nature genetics

Article

https://doi.org/10.1038/s41588-022-01248-z

# DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits

## Supplementary Notes

### Data and Samples

The GTEx v8 data consists of 17,382 RNA-seq samples from 948 post-mortem donors, with genotype data for 838 donors from whole genome sequencing (WGS) available in a phased analysis freeze VCF. The GTEx biospecimen collection, molecular phenotype data production and quality control are described in detail in [4]. The eGTEx project [39] seeks to complement the gene expression traits determined in the GTEx project with other molecular traits across the same tissues and individuals, including methylation.

Here, we have generated and analyzed DNA methylation (DNAm), and analyzed existing [4] gene expression data from 9 tissue types: colon transverse, kidney cortex, lung, muscle skeletal, ovary, prostate, testis, whole blood and breast mammary tissues (Supplementary Table 1). Altogether, we analyzed a total of 987 DNAm and 3,872 gene expression samples - depending on the tissue and analysis - corresponding to 424 and 938 individuals, respectively, as well as genotype data from a total of 830 individuals.

### Reduction of methylomes' dimensionality

Considering the 987 profiled methylomes, β values were logit-transformed to M-values. The dimensionality of the methylome set was reduced to two dimensions with the t-Distributed Stochastic Neighbor Embedding (t-SNE) approach [76] implemented in the *Rtsne* R package (v.0.15) (parameters: perplexity = 20, theta = 0.5, max_iter = 5000, pca=TRUE). We observed a set of samples that did not optimally cluster with their corresponding tissue when projected to the t-SNE dimensions. Hence, we estimated cluster dissimilarity for each sample: for each dimension, t-SNE values were transformed to z-scores, and samples with t-SNE values greater than 2.5 standard deviations in either t-SNE dimension were flagged as potential tissue type mismatches but retained. In total, 12 samples were flagged - 7 ovary, 4 prostate and 1 colon samples. The untransformed t-SNE estimates of the 975 unflagged samples are visualized in  Extended Data Fig. 1a.

### Technical, biological, clinical and environmental factors influencing methylation

We evaluated the extent that technical (DNAm profiling plate, N = 11 plates, coded as 'plate'; DNAm profiling slide, N = 12 slides per plate, coded as 'slide'; DNAm profiling slide row/position/array, N = 8 rows per slide, coded as 'array'), biological, and clinical factors derived from GTEx subject-level annotations [4] influence DNAm globally (Extended Data Fig. 2), including DNAm-derived [77] cell-type abundances (Extended Data Fig. 2, b-c). We considered the per-tissue DNAm PEER-derived sets (see "Mapping of mQTLs and eQTLs" section) and for each tissue, we fitted multivariate linear regression models considering the top 3 PEERs as a response variable and each single measured factor as an independent variable. We estimated the percent of PEER variance explained by measured factors by calculating the adjusted $R^2$ ($R^2$adj), averaged across top three PEERs and weighted by PEER reciprocal rank. We observe a set of variables with tissue-shared patterns, e.g. PEERs are associated with the DNAm profiling plate (cross-tissue mean $R^2$adj = 0.17), with the chip (cross-tissue mean $R^2$adj = 0.25) and to a lesser extent, with the chip row (cross-tissue mean $R^2$adj = 0.09) across tissues. Clinical and biological variables show milder associations overall, like sex and age (cross-tissue mean R2adj = 0.04 and 0.03, respectively). However, their effect can be tissue-dependant: the sex effect is very predominant in breast - a highly sex-dimorphic tissue - where a cluster of sex-driven HIV/AIDS and antropomorphy related variables is also moderately associated with DNAm PEERs. As expected, lung is identified as the tissue with the strongest smoking - current/former versus never smoker - effect ( R2adj = 0.02 ) while several viral antibody factors and ischemic-time related variables show the strongest effect in blood DNAm.

Cell-type abundances were derived with *EpiSCORE* (v.1.0.0) (*function=wRPC*) [78] from tissue-matching DNAm signatures obtained from a pan-tissue DNAm atlas [77]. DNAm-derived cell-type abundances for solid tissues are challenging to benchmark due to the lack of consensus cell-specific DNAm-based markers. Among the eight solid tissue types profiled for DNAm herein, only cell-type abundances for breast and colon were estimated. These tissues were selected based on a) their previously identified large cell-type heterogeneity across GTEx samples, strongly correlated with expression-derived PEERs [79] and b) for breast, the expectation of observable sex-differential cell-type abundances [80]. Cell-type abundances in whole blood were derived with *EpiDISH* (v.2.10.0) (*function=epidish, method=RPC*) from a seven immune-cell DNAm signature [81]. We detect strong associations between DNAm-derived PEERs and epithelial and stromal cells in colon and breast ($R^2$adj = 0.26-0.40), as well as a strong association between neutrophil abundances and blood-derived PEERs ($R^2$adj = 0.43) (Extended Data Fig. 3a). As expected, we observe large inter-individual and inter-sex (in breast) cell-abundance differences (Extended Data Fig. 3b).

.

**Evaluation of tissue similarity based on methylation and gene expression**

Hierarchical clustering of the tissues (see Methods) resulted in the obtention of one clustering tree for the DNAm-based distance and another for the transcription-based distance (Extended Data Figure 1b), as well as bootstrap probabilities (BP, obtained by normal bootstrap resampling), which is a measure of clustering support with respect to the data. BP values range from 0 to 100; higher value indicates stronger support for the clusters. All nodes of the DNAm-based tree exhibited maximum support (BP = 100). Most (7/9) nodes of the transcription-based tree exhibited maximum support, except the ones corresponding to muscle with ovary (BP = 53) and kidney (BP = 97) clusters.

