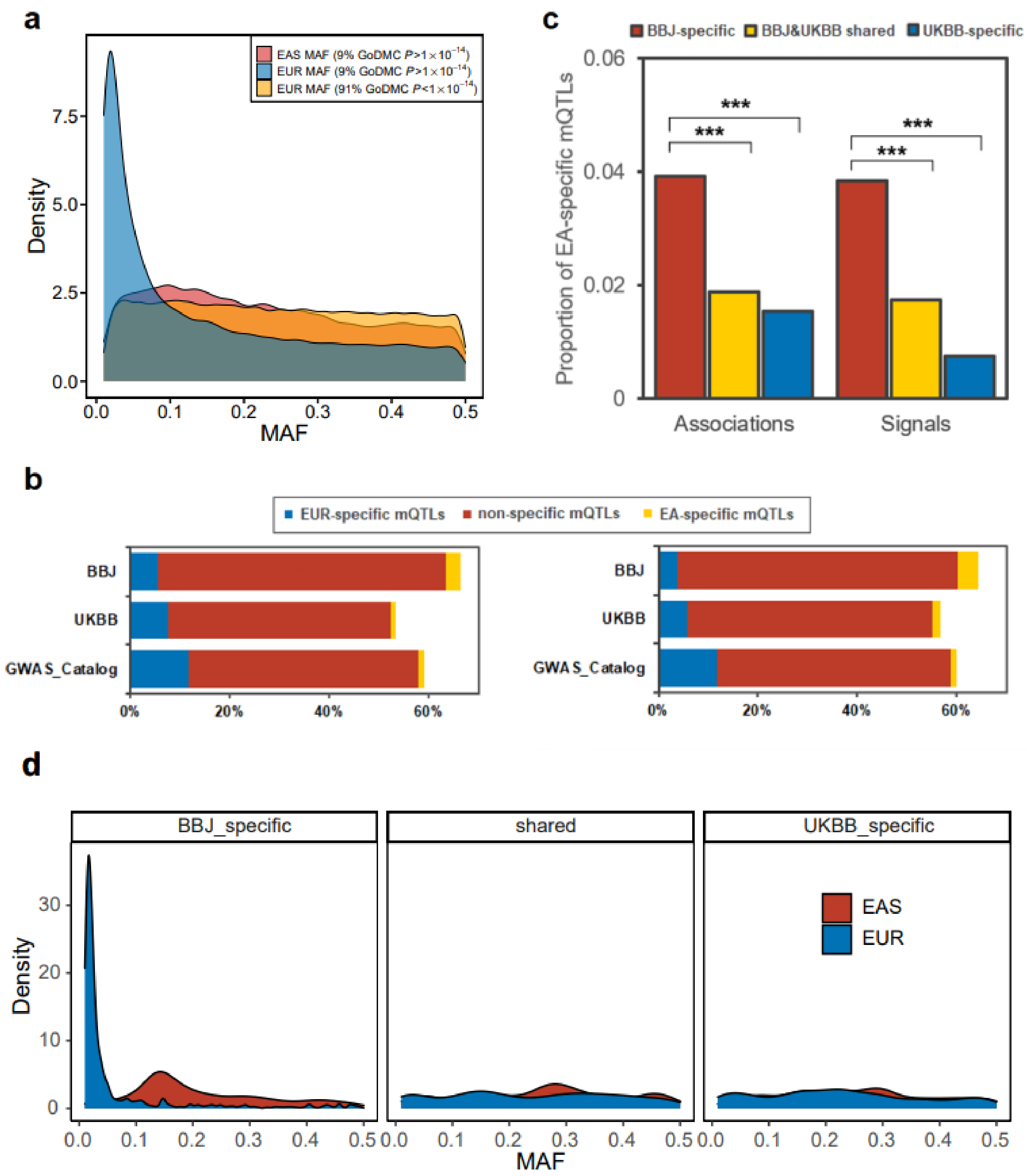


Fig. S10 MAF characteristics and application of EA-specific mQTLs in explaining GWAS signals

