Supplementary Information

Participants

To study the relationship between common genetic variation and DNA methylation (DNAm) we focused on studies of European ancestry with genotype data imputed to the 1000 Genomes reference panel[1] 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 **Supplemental Note 1** and **Table S1**.

Study design mQTL analyses

Initially, 38 independent studies were recruited to contribute data towards a (DNA methylation-quantitative trait loci) mQTL meta-analysis of which 36 studies (**Table S1, Supplemental Note 1**) passed our stringent quality criteria described below. Conventional genome wide association studies (GWAS) meta-analyses involve performing complete GWAS in each study, sharing the summary data and meta-analysing every tested SNP. As a 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 a genome-wide analysis but provided only the associations that surpass a relaxed significance threshold ($p < 1e-5$) 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 the analysis in this study has been performed in two phases.

In Phase 1 of our study we performed mQTL analyses of 420,509 high quality DNAm sites[2] using data from 22 independent European studies to identify putative associations (**Table S1**, **Figure 1A**) at a threshold of $p< 1e-5$. 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.[2] which includes probes with potential cross reaction and probes that could not be mapped to genome. Secondly, 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 (MAF >5%, UK10K) at the CpG dinucleotide or the extension base (for type I probes).
All candidate mQTL associations at $p<1e-5$ were combined to create a unique 'candidate list' of mQTL associations. In total we identified 102,965,711 candidate mQTL associations in *cis* ($p < 1e-5$, +/- 1 Mb from DNAm site) and 710,638,230 candidate mQTL associations in *trans* (>1Mb from DNAm site) in at least one dataset. 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 in at

least two datasets (**Figure S1**). 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) (**Table S1**). The estimates for the candidate list are meta-analysed to obtain the final results (**Figure 1A**).
This two-phase approach has a single objective: to minimise the computational burdens of storing summary data from the complete analysis from every study. However, we have effectively performed a complete search of all candidate mQTL associations, though with likely 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 is that we will have missed some associations that were not at p<1e-5 in any study but when combined across all studies would have surpassed an experiment wide multiple testing correction.

Data preparation

The Genetics of DNA Methylation Consortium (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 covariate preparation, DNAm data preparation, and subsequent analyses. In the first module the data format of the genotype data, DNAm and covariate 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, covariates including cell counts, genotype build and strand, and the number of DNAm outliers were recorded. We also generated matrices with mean and standard deviation (SD) 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 that were used.

Genotype data

Each study performed quality control on genotype data for all autosomes and chromosome X (if available) and imputed to 1000G phase 1 or above using hg19/build37. Dosages were converted to bestguess data without a probability cut-off.
SNPs that failed Hardy Weinberg equilibrium (p<1e-6), had a minor allele frequency (MAF) <0.01, an info score <0.8 or missingness in more than 5% of the participants were removed. We recoded SNPs to CHR:POS:{SNP/INDEL} format and removed duplicate SNPs. We then harmonized the recoded SNPs to the 1000G reference using easyQC[3]. This harmonization script removed SNPs with mismatched alleles and recoded INDEL alleles to I and D.

We performed a gender check to remove participants with discordant gender to the covariate file. We extracted and pruned a set of common HapMap3 SNPs (MAF>0.2, without long-range linkage disequilibrium (LD) regions) before we calculated the first 20 PCs on LD pruned SNPs and excluding regions of high LD from the analysis. We used PLINK.2.0[4] for unrelated participants and GENESIS[5] for related participants to identify ethnic outliers.

Ethnic outliers that deviated 7 SDs 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 HapMap 3 SNPs with MAF>0.01 and calculated genome-wide average identity by state (IBS) using PLINK2.0. Participants with IBS > 0.125 were removed.

DNAm data normalisation 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 normalisation through the R package meffil (see **Table S1**)[6]. 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 where more than 10% of the DNAm sites failed the detection p-value of 0.1 and/or threshold of 3 beads. In addition, mismatched samples were identified by comparing the 65 SNPs on the DNAm array to the genotype array and a gender check. Additional DNAm quality was checked by the methylated versus unmethylated ratio, dye bias using the normalisation control probes and bisulphate control probes. Protocols can be found here: https://github.com/perishky/meffil/wiki. For each DNAm site, we replaced outliers that were 10 SDs from the mean (3 iterations) with the DNAm site mean.

Covariates

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 principal components (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[7]. In addition, we predicted cell counts using the Houseman algorithm implemented in meffil[8]. We performed a PC analysis on the 20,000 most variable autosomal DNAm sites and kept all PCs that cumulatively explained 80% of the variance. We performed a genome wide association analysis on the DNAm PCs and retained the PCs that were not associated with a genotype (p > 1e-7). We kept a maximum of 20 nongenetic PCs for subsequent adjustment.

DNAm data adjustment

We attempted to minimise non-genetic variation in the DNAm data to improve 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[9].

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.

Positive and negative controls

Before we performed the meta-analysis, we checked the number of SNPs and INDELs, sites and individuals analysed and the average mean and SD 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 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[2,10,11,12]. 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 lambdas (lambda >1.1 or lambda <0.9). We noticed deflation of the genomic lambda after adjustment of the index *cis* SNP in datasets with relatedness. However, lambdas were around 1 when not adjusted. 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.

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[2] residualised DNAm sites separately in 22 studies (N=16,907) using the R package Matrix eQTL v2.1.0[13]. 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 $\beta_{jk}$was the effect estimate of each SNP k on each residualised 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) where effect sizes for all putative associations were recalculated by fitting linear models. For putative *cis*-mQTL 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 lowest p-value within that study) and fit this as a covariate in the *trans*-mQTL regressions

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

Impact of two-stage design on power of study

Though the multi-stage study design was performed out of practical necessity, we evaluated the impact it had on statistical power in comparison to the hypothetical situation of analysing 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 one of 22 studies at $p < 1e\text{-}5$. To do this we calculate what is the probability of missing an association as being the product of the probability of missing it in study 1 AND in study 2 AND in study 3 etc.

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

where $f(x; k; \lambda)$ is the probability density function for the non-central chi-square distribution with k degrees of freedom and λ non-centrality parameter based on the postulated variance explained by an mQTL (r^2) and the study sample size n_i and 19.5 denotes the chi-square threshold at p = 1e-5 with one degree of freedom.

For *trans* mQTL associations we calculated the power to detect an association in at least two of 22 studies at p< 1e-5. We calculate what is the probability of missing an association as being 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 etc.