We observe, both for gene expression and DNAm profiles, that testis and blood exhibit a lower degree of similarity relative to other tissue types. Testis is characterized by higher gene expression compared to other tissues, as it has been previously shown [4,82]. In contrast, blood appears to be the most divergent tissue for DNAm; CpGs highly methylated in whole blood and lowly methylated in ovary are prominent features of the tissue-specific DNAm signatures. We observe that tissues tend to cluster, in part, by abundance of shared cell types: myocyte and fibroblast rich tissues cluster together, as well as epithelial-rich tissues.

**Definition of expression quantitative trait methylation (eQTM) mapping sets**

The number of CpG-gene tests performed per tissue varied as a function of the number of genes expressed per tissue; we analyzed a total of 5,350,829 CpG-gene pairs. For CpGs with at least one significant (FDR < 0.05) eQTM (see Methods), we defined as significant all CpG-gene pairs at Bonferroni-adjusted P < 0.05, resulting in a non-redundant 12,652 eQTM set across tissues. This set, defined as ‘single-tissue eQTM set’ was complemented with significant cases derived from a cross-tissue approach to constitute the complete set of significant eQTMs reported, defined as ‘complete eQTM set’. This approach [42] enables joint modeling of cross-tissue effects, and it is implemented in the R package *mashr*. For the cross-tissue analysis, we considered  the set of 12,652 eQTMs, i.e. CpG-gene pairs, significant in at least one tissue derived from the single-tissue eQTM-mapping approach. We applied Fisher's z transformation to Spearman correlation coefficients with the function *fisherz* of the R package *psych* (v.1.8.12), and calculated corresponding standard errors. For all tissues, we selected Fisher-transformed Spearman coefficients and corresponding standard errors for each of the 12,652 eQTMs, as well as for 100,000 randomly selected CpG-gene pairs that were tested across all tissues. Those estimates were used to fit the *mashr* model. The local false sign rate (LFSR) generated by *mashr* was used to identify significant (LFSR < 0.05) eQTMs. The complete set of significant eQTMs, referred to as 'complete eQTM set', was defined by the union of significant cases derived from the single-tissue (FDR < 0.05) and multi-tissue (LFSR < 0.05) eQTM-mapping approaches (FDR < 0.05 or LFSR < 0.05). This resulted in an expansion of the 15,839 eQTM-tissue significant eQTM set - corresponding to 12,652 eQTMs - to 47,783 eQTM-tissue significant cases (Extended Data Fig. 1c, Supplementary Table 2). Across the article, we refer to CpGs with at least one significant eQTM as eCpGs. Considering eQTM-based downstream analyses, the 'single-tissue eQTM set' or the 'complete eQTM set' were used depending on the particular analysis, as noted.

**Replication of eQTMs in external cohorts**

We assessed eQTM replication for all tissues in the FUSION Skeletal Muscle Study cohort, where N = 265 individuals were utilized for eQTM mapping. The FUSION cohort characteristics, and the eQTM-mapping procedure, which is similar to the one employed herein, are described in [20]; we employed available summary statistics. For a particular tissue, we considered for replication analyses CpG-gene pairs from the single-tissue eQTM set that passed P < 0.05 in the tissue tested. Across tested tissues, we observed a high replication rate (cross-tissue average π1 = 0.75), being the highest for muscle (π1 = 0.84), possibly due to muscle-specific eQTMs contribution (Supplementary Table 2).

**Characterization of eQTM tissue-specificity**

The number of eCpGs ranged from 1,542 in testis to 4,568 in ovary (Extended Data Fig. 1c, Supplementary Table 2). The number of significant eQTMs was strongly correlated with per-tissue sample size (Spearman's ρ = 0.86). Similarly to transcriptome- and eQTL-derived patterns observed in previous GTEx analyzes [3,4,82], eQTMs tend to be either tissue-specific or shared across most tissue types (Extended Data Fig. 1d). However, the fraction of eCpGs identified that were detected as eCpGs exclusively in a single tissue is assumed to be a lower bound, as we observe that the abundance of tissue-specific eCpGs is strongly correlated (Spearman's ρ = 0.83) with sample size. Altogether these results indicate that the limited sample size of the sets utilized for eQTM mapping (N < 40 in 4 tissues, Supplementary Table 1) and the limited number of tissues (N = 8) impose limitations in accurately estimating how many eQTMs exist, and are shared, across tissues. This analysis would benefit from a more powered and exhaustive eQTM catalog.

**Characterization of eQTM predictors**

To annotate mCpGs for gene regulatory elements, we extended the span of their genomic location by +/- 100bps, and checked for overlap (>= 1bp) with regulatory regions. For each regulatory element class, a Fisher's exact test was conducted to determine enrichment of CpGs included in the single-tissue eQTM set in gene regulatory elements, and significance was defined at Bonferroni-adjusted P < 0.01. We observe that the association of DNA methylation with expression of nearby genes differs by regulatory elements. Promoters and proximal enhancers are strongly enriched (OR = 4.55 and 4.17, respectively) for eCpGs, as well as insulators (OR = 2.32) and distal enhancers (OR = 1.74). Overall, 54% (3,188/5,898) of eCpGs overlap with gene regulatory elements.

