nature genetics

ARTICLES

https://doi.org/10.1038/s41588-021-00923-x

Check for updates

Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation

Characterizing genetic influences on DNA methylation (DNAm) provides an opportunity to understand mechanisms underpinning gene regulation and disease. In the present study, we describe results of DNAm quantitative trait locus (mQTL) analyses on 32,851 participants, identifying genetic variants associated with DNAm at 420,509 DNAm sites in blood. We present a database of >270,000 independent mQTLs, of which 8.5% comprise long-range (*trans*) associations. Identified mQTL associations explain 15–17% of the additive genetic variance of DNAm. We show that the genetic architecture of DNAm levels is highly polygenic. Using shared genetic control between distal DNAm sites, we constructed networks, identifying 405 discrete genomic communities enriched for genomic annotations and complex traits. Shared genetic variants are associated with both DNAm levels and complex diseases, but only in a minority of cases do these associations reflect causal relationships from DNAm to trait or vice versa, indicating a more complex genotype–phenotype map than previously anticipated.

The role of common interindividual variation in DNAm on disease mechanisms has not yet been well characterized. It has, however, been hypothesized that DNAm serves as a viable biomarker for risk stratification, early disease detection, and the prediction of disease prognosis and progression[1]. As genetic influences on DNAm in blood are widespread[2–4], a powerful avenue for studying the functional consequences of differences in DNAm levels is to map genetic differences associated with population-level variation, identifying mQTLs that include both local (*cis*-mQTL) and distal (*trans*-mQTL) effects. We can harness mQTLs as natural experiments, allowing the observation of randomly perturbed DNAm levels in a manner that is not confounded with environmental factors[5,6]. In this regard, mapping even very small genetic effects on DNAm is valuable for gaining power to evaluate whether its variation has a substantial causal role in disease and other biological processes.

To date, only a small fraction of the total genetic variation estimated to influence DNAm across the genome has been identified[7], and the proportion of *trans* heritability explained by *trans*-mQTLs (defined as variants >1 Mb from the DNAm site) is much smaller than the proportion of *cis* heritability explained by *cis*-mQTLs. Therefore, most genetic effects are likely to act in *trans*, with small effect sizes[5,7–9], while being potentially biologically informative[8,10]. Much larger sample sizes are required to map associations involving small genetic effects to permit greater understanding of the genetic architecture and the biological processes underlying DNAm[7]. To this end, we established the Genetics of DNA Methylation Consortium (GoDMC), an international collaboration of human epidemiological studies that comprises >30,000 study participants with genetic and DNAm data.

We use this resource to develop a comprehensive catalog of *cis*- and *trans*-mQTLs, which enables us to examine the genetic architecture of DNAm. By constructing networks of multiple *cis*- and *trans*-mQTLs, we learn about their collective impact on pathways and complex traits. Finally, we interrogate the potential role of DNAm in disease mechanisms by mapping the causal relationships between variable DNAm and 116 complex traits and diseases in a bidirectional manner. A database of our results is available as a resource to the community at http://mqtldb.godmc.org.uk.

Results

Genetic variants influence 45% of tested DNAm sites. To map genetic influences on DNAm, we established an analysis workflow that enabled standardized meta-analysis and data integration across 36 population-based and disease datasets. Using a two-phase discovery study design, we analyzed ~10 million genotypes imputed to the 1000 Genomes reference panel[11] and quantified DNAm at 420,509 sites using Illumina HumanMethylation BeadChips in whole blood derived from 27,750 European participants (Fig. 1a, Supplementary Figs. 1–4, Extended Data Fig. 1, Supplementary Tables 1 and 2, and Supplementary Note). The microarray technology used in most cohorts limited us to the analysis of only 1.5% of the ~28 million DNAm sites across the genome[12], including 96% of CpG islands and CpG shores and 99% of RefSeq genes[13], and all inferences relate only to these sites.

Using linkage disequilibrium (LD) clumping, we identified 248,607 independent *cis*-mQTL associations ($P < 1 \times 10^{-8}$, <1 Mb from the DNAm site; Supplementary Fig. 3) with a median distance between single nucleotide polymorphisms (SNPs) and DNAm sites of 36 kb (interquartile range (IQR) = 118 kb; Extended Data Fig. 2). We found 23,117 independent *trans*-mQTL associations (using a conservative threshold of $P < 1 \times 10^{-14}$ (ref. [7]); Supplementary Fig. 3 and Supplementary Note). These mQTLs involved 190,102 DNAm sites, representing 45.2% of all those tested (Fig. 1b) which is a 1.9× increase of sites with a *cis* association ($P < 1 \times 10^{-8}$) and a 10× increase of sites with a *trans* association ($P < 1 \times 10^{-14}$) over a previous study with a sample size that was 7× smaller[8]. As expected, mQTL effect sizes for each DNAm site (the maximum absolute additive change in DNAm level measured in s.d. per allele) were lower for sites with a *trans* association (compared with sites with a *cis* association (per allele s.d. change = −0.02 (s.e. = 0.002, $P = 2.1 \times 10^{-14}$)); Extended Data Fig. 3). The differential improvement in yield between *cis* and *trans* associations is revealing in terms of the genetic architecture—relatively small sample sizes are sufficient to uncover most large *cis* effects, whereas much larger sample sizes are required to identify the polygenic *trans* component.

Most *trans* associations (80%) were interchromosomal. Of the intrachromosomal *trans* associations, 34% were >5 Mb from the DNAm site (Extended Data Fig. 2a). We found a substantially lower number of interchromosomal *trans* associations per 5-Mb region

A full list of affiliations appears at the end of the paper.

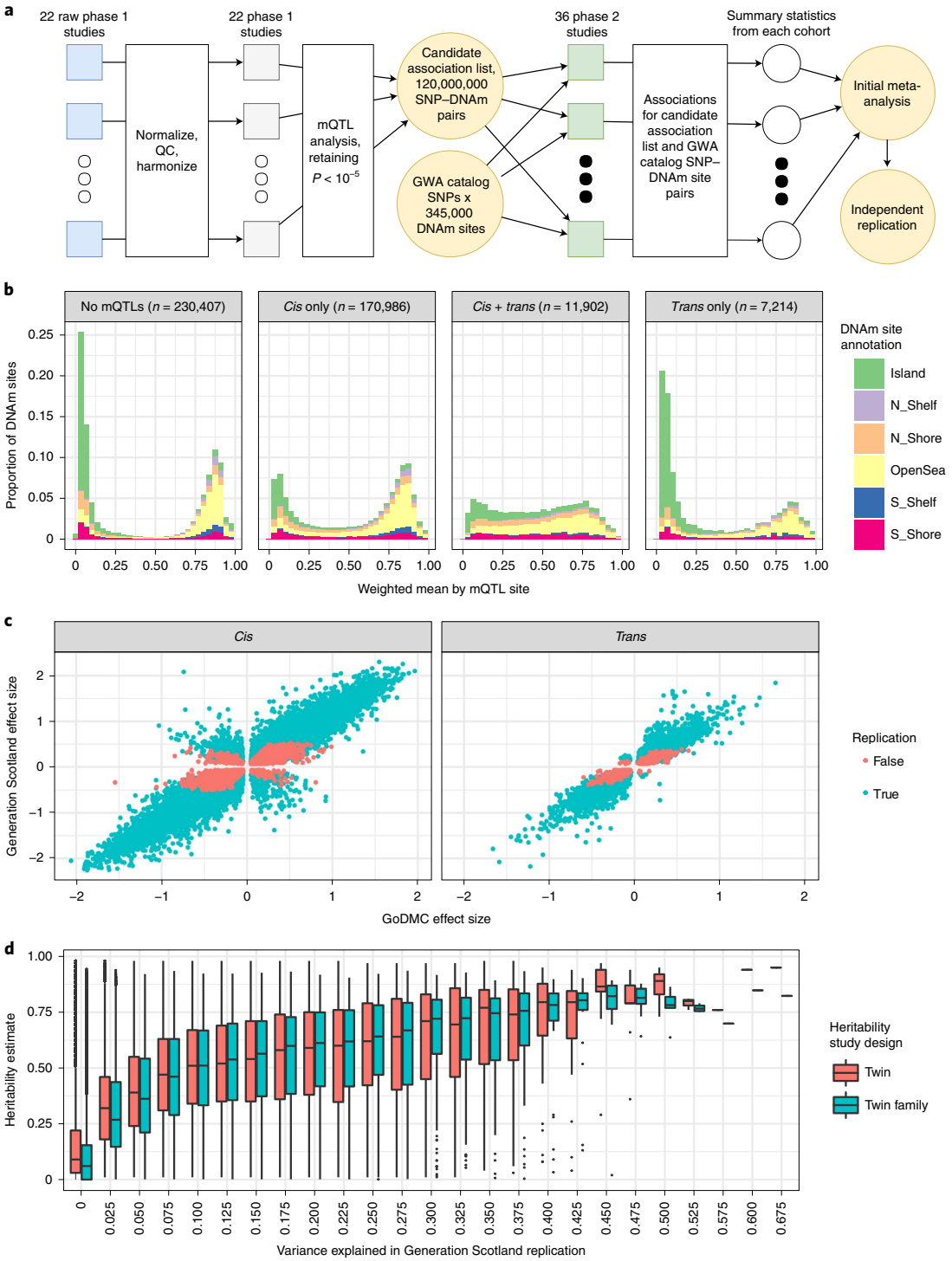


Fig. 1 | Discovery and replication of mQTLs. a, Study design. In the first phase, 22 cohorts performed a complete mQTL analysis of up to 480,000 sites against up to 12 million variants, retaining their results for $P < 1 \times 10^{-5}$. In the second phase, 120 million SNP–DNAm site pairs selected from the first phase, and GWA catalog SNPs against 345,000 DNAm sites, were tested in 36 studies (including 20 phase 1 studies) and meta-analyzed. QC, quality control. **b**, Distributions of the weighted mean of DNAm across 36 cohorts for *cis*-only, *cis* + *trans* and *trans*-only sites. The weighted mean DNAm level across 36 studies was defined as low (<20%), intermediate (20–80%) or high (>80%). Plots are colored with respect to the genomic annotation. *Cis*-only sites showed a bimodal distribution of DNAm. *Cis* + *trans* sites showed intermediate levels of DNAm. *Trans*-only sites showed low levels of DNAm. **c**, Discovery and replication effect size estimates between GoDMC ($n = 27{,}750$) and Generation Scotland ($n = 5{,}101$) for 169,656 mQTL associations. The regression coefficient is 1.13 (s.e. = 0.0007). **d**, Relationship between DNAm site heritability estimates and DNAm variance explained in Generation Scotland. The center line of a boxplot corresponds to the median value. The lower and upper box limits indicate the first and third quartiles (the 25th and 75th percentiles). The length of the whiskers corresponds to values up to 1.5× the IQR in either direction. The regression coefficient for the twin family study was 3.16 (s.e. = 0.008) and for the twin study 2.91 (s.e. = 0.008) across 403,353 DNAm sites. The variances explained for DNAm sites with missing r^2 ($n = 277{,}428$) and/or $h^2 = 0$ (twin family: $n = 80{,}726$; twins: $n = 34{,}537$) were set to 0.

(1.59) than intrachromosomal associations (>1 Mb: 7.95; >6 Mb: 4.81, excluding chromosome 6).

Next, using conditional analysis[14] we explored the potential for multiple independent SNPs operating within the locus of each mQTL, identifying 758,130 putative independent variants. Each DNAm site, for which a mQTL in *cis* had been detected, had a median of two independent variants (IQR = 4 variants; Supplementary Fig. 5). For all subsequent analyses, we used index SNPs from clumping procedures to be conservative and unbiased due to the nonindependence of genetic variants.

We sought to replicate the mQTLs in the Generation Scotland cohort ($n = 5{,}101$) using an independent analysis pipeline. Replication data were available for 188,017 of our discovery mQTLs (137,709 sites). We found a strong correlation of effect sizes for both *cis* and *trans* effects (Pearson's $r = 0.97$, $n = 155{,}191$ and 0.96, $n = 14{,}465$, at $P < 1 \times 10^{-3}$, respectively; Fig. 1c); 99.6% of the associations had a consistent direction of effect (Supplementary Note). At Bonferroni's corrected threshold of 0.05/188,017, 142,727 of the discovery mQTLs replicated in the Generation Scotland cohort (76%); the replication rates for *cis*- and *trans*-mQTLs were 76% and 79%, respectively. To evaluate whether our replication rate was in line with expectations given the smaller replication sample size, we estimated that, under the assumption that the discovery mQTLs are true positives, 171,824 mQTLs would be expected to replicate at a nominal threshold of $P < 1 \times 10^{-3}$; we found that the actual number of mQTLs replicating at this level was 169,656, indicating that most of our discovery mQTLs are likely to be true positives (Supplementary Data 1 and Supplementary Note). Our findings indicate that there is little between-study heterogeneity in our analysis and that genetic effects on DNAm are relatively stable across samples of European ancestry (Extended Data Fig. 1 and Supplementary Table 2).

Overall, the variance explained by replicated genetic effects on DNAm was small. For 99% of the associations in *cis* and *trans*, mQTLs explained <21% and <16% of the variation in DNAm, respectively (Supplementary Fig. 6). Aggregating across all 420,509 tested DNAm sites, our replicated mQTL associations explain 1.3% of the total assayed variation in DNAm, 8% of this being due to *trans* associations. Restricting to sites that have at least one *cis* or *trans* effect, however, we explain 4.2% and 2.5% of the DNAm variance, respectively.

We then investigated how much of the heritability of variable DNAm can be explained by our mQTL associations using family-based heritability studies of DNAm[2,15]. We found a strong positive relationship between variance explained by replication mQTL estimates (127,680 sites in Generation Scotland) and heritability for both studies (family: Pearson's $r = 0.41$ across 121,582 available sites; twin: Pearson's $r = 0.37$ across 118,955 available sites; Fig. 1d and Supplementary Data 2). The mQTLs that we identified explain 15–17% of the additive genetic variance of DNAm (Supplementary Fig. 7). Finally, there were strong positive relationships between the heritability of DNAm levels at a DNAm site and the number of independent mQTLs (Supplementary Fig. 8), heritability and effect size (Supplementary Fig. 9), variance explained and the number of independent mQTLs (Supplementary Fig. 10), and variance explained and distribution of DNAm levels (Supplementary Fig. 11). Overall, our results support a mixed genetic architecture of polygenic genome-wide effects and larger *cis* effects.

Our mQTL coverage was limited by the computational necessity of a multiple-stage study design (Extended Data Fig. 4a). The discovered mQTLs with $r^2 < 1\%$ are probably a small fraction of all the mQTLs in this category expected to exist. Across these DNAm sites, and within the range of mQTLs detected in our study ($r^2 > 0.22\%$), we estimate that there are twice as many *cis*-mQTLs and 22.5 times more *trans*-mQTLs yet to be discovered (Extended Data Fig. 4b). This would probably not explain all estimated heritability, indicating that a substantial set of the heritability is due to either causal variants with smaller effects or rare variants.

***Cis*- and *trans*-mQTLs operate through distinct mechanisms.** To infer biological properties of *trans* features that were independent of any incidental *cis* effects[7,8,16–18], we categorized mQTLs into those only associated with DNAm in *cis* ($n = 157{,}095$, 69.9%), those only associated with DNAm in *trans* ($n = 794$, 0.35%) and those associated with DNAm in both *cis* and *trans* ($n = 66{,}759$, 29.7%). Similarly, of the 190,102 DNAm sites influenced by an SNP, 170,986 DNAm sites (89.9%) were *cis* only, 11,902 DNAm sites (6.3%) were *cis* + *trans* and 7,214 DNAm sites (3.8%) were *trans* only.

We first compared the distributions of DNAm levels (weighted mean DNAm level across 36 studies (Fig. 1b)). We then performed enrichment analyses on the mQTL SNPs and DNAm sites using 25 combinatorial chromatin states from 127 cell types[19] and gene annotations (Fig. 2a, Supplementary Figs. 12–15 and Supplementary Tables 3–6). Consistent with previous studies[7,8,18], we found that *cis*-only sites are represented in high (32%), low (28%) and intermediate (40%) DNAm levels, and these sites are mainly enriched for enhancer chromatin states (mean odds ratio (OR) = 1.37), CpG islands (OR = 1.25) and shores (OR = 1.26). For *cis* + *trans* sites, we found that most (66%) have intermediate DNAm levels. By replicating this finding in two isolated white-blood-cell subsets (Supplementary Fig. 16), we showed that this is due to cell-to-cell variability[19,20] or subcell-type differences. In line with the observation that intermediate levels of DNAm are found at distal regulatory sequences[21,22], these *cis* + *trans* sites were enriched for enhancer (mean OR = 1.65) and promoter states (mean OR = 1.41). However, for *trans*-only sites, we found a pattern of low DNAm (for 55% of sites) and enrichments for promoter states (mean OR = 1.39), especially TssA (active transcription start site (TSS)) promoter state (mean OR = 2.03). These enrichment patterns were consistent if we restricted to only interchromosomal associations (Supplementary Note and Supplementary Fig. 17).

Analyzing the differences in properties for the SNP categories, we found that *cis*-only and *cis* + *trans* SNPs were enriched for active chromatin states and genic regions, whereas *trans*-only SNPs were enriched for intergenic regions and the heterochromatin state (Fig. 2a, Supplementary Figs. 14 and 15, and Supplementary Tables 5 and 6). Overall, these results highlight that a complex relationship between molecular features underlies the mQTL categories and the biological contexts are substantially different between *cis* and *trans* features.