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

where $f(x; k; \lambda)$is the probability density function for the non-central chi-square distribution with $k$degrees of freedom and λ non-centrality 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 chi-square threshold at p = 1e-5 with one degree of freedom.

We found that we have no loss of power (<1%) for loci that explain more than 1.2% or less than 0.1% of the variance. Within these bounds >80% of power is lost for *cis*-mQTL with r^2 0.16% to 0.38%. For *trans*-mQTL, power suffers slightly more because of requiring detection by at least two studies in the first stage (r^2 0.27% to 0.64%) (**Figure S48**).

Meta-analyses

We used the SNP effect estimates and standard errors for each SNP-DNAm site pair in the candidate list in the meta-analyses. Inverse variance fixed effects (FE) meta-analyses of the 36 studies was performed using METAL[14]. We modified METAL (https://github.com/explodecomputer/random-metal) to incorporate the DerSimonian and Laird random effect (RE) models[15] and multiplicative random effects (MRE) models[16]. These results are available here: 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 (**Figure S3**). In addition, the average I^2 heterogeneity estimate for the effect size direction categories was 44% (min=0%, max 100%). For categories with more than 100 associations, average I^2 was 49% (min=36%, max 61%) (**Figure S3**). We also explored whether the number of phase 1 studies was correlated to I^2 and tau^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 didn't show more heterogeneity than mQTL associations found in a high number of phase 1 studies.

To explore heterogeneity further, we meta-analysed 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 tau^2 and smaller effect sizes and DNAm site SDs (**Figure S4, Figure S5**). 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 tau^2 were positively correlated to effect sizes (**Figure S3D**) we deem the use of FE meta-analysis to be appropriate for reducing false negative rates.

Clumping analysis

To obtain a set of independent mQTL we performed clumping using 503 European participants from the 1000 Genomes dataset. We used an R-square threshold of 0.0001 and a clumping radius of 5Mb. The method was applied to each DNAm site GWA separately. For the follow up analyses (unless otherwise stated) we used clumped mQTL results, applying a *cis* p-value threshold of 1e-8 and a *trans* p-value threshold of 1e-14.

Window size

To compare whether the intra-chromosomal *trans* mQTL regions containing more *trans* mQTL associations than inter-chromosomal regions, we calculated the rate of inter and intrachromosomal trans mQTL associations. The rate of *trans* mQTL associations was defined as the number of *trans* mQTL associations divided by the number of 5Mb blocks on the chromosome. Intrachromosomal *trans* mQTL associations were either defined as >1 Mb or >6 Mb from the DNAm site. Due to long range associations around the *HLA* region, chromosome 6 was removed from this analysis.

Conditional analysis

To test for multiple independent SNPs operating within the locus of each mQTL, we used GCTA software to perform cojo-select analysis using an 10Mb window size[17]. Because only summary data was available we used an external LD reference panel, selecting the ALSPAC children dataset (n=8,092) imputed to HRC panel[18]. We retained SNPs that had conditional p-values of 1e-8 for *cis* effects and 1e-14 for *trans* effects.
Limitations of this analysis are numerous. First, because of the two-stage design we do not have complete coverage of the common variants in the region of a putative mQTL. For some sites the SNP density is low in the region. Second, meta-analysis combines the association patterns from multiple studies, each with their own LD patterns. Though our analysis was restricted to European participants which matched the reference panel, it is likely that the reference panel's LD patterns do not perfectly match the aggregated latent LD patterns amongst the studies contributing to the meta-analysis. As a consequence, accounting for a particular variant's effect based on its LD may be unreliable. Third, a small number of sites analysed yielded unrealistically large numbers of conditionally independent mQTL and we advise caution in interpreting these results. For example, 3 sites outside the *MHC* region had more than 50 independent *cis* associations per site. While the median estimate of 2 independent SNPs per site is reasonable and in line with results from other 'omic datasets, it is clear that the reliability of the analysis cannot be guaranteed for any particular site analysed.

Comparing study-wide heterogeneity

As we used studies with different designs, we investigated systematic patterns of heterogeneity by analysing between-study heterogeneity, by aggregating the within-study heterogeneity information across multiple genetic variants[19]. This method can dissect genomic heterogeneity patterns to flag underperforming studies in which the rate of discovery is smaller than expected given the sample size. These studies could compromise the power of the meta-analysis[19]. Following Magosi *et al.* 2017[19], to calculate *M* statistics for each study, we extracted 337 mQTL SNPs on chromosome 20 with a $p < 1e\text{-}14$ from each study and obtained the mean of the Cochran's Q estimates across all associations (**Figure S2**). To explore the impact of technical and biological influences on the magnitude of *M* we performed random-effects meta-regression[20], examining the following technical and biological factors as sources of systematic heterogeneity: number of SNPs, number of sites, sample size, relatedness yes/no, DNAm array (450k versus EPIC), normalization method, average MAF across all SNPs, average info score across all SNPs, lambda, cord vs peripheral blood, ancestry (UK vs non UK, The Netherlands vs non The Netherlands, Spain vs non Spain, Finland vs non Finland, northern versus southern countries), case control versus population-based, number of males and age. We found that some of the variability in average effect size was associated with the number of SNPs ($p=0.0169$, $N=36$) but not with other technical variables. We further observed that the M statistic was highly correlated to the average SD of the DNAm site.

Out-of-sample replication of discovered mQTL

To validate the discovery associations, we used the Generation Scotland (GS) dataset of 5,101 participants[21] **(Supplementary Note 1)**, which was generated using an entirely different pipeline to the one described for the main discovery meta-analysis.

The analysis model included two genomic relationship matrices, G (genomic relationship matrix) and K (kinship relationship matrix), and three environmental relationship matrices, F (environmental matrix representing nuclear-family-member relationships), S (environmental matrix representing full-sibling relationships) and C (environmental matrix representing couple relationships). These five matrices (as random effects), together with covariates (i.e., age, age^2, gender, cell counts for granulocytes, B-lymphocytes, natural killer cells, CD4+ T-lymphocytes and CD8+ T-lymphocytes, season of the visit, appointment time of the day, appointment day of the week) as fixed effects, were fitted simultaneously in a mixed linear model for each site. The resulting residuals were inverse rank transformed prior to GWAS analysis in a simple linear model using REGSCAN v0.5.