a, The distribution of MAFs in East Asians (NSPT data) for the 0.25 million (9%) EA-specific mQTLs (red), MAFs in Europeans (European data from the 1000 Genomes Phase 3) for the 0.25 million (9%) EA-specific mQTLs (blue), and MAFs in Europeans (European data from the 1000 Genomes Phase 3) for the 2.40 million (91%) shared mQTLs (orange). **b,** Non-specific mQTLs are more consistent with GWAS signals (left panel) from BBJ than UKBB and the GWAS Catalog, the same for the associations (right panel) reported in the GWAS Catalog, 107 overlapping GWASs in BBJ and UKBB. **c,** The proportion of EA-specific mQTLs that met GWAS signals and associations that were specific and shared in 107 common traits in BBJ and UKBB (restricted to common variants shared in BBJ and UKBB). The original *P* value from two-tailed Fisher's Exact test is given. **d**, Higher proportions of EA-specific mQTLs overlap with BBJ-specific GWAS signals than with UKBB-specific GWAS signals, largely explained by allele frequency differences.

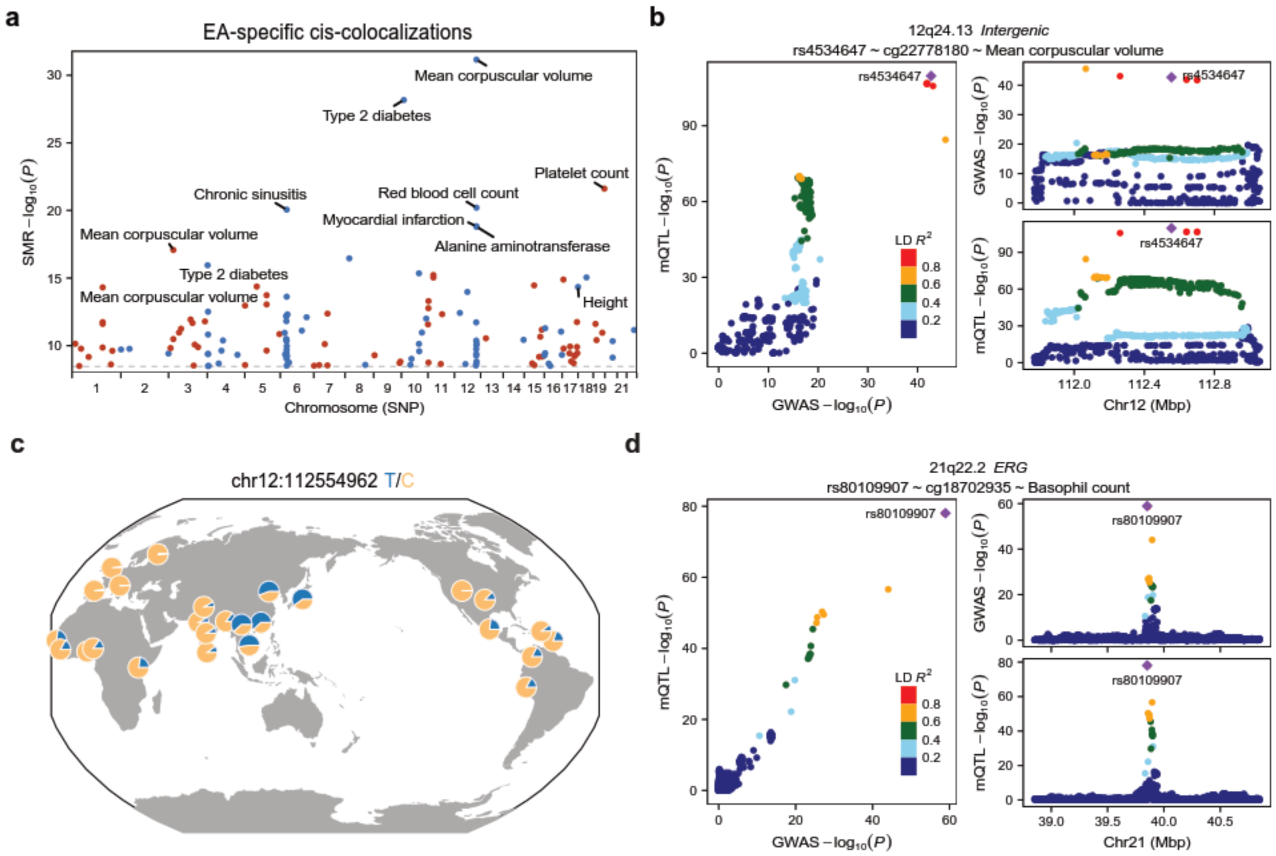


Fig.S11 cis-/trans- East Asian specific colocalizations and examples

a, Manhattan plot showing 144 East Asian-specific cis-colocalizations (96 loci and 38 traits) with -$\log_{10}(P)$ of SMR test on the y-axis and a Bonferroni-corrected threshold of $P < 3.5\times10^{-9}$ indicated by the grey dashed line. **b**, The most significant East Asian-specific cis-colocalization on chr12q24.13, where an intergenic variant rs4534647 (purple point) was cis-associated with cg22778180 in the first intron of *MAPKAPK5* (lower right) and simultaneously associated with mean corpuscular volume (upper right). **c**, The geographic distribution of rs4534647 allele frequencies in different populations (1000 Genomes Phase 3) by the Geography of Genetic Variants (GGV) browser (https://popgen.uchicago.edu/ggv)[58]. **d**, The most significant East Asian-specific trans-colocalization on chr21q22.2, where a variant rs80109907 (purple point) in the intron of *ERG* was trans-associated with cg18702935 on chr13q14.2 (lower right) and simultaneously associated with basophil count (upper right).