We found that these inferences were often shared across other tissues. DNAm sites with low or intermediate DNAm levels have similar DNAm distributions in 12 tissues (Supplementary Figs. 18–20) with stronger enrichments in blood datasets for the enhancer states, indicating some level of tissue specificity for mQTLs in these regions (Supplementary Figs. 12, 14 and 21).

To investigate whether mQTLs are tissue specific, we compared the correlation of effect estimates of *cis*- and *trans*-mQTLs in blood against adipose tissue ($n = 603$)[23] and brain ($n = 170$)[9] (Supplementary Note and Extended Data Fig. 5). We found a larger extent of QTL sharing of blood and adipose tissue compared with blood and brain, which might be explained by shared cell types, in line with *cis*-expression (e)QTL findings[24]. Generally, the between-tissue effect correlations were high, consistent with a recent comparison of *cis*-mQTL effects between brain and blood[25]. However, we found that the highest correlations were for associations involving *trans*-only sites (adipose $r_b = 0.92$ (s.e. = 0.004); brain $r_b = 0.88$ (s.e. = 0.009)) despite having on average smaller effect sizes than *cis*-only associations, implying that they are less tissue specific than *cis* effects (adipose $r_b = 0.73$ (s.e. = 0.002); brain $r_b = 0.59$ (s.e. = 0.004)), which is in line with the notion that DNAm of promoters is less tissue specific. Stratifying the mQTL categories to low, intermediate and high DNAm showed that the brain–blood correlations are the lowest for intermediate DNAm categories and adipose–blood correlations are lowest for high DNAm categories, which may suggest cellular heterogeneity for high DNAm levels

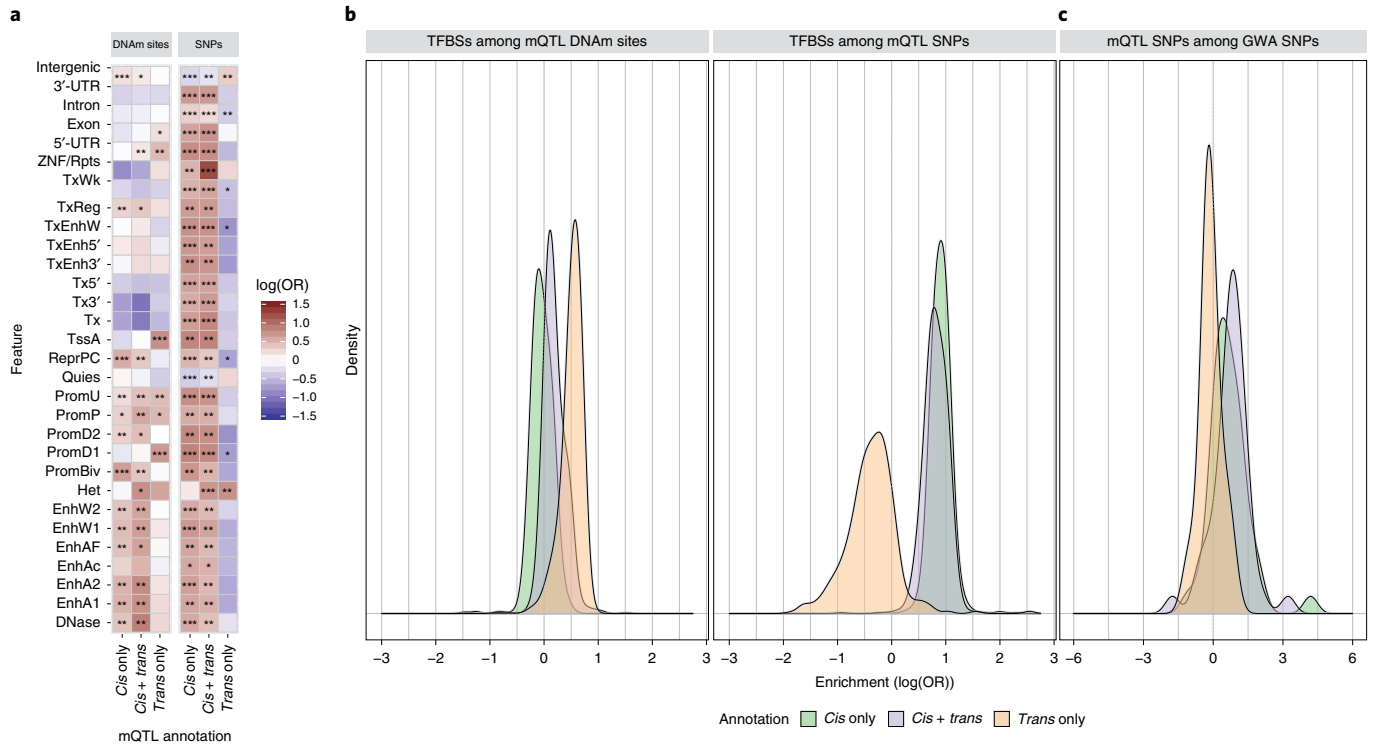


Fig. 2 | *Cis*- and *trans*-mQTLs operate through distinct mechanisms. a, Distributions of enrichments for chromatin states and gene annotations among mQTL sites and SNPs. Enrichment analyses were performed using 25 combinatorial chromatin states from 127 cell types (including 27 blood cell types) and gene annotations. The heatmap represents the distribution of ORs for *cis*-only, *trans*-only or *cis* + *trans* sites and SNPs. For the enrichment of chromatin states, ORs were averaged across cell types. The following chromatin states were analyzed: TssA; PromU, promoter upstream of TSS; PromD1, promoter downstream of TSS with DNase; PromD2, promoter downstream of TSS; Tx5′, transcription 5′; Tx, transcription; Tx3′, transcription 3′; TxWk, weak transcription; TxReg, transcription regulatory; TxEnh5′, transcription 5′-enhancer; TxEnh3′, transcription 3′-enhancer; TxEnhW, transcription weak enhancer; EnhA1, active enhancer 1; EnhA2, active enhancer 2; EnhAF, active enhancer flank; EnhW1, weak enhancer 1; EnhW2, weak enhancer 2; EnhAc, enhancer acetylation only; DNase, DNase only; ZNF/Rpts, *ZNF* genes and repeats; Het, heterochromatin; PromP, poised promoter; PromBiv, bivalent promoter; ReprPC, repressed polycomb, Quies, quiescent/low. The significance was categorized as: *FDR < 0.001; **FDR < 1×10^{-10}; ***FDR < 1×10^{-50} **b**, Distributions of enrichment for occupancy of TFBSs among mQTL sites and SNPs. Each density curve represents the distribution of ORs for *cis*-only, *trans*-only or *cis* + *trans* sites (left) and SNPs (right). **c**, Distributions of enrichment of mQTLs among 41 complex traits and diseases. Each density curve represents the distribution of ORs for *cis*-only, *trans*-only or *cis* + *trans* SNPs.

(Extended Data Fig. 5). These results show the value of large sample sizes in blood to detect *trans*-mQTLs regardless of the tissue.

***Trans*-mQTL SNPs and DNAm exhibit patterned TF binding.** Recent studies have uncovered multiple types of transcription factor (TF)–DNA interactions influenced by DNAm, including the binding of DNAm-sensitive TFs[26–28] and cooperativity between TFs[27,29]. To gain insights into how SNPs induce long-range DNAm changes, we mapped enrichments for DNAm sites and SNPs across binding sites for 171 TFs in 27 cell types[30,31]. We found strong enrichments for most TFs and cell types among DNAm sites with a *trans* association (*cis* + *trans*: 55%; *trans* only: 80%; *cis* only: 18%) and among *cis*-acting SNPs (*cis* only: 96%, *cis* + *trans*: 91%, *trans* only: 1%; Fig. 2b, Supplementary Tables 7 and 8, and Supplementary Figs. 22 and 23). Consistent with the observation that *trans*-only DNAm sites are enriched for CpG islands (Supplementary Fig. 13), DNAm sites that overlap TF-binding sites (TFBSs) were relatively hypomethylated (weighted mean DNAm levels = 21% versus 52%, $P < 2.2 \times 10^{-16}$; Supplementary Fig. 24).

Next, we hypothesized that, if a *trans*-mQTL is driven by TF activity[8,10], then particular TF–TF pairs may exhibit preferential enrichment[32]. An mQTL has a pair of TFBS annotations[31], one for the SNP and one for the DNAm site. We evaluated whether the annotation pairs among 18,584 interchromosomal *trans*-mQTLs were associated with TF binding in a nonrandom pattern (Supplementary Note and Extended Data Fig. 6a,b). We found that 6.1% (22,962 of 378,225) of possible pairwise combinations of SNP–DNAm site annotations were more over- or underrepresented than expected by chance after strict multiple testing correction (Supplementary Note, Supplementary Table 9 and Extended Data Fig. 6c).

After accounting for abundance and other characteristics, the strongest pairwise enrichments involved sites close to TFBSs for proteins in the cohesin complex, for example, CTCF, SMC3 and RAD21, as well as TFs such as GATA2 related to cohesin[33]. Bipartite analysis showed that these clustered due to being related to similar sets of SNP annotations (Extended Data Fig. 6d). Other clusters were also found, for example, sites close to TFBSs for interferon regulatory factor 1 (*IRF1*), a gene for which *trans*-acting regulatory networks[34], and enrichment among causally interacting chromatin accessibility QTLs[35], have been previously reported to be more likely to be influenced by SNPs close to TFBSs for EZH2, SMC3, ATF3, BCL3, TR4 and MAX.

Next, we compared the locations of interchromosomal *trans*-mQTLs ($n = 18{,}584$) to known regions of chromatin interactions[36] as an alternative mechanism for *trans* coordination[8,37]. We found 1,175 overlaps for 637 SNP–DNAm site pairs (3.4%), where the LD region of the mQTL SNP and the corresponding site overlapped with any interacting regions (525 SNPs, 602 sites), compared

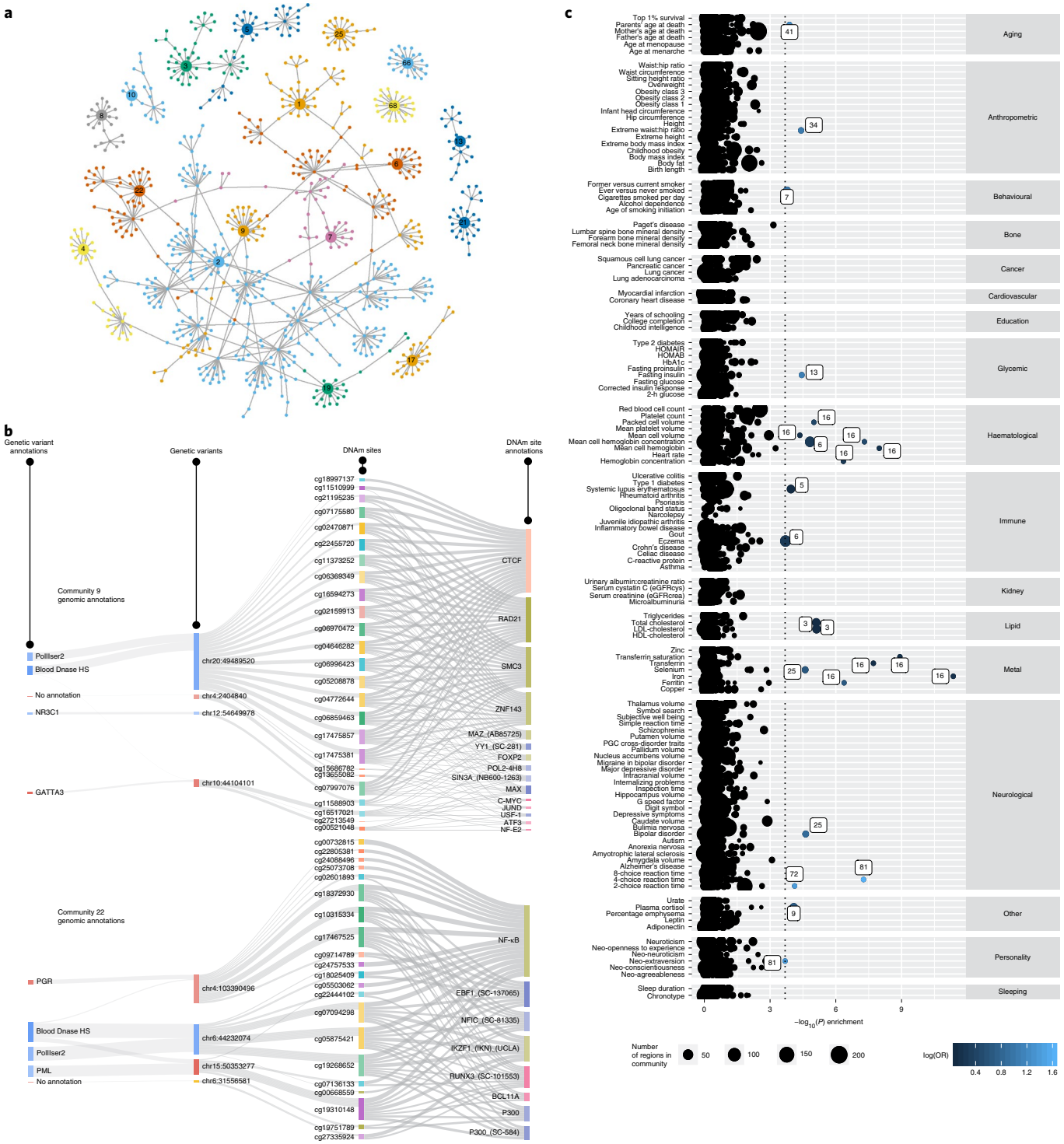


Fig. 3 | Communities constructed from *trans*-mQTLs. a, A network depicting all communities in which there were ≥20 sites. Random walks were used to generate communities (colors), so occasionally a DNA site connects different communities. **b**, The interrelationship of genomic annotations, mQTLs and communities. Communities 9 and 22 comprised DNAm sites that are related through shared genetic factors. The Sankey plots show the genomic annotations for the genetic variants (left) and the DNAm sites (right). The DNAm sites comprising these communities are enriched for TFBSs related to the cohesin complex and NF-κB, respectively. **c,** Enrichment of GWA traits among community SNPs. The genomic loci for each of the 56 largest communities were tested for enrichment of low *P* values in 133 complex trait GWASs (*y* axis) against a null background of community SNPs. The *x* axis depicts the two-sided $-\log_{10}(P \text{ value})$ for enrichment, with the 5% FDR shown by the vertical dotted line. Colors represent log(OR) values. Enrichments were particularly strong for blood-related phenotypes (including circulating metal levels).

with a mean of 473 SNP–DNAm site pairs in 1,000 permuted datasets (OR = 1.36, Fisher's $P = 6.5 \times 10^{-7}$ and empirical $P < 1 \times 10^{-3}$; Supplementary Fig. 25). To summarize, our results show that *trans*-mQTLs are in part driven by long-range cooperative TF interactions and that, for a small proportion of interchromosomal *trans*-mQTLs, the spatial distance in vivo is likely to be small.

***Trans*-mQTL effects form DNAm communities.** Genetic variation can perturb chromatin activity[32,35,37], DNAm[8] or gene expression[38] across multiple sites in *cis* and *trans*, revealing coordinated activity between regulatory elements and genes. We observed that there were 1,728,873 instances where a SNP acting in *trans* also associated with a *cis*-DNAm site (before LD pruning). Genetic co-localization analysis indicated that 278,051 of these instances were due to the *cis* and *trans* sites sharing a genetic factor, representing 3,573 independent *cis*–*trans* genomic region pairs, of which 3,270 were interchromosomal (Supplementary Table 10; see Supplementary Note for sensitivity analysis for the co-localization method used in the context of the two-stage mQTL discovery design). These pairs consisted of 1,755 independent SNPs and 5,109 independent DNAm sites across the genome, indicating that some sites with *cis* associations shared genetic factors with multiple sites with *trans* associations, revealing distal coordination between mQTLs. From the *cis*–*trans* pairs we constructed a network linking these genomic regions which elucidated 405 'communities' of genomic regions that were substantially connected (Supplementary Note); 56 of these communities comprised ≥10 sites, and the largest community comprised 253 sites (Fig. 3a).

We hypothesized that *cis* sites were causally influencing multiple *trans* sites within their communities. We evaluated whether the estimated causal effect (obtained from the *trans*-mQTL effect divided by the *cis*-mQTL effect, that is, Wald's ratio) of the *cis* site on the *trans* site was consistent with the observational correlation between the *cis* and *trans* sites. Although there was an association, the relationship was weak (Pearson's $r=0.096$, $P=1.73\times10^{-6}$; Supplementary Fig. 26), indicating that changes in *cis* sites causing changes in *trans* sites are probably not the predominant mechanism. We did observe that the *cis*–*trans* DNAm levels were more strongly correlated than we would expect by chance (Supplementary Fig. 27), suggesting that they are jointly regulated without generally being causally related.

Next, we evaluated whether DNAm sites within each community were enriched for regulatory annotations and/or gene ontologies (Supplementary Tables 11–14 and Supplementary Figs. 28 and 29). Multiple communities showed enrichments (false discovery rate (FDR) < 0.001); community 9 DNAm sites were strongly enriched for TFBS annotations relating to the cohesin complex in multiple cell types, community 22 DNAm sites were enriched for nuclear factor κ-light-chain-enhancer of activated B cells (NF-κB) and EBF1 in B lymphocytes, and community 76 DNAm sites were enriched for EZH2 and SUZ12, and bivalent promotor and repressed polycomb states (Fig. 3b). Community 2 (comprising 253 sites) was enriched for active enhancer state in three cell types and for lymphocyte activation (Gene Ontology (GO), accession no. GO:0046649; FDR = 0.016) and multiple KEGG pathways including the JAK–STAT signaling pathway (accession no. I04630; $FDR=8.53\times10^{-7}$; Supplementary Tables 13 and 14).