Because the replication sample size is considerably smaller than the discovery, we expect the replication rate to be relatively low. However, our intention for using the dataset is instead to evaluate whether the rate of replication is in line with expectation given the discovery effect sizes and the replication sample size and multiple testing correction. If we find that it is, it gives us confidence that the discovery mQTL as obtained through our pipeline are en masse reliable.

Observed vs expected replication rate

The summary results based on 5,101 participants in the GS dataset had been pre-calculated and all SNP-DNAm site pairs were stored for which p <1e-3. We filtered these pre-calculated SNP-DNAm site pairs and kept only biallelic SNPs across 22 autosomes with MAF>0.01 and info scores>0.8 and removed 5,910 DNAm sites from the dataset if: (i) they had more than 5 participants with a bead count <3; or (ii) >= 1% participants had a detection p-value of > 0.05. Following the methods outlined in[22] for each i^{th} mQTL we calculated the expected replication rates at α-level significance as a function of the absolute effect size estimate $|\beta_i|$ assuming that it is unbiased, and the expected standard error in the replication dataset $\sigma_{GS,i}$ to be

$$p(replication_i) = \phi\left(-\frac{|\beta_i|}{\sigma_{GS,i}}\right) + \phi^{-1}\left(\frac{\alpha}{2}\right) + \left[1 - \phi\left(-\frac{|\beta_i|}{\sigma_{GS,i}}\right) - \phi^{-1}\left(\frac{\alpha}{2}\right)\right]$$

where ∅ is the standard normal cumulative distribution function. Therefore, the expected replication rate for all M mQTL associations that were tested is

$$\sum_{i}^{M} p(replication_i)$$

The expected replication standard error for an mQTL i was calculated as was calculated as

$$\sigma_{GS,i} = \sqrt{\frac{2p_i(1-p_i)\beta_i^2 + 1}{2p_i(1-p_i)(n-2)}}$$

where $n = 5{,}101$ is the replication sample size, and the allele frequency of the SNP in the outcome is p_i.

For the replication analysis, 169,656 SNP-DNAm pairs were present in the pre-calculated GS dataset (i.e. p <=1e-3, MAF>0.01, info score>0.8). An additional 18,361 SNP-DNAm pairs (population-wide MAF>0.01) were potentially testable in the dataset but were not present due to their p-value being >1e-3 or having been marked as 'NA' by REGSCAN. A total of 188,017 (169,656 + 18,361) has been used for multiple testing correction and downstream analyses. Correspondingly, we expected (based on the methods outlined above) 171,824 mQTL to replicate at p < 1e-3. This very strong agreement between expected and observed rates indicates that our discovery mQTL are, en masse, true positive associations.

By contrast, at the stringent replication threshold of 0.05/188,017, we found 142,727 to replicate which was 6.7% higher than the expected 133,734. At this threshold, 76% of cis mQTL and 79% mQTL associations replicated. This suggests that there are some mQTL with small effects that are replicating at a slightly higher rate than would be expected. There are several possible reasons that could explain this, for example:

1. SNP effects are relatively smaller in the discovery data because the replication dataset is correcting for more residual variance.
2. Meta-analysis incurs heterogeneity which dilutes some effect estimates in the discovery data.

Counterbalancing these factors are other factors that could lead to the observed replication rate to be lower than the expected replication rate, for example

1. The GS replication and GoDMC discovery used different DNAm normalization and adjustment pipelines, which may lead to different patterns of heterogeneity between the studies. This will reduce replication rates.
2. The GS dataset is a family-based study and the GoDMC meta-analysis comprise family and population-based studies. We observed systematic differences in the mQTL yield between population and family-based studies in the discovery stage, so there may be systematic differences between GS and GoDMC due to this also. This will reduce replication rates.
3. There will be a winner's curse effect inflating the effect estimates of the GoDMC discovery data. The expected replication rates are based on the assumption that the effect estimates are unbiased, and the expected replication rate based on upward-biased effect estimates will be higher than the true estimate.

The processes that lead to whether or not our replication rate is in agreement with expectation are complex. But overall, we find that there is broad agreement between our expected and observed replication rates, indicating that en masse the discovery mQTL are likely to be true positive associations.

Concordance of effect sign between discovery and replication

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[22]. 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 I^2 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.

Variance explained by mQTL

To estimate the overall proportion of the DNAm variance explained by the discovered mQTL, we summed r^2 estimates from the replication dataset (coefficients of determination of the regression of the SNP on the inverse-normal rank-transformed residual) for each DNAm site and divided it by the number of tested DNAm sites (n=420,509). To estimate the overall proportion of the estimated additive genetic variance explained by discovered mQTL, we divided the r^2 estimates from the replication dataset by the h^2 estimates from a family study[23] and a twin study[24]. To be conservative we disallowed any specific DNAm sites to explain more than 100% of the genetic variance and for DNAm sites with h^2 of 0, we set r^2 to 0. This can arise when h^2 point estimates are underestimated due to large standard errors, and the mQTL apparently explains more genetic variance than is estimated to exist.

Enrichment analyses of regulatory annotations

To assess the relative enrichment of regulatory annotations amongst mQTL SNPs, we used GARFIELD.[25] Variants with missing scores were set to 0 and were excluded from the mQTL GWA files. The minimal GWA p-value for each mQTL SNP across all DNAm sites was used. P-values for SNPs with missing mQTL association statistics were set to 1. GARFIELD selects an independent set of SNPs by sequentially removing variants with r^2>0.1 within 1Mb window from the most significantly associated mQTL variant and it annotates each variant with a regulatory feature if either the variant, or a correlated variant, overlaps the feature (defined as r^2>0.8). It calculates statistical significance by using a glm model at different GWA p-value thresholds (p<1e-10 to p<1e-14) while variants are matched by MAF, distance to the nearest transcription start site (TSS), number of LD proxies, CpG and GC content. CpG and GC content were calculated for a region 500bp up and downstream of the variant using the BSgenome.Hsapiens.UCSC.hg19 R package.