To assess the relative contribution of molecular signatures to eQTMs linked to different gene regulatory elements, we stratified eQTM tests by CpG-overlap with gene regulatory element classes. We considered one eQTM test per CpG per tissue, corresponding to the CpG-gene test with the smallest p-value. For each element class, in each tissue, a logistic regression model of eQTM likelihood was built to predict whether an eQTM test was significant given the CpG-Gene TSS distance (in Kb), direction of the eQTM effect ('1' for negative correlation between methylation and expression, '0' otherwise), the tissue-averaged methylation (in M-value units) and gene expression abundance (in log2(TPM+1) units), as predictors. For all predictors in all logit models, multicollinearity was tested with the R package *vif* implemented in the R package car (v.3.0). None of the predictors displayed problematic variance inflation; *vif* score for any predictor was below 1.5. Cross-tissues meta-effect was evaluated by modeling single-tissue effect estimates (log of odds ratio) with a random-effects model (*rma* function, *metafor* R package, v.2.0.0). Summary statistics relative to the characterization of eQTMs are provided in Supplementary Table 2. This analysis revealed that eCpGs were enriched (Fisher's exact test FDR < 0.01) in gene regulatory elements, but showed distinctive signatures by regulatory element class (Extended Data Fig. 1e). Proximity of CpG to gene transcription start site (TSS) increases the likelihood of the eQTM association ubiquitously across regulatory elements, but high CpG methylation and corresponding negative correlation with gene expression are only predictive of eQTMs linked to promoters and proximal enhancers. Distal enhancer and insulator eQTMs are enriched for low-methylated eCpGs, and low gene expression appears to be predictive of insulator-linked eQTMs exclusively. These results suggest that DNAm in CpG sites is associated with transcription of *cis*-genes through their regulatory elements, as described [41]. However, they also indicate heterogeneity of the biological mechanisms driving eQTMs, that depend on the class of regulatory elements involved, compatible with patterns observed in blood cells [26,40].

Additionally, to evaluate the contribution of e/mQTLs to eQTMs, we aligned our e/mQTL colocalization results (see Characterization of mQTL-eQTL shared signal) with eQTMs identified at FDR < 0.05, and observed that 37% of eQTMs correspond to e/mQTL colocalizations, highlighting a substantial contribution of genetics to CpG-gene expression correlation.

**Characterization of eQTM CpG-gene distances and pleiotropy**

The per-tissue average and median CpG site and gene transcription start site (TSS) distance is 31-41Kb and 119-157Kb, respectively, and is positively correlated with sample size (median TSS-eCpG dist: Spearman's $\rho = 0.86$, mean TSS-eCpG dist: Spearman's $\rho = 0.67$). Per tissue, 22-31% eCpGs correlate to >1 gene, this estimate is also positively correlated with sample size (Spearman's $\rho = 0.72$). These results suggest that the limited sample size of the sets utilized for eQTM mapping impose limitations in the precision of these estimates. This analysis would benefit from a more powered and exhaustive eQTM catalog.

### Comparison of empirical associations of DNA methylation with gene expression to array annotations

To investigate how accurately the CpG-gene assignments provided by Illumina reflect the eQTM results observed in GTEx data, we contrasted our eQTM findings with the EPIC array CpG annotation file provided by Illumina. We observed that while 76% (4,489/5,898) of the eCpGs we identify have an assigned gene provided in the annotation file, for only 45% (2,641/5,898) of these eCpGs does the annotated gene match a CpG-gene association detected through eQTM mapping. Moreover, only 22% (2,828/12,652) of the eQTMs we identify in at least one tissue match any annotated CpG-gene pair. Per tissue, most (48-62%) of eCpGs are correlated to the corresponding closest gene (among genes expressed in that tissue). Overall, our eQTM results can enhance our ability to assign CpGs to gene(s) with which they are biologically linked, therefore facilitating the interpretation of methylation-derived analyses, as methylation patterns are often interpreted in light of their predicted impact on gene regulation.

**Mapping of mQTLs and eQTLs**

To define mQTLs we analyzed all samples with available methylation and genotype data (Supplementary Table 1), comprising a total of 856 samples, from 42 - muscle skeletal - to 190 - lung - per tissue, derived from 367 subjects [4] and interrogated a total of 754,054 CpGs and 10,296,879 variants across all tissues. Genotype data was derived from WGS, data generation, variant calling and quality control is explained in detail elsewhere [4]. For each variant-CpG pair, we fit a linear regression model separately in each tissue, and tested for significance of genotype - quantified as allele counts, assuming additive effect - on methylation estimates while adjusting for additional known and unknown factors:

$$Y = \beta_0 + \beta_G \, Genotype + \boldsymbol{\beta}_{(1\ldots m)} \, \boldsymbol{C} + \boldsymbol{\beta}_{(1\ldots n)} \, \boldsymbol{PEER} + \varepsilon$$

where,

*Y* is the inverse-normal-transformed DNAm levels

$\beta_0$ is the intercept

$\beta$ are the corresponding effect sizes. $\beta_G$ is the effect size of genotype on DNAm.

*C* represents a subset of covariates that were used in *cis*-eQTL mapping [4]. These covariates include 5 genotype principal components, 2 covariates derived from the generation of genotype data by whole genome sequencing - described in [4] - and biological sex status.

*PEER* represents PEER factors [83] derived from DNAm. The number of PEER factors was selected to maximize mQTL discovery, across two sample size bins: tissues with < 50 samples and tissues with ≥ 100 samples. The optimization was performed similarly to [4], and resulted in the selection of 5 and 20 PEER factors, respectively, for the two sample size bins. In the optimization step, PEERs were calculated from inverse-normalized DNAm β values from CpGs in chromosome 1 (~70K CpGs) and significant mQTLs were defined at nominal P < 1e-05. To correct for multiple testing of variants per CpG, we permuted DNAm estimates 1,000 times, adjusting p-values with a beta distribution approximation [84]. Genome-wide CpG multiple testing correction was performed on top-significant CpG-variant beta-adjusted p-values using Storey *qvalue* [85]. The set of significant mQTL CpGs (mCpGs) identified at FDR < 0.05 was defined as 'single-tissue mQTL set', and complemented by significant cases derived from cross-tissue QTL mapping (see 'Definition of QTL sets').