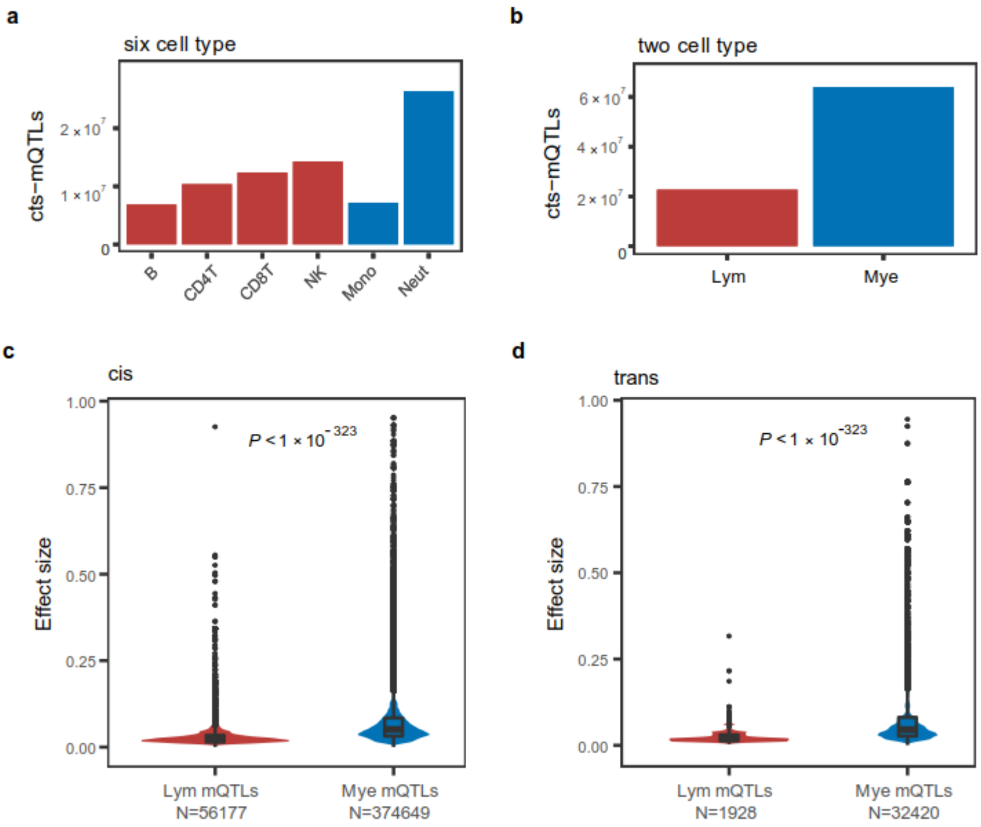


Fig. S12 The detection and effect sizes of cell-lineage specific mQTLs

a&b, The number of mQTLs detected in each cell-type of 6 immune cell types (**a**) and 2-lineages (lymphoid and myeloid **(b**) by application of CellDMC. Only cell-type mQTLs with two-tailed t-test *P*-value <10^{-8} are counted. **c&d**, The distribution of effect sizes of lymphoid- and myeloid-lineage specific cis-mQTLs (**c**) and trans-mQTLs (**d**). *P* value from a two-tailed Wilcoxon rank sum test is given.