We assessed the relative enrichment of 25 combinatorial chromatin states amongst mQTL using data on genomic segmentations for 127 cell types from the Epigenome Roadmap[26] and ENCODE project[27]. We calculated enrichments for each of the 171 transcription factor binding sites (TFBS) from the ENCODE[27] and CODEX projects[28]. We downloaded DNAm site to gene annotations from https://zwdzwd.github.io/InfiniumAnnotation. We used a mQTL pvalue threshold of 1e-14 in these analyses. To correct for multiple testing on the number of different annotations, GARFIELD estimates the effective number of independent annotations by using the eigenvalues of the correlation matrix of the binary annotation overlap matrix and then applies a Bonferroni correction at the 95% significance level. For the segmentation states, we defined a pvalue of 1.23e-3 as significant. For the TFBS, we considered a pvalue of 1.06e-4 as significant.
We then used Locus Overlap Analysis (LOLA)[29] to identify overlap of the mQTL sites with various functional and regulatory features including gene annotations, chromatin states and TFBS. This analysis has been performed against a background set of sites from the HumanMethylation450 array which was matched on CpG and CG content. LOLA uses Fisher's exact test with false discovery rate (FDR) correction to assess the significance of overlap in each pairwise comparison. We considered an enrichment FDR pvalue of 0.001 as significant. To investigate the effect of the chosen window size on the results of the analysis, we conducted sensitivity analyses on DNAm site enrichments for chromatin state (**Figure S20**), gene annotations (**Figure S50**) and transcription factors (**Figure S51**). We defined *trans* associations as inter-chromosomal, conducted enrichment analyses and compared enrichment odds ratios to the original analyses where we defined *trans* associations as > 1Mb. There were no differences in odds ratios for the gene annotations, transcription factors and 24/25 chromatin states.

Cross-tissue and cross-cell type DNAm

To explore the relationship between DNAm levels and tissue specificity of mQTL sites, we calculated the weighted mean of each DNAm level in blood across 36 studies and categorised a DNAm site in low (0-20%), intermediate (20-80%) or high (>80%) methylation. To understand whether cell type differences were underlying intermediate DNA in *cis+trans* mQTL sites, we downloaded T-cell DNAm profiles (GSE56581, N=214) and monocyte DNAm profiles (GSE56046, N=2,002). We extracted *cis+trans* blood DNAm sites from the

cellular subsets and plotted their mean DNAm levels across all samples within each cell type for each DNAm site.
To investigate tissue specificity, we downloaded DNAm profiles on 12 different tissues from 16 individuals (GSE78743, **Supplemental Note 1**). We extracted DNAm sites with a mQTL in blood and plotted the mean DNAm levels across all samples within each tissue type for each DNAm site. Heatmaps were plotted using the heatmap.2 function in the Rpackage ggplots (gplots_3.0.1).

Correlation of mQTL between tissues

To assess the extent of tissue specificity in mQTL we re-analysed adipose and brain tissue mQTL data using the analysis approach applied here to estimate blood mQTL **(Supplemental Note 1)**. For brain mQTL analysis, 170 fetal brain samples with high quality DNAm data and imputed SNP data were available. Using scripts from the GoDMC pipeline, DNAm data were adjusted for age (in weeks post conception), sex and the first five genetic principal components. These data were than rank normal transformed, so that the mQTL units would be consistent. All candidate genome-wide SNP-methylation probe pairs from the blood analyses were tested using the R package MatrixEQTL[13] with a linear model. For adipose mQTL analysis, SNP-DNAm site pairs identified in this study with blood data were tested in adipose samples derived from 603 twins[30] using scripts from the GoDMC pipeline.

To assess systematic between study heterogeneity we calculated M statistics[19] for each of the 38 studies (36 blood studies, 1 adipose study and 1 brain study) using 337 mQTL SNPs on chromosome 20 with a p < 1e-14 (**Figure S52**).
For each mQTL category, the correlation of genetic effects between tissues (r_b) were estimated using the r_b method[31] where we used the blood mQTL as reference. We set θ to 0 as the sample overlap between the blood and brain samples was 0. The individuals who donated adipose samples are also included in the GoDMC blood mQTL analysis. As the proportion of sample overlap is very small (603/27750) and null SNPs in mQTL regions were not available in GoDMC, we set θ to 0. For each mQTL category we only included the strongest mQTL for each DNAm site. To be consistent across mQTL categories, mQTL were filtered on p<1e-14. Sensitivity analysis showed that a pvalue threshold of 1e-8 didn't change the results. To explore the relationship between DNAm levels and tissue specificity of mQTL sites, we categorised the blood mQTL sites in low (0-20%), intermediate (20-80%) or high (>80%) methylation and re-calculated r_b.

Chromosome interaction overlaps

We tested whether interchromosomal *trans*-mQTL were enriched for overlap with chromosomal interactions using public Hi-C data[32]. The pipeline used to construct the Hi-C contact matrices uses BWA[33] to map each read end to the b37 reference genome; remove duplicate and near-duplicate reads; remove reads that map to the same fragment; and filters the remaining reads based on a mapping quality score. Filtering of abnormal alignments of each read pair to the genome included: removal of ambiguous chimeric read pairs (where "subsequences" of a read aligns to different parts of the chromosome) and unalignable read pairs (where at least one end cannot be successfully aligned). In addition, the reads were filtered for duplicates (considered duplicates if reads lie at closely corresponding positions; i.e. within 4bp of one another). We used Hi-C data mapped at the mapping quality score of

MAPQE30. This means that the chances that an alignment is erroneous is at most 1 in 1000. Hi-C data was normalised using Knight-Ruiz normalisation on interchromosomal 1kb resolution and a quality threshold of E30 using the GM12878 LCL sample as previously described[32]. Positions of the *trans*-mQTL sites were matched to their relevant 1kb blocks in the interaction data. *Trans*-mQTL SNP LD blocks were also matched. Interchromosomal *trans*-mQTL SNP-DNAm site pairs were confirmed as overlapping interactions when either the site or SNP resided in the bait and their respective SNP or site resided in the "other end". In order to test enrichment, we generated 1000 permutation datasets of broken interchromosomal *trans*-mQTL SNP-DNAm site pairs. A permutation p-value was calculated based on how often the number of overlaps within the permutation data exceeded those of the real data. In addition, we performed a Fisher's Exact Test to compare the number of overlaps and non-overlapping pairs in our real data with the 1000 permutations.

Two-dimensional enrichments

We hypothesised that SNPs that influence sites falling under a particular annotation will themselves have non-random annotations. In other words, the annotation of the site might be correlated with the annotation of the SNP in an mQTL pair. We restricted this analysis to *trans*-mQTL only, to avoid the problems of within-locus correlations.