To identify independent mQTLs, we started from the set of mCpGs discovered in the first pass of association analysis (complete mQTL set: FDR < 0.05 or LFSR < 0.05). Then, the maximum beta-adjusted p-value (correcting for multiple testing across the variants) over these CpGs was taken as the CpG-level threshold. The next stage proceeded iteratively for each CpG and threshold. A *cis*-scan of the window was performed in each iteration, using 1,000 permutations and correcting for all previously discovered variants. If the beta-adjusted p-value for the most significant CpG-variant, i.e. best association, was not significant at the CpG-level threshold, the forward stage was complete and the procedure moved on to the backward step. If this p-value was significant, the best association was added to the list of discovered mQTLs as an independent signal and the forward step proceeded to the next iteration. Once the forward stage was complete for a given CpG, a list of associated variants was produced which we refer to as forward signals. The backward stage consisted of testing each forward signal separately, controlling for all other discovered signals. To do this, for each forward signal we ran a *cis* scan over all variants in the window using *FastQTL*, fitting all other discovered signals as covariates. If no variant was significant at the CpG-level threshold the signal being tested was dropped, otherwise the best association from the scan was chosen as the variant that represented the signal best in the full model.

We define eQTLs as *cis*-gene variants with a significant genotype effect on gene expression, utilizing a single-tissue approach analogous to the mQTL-mapping one. We included the same covariates and variant set (±1Mb from gene transcription start site, MAF < 0.01) employed for eQTL mapping in [4]. A total of 3,438 samples was considered, from 73 - kidney cortex - to 706 - muscle skeletal - samples per tissue, from a total of 829 subjects. Analogously to mQTLs, we identified multiple independent eQTLs, and the complete set of significant eQTLs was obtained by complementing the single-tissue mQTL set with significant cases derived from the cross-tissue approach (see 'Definition of QTL sets').

### Definition of mQTL and eQTL sets

To overcome QTL-mapping limited power due to per-tissue available sample sizes, and to determine QTL tissue-specific patterns, we used an approach to perform a cross-tissue QTL analysis by leveraging QTL signal across tissues [42], implemented in the R package *mashr*. Considering the set of 286,153 mCpGs significant in at least one tissue derived from the single-tissue mQTL-mapping approach, i.e. the 'single-tissue mQTL set', for every top mQTL per CpG per tissue, mQTL effect sizes, corresponding standard errors and 301,801 randomly selected variant-CpG pairs that were tested across all tissues were used to fit the *mashr* model. The *mashr* version employed herein (0.2.6) sets missing effect size values to 0 and corresponding standard error to 1,000,000. The local false sign rate (LFSR) generated by *mashr* was used to define significant (LFSR < 0.05) mQTLs.

The 'complete mQTL set' set was defined by the union of significant cases derived from the single-tissue (FDR < 0.05) and cross-tissue (LFSR < 0.05) mQTL-mapping approaches (FDR < 0.05 or LFSR < 0.05). This resulted in an expansion of the 607,987 mCpG-tissue significant mQTL set - corresponding to 286,152 mCpGs - to 1,385,225 mCpG-tissue significant cases. An equivalent approach was employed to perform cross-tissue eQTL meta-analysis. Considering mQTL- or eQTL-based (e/mQTL) analyses, 'single-tissue e/mQTL set' or the 'complete e/mQTL set' are used depending on the particular analysis, as noted.

### Replication of mQTLs in external cohorts

We assessed mQTL replication, for single-tissue and complete mQTL sets, in the ROSMAP brain cohort (N = 543, A = HumanMethylation 450K) [23] and the FUSION Skeletal Muscle Study cohort (N = 282, A = HumanMethylation EPIC) [20]; 'N' and 'A' define the number of individuals utilized for mQTL-mapping and the Illumina array used to profile methylation, respectively. In all cases, we tested for replication the mCpG lead variants (best variant per mCpG) from the single-tissue mQTL set. The FUSION and ROSMAP mQTL-mapping procedure and the cohort characteristics are described in [20,23]; we employed available summary statistics. In brief, DNAm β values were logit-transformed and a linear model was fit with genotype, top genotype PCs and methylation PEER factors in a ±1000 Kb window from the CpG locus (FUSION). For ROSMAP, the revised mQTL statistics (http://mostafavilab.stat.ubc.ca/xQTLServe/) were utilized, where a ±50 Kb window was employed and DNAm estimates were adjusted for potential batch effects by regressing them out. Replication was assessed by means of π1, which measures the estimated true positive rate [85]. We observed high GTEx mQTL true positive rate values in the external cohorts, especially for tissue-matched replication datasets (π1= 0.91). The average π1 across tissues and cohorts, for the complete mQTL set, is π1=0.85 (Supplementary Table 3).