Regulatory features within a network may share a set of biological features that are related to complex traits. We performed enrichment analysis to evaluate whether the loci tagged by DNAm sites in a community were related to each of 133 complex traits (Supplementary Table 15), accounting for nonrandom genomic properties of the selected loci. Restricting the analysis to only the 56 communities with ≥10 sites, we found 11 communities that tagged genomic loci enriched for small *P* values with 22 complex traits (FDR < 0.05; Fig. 3c and Supplementary Table 16). Blood-related phenotypes were overrepresented (11 of 23 enrichments being related to metal levels or hematological measures; binomial test $P=4.2\times10^{-5}$). Among the communities enriched for genome-wide association study (GWAS) signals, community 16 was highly associated with iron and hemoglobin traits, and community 9 was associated with plasma cortisol ($P=8.27\times10^{-5}$). Finally, we performed enrichment analysis on 36 blood cell count traits[39]. We found that community 16 was enriched for hematocrit ($P=4.34\times10^{-10}$) and hemoglobin concentration ($P=1.99\times10^{-8}$) and community 5 was enriched for reticulocyte traits ($P=1.67\times10^{-6}$; Supplementary Fig. 30). The enrichments found for these DNAm communities indicate that a potentially valuable utility of mapping *trans*-mQTL is to indicate how distal regions of the genome are functionally related.

DNAm and complex traits share genetic factors. Most GWA loci map to noncoding regions[40] and *cis*-mQTLs are enriched among GWAs[17,41,42]. In the present study, we investigated the value of the large number of mQTLs, especially *trans*-mQTLs, for annotating the functional consequences of GWA loci. We first compared distributions of enrichment of *cis*- and *trans*-mQTL categories among 41 complex traits. After accounting for nonrandom genomic distribution of mQTLs[43] and multiple testing, we identified enrichments for 35% of the complex traits, particularly for studies with a larger number of GWA signals (Supplementary Fig. 31, Supplementary Table 17 and Supplementary Note). The distribution of enrichment effect estimates (ORs) of *trans*-mQTLs was substantially closer to the null or in depletion when compared with mQTLs that included *cis* effects (Fig. 2c). These enrichments correspond to the results reported earlier, in which *trans*-SNPs were typically depleted for enhancer and promoter regions, whereas complex trait loci are enriched for coding and regulatory regions[44].

Although the mQTL discovery pipeline adjusted for predicted cell types[45,46] and nongenetic DNAm principal components (PCs), there is a possibility that residual cell-type heterogeneity remains. We performed another set of GWAS enrichment analyses, this time using 36 blood cell traits[39], and found enrichments. These were strongest among *cis* + *trans* mQTLs, as seen in the previous enrichments (Supplementary Fig. 32). For 98.9–100% of the mQTLs, mQTL SNPs explained more variation in DNAm than variation in blood cell counts, suggesting a causal chain of mQTLs to blood trait[47]. Alternatively, a systematic measurement error difference could explain these observations, where DNAm captures blood cell counts more accurately than conventional measures.

We next searched for instances of specific DNAm sites sharing the same genetic factors against each of 116 complex traits and diseases, and initially found 23,139 instances of an mQTL strongly associating with a complex trait (Fig. 4). To evaluate the extent to which these were due to shared genetic factors (and not, for example, LD between independent causal variants), we performed genetic co-localization analysis[48] (Supplementary Tables 15 and 18). Excluding genetic variants in the *MHC* region, we found 1,373 potential examples in which at least one DNAm site putatively shared a genetic factor with at least 1 of 71 traits (including 19 diseases). Those DNAm sites that had a shared genetic factor with a trait were 6.9× more likely to be present in a community compared with any other DNAm site with a known mQTL (Fisher's exact test 95% confidence interval 4.8–9.7, $P=9.2\times10^{-19}$). Next, we evaluated how often the DNAm site that co-localized with a known GWAS hit was the closest DNAm site to the lead GWAS variant by physical distance. Notably, in only 18.1% of the cases where a GWAS signal and an assayed 450,000 DNAm site co-localized was that DNAm site the closest DNAm site to the signal. This finding is similar to results found for gene expression[49], but the converse has been found for protein levels[50].

It has previously been difficult to conclude whether genetic co-localization between DNAm and complex traits indicates (1) a causal relationship where the DNAm level is on the pathway from genetic variant to trait (vertical pleiotropy) or (2) a noncausal relationship where the variant influences the trait and DNAm independently through different pathways (horizontal pleiotropy)[51]. In Mendelian randomization (MR), it is reasoned that, under a causal model, multiple independent genetic variants influencing DNAm should exhibit consistent causal effects on the complex trait[52]. Among the putative co-localizing signals, 440 (32%) involved a

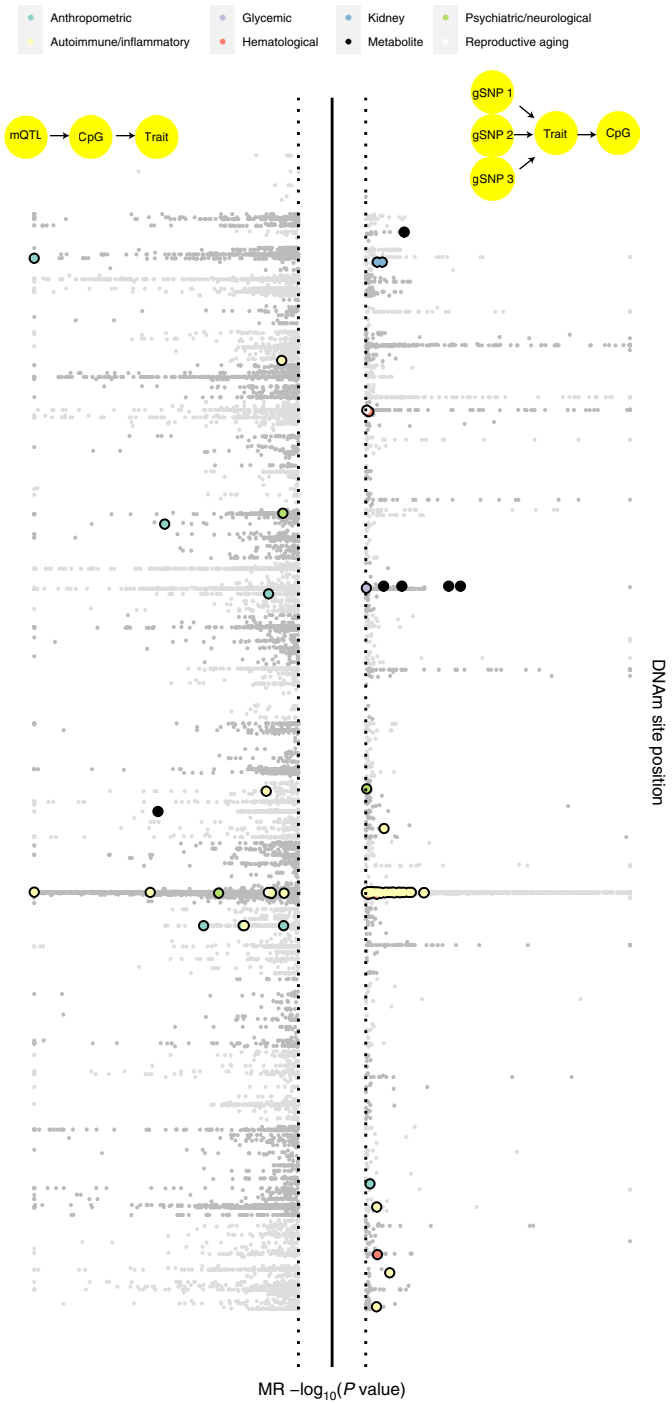


Fig. 4 | Identifying putative causal relationships between sites and traits using bidirectional MR. Aggregated results from a systematic bidirectional MR analysis between DNAm sites and 116 complex traits. The *y* axis represents the two-sided *P* value from MR analysis. The top plot depicts results from tests of DNAm sites co-localizing with complex traits. The light-gray points represent MR estimates that either did not surpass multiple testing or shared small *P* values at both the DNAm site and the complex trait, but had weak evidence of co-localization. Bold, colored points are those that showed strong evidence for co-localization (posterior probability > 0.8 for H4—one shared SNP for DNAm and trait). The bottom plot shows the two-sided $-\log_{10}$(*P* values) from MR analysis of risk factor or genetic liability of disease at DNAm levels. Extensive follow-up was performed on DNAm site–trait pairs with putative associations, and those that pass filters are plotted in bold and colored according to the trait category. A substantial number of MR results in both directions exhibited very strong effects but failed to withstand sensitivity analyses.

DNAm site that had at least one other independent mQTL. We cannot determine with certainty the causal relationship of any specific site with a trait. To test whether there was a general trend for DNAm sites causally influencing a trait, we evaluated whether the MR effect estimate based on the co-localizing signals was consistent with that obtained based on the secondary signals. There were substantially larger genetic effects of the secondary mQTLs on respective traits than expected by chance (70 with $P<0.05$, binomial test $P=2.4\times10^{-16}$). However, only 41 (59%) of these had effect estimates in the same direction as the primary co-localizing variant, which is not substantially better than chance (binomial test $P=0.19$). Of the 41 mQTLs, 12 were located in the *HLA* region. Of the remaining mQTLs, 27 were associated with anthropometric (*ESR1* and birth weight), immune response (*IRF5* and systemic lupus erythematosus) and lipid traits (*TBL2* and triglycerides). We then performed systematic co-localization analysis of all mQTLs against 36 blood cell traits[39]. In the present study, we discovered 94,738 instances of a DNAm site and a blood cell trait sharing a causal variant. In 28,138 instances, the co-localizing DNAm site had an independent secondary mQTL, and with these associations we again tested for a general trend of DNAm sites causally influencing the blood trait. The association between independent signals was very weak ($R^2=0.008$). Together, across the sites that were analyzable in this manner, these results indicate that those blood-measured DNAm sites that have shared genetic factors with traits cannot be typically thought of as mediating the genetic association with the trait (Extended Data Fig. 7 and Supplementary Table 19). Instead, if DNAm is a co-regulatory phenomenon then the co-localizing signals between DNAm sites and complex traits may be due to a common cause, for example, genetic variants primarily acting on TF binding[8,10].

The influence of traits on DNAm variation. Previous studies have not been adequately powered to estimate the causal influences of complex traits on DNAm variation through MR, because the sample size of the outcome variable (DNAm) is a predominant factor in statistical power[48,53]. We systematically analyzed 109 traits for causal effects on DNAm using two-sample MR[54,55], where each trait was instrumented using SNPs obtained from their respective, previously published GWAS (Supplementary Note and Supplementary Table 15). Included among the traits were 35 disease traits, which when used as exposure variables in MR must be interpreted in terms of the influence of liability rather than presence/absence of disease. The sample size used to estimate SNP effects in DNAm was up to 27,750 (Fig. 4).

We initially identified 4,785 associations where risk factors or genetic liability to disease influences DNAm levels (multiple testing threshold $P<1.4\times10^{-7}$). However, causal inference on -omic variables can lead to false positives due to violations in the MR assumptions. We developed a filtering process involving a new causal inference method to help protect against these invalid associations (Supplementary Note and Supplementary Fig. 33). This left 85 associations (involving 84 DNAm sites) in which DNAm sites were putatively influenced by 13 traits (9 risk factors or 4 diseases) (Supplementary Table 20). Further filtering, which would exclude traits that were predominantly instrumented by variants in the *HLA* region or driven by one SNP, would reduce the total number of associations substantially from 84 to 19. We replicated five associations for triglycerides influencing DNAm sites near *CPTA1* and *ABCG1* (ref. [56]) and found associations for transferrin saturation/iron influencing DNAm sites near *HFE*.

We next evaluated whether there was evidence for small, widespread changes in DNAm levels in response to complex trait variation, by calculating the genomic control inflation factor (GC_{in}) for the *P* values obtained from the MR analyses of each trait against all DNAm sites. Five traits (fasting glucose, age at menarche, cigarettes smoked per day, immunoglobulin (Ig) G index levels and serum

creatinine) showed GC_{in} values >1.05 (Extended Data Fig. 8). GC_{in} calculations were performed at each chromosome singly for each trait (Supplementary Fig. 34) and in a leave-one-chromosome-out analysis (Supplementary Fig. 35). The GC_{in} remained consistent (except for IgG index levels), indicating that the traits have small but widespread influences on DNAm levels across the genome.

Although most of the traits ($n = 105$, 96%) tested did not appear to induce genome-wide enrichment, this does not rule out the possibility that they have many localized small effects. For example, the smallest MR P value for the analysis of body mass index on DNAm levels was 2.27×10^{-6}, which did not withstand genome-wide multiple testing correction, and the GC_{in} was 0.95. However, restricting GC_{in} to 187 sites known to associate with body mass index from previous epigenome-wide association studies (EWASs)[20] indicated a strong enrichment of low P values (median $GC_{in} = 3.95$). A similar pattern was found for triglycerides, in which genome-wide median $GC_{in} = 0.94$ but the ten sites known to associate with triglycerides from previous EWAS[57] had an MR P value of 8.3×10^{-70} (Fisher's combined probability test). These results indicate that traits causally influencing DNAm levels in blood is the most likely mechanism to give rise to these EWAS hits. It also indicates that the general finding that there were very few filtered putative causal effects of risk factors, or genetic liability to disease on DNAm, could be due to true positives being generally very small, even to the extent that our sample size of up to 27,750 individuals was insufficient to find them.

Discussion

A map of hundreds of thousands of genetic associations has enabled new biological insights related to DNAm variation. Using a rigorous analytical framework enabled us to minimize heterogeneity and expand sample sizes for large -omic data. This revealed a genetic architecture of DNAm that is polygenic. Given the diverse ranges of age, sex proportions and geographical origins between the cohorts in this analysis, the minimal extent of heterogeneity across datasets indicates that genetic effects on DNAm are relatively stable across contexts, at least when restricted to European ancestries. We show that *cis*- and *trans*-mQTLs operate through distinct mechanisms, because their genomic properties are distinct. A driver of long-range associations may be co-regulated through TF binding and nuclear organization.

Although we found substantial sharing of genetic signals between DNAm sites and complex traits, we were able to demonstrate that this was not predominantly due to DNAm variation being on the causal path from genotype to phenotype. Although our results were restricted to 1.5% of the DNAm sites in the genome, and limited by the two-phase design, these findings have several implications, especially in the context of EWAS studies that are often based on the same tissue and DNAm array. First, we anticipate that some previously reported EWAS associations are probably due to reverse causation, for example, the risk factor or genetic liability to disease state itself alters DNAm and not vice versa, or to confounding. Second, the genetic effects on DNAm that overlap with complex traits probably primarily influence other regulatory factors, which in turn influence complex traits and DNAm through diverging pathways. Third, DNAm might be on the causal pathway in a disease-relevant cell type or context. Fourth, if the path from genotype to complex traits is nonlinear, for example, involving the statistical interactions between different regulatory features[16], then our results indicate that large individual-level multi-omic datasets will be required to dissect such mechanisms. Higher-density DNAm microarrays[12] or low-cost sequencing technologies[58] will expedite detailed interrogations of enhancer and other regulatory regions. Given our projection of mQTL yields expected for future studies, pleiotropy involving mQTLs is likely to be increasingly important to model when interpreting genotype-trait pathways.

Overall, our data and results present a comprehensive atlas of genetic effects on DNAm. We expect that this atlas will be of use to the scientific community for studies of genome regulation and causality analysis, and that it will contribute to the control of confounding in EWAS.

Online content

Any methods, additional references, Nature Research reporting summaries, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00923-x.

Received: 13 October 2020; Accepted: 12 July 2021;
Published online: 6 September 2021

References

1. Petronis, A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* **465**, 721–727 (2010).
2. van Dongen, J. et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat. Commun.* **7**, 11115 (2016).
3. Hannon, E. et al. Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS Genet.* **14**, e1007544 (2018).
4. Kerkel, K. et al. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat. Genet.* **40**, 904–908 (2008).
5. Schadt, E. E. et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
6. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–R98 (2014).
7. Gaunt, T. R. et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.* **17**, 61 (2016).
8. Bonder, M. J. et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).
9. Hannon, E. et al. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat. Neurosci.* **19**, 48–54 (2016).
10. Hop, P. J. et al. Genome-wide identification of genes regulating DNA methylation using genetic anchors for causal inference. *Genome Biol.* **21**, 220 (2020).
11. Abecasis, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
12. Pidsley, R. et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).
13. Bibikova, M. et al. High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).
14. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
15. Shah, S. et al. Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res.* **24**, 1725–1733 (2014).
16. Gutierrez-Arcelus, M. et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2**, e00523 (2013).
17. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414.e24 (2016).
18. McRae, A. F. et al. Identification of 55,000 replicated DNA methylation QTL. *Sci. Rep.* **8**, 17605 (2018).
19. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
20. Wahl, S. et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).
21. Elliott, G. et al. Intermediate DNA methylation is a conserved signature of genome regulation. *Nat. Commun.* **6**, 6363 (2015).
22. Feldmann, A. et al. Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genet.* **9**, e1003994 (2013).
23. Grundberg, E. et al. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.* **93**, 876–890 (2013).
24. Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
25. Qi, T. et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.* **9**, 2282 (2018).

26. Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).
27. Domcke, S. et al. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575–579 (2015).
28. Baubec, T. et al. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature* **520**, 243–247 (2015).
29. Ginno, P. A. et al. A genome-scale map of DNA methylation turnover identifies site-specific dependencies of DNMT and TET activity. *Nat. Commun.* **11**, 2680 (2020).
30. Sánchez-Castillo, M. et al. CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res.* **43**, D1117–D1123 (2015).
31. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
32. Waszak, S. M. et al. Population variation and genetic control of modular chromatin architecture in humans. *Cell* **162**, 1039–1050 (2015).
33. Viny, A. D. et al. Dose-dependent role of the cohesin complex in normal and malignant hematopoiesis. *J. Exp. Med.* **212**, 1819–1832 (2015).
34. Battle, A. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
35. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat. Genet.* **51**, 128–137 (2019).
36. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
37. Delaneau, O. et al. Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* **364**, eaat8266 (2019).
38. Võsa, U. et al. Large-scale *cis*- and *trans*-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* https://doi.org/10.1038/s41588-021-00913-z (2021).
39. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429.e19 (2016).
40. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
41. Tachmazidou, I. et al. Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. *Am. J. Hum. Genet.* **100**, 865–884 (2017).
42. Kato, N. et al. *Trans*-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat. Genet.* **47**, 1282–1293 (2015).
43. Iotchkova, V. et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat. Genet.* **51**, 343–353 (2019).
44. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
45. Reinius, L. E. et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS ONE* **7**, e41361 (2012).
46. Houseman, E. A. et al. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinform.* **9**, 365 (2008).
47. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* **13**, e1007081 (2017).
48. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
49. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
50. Zheng, J. et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat Genet.* **52**, 1122–1131 (2020).
51. Richardson, T. G. et al. Systematic Mendelian randomization framework elucidates hundreds of CpG sites which may mediate the influence of genetic variants on disease. *Hum. Mol. Genet.* **27**, 3293–3304 (2018).
52. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* **27**, R195–R208 (2018).
53. Brion, M. J., Shakhbazov, K. & Visscher, P. M. Calculating statistical power in Mendelian randomization studies. *Int. J. Epidemiol.* **42**, 1497–1501 (2013).
54. Pierce, B. L. & Burgess, S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* **178**, 1177–1184 (2013).
55. Hemani, G. et al. The MR-base platform supports systematic causal inference across the human phenome. *Elife* **7**, e34408 (2018).
56. Dekkers, K. F. et al. Blood lipids influence DNA methylation in circulating cells. *Genome Biol.* **17**, 138 (2016).
57. Braun, K. V. E. et al. Epigenome-wide association study (EWAS) on lipids: the Rotterdam study. *Clin. Epigenet.* **9**, 15 (2017).
58. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Josine L. Min[1,2,98] ✉, Gibran Hemani[1,2,98], Eilis Hannon[3], Koen F. Dekkers[4], Juan Castillo-Fernandez[5], René Luijk[4], Elena Carnero-Montoro[5,6], Daniel J. Lawson[1,2], Kimberley Burrows[1,2], Matthew Suderman[1,2], Andrew D. Bretherick[7], Tom G. Richardson[1,2], Johanna Klughammer[8], Valentina Iotchkova[9], Gemma Sharp[1,2], Ahmad Al Khleifat[10], Aleksey Shatunov[10], Alfredo Iacoangeli[10,11], Wendy L. McArdle[2], Karen M. Ho[2], Ashish Kumar[12,13,14], Cilla Söderhäll[15], Carolina Soriano-Tárraga[16], Eva Giralt-Steinhauer[16], Nabila Kazmi[1,2], Dan Mason[17], Allan F. McRae[18], David L. Corcoran[19], Karen Sugden[19,20], Silva Kasela[21], Alexia Cardona[22,23], Felix R. Day[22], Giovanni Cugliari[24,25], Clara Viberti[24,25], Simonetta Guarrera[24,25], Michael Lerro[26], Richa Gupta[27,28], Sailalitha Bollepalli[27,28], Pooja Mandaviya[29], Yanni Zeng[7,30,31], Toni-Kim Clarke[32], Rosie M. Walker[33,34], Vanessa Schmoll[35], Darina Czamara[35], Carlos Ruiz-Arenas[36,37,38], Faisal I. Rezwan[39], Riccardo E. Marioni[33,34], Tian Lin[18], Yvonne Awaloff[35], Marine Germain[40], Dylan Aïssi[41], Ramona Zwamborn[42], Kristel van Eijk[42], Annelot Dekker[42], Jenny van Dongen[43], Jouke-Jan Hottenga[43], Gonneke Willemsen[43], Cheng-Jian Xu[44,45], Guillermo Barturen[6], Francesc Català-Moll[46], Martin Kerick[47], Carol Wang[48], Phillip Melton[49,50,51], Hannah R. Elliott[1,2], Jean Shin[52], Manon Bernard[52], Idil Yet[5,53], Melissa Smart[54], Tyler Gorrie-Stone[55], BIOS Consortium[*,], Chris Shaw[10,56], Ammar Al Chalabi[10,56,57], Susan M. Ring[1,2], Göran Pershagen[12], Erik Melén[12,58],**

Jordi Jiménez-Conde[16], Jaume Roquer[16], Deborah A. Lawlor[1,2], John Wright[17], Nicholas G. Martin[59], Grant W. Montgomery[18], Terrie E. Moffitt[19,20,60,61], Richie Poulton[62], Tõnu Esko[21,63], Lili Milani[21], Andres Metspalu[21], John R. B. Perry[22], Ken K. Ong[22], Nicholas J. Wareham[22], Giuseppe Matullo[24,25], Carlotta Sacerdote[25,64], Salvatore Panico[65], Avshalom Caspi[19,20,60,61], Louise Arseneault[61], France Gagnon[26], Miina Ollikainen[27,28], Jaakko Kaprio[27,28], Janine F. Felix[66,67], Fernando Rivadeneira[29], Henning Tiemeier[68,69], Marinus H. van IJzendoorn[70,71], André G. Uitterlinden[29], Vincent W. V. Jaddoe[66,67], Chris Haley[7], Andrew M. McIntosh[32,34], Kathryn L. Evans[33,34], Alison Murray[72], Katri Räikkönen[73], Jari Lahti[73], Ellen A. Nohr[74,75], Thorkild I. A. Sørensen[1,2,76,77], Torben Hansen[76], Camilla S. Morgen[76,78], Elisabeth B. Binder[35,79], Susanne Lucae[35], Juan Ramon Gonzalez[36,37,38], Mariona Bustamante[36,37,38,80], Jordi Sunyer[36,37,38,81], John W. Holloway[82,83], Wilfried Karmaus[84], Hongmei Zhang[84], Ian J. Deary[34], Naomi R. Wray[18,85], John M. Starr[34,86], Marian Beekman[4], Diana van Heemst[87], P. Eline Slagboom[4], Pierre-Emmanuel Morange[88], David-Alexandre Trégouët[40], Jan H. Veldink[42], Gareth E. Davies[89], Eco J. C. de Geus[43], Dorret I. Boomsma[43], Judith M. Vonk[90], Bert Brunekreef[91,92], Gerard H. Koppelman[44], Marta E. Alarcón-Riquelme[6,12], Rae-Chi Huang[93], Craig E. Pennell[48], Joyce van Meurs[29], M. Arfan Ikram[94], Alun D. Hughes[95], Therese Tillin[95], Nish Chaturvedi[95], Zdenka Pausova[52], Tomas Paus[96], Timothy D. Spector[5], Meena Kumari[54], Leonard C. Schalkwyk[55], Peter M. Visscher[18,85], George Davey Smith[1,2], Christoph Bock[8,97], Tom R. Gaunt[1,2], Jordana T. Bell[5,99], Bastiaan T. Heijmans[4,99], Jonathan Mill[3,99] and Caroline L. Relton[1,2,99]

[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK. [3]University of Exeter Medical School, College of Medicine and Health, University of Exeter, Exeter, UK. [4]Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands. [5]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [6]Pfizer-University of Granada-Andalusian Government Center for Genomics and Oncological Research, Granada, Spain. [7]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [8]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria. [9]MRC Weatherall Institute of Molecular Medicine, Oxford, UK. [10]Department of Basic and Clinical Neuroscience, Maurice Wohl Clinical Neuroscience Institute, London, UK. [11]Department of Biostatistics and Health Informatics, King's College London, London, UK. [12]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [13]Chronic Disease Epidemiology unit, Swiss Tropical and Public Health Institute, Basel, Switzerland. [14]University of Basel, Basel, Switzerland. [15]Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden. [16]Neurology Department, Hospital del Mar, Institut Hospital del Mar d'Investigacions Mèdiques, Barcelona, Spain. [17]Bradford Institute for Health Research, Bradford, UK. [18]Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia. [19]Center for Genomic and Computational Biology, Duke University, Durham, NC, USA. [20]Department of Psychology and Neuroscience, Duke University, Durham, NC, USA. [21]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. [22]MRC Epidemiology Unit, School of Clinical Medicine, Institute of Metabolic Science, University of Cambridge, Cambridge, UK. [23]Department of Genetics, University of Cambridge, Cambridge, UK. [24]Department of Medical Sciences, University of Turin, Turin, Italy. [25]Italian Institute for Genomic Medicine, Turin, Italy. [26]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada. [27]Institute for Molecular Medicine, University of Helsinki, Helsinki, Finland. [28]Department of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [29]Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, the Netherlands. [30]Faculty of Forensic Medicine, Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China. [31]Guangdong Province Key Laboratory of Brain Function and Disease, Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China. [32]Division of Psychiatry, Royal Edinburgh Hospital, University of Edinburgh, Edinburgh, UK. [33]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, Western General Hospital, University of Edinburgh, Edinburgh, UK. [34]Centre for Cognitive Ageing and Cognitive Epidemiology, Department of Psychology, University of Edinburgh, Edinburgh, UK. [35]Department of Translational Research in Psychiatry, Max-Planck-Institute of Psychiatry, Munich, Germany. [36]ISGlobal, Barcelona Global Health Institute, Barcelona, Spain. [37]Universitat Pompeu Fabra, Barcelona, Spain. [38]CIBER Epidemiología y Salud Pública, Madrid, Spain. [39]Department of Computer Science, Aberystwyth University, Aberystwyth, UK. [40]INSERM UMR_S 1219, Bordeaux Population Health Center, University of Bordeaux, Bordeaux, France. [41]Department of General and Interventional Cardiology, University Heart Center Hamburg, Hamburg, Germany. [42]Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands. [43]Department of Biological Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands. [44]University of Groningen, University Medical Center Groningen, Department of Pediatric Pulmonology and Pediatric Allergology, Beatrix Children's Hospital, GRIAC Research Institute Groningen, Groningen, the Netherlands. [45]CiiM and TWINCORE, Hannover Medical School and Helmholtz Centre for Infection Research, Hannover, Germany. [46]Chromatin and Disease Group, Cancer Epigenetics and Biology Programme, Bellvitge Biomedical Research Institute, Barcelona, Spain. [47]Instituto de Parasitología y Biomedicina López Neyra, CSIC, Granada, Spain. [48]School of Medicine and Public Health, College of Health, Medicine and Wellbeing, University of Newcastle, Newcastle, Australia. [49]Menzies Institute for Medical Research, College of Health and Medicine, University of Tasmania, Hobart, Australia.

[50]School of Global Population Health, Faculty of Health and Medical Sciences, University of Western Australia, Perth, Australia. [51]School of Pharmacy and Biomedical Sciences, Faculty of Health Sciences, Curtin University, Perth, Australia. [52]The Hospital for Sick Children, University of Toronto, Toronto, Canada. [53]Department of Bioinformatics, Institute of Health Sciences, Hacettepe University, Ankara, Turkey. [54]Institute for Social and Economic Research, University of Essex, Colchester, UK. [55]School of Life Sciences, University of Essex, Colchester, UK. [56]Department of Neurology, King's College Hospital, London, UK. [57]United Kingdom Dementia Research Institute, King's College London, London, UK. [58]Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden. [59]QIMR Berghofer Medical Research Institute, Brisbane, Australia. [60]Department of Psychiatry and Behavioral Sciences, Duke University Medical School, Durham, NC, USA. [61]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [62]Dunedin Multidisciplinary Health and Development Research Unit, Department of Psychology, University of Otago, Dunedin, New Zealand. [63]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA. [64]Piemonte Centre for Cancer Prevention, Turin, Italy. [65]Dipartimento Di Medicina Clinica E Chirurgia, Federico II University, Naples, Italy. [66]The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands. [67]Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands. [68]Department of Child and Adolescent Psychiatry, Erasmus Medical Center, Rotterdam, the Netherlands. [69]Department of Social and Behavioral Science, Harvard TH Chan School of Public Health, Boston, MA, USA. [70]Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, the Netherlands. [71]Department of Clinical, Educational and Health Psychology, Division on Psychology and Language Sciences, Faculty of Brain Sciences, University College London, London, UK. [72]Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK. [73]Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [74]Research Unit for Gynaecology and Obstetrics, Institute of Clinical research, University of Southern Denmark, Odense, Denmark. [75]Centre of Women's, Family and Child Health, University of South-Eastern Norway, Kongsberg, Norway. [76]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [77]Department of Public Health (Section of Epidemiology), Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [78]The National Institute of Public Health, University of Southern Denmark, Copenhagen, Denmark. [79]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA, USA. [80]Center for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain. [81]Hospital del Mar Medical Research Institute, Barcelona, Spain. [82]Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK. [83]Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, UK. [84]Division of Epidemiology, Biostatistics, and Environmental Health Sciences, School of Public Health, University of Memphis, Memphis, TN, USA. [85]Queensland Brain Institute, University of Queensland, Brisbane, Australia. [86]Alzheimer Scotland Dementia Research Centre, University of Edinburgh, Edinburgh, UK. [87]Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, the Netherlands. [88]C2VN, Aix-Marseille University, INSERM, INRAE, Marseille, France. [89]Avera Institute for Human Genetics, Sioux Falls, SD, USA. [90]University of Groningen, University Medical Center Groningen, Department of Epidemiology, GRIAC Research Institute Groningen, Groningen, the Netherlands. [91]Institute for Risk Assessment Sciences, Universiteit Utrecht, Utrecht, the Netherlands. [92]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands. [93]Telethon Kids Institute, University of Western Australia, Perth, Australia. [94]Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands. [95]UCL Institute of Cardiovascular Science, London, UK. [96]Departments of Psychology and Psychiatry, University of Toronto, Toronto, Canada. [97]Institute of Artificial Intelligence and Decision Support, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria. [98]These authors contributed equally: Josine L. Min, Gibran Hemani. [99]These authors jointly supervised this work: Jordana T. Bell, Bastiaan T. Heijmans, Jonathan Mill, Caroline L. Relton. *A list of authors and their affiliations appears at the end of the paper. **A full list of members and their affiliations appears in the Supplementary Information. ✉e-mail: josine.min@bristol.ac.uk

BIOS Consortium

Marian Beekman[4], Dorret I. Boomsma[43], Jenny van Dongen[43], Diana van Heemst[87], Bastiaan T. Heijmans[4,99], Jouke-Jan Hottenga[43], René Luijk[4], Joyce van Meurs[29], P. Eline Slagboom[4], André G. Uitterlinden[29] and Jan H. Veldink[42]

A full list of members and their affiliations appears in the Supplementary Information.

Methods

Study design overview. Initially, 38 independent studies were recruited to contribute data toward an mQTL meta-analysis, of which 36 studies (Supplementary Table 1 and Supplementary Note) passed our stringent quality criteria, described below. Conventional GWAS meta-analyses involve performing complete GWAS in each study, sharing the summary data and meta-analyzing every tested SNP. As an mQTL analysis involves ~450,000 GWAS analyses, it is difficult to store and share the complete summary data from 38 studies. To circumvent this problem, each study performed GWAS analyses, but provided only the associations that surpass a relaxed significance threshold ($P < 1 \times 10^{-5}$) in their study. Due to sampling variation, the exact mQTL associations reported would differ between studies, meaning that the number of studies contributing to the meta-analysis would be highly variable and could be as low as two studies. This would introduce two problems: first, publication bias arises if it is in fact a null association because the studies demonstrating null effects would not contribute to counteract the inflated effects from those that do happen to surpass the threshold. Second, the precision of the effect estimate is limited by the number of studies that happen to contribute data on that association. To mitigate both problems in the analysis, the present study was performed in two phases.

In phase 1 of our study we performed mQTL analyses of 420,509 high-quality DNAm sites[59] using data from 22 independent European studies to identify putative associations (Supplementary Table 1 and Fig. 1a) at a threshold of $P < 1 \times 10^{-5}$. We used two approaches to exclude DNAm sites from our analyses. First, we excluded 50,186 DNAm sites that were masked by Zhou et al.[59] which includes probes with potential cross-reaction and probes that could not be mapped to a genome. Second, we removed an additional 14,882 probes, including multi-mapping probes (bisulfite-converted sequences allowing two mismatches at any position mapped to the hg19 primary assembly) and probes with variants (minor allele frequency (MAF) >5%, UK10K) at the CpG dinucleotide or the extension base (for type I probes).

All candidate mQTL associations at $P < 1 \times 10^{-5}$ were combined to create a unique 'candidate list' of mQTL associations. In total we identified 102,965,711 candidate mQTL associations in *cis* ($P < 1 \times 10^{-5}$, ±1 Mb from DNAm site) and 710,638,230 candidate mQTL associations in *trans* (>1 Mb from DNAm site) in at least one dataset. In at least two datasets, 59% of the candidate mQTL associations in *cis* ($n = 61{,}103{,}065$) and 2.4% of the associations in *trans* ($n = 17{,}246{,}702$) were found (Supplementary Fig. 1). To reduce the computational burden, we included *cis* associations found in at least one dataset and *trans* associations in at least two datasets. The candidate list ($n = 120{,}212{,}413$) was then sent back to all studies, and the association estimates were obtained for every mQTL association on the candidate list. In phase 2 of our study, we performed association tests for each of the candidate mQTL associations in 20 studies from phase 1 and 16 additional studies with European ancestry (total $n = 27{,}750$; Supplementary Table 1). The estimates for the candidate list were meta-analyzed to obtain the final results (Fig. 1a).