We used LOLA[29] to annotate SNPs and sites to overlapping TFBS using all 615 datasets (171 TFBS, 20 cell types, 25 tissues, 27 treatments) in the ENCODE database.[34] We constructed a matrix T of all *site* TFBS against all *SNP* TFBS. Under a simple null model, we expect that the number of mQTL that comprise a SNP with the first SNP-annotated TFBS s_1 and a site with the first site-annotated TFBS c_1 will be $E(t_{s_1,c_1}) = \left(n_{s_1} / \sum_{i=1}^{S} n_{s_i}\right)\left(n_{c_1} / \sum_{j=1}^{C} n_{c_j}\right)$ where n is the count of SNPs or site that have a particular TFBS annotation, S is the total number of SNP TFBS and C is the total number of site TFBS. Whether T is systematically non-random can then be evaluated by comparing it against a null matrix T_{null} in which each element is the estimate of the expected value, using Fisher's exact test.

We also performed permutation analysis to identify particular elements that were deviating from expectation more than by chance. We took the SNP-DNAm site pairs and shuffled them such that the SNPs and sites were no longer matched based on genetic association, instead they were random matchings. We constructed $p = 1..500$ new T_p matrices that were based on randomly shuffled SNP-DNAm site pairings. For each element in T we now have a permuted distribution $\mathbf{t}_{s_i,c_j}$. To identify those annotation pairs that are substantially over- or under-represented compared to chance, we need to ensure that we have accounted for a large number of multiple testing comparisons ($S \times C = 3009 \times 2478 = 7456302$). To identify those annotation pairs that are substantially over- or under-represented compared to chance, we need to ensure that we have accounted for a large number of multiple testing comparisons ($S \times C = 3009 \times 2478 = 7456302$). To do this we estimated how many SDs from the mean the most extreme values found for $\mathbf{t}_{s_i,c_j}$ were and used this as a threshold for significant enrichment.

DNAm communities

Constructing DNAm communities

Many of the *trans*-mQTL comprised SNPs that associated with other DNAm *cis*- and *trans*-sites. We defined *cis-trans* DNAm site pairs (CTDPs) as those in which two DNAm sites shared a causal variant and *trans* sites as those that were more than 1 Mb apart.

To identify CTDPs we did the following

1. For a particular *trans*-acting SNP ($p < 1e-10$), search for whether it associates with any other SNPs in *cis* ($p < 1e-8$, 1Mb radius).
2. For putative pairs from (1), obtain all SNPs within a 1Mb radius that are available for both the *cis* and *trans* site.
3. Perform colocalization analysis using the coloc.abf function in the R/coloc package with default parameters[35].
4. Retain putative pairs if the posterior probability for colocalization is >0.8.

Once a list of CTDPs had been created we next removed possible duplicate representation due to linkage disequilibrium. To do this we used a greedy algorithm whereby for each *cis-trans* chromosome pair we started by identifying the most connected (sentinel) *cis*-SNP and removed any CTDPs for which the *cis* and *trans* SNPs were within 2.5Mb of the sentinel CTDP. We then followed on to the next most connected *cis*-SNP, and so on until pruning was exhausted.

Because one site might have shared genetic factors with several other sites, we attempted to create DNAm site communities. To do this we used the R/igraph package to construct an unweighted and directed graph of site sharing, and then used the Walktrap community finding algorithm with random walks of a maximum of 20 steps to identify communities[36]. Here, communities are defined as subgraphs that are connected such that random short walks tend to stay within that community.

To test if the *cis-trans* site pairs arose because of cross-hybridisation, we checked if the sites were more likely to be found to have non-unique probe subsequences of length 25bp. We found that our community sites were strongly *depleted* for probes that were liable to cross-hybridise (OR=0.7, p=1.0e-5), likely due to the stringent exclusion criteria used for retaining sites in the mQTL discovery phase.

Community enrichment analyses

To test whether there was enrichment for genomic annotations (TFBS, chromatin state), we used LOLA[29]. This analysis has been performed against a background set of 5,109 sites sharing a causal variant with at least one other site and was matched on CpG and CG content. We used the gometh function in missMethyl (v1.12.0)[37] to calculate enrichments for Gene Ontology terms and Kyoto Encyclopedia of Genes and Genomes pathways.

To test if a DNAm site community coordinated genomic regions that were relevant to specific traits or diseases, we performed enrichment analysis using GWAS summary data. Note that we are not specifically testing whether the SNPs that influence the DNAm sites are of

relevance to GWAS (there are relatively few mQTL for each community making enrichment analysis difficult), rather we are testing if the genomic regions marked by DNAm sites in a single community coordinate to have low p-values for a specific trait. We denoted each DNAm site in a community as a marker of a genomic location, and we wanted to test if SNPs in those regions within a community were enriched for low p-values more expected by chance. To do this, for each DNAm site community we:

1. Identified a representative variant for each DNAm in the community. This was simply the single closest 1000 Genomes common variant to the DNAm site.
2. We tested if the -log10 p-values of the representative SNPs for the community were substantially larger than the -log10 p-values of representative SNPs for all other communities.

Hence our background SNPs are the representative regions for all other communities. We chose this background because it is likely to have the same ascertainment properties as those in the target community (higher MAF, higher LD proxies, distance to TSS, etc).

To perform (2) we used a model similar to that implemented in GARFIELD, in which we perform logistic regression of the -log10 p-value of the representative SNPs against a binary variable denoting whether the representative SNP was for the community being tested (y=1) or the background (y=0).

We also performed sensitivity analyses, evaluating if the regions of the communities were enriched for higher MAF, higher numbers of LD proxies, closer distance to TSS, CG content, and CpG density. The distribution of p-values for these enrichments were depleted for low p-values across all communities, and also specifically for communities that indicated enrichments for GWAS traits, indicating that enrichments are not likely driven by confounding with genomic context.

GWAS of complex traits and diseases

The GARFIELD software (see above) was used to test for over-representation of *cis* and *trans* mQTL in GWAS associated variants. Publicly available GWAS results were downloaded from a range of sources and formatted for use as annotation categories. In total 41 GWAS (37 traits) were tested, these were selected such that they had i) > 100,000 variants after LD pruning, ii) > 5 significant associations and iii) > 0 overlapping GWAS significant variants and *trans* mQTL. SNPs associated with complex traits defined as those associated at genome-wide significance ($p < 5e-8$). All variants from the UK10K/1000G were used as the background set of genetic variants. In the enrichment, variants were matched by MAF, distance to the nearest TSS, number of LD proxies, CpG and GC content. We considered a p-value of 0.05 divided by number of traits*number of SNP categories as significant ($p<4.5e-4$).
Next, we tested for over-representation of *cis* and *trans* mQTL in GWAS associated variants from 36 blood related traits[38] using the same procedure as described above.