## Characterization of tissue specificity patterns of mQTLs and eQTLs

The number of mCpGs detected per tissue was strongly correlated with per-tissue sample size (Spearman's ρ = 0.92). The overall tissue specificity of mQTLs follows a skewed U-shaped curve, i.e. for a particular CpG, genetic regulation of DNAm tends to be either highly tissue-specific or highly shared across tissue types (Extended Data Fig. 4a). The fraction of mCpGs identified that were detected as mCpGs exclusively in a single tissue (Extended Data Fig. 4b) is assumed to be a lower bound, as we observe that the abundance of tissue-specific mCpGs is strongly correlated (Spearman's ρ = 0.80) with sample size, indicating power limitations to detect tissue-specific QTLs in low-sampled tissue sets. This assumption is compatible with the larger tissue-specific eGene fractions (Extended Data Fig. 4b) observed for eQTLs, mapped in larger sample sets. Differential tissue-sharing counts distribution of comparing eQTLs to mQTLs was tested by means of a Wilcoxon rank-sum test, and the null hypothesis was rejected (P = 2.9e-03). Cross-tissue similarities based on mQTL and eQTL effect size magnitude were assessed with *mashr::get_pairwise_sharing* function with default parameters. With this approach, we computed the proportion of significant mQTL and eQTL signals shared by magnitude in each pair of tissues, based on *mashr* mQTL posterior effect sizes for single-tissue (FDR < 0.05) mCpGs/eGenes with LFSR < 0.05 in at least one of the two pairwise tissues (Fig. 2b). Differential mQTL proportion of signals shared across tissues by different regulatory regions (distal and proximal enhancers, and promoters) was tested by means of a paired Wilcoxon rank-sum test, and the null hypothesis was rejected (P < 1e-05) for all comparisons (Extended Data Fig. 4c). The set of *mashr* mQTL posterior effect sizes was also utilized to identify, for each mCpG, the number of tissues with mQTL effect size within a factor of 2 of the tissue with the largest mQTL effect size, in absolute number. Differential eQTL versus mQTL proportion of signals shared across tissues was tested by means of a Wilcoxon signed rank test, and the null hypothesis was rejected (P = 1.2e-02). That implies that, compared to eQTLs, the proportions of mQTLs with tissue-shared effect sizes appear to be significantly higher. These patterns could be due to more stable cross-tissue DNAm QTL effects compared to expression QTLs, to substantially lower mQTL sample sizes (compared to eQTL sample sizes), to biased (promoter-enriched) CpG site interrogation or a combination of all.

## Validation of tissue specificity patterns of mQTLs in external cohorts

To validate mQTL tissue specific patterns, we investigated GTEx tissue-specific mQTLs (defined as detected in one out of nine tissues) in the FUSION Skeletal Muscle Study [20], the ROSMAP brain [23] and the GoDMC blood studies [86] mQTL datasets, by evaluating magnitude and correlation of corresponding mQTL effect sizes. We observe that mQTL effect sizes derived from external cohorts are significantly (one-sided Wilcox test P < 1e-06) larger (Extended Data Fig. 4d), and more correlated (Extended Data Fig. 4e), when matching the corresponding GTEx mQTL tissue, or tissues with shared cell types. Of note, cross-cohort differences can be derived, in part, from mQTL mapping approaches. For example, mQTL mapping cis window sizes are 50, 500 and 1000kb for ROSMAP, GoDMC and FUSION cohorts, respectively, and GoDMC only includes mQTLs with nominally significant p-value in at least one included study.

### Representation of GTEx mQTLs in external cohorts

To evaluate the presence of mQTLs detected in GTEx in other mQTL resources, we assessed the overlap of GTEx mCpGs (FDR < 0.05) in the FUSION Skeletal Muscle Study [20], the ROSMAP brain [23] and the GoDMC blood studies [86] (Extended Data Figure 4). We considered that a mCpG identified in GTEx was also identified in an external mQTL map if the p-value of the lead variant for that mCpG in the external cohort was smaller than 1e-05 ( P < 1e-05 ). We also evaluated a more relaxed threshold ( P < 1e-03 ) and characterized the fraction of identified mCpGs attributable to EPIC and 450K arrays. At P < 1e-05, we identify 117,941 mCpGs exclusively in GTEx, 74% (86,694/117,941) of which are not captured by the 450K - but only by the EPIC - Illumina array. Considering 141,800 GTEx mCpGs included both in EPIC and 450K arrays, 50,376 of them (36%) are not among the 193,726 mCpGs - detected at P < 1e-05 - in GoDMC; 21% when considering only blood-derived mCpGs.

### Functional genomic characterization of mQTLs and eQTLs

We observed eQTLs to be more strongly enriched in open chromatin sites than mQTLs (Extended Data Fig. 6a). Additionally, mQTLs appear to be depleted in transcribed genes and genic enhancers but enriched in distal, active enhancers (Extended Data Fig. 6b).

### mQTL-eQTL colocalization

We investigated the associations between mQTLs and eQTLs (single-tissue QTL set: FDR < 0.05) by means of QTL effect size colocalization with *coloc* [87] using default priors. For both QTL types, we considered unconditional QTL mappings, i.e. agnostic to multiple independent QTLs, due to computational limitations of performing colocalization on the complete combinatorial space considering multiple independent QTLs for both QTL types. A mQTL locus was defined as overlapping with an eQTL locus, and subsequently tested for colocalization, if the mQTL-eQTL region a) had at least 50 variants in common and b) included potentially causal mQTL and eQTL variants. That is, it included at least one fine-mapped and/or conditional QTL mapping lead variant, for both mQTL and eQTL signals. Fine-mapped QTL variants were estimated with *dap-g* [88] (v.c805e3cbedd2f5b16f1464d07207af4183ea73dd) and QTL credible sets were defined at 90% confidence.

To classify mQTL loci (mCpGs) into the mutually exclusive colocalization categories depicted in Fig. 2e, we first annotated mQTL loci that did not overlap any eQTL locus in any tissue. For the remaining eQTL-overlapping mQTL loci set, which was tested for eQTL colocalization, a mQTL locus was annotated as involved in a eQTL-mQTL colocalization if it colocalized (PP4 > 0.5) with at least one eQTL in at least one tissue. For the remaining set, a mQTL locus was annotated as independent to eQTL signal if it exhibited a PP3 > 0.5 for at least one eQTL colocalization test in at least one tissue. The remaining set was considered to exhibit inconclusive colocalization signal (PP0 + PP1 + PP2 > 0.5).