This two-phase approach had a single objective: to minimize the computational burdens of storing summary data from the complete analysis of every study. However, we effectively performed a complete search of all candidate mQTL associations, although with probable loss of coverage. The significant results obtained from the meta-analysis are identical to what would have been identified had we performed a meta-analysis on every candidate mQTL association. The only difference between a complete scan and our scan was that we would have missed some associations that were not at $P < 1 \times 10^{-5}$ in any study, but when combined across all studies would have surpassed an experiment-wide multiple testing correction.

Data preparation. *Participants.* To study the relationship between common genetic variation and DNAm, we focused on studies of European ancestry with genotype data imputed to the 1000 Genomes Project reference panel[11] and DNAm profiles quantified from bisulfite-converted genomic whole-blood DNA using the Infinium HumanMethylation BeadChip (HumanMethylation450 or EPIC arrays). Details of the studies for discovery and replication are provided in Supplementary Table 1 and Supplementary Note.

The GoDMC pipeline. To facilitate the harmonization of the large volume of data, we developed a GoDMC pipeline that was split into several modules, each focusing on the separate tasks of data checking, genotype preparation, phenotype and co-variate preparation, DNAm data preparation and subsequent analyses. In the first module, the data format of the genotype data, DNAm and co-variate data was checked. In addition, the number of individuals with DNAm and genotype data (requirement of $n > 100$), the number of SNPs, the number of sites, co-variates including cell counts, genotype build and strand, and the number of DNAm outliers were recorded. We also generated matrices with mean and s.d. by DNAm site and study descriptives. The entire pipeline can be viewed at https://github.com/MRCIEU/godmc, and the following text describes the procedures used.

Genotype data. Each study performed quality control on genotype data for all autosomes and chromosome X (if available) and imputed to 1000 Genome phase 1 or above using hg19/build37. Dosages were converted to best-guess data without a probability cut-off. SNPs that failed the Hardy–Weinberg equilibrium ($P < 1 \times 10^{-6}$), had an MAF < 0.01, an info score <0.8 or missingness in >5% of the participants were removed. We recoded SNPs to CHR:POS[11] format and removed duplicate SNPs. We then harmonized the recoded SNPs to the 1000 Genomes Project reference using easyQC_v.9.2 (ref. [60]). This harmonization script removed SNPs with mismatched alleles and recoded indel alleles to I and D.

We performed a sex check and removed participants discordant to the co-variate file. We extracted and pruned a set of common HapMap3 SNPs (MAF > 0.2) without long-range LD regions before we calculated the first 20 genetic PCs on LD-pruned SNPs and excluded regions of high LD from the analysis. We used PLINK2.0 (ref. [61]) for unrelated participants and GENESIS[62] for related participants to identify ancestry outliers. Ancestry outliers that deviated by 7 s.d. from the mean were removed. After outlier removal we recalculated genetic PCs for use in subsequent analyses. To identify relatedness in unrelated datasets, we pruned the genotype data to a set of independent HapMap3 SNPs with MAF > 0.01 and calculated genome-wide average identity by state using PLINK2.0. Participants with identity by state >0.125 were removed.

DNAm data normalization and quality control. DNAm was measured in whole blood or cord blood using HumanMethylation450 or EPIC arrays in at least 100 European individuals. Each study performed normalization and quality control on the DNAm data independently, with most studies using functional normalization through the R package meffil v.0.1.0 (ref. [63]) (Supplementary Table 1). Briefly, meffil has been designed to preprocess raw idat files to a normalization matrix for large sample sizes without large computational memory requirements, and to perform quality control in an automated way where the analyst can adjust default parameters easily. Sample quality control included removal of participants in whom >10% of the DNAm sites failed the detection *P* value of 0.1 and/or threshold of three beads. In addition, mismatched samples were identified by comparing the 65 SNPs on the DNAm array with the genotype array and a sex check. Additional DNAm quality was checked by the methylated versus the unmethylated ratio, dye bias using the normalization control probes and bisulfite control probes. Protocols can be found at the following website: https://github.com/perishky/meffil/wiki. For each DNAm site, we replaced outliers that were 10 s.d. from the mean (three iterations) with the DNAm site being the mean.

Co-variates. We used sex, age at measurement, batch variables (slide, plate, row if available), smoking (if available) and recorded cell counts to adjust for possible confounding and to reduce residual variation. Additional confounders (genetic PCs, nongenetic DNAm PCs and, where necessary, predicted smoking and cell counts) were calculated using the GoDMC pipeline. After quality control and normalization of the DNAm data, we predicted smoking status by using previously reported DNAm associations with smoking[64]. In addition, we predicted cell counts using the Houseman algorithm[46] implemented in meffil v.0.1.0 (ref. [63]). We performed a principal component analysis on the 20,000 most variable autosomal DNAm sites and kept all PCs that cumulatively explained 80% of the variance. We performed GWASs on the DNAm PCs and retained the PCs that were not associated with a genotype ($P > 1 \times 10^{-7}$). We kept a maximum of 20 nongenetic PCs for subsequent adjustment.

DNAm data adjustment. We attempted to minimize nongenetic variation in the DNAm data to improve the power for mQTL detection. We adjusted datasets with predominant family structures (pedigrees, twin studies) and population-based studies in slightly different ways. For unrelated participants we regressed out age, sex, predicted cell counts, predicted smoking and genetic PCs (adjustment 1). For related participants, we did the same, except also fitting the genetic kinship matrix using the method described in GRAMMAR[65].

We took the residuals from the first adjustment forward to regress out the nongenetic DNAm PCs on the adjusted DNAm β values (adjustment 2). The residuals from these analyses were rank transformed and centered to have mean 0 and variance 1.

Positive and negative controls. Before we performed the meta-analysis, we checked the number of SNPs and indels, sites and individuals analyzed, and the average mean and s.d. for each DNAm site to identify possible inconsistencies. Each of the 38 studies conducted a GWAS of cg07959070. We chose this DNAm site as a positive control because it showed a strong *cis*-mQTL in several datasets on chr22 and there hasn't been a proposal to exclude it from the analyses by probe annotation efforts[59,66–68]. To identify possible errors, we checked the *cis* association on chromosome 22 ($P < 0.001$) for this DNAm site. In addition, we checked quantile–quantile and Manhattan plots for this DNAm site. We also used this control to identify studies with deflated or inflated λ values ($\lambda > 1.1$ or $\lambda < 0.9$). We noticed deflation of the genomic λ after adjustment of the index *cis*-SNP in datasets with relatedness. However, λ values were around 1 when not adjusted. After inspection one study was removed from the analysis due to deflation and one study due to a lack of the positive control association signal, leaving 36 studies for the final meta-analysis.

Association analyses. *Phase 1: creating the candidate list of associations.* We performed a fast, comprehensive analysis of all *cis* and *trans* associations on

420,509 reliable[59] residualized DNAm sites separately in 22 studies ($n=16{,}907$) using the R package Matrix eQTL v.2.1.0 (ref. [69]). For each DNAm site, j, the residual value, y_{ji}, was regressed against each SNP, k:

$$y_{ji} = \alpha_{jk} + \beta_{jk}x_{ki} + e_{jki}$$

where genotype values x_{ki} were coded as allele counts $\{0, 1, 2\}$, α_{jk} was the intercept term and β_{jk} was the effect estimate of each SNP k on each residualized DNAm site j.

Phase 2: obtaining summary data from all studies for meta-analysis. This candidate list was sent to 36 studies ($n=27{,}750$), in which effect sizes for all putative associations were recalculated by fitting linear models. For putative *cis*-mQTLs we performed linear regression as in phase 1. To improve statistical power to estimate the *trans*-mQTL effects we recorded the top *cis*-SNP x_c, for each DNAm site (based on the lowest P value within that study) and fit this as a co-variate in the *trans*-mQTL regressions:

$$y_{ji} = \alpha_{jk} + \beta_{jc}x_{ci} + \beta_{jk}x_{ki} + e_{jki}$$

Evaluation of DNAm data adjustment. As adjustment for nongenetic DNAm PCs might have substantial benefits on power or an adverse effect by inducing collider bias[70], we explored the impact by comparing mQTLs not adjusted for nongenetic PCs with mQTLs adjusted for nongenetic PCs in ARIES. Specifically, we found 80,890 clumped mQTL associations in the PC-adjusted dataset and 74,402 clumped mQTL associations in the PC-unadjusted dataset. Pearson's correlation between effect sizes of the PC-unadjusted clumped mQTLs versus PC-adjusted mQTLs (*cis* $r=0.998$; *trans* $r=0.998$) and PC-adjusted clumped mQTLs (*cis* $r=0.997$; *trans* $r=0.997$) versus PC-unadjusted mQTLs was very high (Supplementary Fig. 36). These results suggest that, if collider bias is impacting the results, it is extremely small. The simplest explanation for the minimal difference in effect sizes and slightly higher mQTL yield among the PC-adjusted mQTLs is that reduced residual variance has improved the power.

Impact of two-stage design on power of study. Although the multi-stage study design was performed out of practical necessity, we evaluated the impact it had on statistical power compared with the hypothetical situation of analyzing all the data together in a standard one-stage mQTL design.

For *cis*-mQTL associations, we calculated the power of detecting an association in at least 1 of 22 studies at $P<1\times10^{-5}$. To do this, we calculated the probability of missing an association as the product of the probability of missing it in study 1 and study 2 and study 3, and so on:

$$P\,(\text{miss}) = \prod_{i=1}^{M=22} 1 - f\left(x = 19.5;\ k = 1,\ \lambda = n_i r^2\right)$$

where $f(x;k;\lambda)$ is the probability density function for the noncentral χ^2 distribution with k degrees of freedom, and λ the noncentrality parameter based on the postulated variance explained by an mQTL (r^2) and the study sample size n_i, and 19.5 denotes the χ^2 threshold at $P=1\times10^{-5}$ with one degree of freedom.

For *trans*-mQTL associations we calculated the power to detect an association in at least 2 of 22 studies at $P<1\times10^{-5}$. We calculated the probability of missing an association as the product of the probability of missing it in both study 1 and study 2, and in study 1 and study 3, and in study 1 and study 4, and so on:

$$P\,(\text{miss}) = \prod_{i=1}^{M=22}\prod_{j=1}^{i=1} 1 - f\left(x = 19.5;k = 1, \lambda = n_i r^2\right) f\left(x = 19.5;k = 1, \lambda = n_i r^2\right)$$

where $f(x;k;\lambda)$ is the probability density function for the noncentral χ^2 distribution with k degrees of freedom, and λ the noncentrality parameter based on the postulated variance explained by an mQTL (r^2) and the study sample sizes n_i and n_j, and 19.5 denotes the χ^2 threshold at $P=1\times10^{-5}$ with one degree of freedom.

We found that we have no loss of power (<1%) for loci that explain >1.2% or <0.1% of the variance. Within these bounds >80% of power is lost for *cis*-mQTLs with $r^2=0.16$–0.38%. For *trans*-mQTLs, power suffers slightly more because of requiring detection by at least two studies in the first stage ($r^2=0.27$–0.64%; Extended Data Fig. 4a).

Meta-analyses. We used the SNP effect estimates and s.e. for each SNP–DNAm site pair in the candidate list in the meta-analyses. Inverse-variance fixed effect (FE) meta-analyses of the 36 studies were performed using METAL[71]. We modified METAL (https://github.com/explodecomputer/random-metal) to incorporate the DerSimonian and Laird random effect (RE) models[72] and multiplicative random effect (MRE) models[73]. These results are available at: http://mqtldb.godmc.org.uk. We also inspected the meta-analysis and conditional analysis (see below) logfiles and removed any SNPs that had inconsistent allele codes between studies, which were in almost all cases multi-allelic SNPs.

We inspected our results by counting the number of associations against the direction of the effect size (+ or −) for each study. A high number of associations was found if the direction of the effect sizes agreed across studies (Supplementary Fig. 2a). In addition, the average I^2 heterogeneity estimate for the effect size direction categories was 44% (min. = 0%, max. = 100%). For categories with >100 associations, average I^2 was 49% (min. = 36%, max. = 61%; Supplementary Fig. 2b). We also explored whether the number of phase 1 studies was correlated to I^2 and τ^2. We found a nonsignificant correlation ($r=0.002$, $P=0.23$; $r=-0.001$, $P=0.32$), indicating that mQTL associations found in a low number of phase 1 studies did not show more heterogeneity than mQTL associations found in a high number of phase 1 studies.

To explore heterogeneity further, we meta-analyzed our SNP–DNAm pairs using FE, RE and MRE models, and found that associations that were dropped in MRE analyses showed higher I^2 and τ^2 and smaller effect sizes and DNAm site s.d.s (Supplementary Figs. 3 and 4).

Further inspection showed that *trans*-only sites had higher I^2 heterogeneity statistics than associations from *cis*-only or *cis*+*trans* sites (mean I^2 values of 53%, 46% and 39%, respectively). However, as I^2 and τ^2 were positively correlated to effect sizes (Supplementary Fig. 2c), we deemed the use of FE meta-analysis to be appropriate for reducing false-negative rates.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

A database of our results is available as a resource to the community at http://mqtldb.godmc.org.uk. The individual-level genotype and DNAm data are available by request from each individual study or can be downloaded from Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo), European Genome–Phenome Archive (EGA, https://ega-archive.org) or Array Express (https://www.ebi.ac.uk/arrayexpress). As the consent for most studies requires the data to be under managed access, the individual-level genotype and DNAm data are not available from a public repository unless stated.

ALS BATCH1 and -2 data are available to researchers by request as outlined in the Project MinE access policy. ARIES data are available to researchers by request from the Avon Longitudinal Study of Parents and Children Executive Committee (http://www.bristol.ac.uk/alspac/researchers/access) as outlined in the study's access policy http://www.bristol.ac.uk/media-library/sites/alspac/documents/researchers/data-access/ALSPAC_Access_Policy.pdf. BAMSE data are available from the GABRIEL consortium as well as on request in EGA, under accession no. EGAC00001000786. BASICMAR DNAm data are available under accession no. GSE69138. Born-in-Bradford data are available to researchers who submit an expression of interest to the Born-in-Bradford Executive Group (https://borninbradford.nhs.uk/research). BSGS DNAm data are available under accession no. GSE56105. GOYA data are available by request from DNBC: https://www.dnbc.dk. Dunedin data are available via a managed access system (contact: ac115@duke.edu). E-Risk DNAm data are available under accession no. GSE105018. Estonian biobank (ECGUT) data can be accessed on ethical approval by submitting a data release request to the Estonian Genome Center, University of Tartu (http://www.geenivaramu.ee/en/access-biopank/data-access). EPIC-Norfolk data can be accessed by contacting the study management committee: http://www.srl.cam.ac.uk/epic/contact. Requests for EPICOR data accession may be sent to Professor Giuseppe Matullo (giuseppe.matullo@unito.it). FTC data can be accessed on approval from the Data Access Committee of the Institute for Molecular Medicine Finland FIMM (fimm-dac@helsinki.fi). Requests for Generation R data access are evaluated by the Generation R Management Team. Researchers can obtain a de-identified GLAKU dataset after having obtained an approval from the GLAKU Study Board. GSK DNAm data are available under accession no. GSE125105. INMA data are available by request from the INfancia y Medio Ambiente Executive Committee for researchers who meet the criteria for access to confidential data. IOW F2 data are available by request from Isle of Wight Third Generation Study. Please contact Mr Stephen Potter (stephen.potter@iow.nhs.uk). LLS DNAm data were submitted to the EGA under accession no. EGAS00001001077. LBC1921 and LBC1936 data are available on request from the Lothian Birth Cohort Study, Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh (I.Deary@ed.ac.uk). DNAm from MARTHA participants are available under accession no. E-MTAB-3127. NTR DNAm data are available on request in EGA, under the accession no. EGAD00010000887. PIAMA data are available on request. Requests can be submitted to the PIAMA Principal Investigators (https://piama.iras.uu.nl/english). PRECISESADS data are available through ELIXIR at https://doi.org/10.17881/th9v-xt85. Collaboration in data analysis of PREDO is possible through specific research proposals sent to the PREDO Study Board (predo.study@helsinki.fi) or primary investigators Katri Räikkönen (katri.raikkonen@helsinki.fi) or Hannele Laivuori (hannele.laivuori@helsinki.fi). Data are available on request at Project MinE (https://www.projectmine.com). Raine data are available on request (https://ross.rainestudy.org.au). Requests for the data accession of the Rotterdam Study may be sent to Frank van Rooij (f.vanrooij@erasmusmc.nl). SABRE data are available by request from SABRE (https://www.sabrestudy.org). SCZ1 DNAm data are

available under accession no. GSE80417. SCZ2 DNAm data are available under accession no. GSE84727. SYS data are available on request addressed to Dr. Zdenka Pausova (zdenka.pausova@sickkids.ca) and Dr. Tomas Paus (tpausresearch@gmail.com). Further details about the protocol can be found at http://www.saguenay-youth-study.org. TwinsUK DNAm data are available in the GEO under accession nos. GSE62992 and GSE121633. TwinsUK adipose DNAm data are stored in EGA under the accession no. E-MTAB-1866. Access to additional individual-level genotype and phenotype data can be applied for through the TwinsUK data access committee: http://twinsuk.ac.uk/resources-for-researchers/access-our-data. Individual-level DNAm and genetic data from the UK Household Longitudinal Study are available on application through the EGA under accession no. EGAS00001001232. Nonidentifiable Generation Scotland data will be made available to researchers through the GS:SFHS Access Committee. MESA DNAm data are available under accession nos. GSE56046 and GSE56581. Tissue DNAm data are available from accession no. GSE78743. Brain DNAm data can be found under accession no. GSE58885.