Colocalisation analysis with complex traits

We hypothesised that some sites will have shared causal variants with complex traits. We used the MR-Base database to pull down traits that had an association with an mQTL SNP with p < 1e-5. For each putative DNAm site-trait pair we then performed colocalisation analysis using the coloc.abf function in the R/coloc package, using default parameters.[35]

For several putative DNAm site-trait pairs in which we had identified a shared causal variant, those sites also had additional mQTL. We hypothesised that if the site was causal for the trait (giving rise to the initially detected shared causal variant), then any other SNPs influencing the site should also associate with the mQTL. Furthermore, we expect that the association should have a consistent sign, that is if the effect allele for one SNP had a positive effect on the site and a negative effect on the trait, then the effect allele at other mQTL should have opposite signs for the site and trait also. For sites that had a single extra independent mQTL, we compared the Wald ratio ($\beta_{trait}/\beta_{site}$)[39] for the original mQTL with the Wald ratio for the subsequent mQTL. If there were multiple additional independent mQTL then the Wald ratios were meta-analysed using the inverse variance weighted (IVW) method[40].

Mendelian randomization analysis of influences of traits on DNAm sites

We estimated the causal effects of 116 complex trait levels[41] (**Table S18**) on 345,109 sites. These sites were selected on the basis of being in the 90% most variable DNAm sites in at least 20 of the studies contributing to the meta-analysis. To estimate the causal effects, we used two-sample Mendelian randomization (2SMR)[41]. The procedure for 2SMR is described briefly here, but there has been much more extensive treatment elsewhere[42,43]. For a particular trait, genetic instruments are obtained by clumping complete GWAS summary data with parameters of LD r^2 = 0.001 and 10 Mb LD windows, retaining only SNPs with p < 5e-8. For a particular trait-DNAm site association we obtain the effects of each instrument on the site and harmonise to ensure that the effect estimates for the trait and the site are based on the same effect allele. If there is only one instrumental variable, we use the Wald ratio to obtain an estimate of the causal effect. If there are multiple instruments we use the IVW estimate using modified 2nd order weights[44] to avoid having to rely on the no measurement error in the exposure (NOME) assumption. For IVW estimates we estimate the heterogeneity using Cochran's Q statistic[45]. We use the contribution of each SNP using modified 2nd order weights to the Cochran's Q statistic, q_i as an indication of being an outlier, which is chi-square distributed with one degree of freedom. Outliers are determined by having p-values smaller than 0.05 divided by the number of instrumenting SNPs for that particular analysis.

Sensitivity analyses

Amongst the IVW analyses that had p < 1.4e-7, 81 were instrumented by SNPs only in the *MHC* region and are likely unreliable due to non-specificity of the instruments. 144 involved 14 traits that were instrumented by both *MHC* and *non-MHC* SNPs. In order to evaluate the reliability of these results we compared the causal effect estimates of SNPs outside the *MHC* region with causal effect estimates from SNPs within the *MHC* region. The agreement of causal estimates from *MHC* and *non-MHC* SNPs was very high for ulcerative colitis, rheumatoid arthritis, juvenile idiopathic arthritis, LDL cholesterol, percent emphysema, birth weight, red blood cell count, multiple sclerosis and coeliac disease, all showing that over

85% of the causal directions were concordant. This provides confidence that DNAm levels are likely to be substantially causally influenced by natural variation in complex traits. The remaining traits had associations that were not generally consistent when comparing instruments from the *MHC* region against elsewhere. The *MHC* region is an obvious potential source of bias in MR analysis because it has known large effects on many traits across an extended region. But we also saw similar biases arising in MR due to other regions. For example, we found many associations between age of menarche and sites because one of the instruments for age of menarche is within the *CHRNA5* locus. Observing these inconsistencies led us to develop the MR Sign Concordance Test (MR-SCT), which attempts to reduce the possibility of the weight of a single or few instruments biasing the IVW estimate without requiring knowledge of potentially problematic regions. The rationale and details of the method are outlined in the **Supplementary Note 2**.

We developed a decision tree to prioritise significant IVW associations that were most likely due to causal relationships, rather than violations of assumptions in MR (**Figure S39**). The principles behind the decision tree are as follows:

1. When there is a single genetic instrument for a trait it is difficult to prove that the SNP influences the site through the trait, rather than the SNP influencing both the trait and the site through independent pathways. We could not perform genetic colocalization analysis for these cases because we only had a single SNP within the region, rather than all SNPs surrounding the instrument - anything else would have been computationally prohibitive in a meta-analysis setting. These single instrument tests are filtered based on p-values for the Wald ratios, but caution is made that further analysis is required to investigate possible causality.
2. If a trait has more than one independent instrument, then we can apply the logic that agreement across multiple instruments reduces the likelihood that horizontal pleiotropy is driving the overall estimate. There are tests that can be applied to evaluate the extent to which there is agreement amongst the methods. Cochran's Q statistic tests for heterogeneity amongst the Wald ratios obtained from each instrument. High heterogeneity indicates that at least one instrument could be invalid (e.g. due to horizontal pleiotropy). In this instance, further analyses are required to evaluate different models of horizontal pleiotropy. We used the MR-simple-median method[46] because it is unweighted, meaning it will not be liable to the types of problems seen with the *MHC* region disproportionately affecting estimates. If there was a consistent effect using the simple-median estimator, then this indicated that the putative association was more likely to be reliable. In conjunction with this approach we developed MR SCT to operate alongside the heterogeneity statistics, with rationale outlined in the **Supplementary Note 2**.

Influences of traits on many sites

To evaluate if a single trait has influences on many sites, we reasoned that the p-values across all sites would be slightly lower than expected by chance, due to the null hypothesis of no association being consistently false. To test this, we used the genomic inflation estimator of calculating the ratio between the expected and observed median chi-square statistics (GC_{in}).

The median test statistic should not be influenced by a single region (e.g. the *MHC*) leading to apparent inflation. But to be sure of this we also performed two sensitivity analyses: single-chromosome analyses where we looked for consistency of the GC_{in} estimate when restricting to sites only on each individual chromosome; and leave-one-chromosome-out analyses, where we tested across the whole genome but sequentially excluding one chromosome at a time, to see if the estimate shifted dramatically upon the exclusion of a particular chromosome.