Across tissues, 93% (266,239/286,152) of mQTL loci do overlap with an eQTL. Using a moderately permissive threshold for the posterior probability of sharing the same causal variant (PP4 > 0.5), only 21% of mQTL loci are suggestively colocalized (PP4 > 0.5) with at least one eQTL, whereas for 38% of cases there is evidence of independent variants driving mQTL and eQTL signals (PP3 > 0.5). For the remaining 34% of the cases analyzed, we lack adequate power to conduct meaningful colocalization analyses (PP0 + PP1 + PP2 > 0.5). Our results indicate that a considerable fraction of mQTLs do not show clear associations with local gene expression in the same tissue type, but we acknowledge that limited power for e/mQTL detection, allelic heterogeneity, and colocalization assumptions may limit our ability to accurately estimate this fraction.

### mQTL-eQTL concordance in direction of effects across regulatory regions

A mQTL-eQTL pair was defined as concordant if the mQTL sign of the top-colocalized e/mVariant matched the corresponding eQTL sign and discordant otherwise. An exact binomial test was conducted to assess whether the proportion of discordant/concordant cases differed significantly from 50%, and the null hypothesis (proportion = 53%, P < 2.2e-16) was rejected. Subsequently, we assessed whether the discordance/concordance rate varied as a function of mCpG location in gene regulatory regions, considering promoters and proximal enhancers jointly, distal enhancers and insulators. Gene regulatory element annotations were derived from ENCODE5 cCREs catalog (see eQTM section above). To annotate the mCpGs, we extended the span of their genomic location by +/- 100bps, and checked for overlap (>= 1bp) with regulatory regions. A mCpG was annotated with promoter/proximal enhancer status if, in addition to overlapping with an ENCODE-predicted promoter or proximal enhancer, it overlapped the 2kb region upstream from the TSS of the corresponding colocalized eGene. To compare the discordance/concordance rate across regulatory regions, we performed a multi-sample test for equality of proportions without continuity correction (*prop.test* function, *stats* R v.3.6.1 base package). Among e/mQTL colocalized variants, the direction of the effect on methylation and expression is often (53%) in the opposite direction (exact binomial test P < 2.2e-16), as previously observed [32]. However, we observe differences in this directionality based on regulatory context (test of equal proportions P < 2.2e-16); the proportion of mCpGs corresponding to opposite e/mQTL effects is larger for mCpGs located in eGene-matching promoters and proximal enhancers (61%) than in distal enhancers (52%) and minority (39%) in insulators. These observations are in line with the view that hypomethylation in proximal gene regulatory regions is associated with active transcription but the association is not consistent in distal regulatory regions [89,90].

### Characterization of mQTL-eQTL regulatory pleiotropy

Regulatory pleiotropy categories are defined in Methods and illustrated in Extended Data Fig. 7a. The most common scenario, comprising 54% of the eQTL-colocalized mCpGs (Extended Data Fig. 7b), corresponds to eQTL-mQTL colocalizations involving multiple mCpGs and a single eGene (Tier 3 in Extended Data Fig. 7a). Overall, we observe a higher mCpGs per eGene than eGenes per mCpG ratio (Extended Data Fig. 7c). The mCpGs per eGene ratio is correlated with sample size (Spearman's $\rho = -0.75$), suggesting that the observed value is a lower bound estimate due to mQTL-mapping power limitations. Across tissues, the largest pleiotropic sets tend to involve mCpGs and eGenes located in the Major Histocompatibility Complex (MHC). Given its complex LD structure and genotype imputation inaccuracies, the MHC is challenging to analyze. To ensure that pleiotropy results are not biased by MHC, we evaluated pleiotropy excluding CpG *cis* loci overlapping the MHC region, as well as corresponding eQTL-mQTL colocalized eGenes. The pleiotropy distributions were not significantly different (test of equal proportions P < 0.05). The largest pleiotropic set was identified in ovary and involved 114 mCpGs and 5 eGenes in the HOXD gene cluster region: *HOXD3, HOXD4, MTX2, HAGLROS* and *RP11-387A1.5*. The pleiotropy of this region was further evaluated by testing for eQTL-mQTL colocalization ovary eGenes defined at LFSR < 0.05, resulting in three additional colocalized eGenes - *HOXD8, HOXD1* and *HAGLR* - considered part of the HOXD pleiotropic set.

**Colocalization of GWAS with QTL signal**

The approach to identify colocalization of GWAS with QTL signal is described in Methods. In total, 6,720 GWAS-GWAS-hit tuples were considered for downstream analyses, from 1 - mothers's age at death, epilepsy, self-reported schizophrenia, intracranial volume, insomnia - to 733 - standing height - GWAS hits depending on the GWAS trait.

To evaluate the conservativeness of selected priors, we compared *coloc* mQTL-GWAS colocalization results to those generated with default priors (p1 = 1e-04, p2 = 1e-04, p12 = 1e-05) for the GWAS with largest amount of signal, i.e., UKB standing height GWAS. We observed that results are strongly correlated (Spearman's ρ = 0.93), but colocalization probabilities derived from *fastenloc*-derived priors tend to be more conservative (Extended Data Fig. 8a). That is, at PP4 > 0.5, considering *fastenloc*-derived priors, we identify 53% less colocalized cases than with the default-priors approach. That is expected, given the higher ratio between *fastenloc*-derived p2 (mQTL association) and p12 (mQTL and GWAS association) priors compared to corresponding default-priors one (Supplementary Table 6).