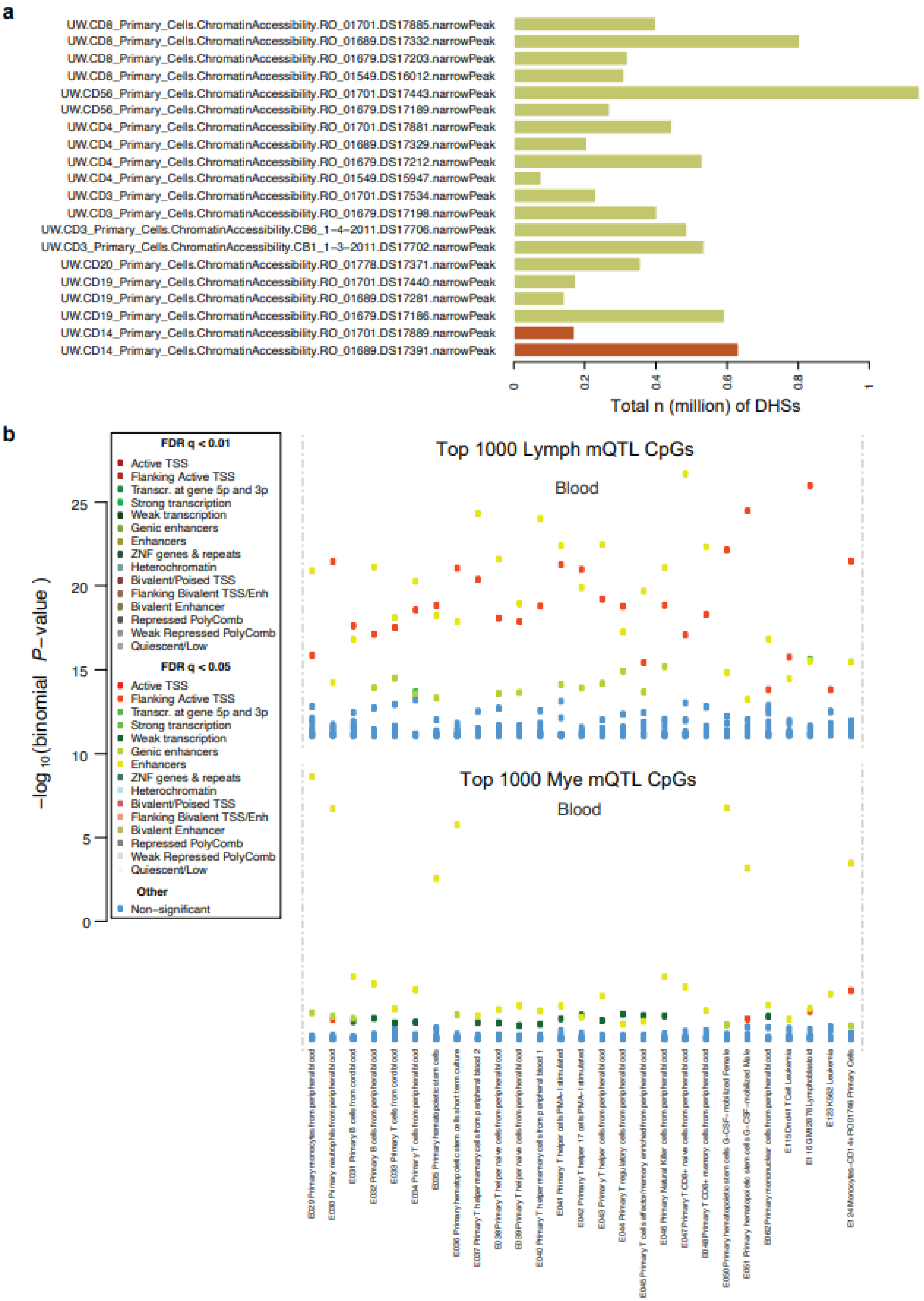


Fig. S13 Mye and Lymph mCpGs enriched in different immune cell types

a, Comparison of DHS numbers per sample suggests that sample quality, rather than immune cell type, is a likely main driver of DHS number. **b**, Comparison of eFORGE2 chromatin state enrichments for different top Mye and Lymph mQTL CpGs indicates that Lymph results are localized in promoters and enhancers shared across different immune cell types (Mye and Lymph, top), while Mye results localize to enhancers present in a few Mye cell types including hematopoietic progenitors and monocytes (bottom). The y-axis labels -log$_{10}$(P) from a two-tailed binominal test.