Selection metrics

Enrichment analysis of selection scores

To assess enrichment amongst mQTL SNPs, we used GARFIELD[25] using a similar approach as described for the regulatory annotations. We generated five annotations reflecting different types of positive selection over different time scales (**Table S24**)[47] including: SDS (UK10K)[48], F_{st} (Global F_{st} (CEU versus YRI versus CHB)), iHS (CEU)[49], XPEHH (CEU versus YRI)[50] and XPEHH (CEU versus CHB)[50]. Scores for these selection metrics were downloaded from: http://hsb.upf.edu/hsb_data/positive_selection_NAR2013/. For each annotation, variants were set to 1 if a variant had a selection score with p<0.01 and to 0 if a variant had a selection score with p>0.01. We removed the *MHC* and *LCT* regions from all our analyses. For the GWA, we used the minimal GWA p-value for each mQTL SNP across all DNAm sites. Variants with missing selection scores were set to 0 and were excluded from the mQTL GWA files. P-values for SNPs with missing mQTL association statistics were set to 1. We used a mQTL p-value of 1e-14 as GWA cut-off p-value and considered an enrichment p-value of 0.05/5 selection metrics = 0.01 as significant. In the enrichment, variants were matched by MAF, distance to the nearest TSS, number of LD proxies, CpG and GC content.
To assess the relationship between effect size of the mQTL SNP and selection scores, we used a linear model to regress selection scores against the strongest absolute mQTL effect size accounting for the number of LD proxies, distance to TSS, CpG and GC frequency.

Enrichment analysis of mQTL SNPs amongst complex traits

To examine enrichment of mQTL SNPs with extreme SDS amongst complex trait GWA signals we generated an annotation where we set variants with a SDS score p<0.01 that were overlapping a *cis only* or a *cis+trans* mQTL LD region to 1 and all other variants to 0. We selected 42 GWA datasets (37 traits) across 11 disease/trait categories including several datasets with extreme anthropometric phenotypes (**Table S20**). Out of the 42 datasets, 19 datasets showed an overlap with at least one *cis acting* SNP overlapping an extreme SDS score. We therefore considered a p-value of 2.6e-3 (0.05/19) traits as significant.

Comparison of genetic variance

For the five traits that were enriched for extreme SDS overlapping mQTL SNPs, we compared the genetic variance for trait associated mQTL or trait associated mQTL with extreme SDS against all trait associated SNPs.
We used the formula below to estimate the genetic variance:

$$Genetic\ variance\ = 2 * \beta^2 * f(1-f)$$

where β is the absolute maximum effect size for each SNP and f is the MAF.

Measure of genomic architecture

Many methods (for example, BayesS[51]) attempt to fit a model of the form:

$$var(\beta) = \sigma^2[f(1-f)]^{-S}$$

to the distribution of effect sizes β(on some trait) as a function of MAF f. If $S = 0$, then effect size is unrelated to frequency and the trait is unlikely to be under selection. If $S > 0$ this is evidence that the trait is under negative selection, as SNPs with large effects are reduced in frequency.

There are two main sources of deviation from this model that both create a form of "censoring", that is, that we cannot observe the true value of β when either β or f are small. The first is that effect sizes that are small cannot be reliably estimated for low frequency variants, due to power. The power to detect a variant is proportional to its total effect, i.e. $\sigma^2 f(1-f)$ and thus, is a function of f, creating bias. This could be modelled with heteroskedasticity.

The second is an explicit censoring in the way we constructed the trait. To consider DNAm as a general trait, we have combined a large number of DNAm sites and considered the effect size of a large number of loci for each. To perform the computation, we omitted many of the effects. This means that the trait under question is not a particular DNAm site but is instead the maximum effect that each SNP has on any DNAm site. This creates a bias in the distribution as small values of β cannot be observed, and the bias will be a function of MAF category. Further, for computational reasons we did not evaluate all possible SNPs on all possible DNAm sites. This cannot be simply modelled as it is taking the form of true censoring.

Additionally, there may be a mixture in the effect size distribution; for example, some SNPs will not affect the trait, others (e.g. *trans* SNP effects) may have small effects and others (e.g. *cis* SNP effects) may have large effects. Modelling this appears complex.

To solve all of these problems, we instead focus only on the upper tail of the SNP-effect distribution. We assume that the data contain at least some SNPs were truly generated by the functional form:

$$p(\beta|f) = N(0, \sigma^2[f(1-f)]^{-S})$$

We then choose two thresholds, $t_1$and $t_2 > t_1$ in terms of the number of SDs into the (two-tailed) distribution and create a summary statistic for a dataset and parameter value set (S, σ) conditional on a set of frequency bins $B = \{b\}_j$ as the number of SNPs in each frequency category that exceed their predicted threshold frequency under the null distribution:

$$S(\{\beta, f\}_i, S, \sigma) = \{\#\beta_i > t_1(f_i), \#\beta_i > t_2(f_i)\}_{f_i \epsilon b_j}$$

Provided that the thresholds are chosen sufficiently large to make all frequency bins unaffected by censoring, then the number in each bin j that are above t_2(called $N_{2,j}$) should be simply related to the number that are above t_1(called $N_{1,j}$):

$$N_{2,j} \sim Binom(N_{1,j}, p)$$

where p is simply computed from the distribution function of the Normal distribution at the chosen thresholds:

$$p = q_1/q_2$$

The difficulty in this approach lies in the fact that different data are discarded depending on the parameters. This makes defining a likelihood difficult, and even defining a loss is difficult since we do not want it to penalise parameter values that retain a high proportion of the data.

We resolve this by loss function defined as the average log-probability under the binomial model, per-accepted datapoint:

$$L(S(\{\beta, f\}_{i,} S, \sigma)) = \sum_{j=1}^{n_B} Log - Binom(N_{2,j}; N_{1,j,} p)/N_{1,j}$$

There might be some concern about this behaving poorly when the model chooses to reject much of the data; however, in practice a large amount of data is needed to make $N_{2,j}/N_{1,j}$ close to p, which is necessary for the contribution from a given bin to be minimised. Therefore, in practice this measure penalises both small retained datasets, as well as those that have a non-uniform assignment of effect sizes into MAF bins.

For our purposes it suffices to use a grid search to find acceptable losses. We use resampling in order to establish an appropriate confidence interval. We formed a confidence interval from all values of S that under bootstrap resampling were ever seen to have lower loss than the inferred value and extended this to halfway to the next point.