Cohort descriptions and further contact details can be found in the Supplementary Note.

For the enrichments, we used chromatin states from the Epigenome Roadmap (https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final), TFBSs from the ENCODE project (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgTfbsUniform) downloaded from the LOLA core database (http://databio.org/regiondb), and gene annotations from https://zwdzwd.github.io/InfiniumAnnotation or GARFIELD (https://www.ebi.ac.uk/birney-srv/GARFIELD). To extract GWA signals for co-localization, we used the MRBase database (https://www.mrbase.org).

Code availability

Datasets were processed using https://github.com/perishky/meffil unless stated otherwise. Individual study analysts used a github pipeline https://github.com/MRCIEU/godmc to conduct the mQTL analysis. We used https://github.com/MRCIEU/godmc_phase1_analysis for the phase 1 analysis, https://github.com/explodecomputer/random-metal for the meta-analyses and https://github.com/MRCIEU/godmc_phase2_analysis for the follow-up analyses.

References

59. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22 (2017).
60. Winkler, T. W. et al. Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
61. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
62. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* **98**, 127–148 (2016).
63. Min, J. L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* **34**, 3983–3989 (2018).
64. Zeilinger, S. et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS ONE* **8**, e63812 (2013).
65. Aulchenko, Y. S., de Koning, D. J. & Haley, C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**, 577–585 (2007).
66. Chen, Y. A. et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).
67. Naeem, H. et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genom.* **15**, 51 (2014).
68. Price, M. E. et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenet. Chromatin* **6**, 4 (2013).
69. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
70. Dahl, A., Guillemot, V., Mefford, J., Aschard, H. & Zaitlen, N. Adjusting for principal components of molecular phenotypes induces replicating false positives. *Genetics* **211**, 1179–1189 (2019).
71. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
72. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control Clin. Trials* **7**, 177–188 (1986).
73. Hedges, L. V. & Olkin, I. *Statistical Methods for Meta-Analysis* 189–203 (Academic Press, 1985).

Acknowledgements

C.L.R., G.D.S., G.S., J.L.M., K.B., M. Suderman, T.G.R. and T.R.G. are supported by the UK Medical Research Council (MRC) Integrative Epidemiology Unit at the University of Bristol (MC_UU_00011/1, MC_UU_00011/4, MC_UU_00011/5). C.L.R. receives support from a Cancer Research UK Programme grant (no. C18281/A191169). G.H. is funded by the Wellcome Trust and the Royal Society (208806/Z/17/Z). E.H. and J.M. were supported by MRC project grants (nos. MR/K013807/1 and MR/R005176/1 to J.M.) and an MRC Clinical Infrastructure award (no. MR/M008924/1 to J.M.). B.T.H. is supported by the Netherlands CardioVascular Research Initiative (the Dutch Heart Foundation, Dutch Federation of University Medical Centres, the Netherlands Organisation for Health Research and Development, and the Royal Netherlands Academy of Sciences) for the GENIUS project 'Generating the best evidence-based pharmaceutical targets for atherosclerosis' (CVON2011-19, CVON2017-20). J.T.B. was supported by the Economic and Social Research Council (grant no. ES/N000404/1). The present study was also supported by JPI HDHL-funded DIMENSION project (administered by the BBSRC UK, grant no. BB/S020845/1 to J.T.B., and by ZonMW the Netherlands, grant no. 529051021 to B.T.H). A.D.B. has been supported by a Wellcome Trust PhD Training Fellowship for Clinicians and the Edinburgh Clinical Academic Track programme (204979/Z/16/Z). J. Klughammer was supported by a DOC fellowship of the Austrian Academy of Sciences. Cohort-specific acknowledgements and funding are presented in the Supplementary Note.

Author contributions

G.H., G.S. and J.L.M. managed the project. A.A.C., A. Caspi, A.D.H., A.G.U, A. Metspalu, A. Murray, A.M.M., B.B., B.T.H., C.H., C.L.R., C.P., C. Sacerdote, C. Shaw, C. Söderhäll, D.A.L., D.v.H., D.I.B., D.-A.T., E.A.N., E.B.B., E.J.C.d.G, E.M., F.G., F.R., G.E.D, G.H.K., G.P., G.W.M., H.R.E., H.T., H.Z., I.J.D., J.F.F., J.H.V., J.J.-C., J. Kaprio, J.L., J.M., J.M.S., J.M.V., J.v.M., J.R., J.R.B.P., J.R.G., J. Shin, J.T.B., J.W., J.W.H., K.K.O., K.L.E., K.R., L.A., L.C.S., L.M., M.A.I., M. Beekman, M. Bustamante, M.E.A.-R., M.H.v.IJ., M. Kerick, M.O., N.C., N.G.M., N.J.W., N.R.W., P.E.S., P.-E.M., P.M.V., R.-C.H., R.P., S.L., S.P., T.D.S., T.E., T.E.M., T.I.A.S, T.P., T.T., V.W.V.J., W.K. and Z.P. designed individual studies and contributed data. A.A.K., A.I., A.S., B.C., C.S.M., H.R.E., J.L.M., K.B., K.M.H., N.K., S.M.R., T.H., R.M.W. and W.L.M. generated and/or quality-controlled data. G.H., J.L.M., M. Suderman, T.R.G. and V.I. designed new statistical or bioinformatics tools. A.D.B., A. Cardona, A.D., A.F.M., A.K., B.T.H., C.B., C.H., C.L.R., C.R.-A., C.S.-T., C.V., C.-J.X., C.W., D.A., D.C., D.J.L., D.L.C., D.M., E.C.-M., E.G.-S., E.H., E.M., F.C.-M., F.I.R., F.R.D., G.B., G.C., G.D.S., G.H., G.H.K., G.M., G.W., I.Y., J.C.-F., J.v.D., J.-J.H., J. Kaprio, J. Klughammer, J.L.M., J.M., J. Sunyer, J.T.B., K.B., K.v.E., K.F.D., K.S., L.C.S., M. Bernard, M. Bustamante, M.H.v.IJ., M.G., M. Kumari, M.L., M. Smart, M. Suderman, N.K., P. Melton, P. Mandaviya, P.M.V., R.E.M., R.G., R.L., R.Z., S.B., S.G., S.K., T.-K.C., T.G.-S., T.G.R., T.I.A.S., T.L., T.R.G., Y.A., Y.Z., V.I. and V.S. analyzed the data and/or provided critical interpretation of results. B.T.H., C.B., C.L.R., J.M., J.T.B. and T.R.G. designed and/or managed the study. A.D.B., B.T.H., C.B., C.L.R., D.J.L., E.C.-M., E.H., G.D.S., G.H., J.C.-F., J. Klughammer, J.L.M., J.M., J.T.B., K.B., K.F.D., M. Suderman, P.M.V., R.L., T.G.R., T.R.G. and V.I. wrote the manuscript.

Competing interests

T.R.G. receives funding from GlaxoSmithKline and Biogen for unrelated research. The other authors declare no competing interests.

Additional information

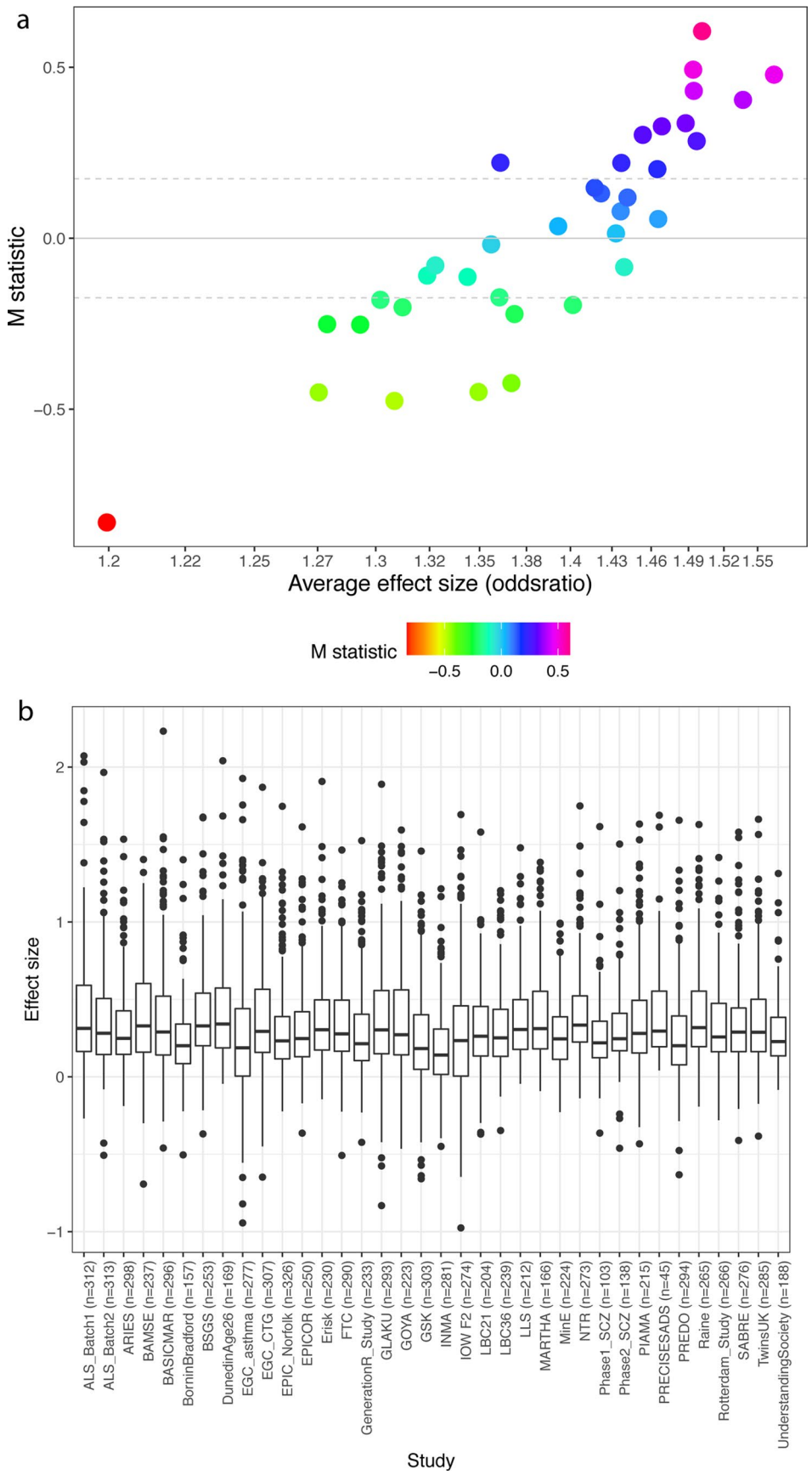
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00923-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00923-x.

Correspondence and requests for materials should be addressed to Josine L. Min.

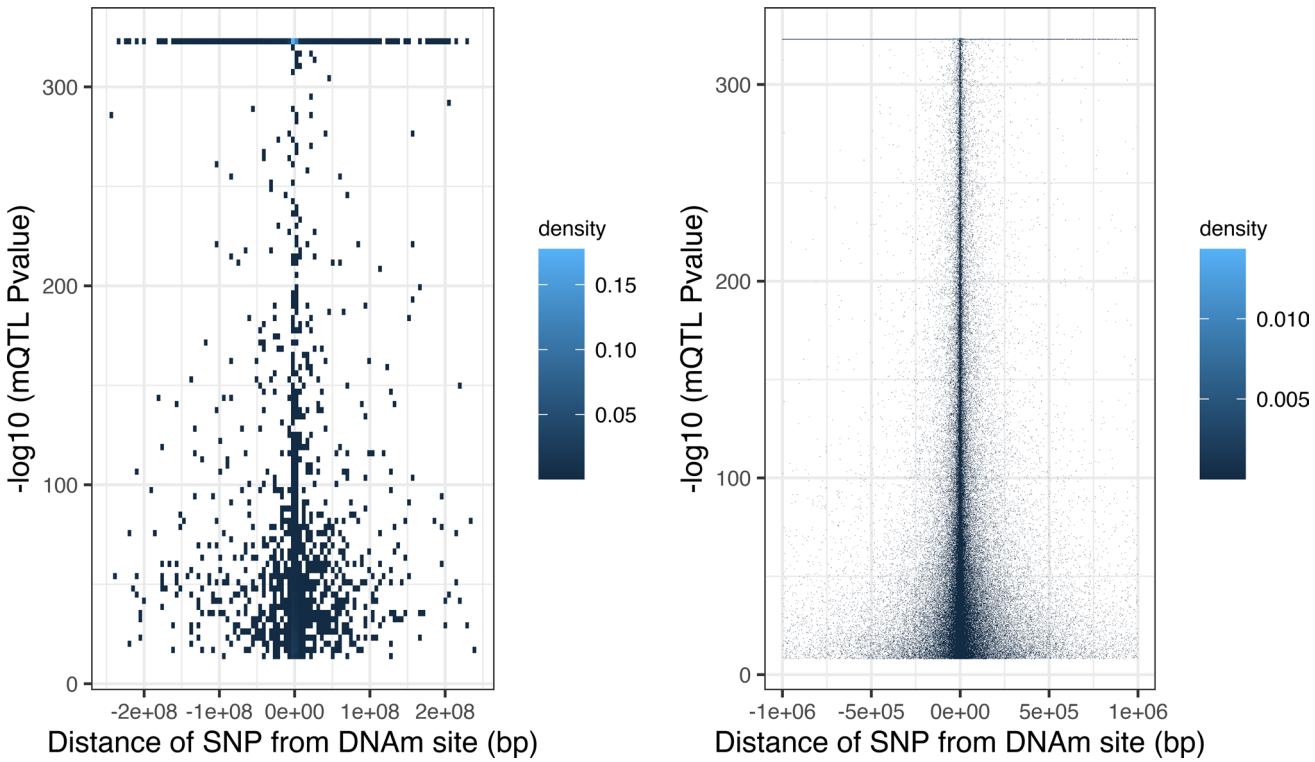
Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer review reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

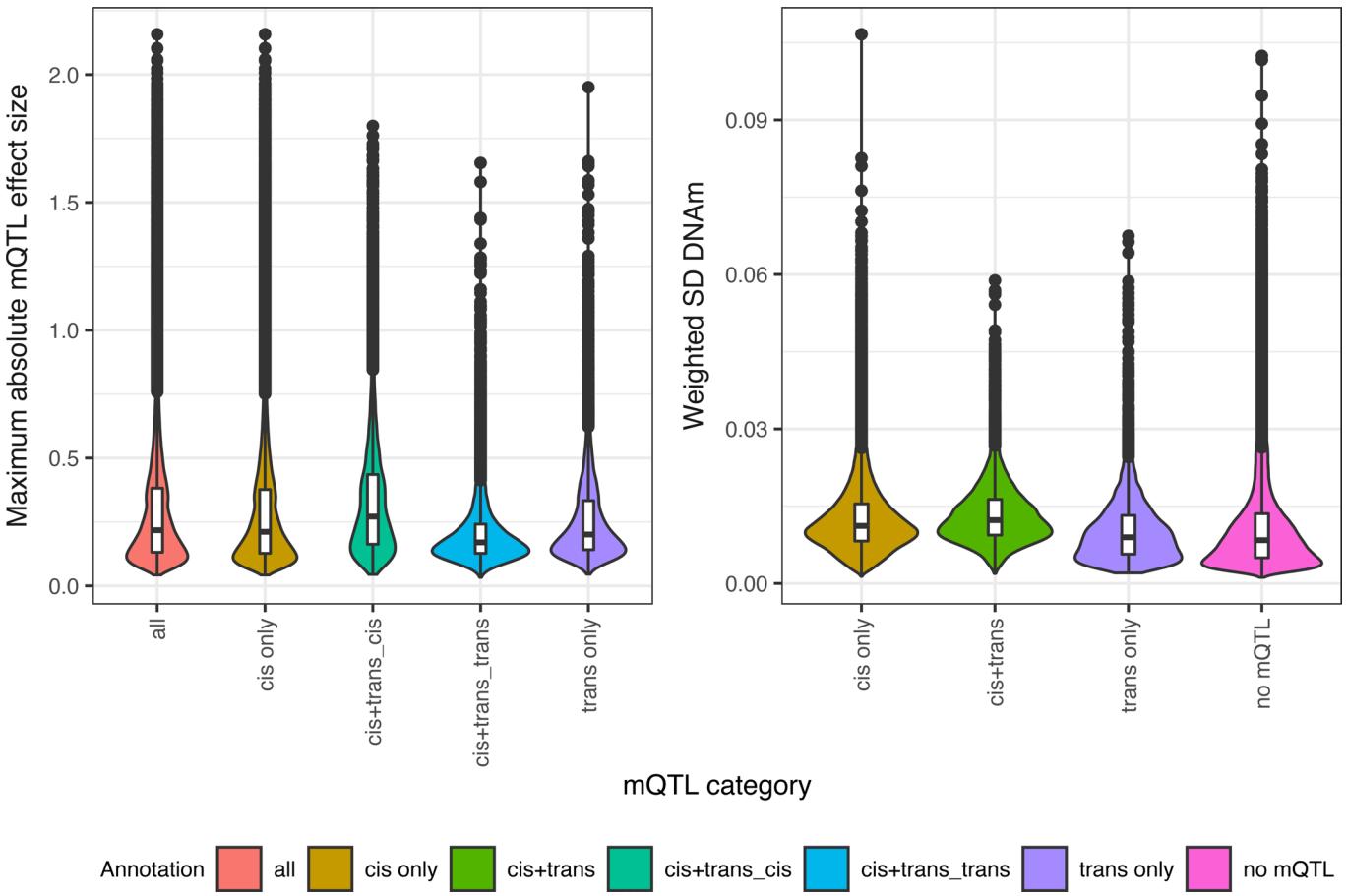


Extended Data Fig. 1 | See next page for caption.