**Evaluation of mQTL-GWAS colocalization approach**

Considering results with suggestive colocalization probability (the intersection set of *coloc* PP4 > 0.1 and *fastenloc* RCP > 0.1), we observe a strong correlation (Spearman's ρ = 0.79) between results from both methods (Extended Data Fig. 8b). We identified as significantly colocalized those GWAS-GWAS hit-mCpG/eGene-tissue tuples with both corresponding *coloc* PP4 > 0.3 and *fastenloc* RCP > 0.3, i.e. the intersection of cases with moderate colocalization signal derived from both methods (RCP > 0.3 and PP4 > 0.3). Across the article, *coloc* PP4 is provided as the reference colocalization probability, unless stated otherwise. We identify 55% of GWAS hits (1,505/2,734) colocalizing with at least one mQTL but with no eQTLs at RCP > 0.3 and PP4 > 0.3; this estimate can range from 44 to 66% depending on the combination of PP4 and RCP thresholds selected. Colocalization cases involving at least one mQTL but no eQTLs are defined as 'mQTL-specific' colocalizations, as opposed to 'eQTL-specific' colocalizations, which comprise cases involving at least one eQTL but no mQTLs. Colocalization cases involving at least one mQTL and one eQTL are defined as 'e/mQTL-shared' colocalizations.

**Scope of mQTL-GWAS colocalizations**

Instances of mQTL-GWAS colocalizations were observed among 81% (67/83) of tested GWASs and involved 41% (2,734/6,720) of GWAS hits, and 3,381 and 940 colocalized (trait-linked) mCpGs and eGenes, respectively. For 4.5% (102/2,254) of GWAS hits involved in mQTL-GWAS colocalizations, the colocalizing signal corresponded to a secondary mQTL. For nine GWAS traits, colocalizations were only detected for mQTLs, including osteoporosis and certain balding and metabolic phenotypes, among other traits. We observe that the effect size of mQTLs involved in mQTL-specific colocalizations is significantly (Wilcoxon rank-sum test P = 0.003) smaller than the effect size of mQTLs involved in e/mQTL-shared colocalizations.

### Signatures of QTL-GWAS colocalizations and trait-linked mCpGs

The fraction of the colocalized e/mQTL-GWAS shared loci for which mQTL-GWAS association shows greater colocalization probability than the eQTL-GWAS association was evaluated considering the colocalization probabilities of both *coloc* and *fastenloc* approaches across tissues and independent QTL signals. That is, max(*coloc* PP4(mQTL)) > max(*coloc* PP4(eQTL)) and max(*fastenloc* RCP(mQTL)) > max(*fastenloc* RCP(eQTL)) in at least one tissue and/or independent QTL colocalization corresponding to the e/mQTL-GWAS shared locus. To evaluate the overlap of trait-linked mCpGs with open chromatin regions, we extended the genomic location span of mCpGs tested for GWAS colocalization by +/- 100bps, and checked for overlap (>= 1bp) with the aggregated set of DNAse-seq derived ENCODE5 open chromatin regions utilized to characterize mQTL signatures. Trait-linked mCpGs were classified as eQTL-shared or mQTL-specific (see Fig. 4). Enrichment significance of eQTL-shared or mQTL-specific trait-linked mCpGs in open chromatin regions was estimated at Fisher's exact test P < 0.05. To evaluate the methylation signatures of mQTL-GWAS colocalizations, for each tissue, we performed a Wilcoxon rank-sum test comparing DNAm levels of mCpGs tested for colocalization to those significantly colocalized (RCP > 0.3 and PP4 > 0.3). For the majority (8/9) of tissues, DNAm levels of colocalized mCpGs were significantly (Wilcoxon P < 0.05) lower than tested ones. We applied an analogous approach to eQTL-GWAS colocalizations, and observed an inverse pattern: for the majority (6/9) of tissues, expression levels of colocalized eGenes were significantly (Wilcoxon P < 0.05) higher than tested ones. Bootstrapped (N = 5,000 replicates) values for DNAm and gene expression means - averaging mCpGs and eGenes within each tissue - for all QTL-GWAS colocalization groups are displayed in Extended Data Fig. 9b; confidence intervals were computed using bootstrapping with replacement.

### Integration of trait-linked genetically-regulated methylated loci with functional maps

To identify genes involved in trait-linked mCpGs that co-located with gene regulatory elements, we integrated mQTL-derived colocalization results with curated promoter- and enhancer-gene target predictions [63,64] and eQTM associations generated herein (Methods). We identified 57% (152,397/267,401) of non-colocalized mCpGs as eCpGs and/or co-located with enhancers, hence gene-linked. In contrast, we identified 68% (1,308/1,911) of mQTL-specific trait-linked mCpGs as gene-linked, across 61 GWASs and 1,129 GWAS hits, and reported findings in Supplementary Table 7. For 35% (400/1,129) of these loci, multiple mCpGs consistently support the same gene candidate(s). Among highly supported (by ≥ 3 mCpGs) cases, we identify poorly or not characterized gene-trait associations. For instance, the topmost supported instance corresponds to the *RUNX1* locus associated with asthma. For the asthma GWASs analyzed, we observe 12 distinct mCpGs linked to *RUNX1* regulatory regions. Given that other members of the *RUNX* transcription factor family are reported to play a role in asthma [91], *RUNX1* is a strong candidate to be involved in the etiology of the trait. Another well-supported case corresponds to the *TMEM72* locus, associated with red blood cell counts, for which we identify 6 mCpGs linked to *TMEM72* regulatory regions. The *TMEM72* transmembrane protein is strongly and differentially expressed in ductal cells of the kidney [92], which plays a major role in red blood cell homeostasis [93].