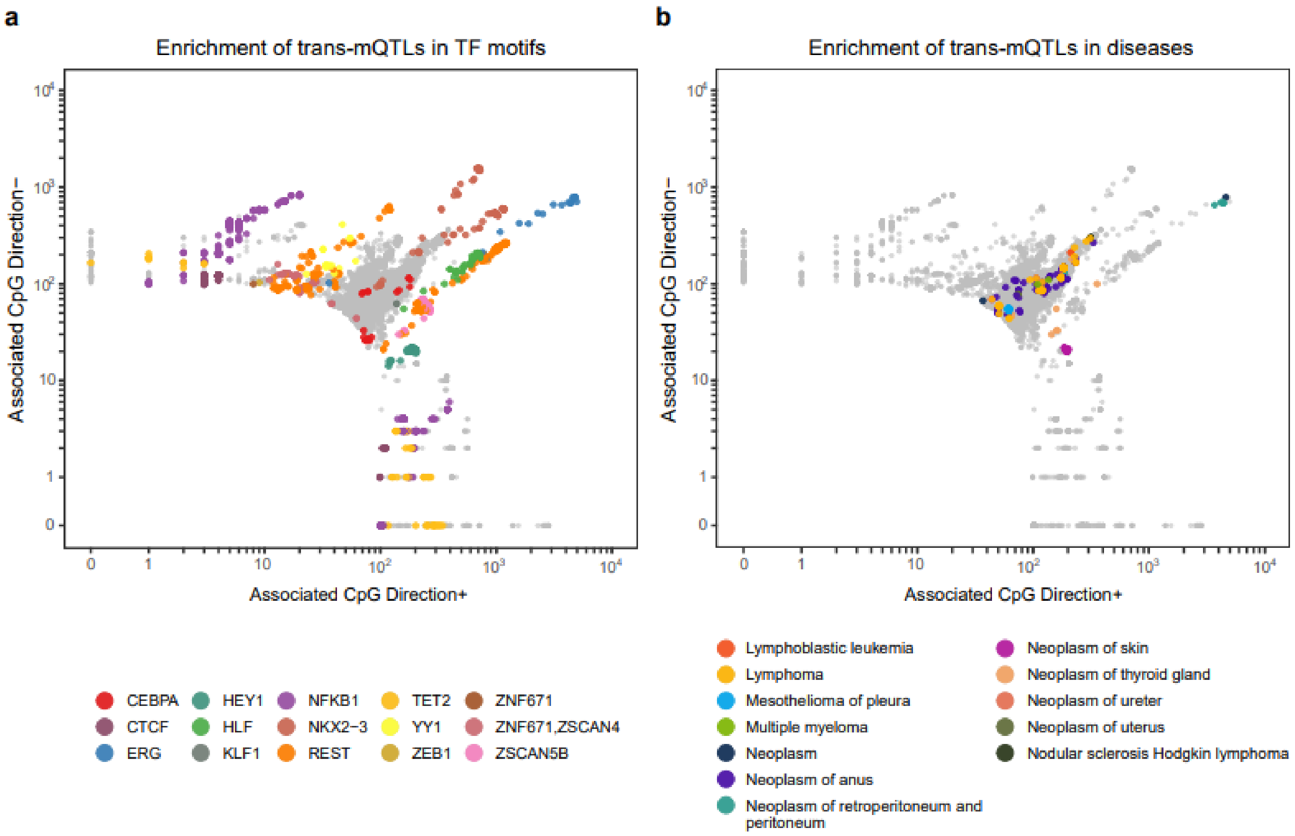


Fig. 14 Enrichment of trans-mQTLs in TFBSs and disease/traits

Enrichment of trans-mQTLs in TFBSs (**a**) and diseases (**b**). One-tailed Wilcoxon rank sum test is applied to compare the TF binding intensity with those of 5,000 randomly selected CpGs. The enrichment of TF-binding motifs is also performed by using Position Weight Matrices (PWMs) from MotifDB (R package PWMEnrich). A union list of trans-mQTLs with the GWAS trait SNP (from GWAS Catalog and Phenoscanner) linked in-cis to a TF and with the corresponding trans-mQTL CpGs enriched for TF-binding sites (as determined either by ChIP-seq or motif-enrichment).