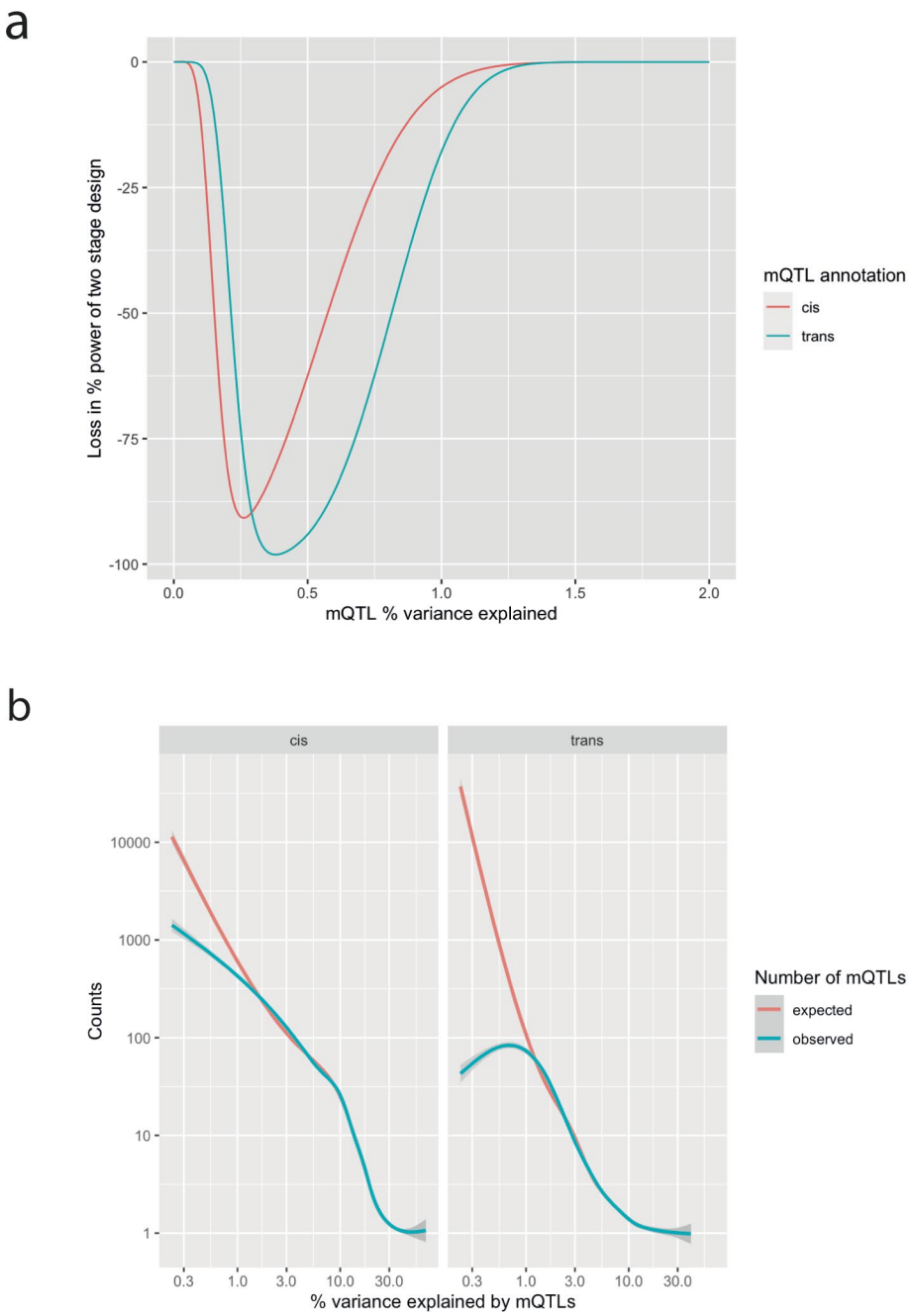
Extended Data Fig. 1 | Quality control of 36 studies. We used 337 independent SNPs on chromosome 20 with a p-value<1e-14. The number of SNPs used for each study are indicated in the bottom plot. **a**, Mstatistic (Magosi et al., *PLoS Genet*., **13**, e1006755 (2017)) for each of the 36 cohorts. **b**, Boxplot of mQTL effect sizes for each of the 36 studies. The center line of a boxplot corresponds to the median value. The lower and upper box limits indicate the first and third quartiles (the 25th and 75th percentiles). The length of the whiskers corresponds to values up to 1.5 times the IQR in either direction.



Extended Data Fig. 2 | Distance of SNP from DNAm site. a, Density plot of the distance of SNP from DNAm site against the -$\log_{10}$ p-value of 4,533 intrachromosomal *trans*-mQTL associations (>1Mb). **b**, Density plot of the distance of SNP from DNAm site against the -$\log_{10}$ p-value of 248,607 *cis*-mQTL associations (<1Mb).



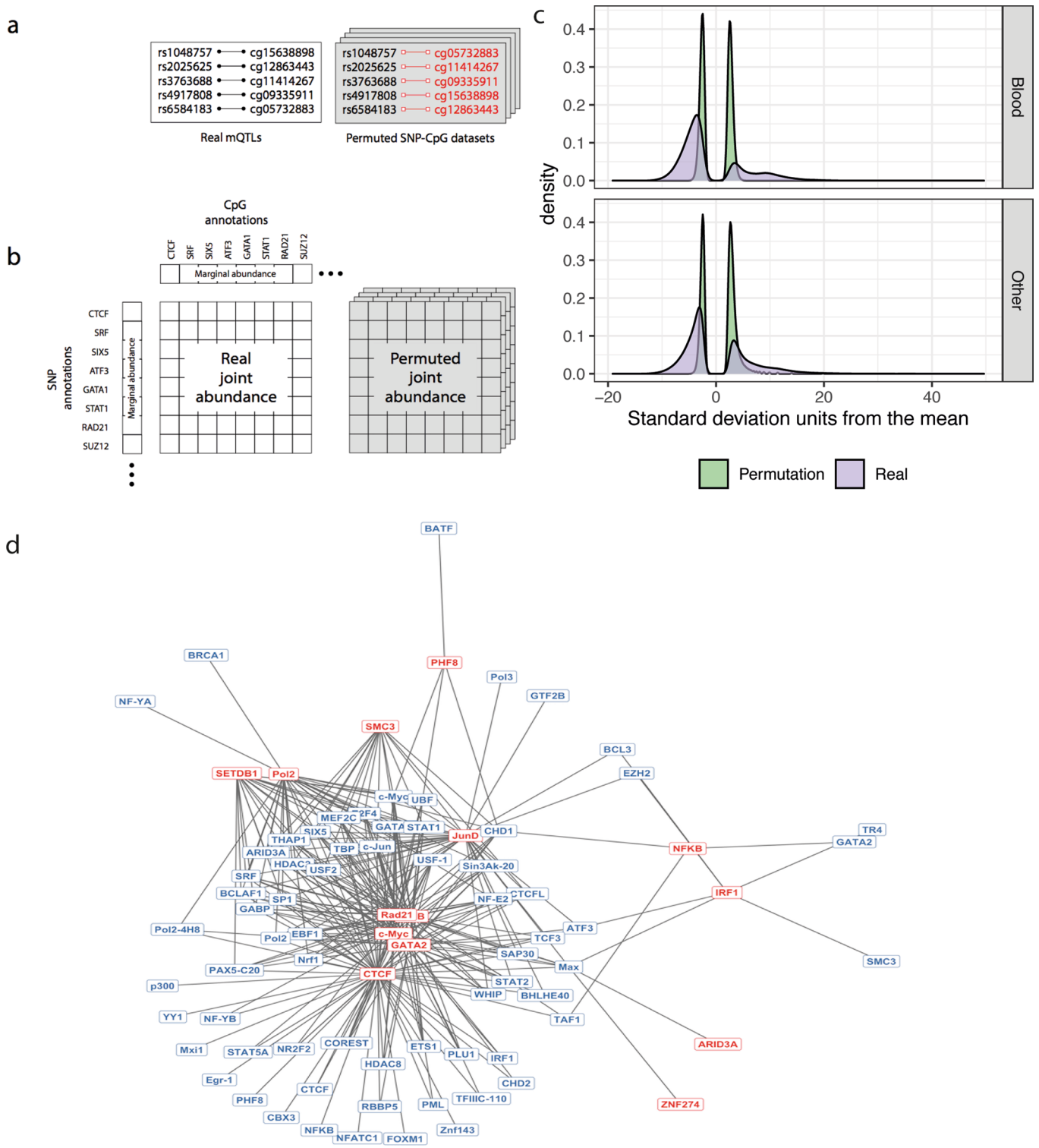
Extended Data Fig. 3 | Effect sizes and weighted standard deviation (SD) for each mQTL category. a, For each DNAm site, the strongest absolute effect size (the maximum absolute additive change in DNAm level measured in SD per allele) was selected. The kernel density estimations of the effect sizes were shown for all sites with a mQTL (n=190,102), sites with *cis only* effects (n=170,986), *cis* effects for sites with *cis* and *trans* effects (n=11,902), *trans* effects for sites with *cis* and *trans* effects (n=11,902) and sites with *trans only* effects (n=7,214). Comparing the strongest effect size for each site in a two-sided linear regression model showed that *cis+trans* sites had larger *cis* effect sizes (per allele SD change = 0.05 (s.e.= 0.002), p<2e-16) as compared to *cis only* sites and weaker *trans* effect sizes (per allele SD change = −0.06 (s.e.= 0.002), p<2e-16) as compared to *trans only* sites. To detect these small *trans* effect sizes at sites with both a *cis* and a *trans* association, it is crucial to regress out the *cis* effect to decrease the residual variance and improve power to detect a *trans* effect. **b**, The violin plots represent kernel density estimates of the weighted SD across 36 cohorts for each DNAm site. The center line of the boxplot in the violin plots corresponds to the median value. The lower and upper box limits indicate the first and third quartiles (the 25th and 75th percentiles). The length of the whiskers corresponds to values up to 1.5 times the IQR in either direction.



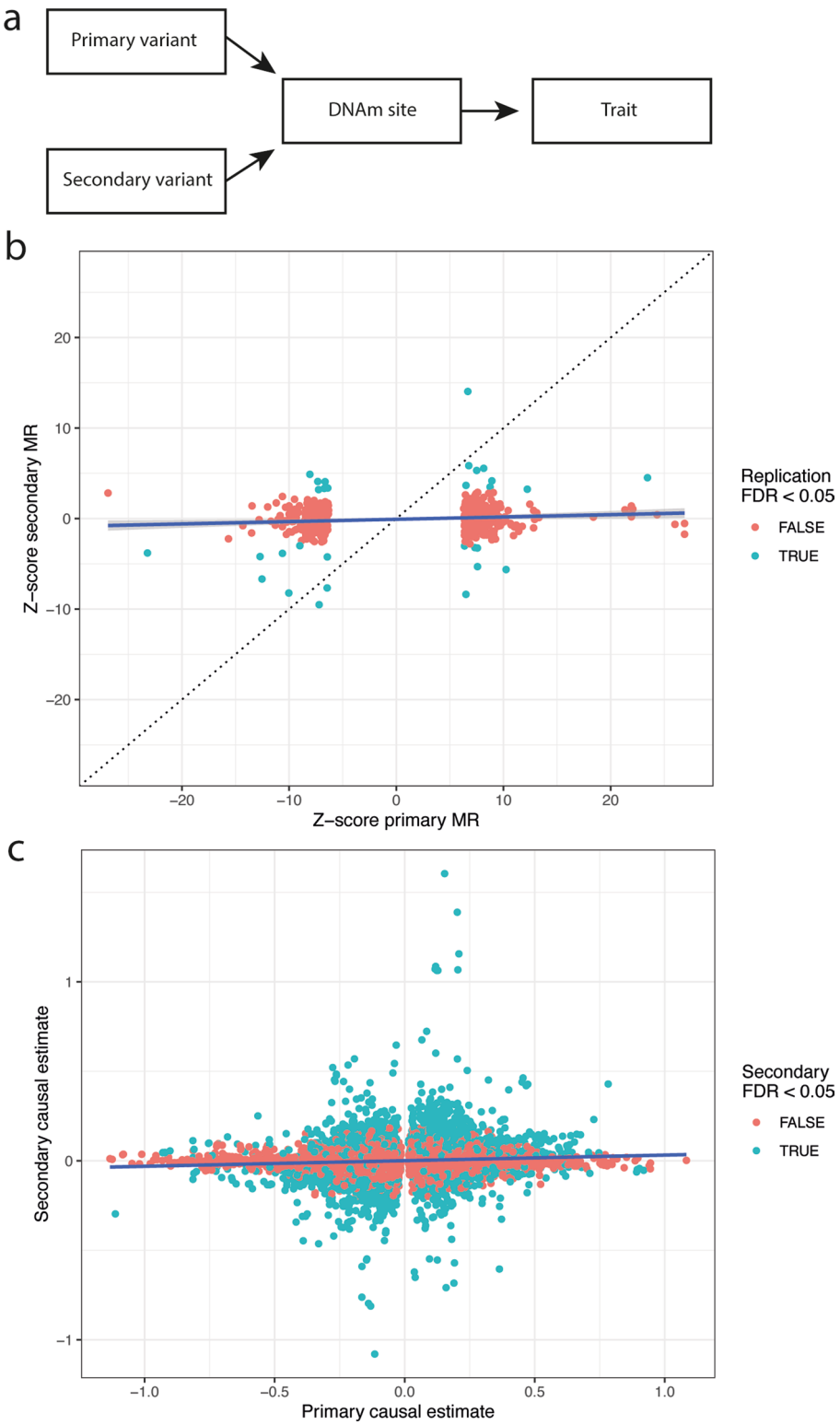
Extended Data Fig. 4 | Impact of the twostage design on mQTL coverage. a, Loss in power in twostage design. We calculated the power of detecting a *cis* association in at least one of the 22 studies at p<1e-5 or a *trans* association in at least two of 22 studies at p<1e-5. **b**, Expected number of mQTLs. Using the number of mQTLs with a particular r^2 value, and the power of detecting mQTLs with that r^2 value, we calculated how many mQTLs would expect to exist with that value.

mQTL sites category	No. mQTLs Blood	No. mQTLs Adipose	No. mQTLs Brain	R_b Adipose	R_b Brain
cis any	170357	140092	108839	0.73 (0.002)	0.59 (0.003)
cis only	159702	131099	101866	0.73 (0.002)	0.59 (0.004)
cis only_high	50367	40793	32041	0.63 (0.005)	0.62 (0.006)
cis only_low	43506	35890	27988	0.76 (0.004)	0.63 (0.006)
cis only_intermediate	65829	54416	41837	0.78 (0.003)	0.57 (0.006)
cis+trans_cis	10655	8993	6973	0.79 (0.008)	0.58 (0.014)
cis+trans_cis_high	1298	1089	799	0.69 (0.031)	0.64 (0.047)
cis+trans_cis_low	2266	1936	1463	0.81 (0.015)	0.67 (0.023)
cis+trans_cis_intermediate	7091	5968	4711	0.81 (0.008)	0.55 (0.018)
cis+trans_trans	10655	9405	6521	0.83 (0.007)	0.64 (0.019)
cis+trans_trans_high	1298	1109	762	0.81 (0.017)	0.73 (0.038)
cis+trans_trans_low	2266	1958	1439	0.85 (0.014)	0.71 (0.029)
cis+trans_trans_intermediate	7091	6338	4320	0.84 (0.01)	0.58 (0.027)
trans any	18582	15921	11478	0.87 (0.004)	0.77 (0.01)
trans only	7927	6516	4957	0.91 (0.004)	0.86 (0.01)
trans only_high	1352	1128	805	0.89 (0.012)	0.87 (0.03)
trans only_low	4155	3324	2721	0.93 (0.005)	0.9 (0.011)
trans only_intermediate	2420	2064	1431	0.89 (0.008)	0.74 (0.026)