### Integration of trait-linked genetically-regulated methylated loci with multi-context eQTL maps

We detect a considerable amount of GWAS hits colocalizing with at least one mQTL but with no eQTLs (Fig. 4). We hypothesize that many of these observed mQTL-specific colocalizations are attributable to one of two scenarios. First, DNAm may be genetically co-regulated with gene expression only in a particular context (e.g. cell type, early developmental state) which causally impacts the trait; but only methylation - not gene expression - QTLs are identifiable beyond the causal context (e.g., non-causal cell type, adulthood), where expression may be absent or regulated differently. This hypothesis is compatible with many mQTLs residing in chromatin regions that are developmentally regulated [28]. Second, it is possible that in bulk tissue samples, mQTLs can be more easily detected than eQTLs, potentially due to low RNA abundance or quality. To identify genes involved in trait-linked mCpGs for which no corresponding eGene is found in GTEx-matching tissues, we integrated mQTL-derived colocalization results with 59 non-GTEx eQTL maps. As shown (see Fig. 4a), considering GTEx-matching tissues, we identified 1,505 mQTL-specific colocalizations, 1,461 of which derived from primary mQTLs. For each of these loci, we performed GWAS-mQTL-eQTL multivariate colocalization integrating the 61 non-GTEx eQTL maps as described in Methods. At prior.2=0.98 and empirical (eQTL-permutation, see Methods) false positive rate < 0.05, we identified 824/1,461 (56%) GWAS-mQTL-eQTL colocalizations for previously defined mQTL-specific loci. As expected, colocalization findings are sensitive to the prior.2 choice: 50-75% of tested loci appear colocalized depending on the prior (prior.2=0.95, 0.98, 0.99, false positive rate not considered). Among context-specific eQTL colocalizations, we identified a *CD24* association with asthma exclusively in T-cell eQTLs (although mCpG colocalizations are found in multiple tissues), eQTL colocalizations for *NSFM* with cholesterol traits in stimulated monocytes and a *JAK2* association with granulocyte abundances and inflammatory bowel diseases (IBDs). *JAK* inhibitors have been recently approved to treat IBDs [94]. Colocalization findings are provided in Supplementary Table 8.

### Experimental design limitations

Across different analyses, results are subjected to experimental design limitations, mainly derived from sample size, partial coverage of the genome, and analysis of tissues in bulk. The number of mCpGs and eQTMs detected per tissue, and the abundance of tissue-specific mCpGs and eQTMs, are strongly correlated with per-tissue sample size (see sections "Characterization of eQTM tissue-specificity", "Characterization of tissue specificity patterns of mQTLs and eQTLs"). Importantly, associations are adjusted for cellular heterogeneity but are mostly representative of abundant cell types within a given tissue type. Observations are also constrained to the targeted array-based DNAm-measuring approach employed: only ~3% of existing CpG sites [48,95], mostly located in putative gene regulatory regions, are analyzed; eQTLs are more comprehensively captured [4]. Hence, identification of TF-mQTL links is limited here by the incomplete, promoter-enriched set of CpG sites profiled; multi-tissue whole-genome mQTL studies are expected to provide a more accurate representation of TF-mQTL links compared to targeted DNAm-measuring approaches. Observations of trait-mQTL links are also limited by array-based DNAm profiling; integration of GWAS signal with DNAm derived from genome-wide sequencing-based technologies can provide a more complete representation of trait-mQTL links and enable haplotypic and allele-specific analyses of DNAm in disease contexts [44,96,97] .

## Supplementary References

76. van der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
77. Zhu, T. *et al.* A pan-tissue DNA methylation atlas enables in silico decomposition of human tissue methylomes at cell-type resolution. *Nat. Methods* **19**, 296–306 (2022).
78. Teschendorff, A. E., Zhu, T., Breeze, C. E. & Beck, S. EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. *Genome Biol.* **21**, 221 (2020).
79. Kim-Hellmuth, S. *et al.* Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, (2020).
80. Oliva, M. *et al.* The impact of sex on gene expression across human tissues. *Science* **369**, (2020).
81. Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics* **18**, 105 (2017).
82. Melé, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
83. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
84. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
85. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9440–9445 (2003).
86. Min, J. L. *et al.* Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat. Genet.* **53**, 1311–1321 (2021).
87. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
88. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *Am. J. Hum. Genet.* **98**, 1114–1129 (2016).
89. Wagner, J. R. *et al.* The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol.* **15**, R37 (2014).
90. Wang, Y. *et al.* Roles of Distal and Genic Methylation in the Development of Prostate Tumorigenesis Revealed by Genome-wide DNA Methylation Analysis. *Sci. Rep.* **6**, 22051 (2016).
91. Yu, Y., Wang, L. & Gu, G. The correlation between Runx3 and bronchial asthma. *Clin. Chim. Acta* **487**, 75–79 (2018).
92. Lindgren, D. *et al.* Cell-Type-Specific Gene Programs of the Normal Human Nephron Define Kidney Cancer Subtypes. *Cell Rep.* **20**, 1476–1489 (2017).
93. Babitt, J. L. & Lin, H. Y. Mechanisms of anemia in CKD. *J. Am. Soc. Nephrol.* **23**, 1631–1634 (2012).

94. Rogler, G. Efficacy of JAK inhibitors in Crohn's Disease. *Journal of Crohn's and Colitis* vol. 14 S746–S754 (2020).
95. Lövkvist, C., Dodd, I. B., Sneppen, K. & Haerter, J. O. DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic Acids Res.* **44**, 5123–5132 (2016).
96. Bell, C. G. *et al.* Obligatory and facilitative allelic variation in the DNA methylome within common disease-associated loci. *Nat. Commun.* **9**, 8 (2018).
97. Abante, J., Fang, Y., Feinberg, A. P. & Goutsias, J. Detection of haplotype-dependent allele-specific DNA methylation in WGBS data. *Nat. Commun.* **11**, 5238 (2020).