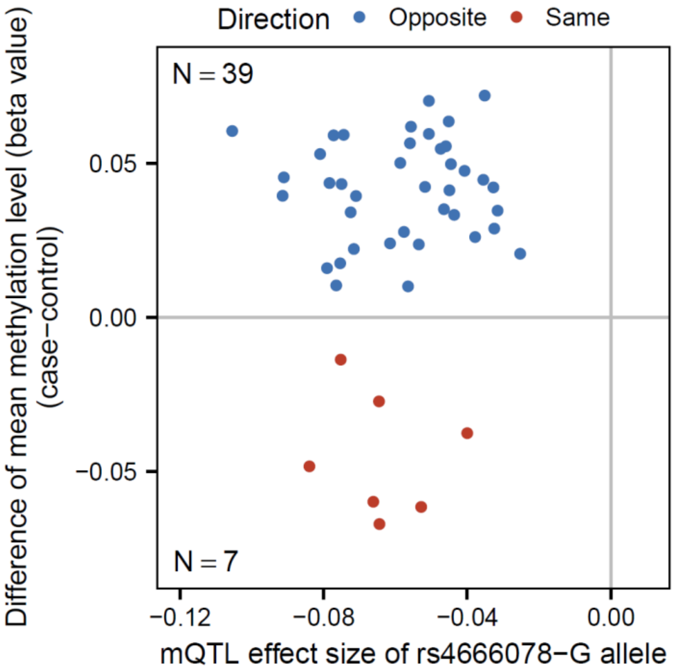


Fig. 15 rs4666078-G allele negatively associated with all of the trans-mCpGs while most mCpGs positively associated with tissue eosinophilia

a, The plot showing the 46 trans-mCpGs that are also the CpGs (differentially methylated positions, DMPs) associated with tissue eosinophilia. The x-axis shows the mQTL effect sizes (calculated based on methylation M values) of rs4666078-G allele. And the y-axis shows the difference of mean methylation level (beta value) between case and control (case-control, FDR < 0.05). Points with different colors show the effects of the two associations are in the same direction (red) or opposite (blue).