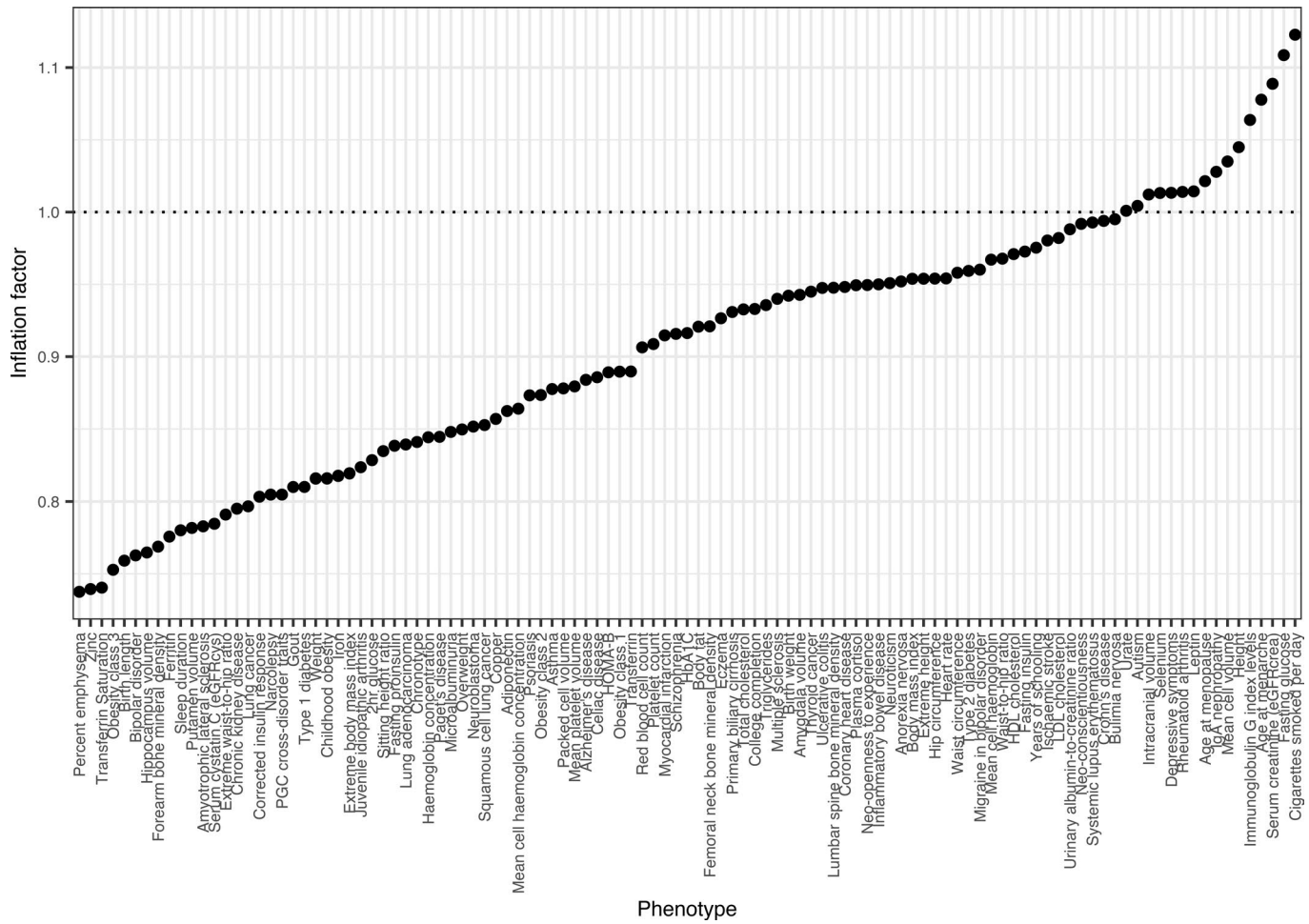
Extended Data Fig. 5 | Correlation of mQTL effects (p<1e-14) between blood and other tissues. For each mQTL category, the correlation of genetic effects between tissues (r_b) were estimated using the r_b method[25] where we used the blood mQTLs as reference. DNAm levels are categorized as low (<0.2), intermediate (0.2–0.8) or high (>0.8).



Extended Data Fig. 6 | 2D enrichment of SNP and DNAm site TFBS annotation. a, To test if the annotations of the SNPs involved in *trans*-mQTLs were specific to the annotations of the DNAm sites that they influence, we compared the real SNP-DNAm site pairs against permuted SNP-DNAm site pairs, where the biological link between SNP and site is severed whilst maintaining the distribution of annotations for the SNPs and sites. We constructed 100 such permuted datasets **b**, SNP and site positions were annotated against genomic features, and we quantified how frequently mQTLs were found for each pair of SNP-DNAm site annotations. This enabled the construction of 2D-annotation matrices for both the real *trans*-mQTL list and the permuted *trans*-mQTL lists. **c**, Distribution of two-dimensional enrichment values of *trans*-mQTLs. There was substantial departure from the null in the real dataset for all tissues indicating that the TFBS of a site depended on the TFBS of the SNP that influenced it. **d**, A bipartite graph of the two-dimensional enrichment for *trans*-mQTLs, SNPs annotations (blue) with $p_{emp} < 0.01$ after multiple testing correction co-occur with particular site annotations (red).



Extended Data Fig. 7 | Correspondence of MR estimates amongst multiple independent instruments. a, To evaluate if a site having a shared causal variant with a trait was potentially due to the site being on the causal pathway to the trait, we reasoned that independent instruments for the site should exhibit consistent effects on the outcome consistent with the original co-localizing variant. **b**, Amongst the putative co-localizing signals, 440 involved a DNAm site that had at least one other independent mQTL. The plot shows the causal effect estimate estimated from the original co-localizing signal against the causal effect estimates obtained from the independent variants (n=440). Grey regions represent the 95% confidence of the slope. **c**, Correspondence of MR estimates amongst multiple independent instruments on 36 blood traits. To evaluate if a site having a shared causal variant with a blood trait was potentially due to the site being on the causal pathway to the trait, we reasoned that independent instruments for the site should exhibit consistent effects on the outcome consistent with the original co-localizing variant. Amongst the putative co-localizing signals, 30% involved a DNAm site that had at least one other independent mQTL. The plot shows the causal effect estimate estimated from the original co-localizing signal against the causal effect estimates obtained from the independent variants. The *HLA* region has been removed and betas are plotted.



Extended Data Fig. 8 | Genomic inflation factors for genome-wide scans of causal effects of traits on DNAm sites. Each trait (x axis) was tested for causal effects against (on average) 317,659 DNAm sites, excluding sites in the *MHC* region. The p-values from IVW MR analysis were used to estimate the genomic inflation for each trait (y-axis). Traits are ordered by genomic inflation factor.

nature research

Corresponding author(s): Josine Min

Last updated by author(s): Jun 3, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Using a two-phase discovery study design, we analyzed ~10 million genotypes imputed to the 1000 Genomes reference panel and quantified DNAm at 420,509 sites using Illumina HumanMethylation BeadChips in whole blood derived from 27,750 European participants (36 studies). Most of the datasets were processed using https://github.com/perishky/meffil (v0.1.0). Individual study analysts used a github pipeline https://github.com/MRCIEU/godmc to conduct the mQTL analysis. Analyses were done with R3.2.0 (https://cran.r-project.org/) or higher using local installations of PLINK2.0 and the Rpackages: lattice, ggplot2, data.table, Matrix eQTL v2.1.0, parallel, matrixStats, plyr, meffil v0.1.0, EasyQC_9.2 and impute. In addition for datasets with related samples, GenABEL, SNPRelate and GENESIS were used. For unrelated participants we regressed out age, sex, predicted cell counts using the Houseman algorithm implemented in meffil v0.1.0, predicted smoking and genetic PCs (adjustment 1). For related participants we did the same except also fitting the genetic kinship matrix using the method described in GRAMMAR (Aulchenko et al. 2007). We took the residuals from the first adjustment forward to regress out the non-genetic DNAm PCs on the adjusted DNAm beta values (adjustment 2). The residuals from these analyses were rank transformed and centered to have mean 0 and variance 1.
Data analysis	We used https://github.com/MRCIEU/godmc_phase1_analysis for the phase1 analysis, https://github.com/explodecomputer/random-metal for the meta analyses and https://github.com/MRCIEU/godmc_phase2_analysis for the followup analyses. Analyses were done with R3.2.0 (https://cran.r-project.org/) or higher, custom shell and perl scripts, PLINK v1.90 (https://www.cog-genomics.org/plink/2.0/), GCTA v1.26.0 (https://cnsgenomics.com/software/gcta/#Overview) or GARFIELD v2 (https://www.ebi.ac.uk/birney-srv/GARFIELD/).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

A database of our results is available as a resource to the community at http://mqtldb.godmc.org.uk. The individual level genotype and DNAm data are available by request from each individual study or can be downloaded from Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/), European Genome-phenome Archive (EGA, https://ega-archive.org/) or Array Express (https://www.ebi.ac.uk/arrayexpress/). As the consents for most studies require the data to be under managed access, the individual level genotype and DNAm data are not available from a public repository unless stated.
ALS BATCH1 & 2 data are available to researchers by request as outlined in the Project MinE access policy. ARIES data are available to researchers by request from the Avon Longitudinal Study of Parents and Children Executive Committee (http://www.bristol.ac.uk/alspac/researchers/access/) as outlined in the study's access policy http://www.bristol.ac.uk/media-library/sites/alspac/documents/researchers/data-access/ALSPAC_Access_Policy.pdf. BAMSE data are available from the GABRIEL consortium as well as from the study portal at http://ki.se/en/imm/medallomics. BASICMAR DNAm data are available under accession number GSE69138. Born in Bradford data are available to researchers who submit an expression of interest to the Born in Bradford Executive Group (https://borninbradford.nhs.uk/research/). BSGS DNAm data are available under accession code GSE56105. GOYA data are available by request from DNBC, https://www.dnbc.dk/. Dunedin data are available via a managed access system (contact: ac115@duke.edu). E-Risk DNAm data are available under accession number GSE105018. Estonian biobank (ECGUT) data can be accessed upon ethical approval by submitting a data release request to the Estonian Genome Center, University of Tartu (http://www.geenivaramu.ee/en/access-biopank/data-access). EPIC-Norfolk data can be accessed by contacting the study management committee http://www.srl.cam.ac.uk/epic/contact/. Requests for EPICOR data accession may be sent to Prof. Giuseppe Matullo (giuseppe.matullo@unito.it). FTC data can be accessed upon approval from the Data Access Committee of the Institute for Molecular Medicine Finland FIMM (fimm-dac@helsinki.fi). Requests for Generation R data access are evaluated by the Generation R Management Team. Researchers can obtain a de-identified GLAKU dataset after having obtained an approval from the GLAKU Study Board. GSK DNAm data are available under accession number GSE125105. INMA data are available by request from the INfancia y Medio Ambiente Executive Committee for researchers who meet the criteria for access to confidential data. IOW F2 data are available by request from Isle of Third Generation Study (http://www.allergyresearch.org.uk/contact-us/). LLS DNAm data were submitted to the EGA under accession EGAS00001001077. LBC1921 and LBC1936 data are available on request from the Lothian Birth Cohort Study, Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh (email: I.Deary@ed.ac.uk). DNAm from MARTHA participants are available under accession number E-MTAB-3127. NTR DNAm data are available upon request in EGA, under the accession code EGAD00010000887. PIAMA data are available upon request. Requests can be submitted to the PIAMA Principal Investigators (https://piama.iras.uu.nl/english/). PRECISESADS data are available through ELIXIR at doi:10.17881/th9v-xt85. Collaboration in data analysis of PREDO is possible through specific research proposals sent to the PREDO Study Board (predo.study@helsinki.fi) or primary investigators Katri Räikkönen [katri.raikkonen@helsinki.fi] or Hannele Laivuori [hannele.laivuori@helsinki.fi]. Data is available upon request at project MinE (https://www.projectmine.com). Raine data are available upon request (https://ross.rainestudy.org.au). Requests for the data accession of the Rotterdam Study may be sent to: Frank van Rooij (f.vanrooij@erasmusmc.nl). SABRE data are available by request from SABRE (https://www.sabrestudy.org). SCZ1 DNAm data are available under accession number GSE80417. SCZ2 DNAm data are available under accession number GSE84727. SYS data are available upon request addressed to Dr Zdenka Pausova [zdenka.pausova@sickkids.ca] and Dr Tomas Paus [tpausresearch@gmail.com]. Further details about the protocol can be found at [http://www.saguenay-youth-study.org/]. TwinsUK DNAm data are available in GEO under accession numbers GSE62992 and GSE121633. TwinsUK adipose DNAm data are stored in EGA under the accession number E-MTAB-1866. Access to additional individual-level genotype and phenotype data can be applied for through the TwinsUK data access committee http://twinsuk.ac.uk/resources-for-researchers/access-our-data/. Individual level DNAm and genetic data from the UK Household Longitudinal Study are available on application through EGA under accession EGAS00001001232. Non-identifiable Generation Scotland data from this study will be made available to researchers through GS:SFHS Access Committee. MESA DNAm data are available under accession GSE56046 and GSE56581. Tissue DNAm data are available from GSE78743. Brain DNAm data can be found under accession number GSE58885.
For the enrichments, we used chromatin states from the Epigenome Roadmap (https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/), transcription factor binding sites from the ENCODE project (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/ downloaded from the LOLA core database (http://databio.org/regiondb) and gene annotations from https://zwdzwd.github.io/InfiniumAnnotation or from GARFIELD (https://www.ebi.ac.uk/birney-srv/GARFIELD/. To extract genome-wide association signals for colocalization, we used the MRBase database (https://www.mrbase.org/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We included 32,851 participants (discovery and replication) in our study. Pre-study power calculations were not performed but we aimed to maximize sample size based on the cohorts available and included 7x more participants than a previous study.
Data exclusions	We removed SNPs, participants or studies for the following pre-established reasons: SNPs: SNPs that failed Hardy Weinberg equilibrium (p<1e-6), had a minor allele frequency <0.01, an info score <0.8 or missingness in more than 5% of the participants were removed. We recoded SNPs to CHR:POS format and removed duplicate SNPs. We then harmonized the recoded SNPs to the 1000G reference using easyQC_v9.260. This harmonization script removed SNPs with mismatched alleles and recoded INDEL alleles to I and D. Participants: We removed participant that were discordant between reported sex and genotype predicted sex. Based on principal

components analysis using pruned HapMap 3 SNPs, ethnic outliers that deviated 7 SDs from the mean were removed. Related participants (identity by state > 0.125) were removed.
Studies: Initially we included 38 discovery studies in our study. Before we performed the meta-analysis, we checked the number of SNPs and INDELs, sites and individuals analyzed and the average mean and SD for each DNAm site to identify possible inconsistencies. Each of the 38 studies conducted a genome-wide association study of cg07959070. We chose this DNAm site as a positive control as it showed a strong cis mQTL in several datasets on chr22 and hasn't been proposed to be excluded from the analyses by probe annotation efforts. To identify possible errors, we checked the cis association on chromosome 22 for this DNAm site and removed any study that showed p>0.001 for the cis mQTL. In addition, we checked quantile-quantile and Manhattan plots for this DNAm site and excluded any study with lambda >1.1 or lambda <0.9. After inspection one study was removed from the analysis due to deflation and one study was removed due to a lack of the positive control association signal, leaving 36 studies for the final meta-analysis.

Replication

We sought to replicate our discovered mQTL in the Generation Scotland cohort (n = 5,101) using an independent analysis pipeline. Replication data were for 188,017 of our discovery mQTL (137,709 sites). We found a strong correlation of effect sizes for both cis and trans effects (Pearson r=0.97, n=155,191 and r=0.96, n=14,465 at p<1e-3, respectively; 99.6% of the associations had a consistent direction of effect. At a Bonferroni corrected threshold of 0.05/188,017, 142,727 of the discovery mQTL replicated in the Generation Scotland cohort (76%); the replication rate for cis and trans mQTL were 76% and 79%, respectively. To evaluate whether our replication rate was in line with expectations given the smaller replication sample size, we estimated that under the assumption that the discovery mQTL are true positives, 171,824 mQTL would be expected to replicate at a nominal threshold of p<1e-3; we found that the actual number of mQTL replicating at this level was 169,656, indicating that the majority of our discovery mQTL are likely to be true positives.
Of the 169,656 associations for which we had effect estimates for both the discovery and replication datasets, there were 702 mQTL that replicated after multiple testing correction (p < 2.7e-7) but the effect size was in a different direction. This is a very small proportion of all mQTL, but we estimated that we would only expect one to replicate in the wrong direction by chance. These mQTL comprised SNPs with relatively equal proportions of allele codes (i.e. not dominated by A-T or G-C SNPs), and had similar sample sizes. However, the average absolute effect size was much smaller than all other mQTL (0.22 vs 0.30), and the average I2 was close to double (85.2 vs 47.4). Whether these associations represent examples where the sign of the direction truly is variable between populations, or if they are statistical artefacts (e.g. due to allele coding issues) is not clear, therefore we have flagged these mQTL as being unreliable.

Randomization

Genotyping and DNA methylation profiling have not been performed in experimental groups. No randomization was carried out. This is a genetic association analysis and the inheritance of alleles is random during meiosis. For each study, we regressed out possible batch effects from the DNA methylation using non-genetic PCs and we used genetic PCs to adjust for genetic confounding.

Blinding

Blinding is not relevant for this study as we used genotypes as exposures and DNA methylation as outcome. Both are array based measurements, measuring multiple genotype and DNA methylation profiles at the same time. Researchers were blinded to the genotypes and DNAm sites of the participants.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

Supplementary Table 1 describes the population characteristics of the 36 studies used in the discovery analysis. The studies included were birth cohorts, case control collections, twin registries, offspring samples and population-based cohorts. Studies comprised of participants aged between a mean of zero and 79 years with 4%-100% men. DNAm was measured in whole blood or cord blood using HumanMethylation450 or EPIC arrays in at least 100 European individuals. Studies were genotyped with a wide range of assays and imputed using 1000G or HRC as reference panel.

Recruitment

For each of the 36 discovery datasets, replication dataset, study with isolated subsets and three tissue datasets a recruitment section has been described in Supplementary Note. There are potential selection biases in the recruitment of samples, and they could have small effects on genetic associations but we conducted sensitivity tests for this as described in Supplementary Table 2 and the Supplementary Note.

Ethics oversight

For each of the 36 discovery datasets, replication dataset, study with isolated subsets and three tissue datasets, ethics and informed consent statements are described in Supplementary Note. Ethical approval for each study was obtained from study-specific research ethics committees.

Note that full information on the approval of the study protocol must also be provided in the manuscript.