Figure S44 illustrates this procedure. **Figure S44D** shows what the distribution of MAF and beta should look like for a well calibrated and pair, with the fitted model clearly following both the tail density and also the maximum observed value. **Figure S44C,E** show poor choices of S. The summary statistics for these situations are shown in **Figure S44F-H**, again confirming that S=0.4 is a good choice. **Figure S44A-B** shows the loss function as a function of S (taking the best values of S for each σ).

Figure S45 shows how different subsets of SNPs behave, comparing *cis+trans* or *cis only* to all SNPs. The inferred S may increase slightly in these cases, which is evidence that these two subsets of SNPs have different distributions which are made more uniform when mixed.

This process should not be considered as a full model for the data, but instead as an informative descriptive statistic describing the distribution. In theory, it is possible that there is a different “true S” for different choices of threshold, if there is a mixture of underlying true effects. This would occur for example if a stringent threshold included only *cis* effects but a more generous threshold included *trans* effects.

Simulations of genomic architecture

To confirm that the method we use to discover genomic architecture is accurate, we performed a simulation study. We simulated N SNPs and retained the fraction r that were largest in terms of variance explained; that is, we retained the Nr SNPs that had the largest value of $\beta f(1-f)$. We then repeated this simulation 10 times for a range of values of N and r, as shown in **Figure S53**.

Statistical uncertainty due to power has the same shape as genetic architecture[31] **(Figure S54A**) because lower minor allele frequency means lower power. To confirm that the signal we see is due to real effects and not power, we compared the effect size to its statistical uncertainty (**Figure S54B**). We noted that there is a factor around 16 when MAF=0.01, with a mean of 38 (with MAF and the ratio modelled linearly on a log-scale). This is negligible with respect to our model, which produces the same qualitative estimates if we instead use the conservative effect sizes sign(beta)*(|beta| - se(beta)) (**Figure S54C-D**). Similarly, we

are above the MAF threshold T for which different qualitative relationship between MAF and effect size is operating[52]. A lack of such threshold is visible in **Figure S54A** and is confirmed by sensitivity analysis in which we repeat our inference model for S using a MAF threshold of 0.05 (**Figure S54E-F**). This again does not qualitatively change the results, though does widen the confidence interval significantly to include S=0.5, because the power to distinguish similar values of S comes in the low frequency region.

Coverage

Using the annotation file from https://zwdzwd.github.io/InfiniumAnnotation we selected 331,884 DNAm sites that were annotated to 18,993 protein-coding genes. For each of the genes, we counted the number of probes on the 450k array and the number of *cis* and *trans* mQTL. Using linear regression, we calculated the relationship between the median number of probes by protein-coding gene on the 450k array and the median number of *cis* and *trans* mQTL. There was a similar linear relationship between 58,356 450k probes with 43,324 *cis* mQTL and 3,330 *trans* mQTL in regions that were not annotated to a protein-coding gene ("non-genic regions").

Genetic architecture

If we know the number of mQTL with a particular r^2value, and the power of detecting mQTL with that value as described above, then we can obtain a rough estimate of how many mQTL would expect to exist with that value regardless of power. Here we use the GS replication mQTL r^2values, and estimate the expected number of mQTL for a particular r^2value as being

$$n_{expected} = \frac{n_{observed}}{1 - p(miss)}$$

References

1. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
2. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22–e22 (2017).
3. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
4. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, (2015).
5. 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).

6. Min, J. L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty476.
7. Zeilinger, S. *et al.* Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS One* **8**, e63812 (2013).
8. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
9. 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).
10. Naeem, H. *et al.* Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics* **15**, 51 (2014).
11. Price, M. E. *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* **6**, 4 (2013).
12. Chen, Y.-A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).
13. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
14. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
15. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177–188 (1986).
16. Hedges, L. V. & Olkin, I. Random Effects Models for Effect Sizes. in *Statistical Methods for Meta-Analysis* 189–203 (1985).
17. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics

identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–75, S1–3 (2012).

18. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. (2015) doi:10.1101/035170.
19. Magosi, L. E., Goel, A., Hopewell, J. C., Farrall, M. & CARDIoGRAMplusC4D Consortium. Identifying systematic heterogeneity patterns in genetic association meta-analysis studies. *PLoS Genet.* **13**, e1006755 (2017).
20. Thompson, S. G. & Higgins, J. P. T. How should meta-regression analyses be undertaken and interpreted? *Stat. Med.* **21**, 1559–1573 (2002).
21. Navrady, L. B. *et al.* Cohort Profile: Stratifying Resilience and Depression Longitudinally (STRADL): a questionnaire follow-up of Generation Scotland: Scottish Family Health Study (GS:SFHS). *Int. J. Epidemiol.* **47**, 13–14g (2017).
22. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
23. McRae, A. F. *et al.* Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biol.* **15**, R73 (2014).
24. van Dongen, J. *et al.* Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat. Commun.* **7**, 11115 (2016).
25. Iotchkova, V. *et al.* GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. (2016) doi:10.1101/085738.
26. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
27. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
28. Sánchez-Castillo, M. *et al.* CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res.* **43**, D1117–23 (2015).
29. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and

regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587–589 (2015).

30. 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).
31. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.* **9**, 2282 (2018).
32. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
33. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* vol. 26 589–595 (2010).
34. Encode, T. & Consortium, P. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
35. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
36. Pons, P. & Latapy, M. Computing Communities in Large Networks Using Random Walks. *J. Graph Algorithms Appl.* **10**, 191–218 (2006).
37. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32**, 286–288 (2016).
38. 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).
39. Wald, A. The Fitting of Straight Lines if Both Variables are Subject to Error. *Ann. Math. Stat.* **11**, 284–300 (1940).
40. Toby Johnson, G. S. U. Efficient Calculation for Multi-SNP Genetic Risk Scores. *citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.398.7674*.
41. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, (2018).
42. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–R98 (2014).

43. Hemani, G., Bowden, J. & Smith, G. D. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* **27**, R195–R208 (2018).
44. Bowden, J. *et al.* Improving the accuracy of two-sample summary data Mendelian randomization: moving beyond the NOME assumption. (2017) doi:10.1101/159442.
45. Cochran, W. G. The comparison of percentages in matched samples. *Biometrika* **37**, 256–266 (1950).
46. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
47. Pybus, M. *et al.* 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res.* **42**, D903–9 (2014).
48. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).
49. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
50. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
51. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).
52. Eyre-Walker, A. Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc. Natl. Acad. Sci. U. S. A.* **107 Suppl 1**, 1752–1756 (2010).