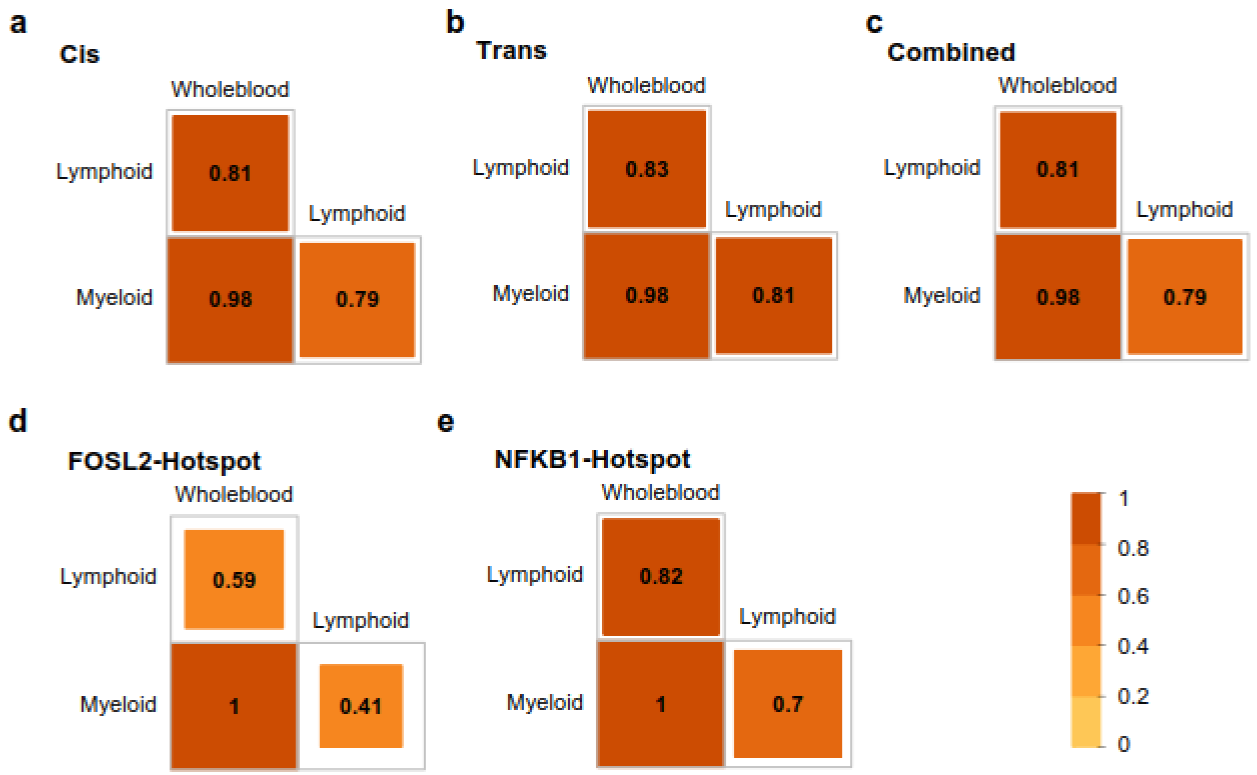


Fig. 16 Sharing effect of all the cell-type mQTLs and those in FOSL2- and NFKB1-mediated hotspots in whole blood and different cell lineages

a-c, Sharing effect estimated by mashr of cell-type mQTLs (**a**), cis- (**b**) and trans-ones (**c**) in whole blood, myeloid and lymphoid lineages. **d&e,** Sharing effect of *FOSL2*-mediated (**d**) and *NFKB1*-mediated (**e**) trans-mQTLs in whole blood, myeloid and lymphoid lineages estimated by mashr with factor = 0.