Genome Biology

RESEARCH

Open Access

Genetic impacts on DNA methylation help elucidate regulatory genomic processes

Sergio Villicaña[1*], Juan Castillo-Fernandez[1], Eilis Hannon[2], Colette Christiansen[1], Pei-Chien Tsai[1], Jane Maddock[3], Diana Kuh[3], Matthew Suderman[4], Christine Power[5], Caroline Relton[4,5], George Ploubidis[6], Andrew Wong[3], Rebecca Hardy[7,8], Alissa Goodman[6], Ken K. Ong[9] and Jordana T. Bell[1*]

*Correspondence: sergio.villicana_munoz@kcl.ac.uk; jordana.bell@kcl.ac.uk

[1] Department of Twin Research and Genetic Epidemiology, King's College London, London, UK
[2] University of Exeter Medical School, Exeter, UK
[3] MRC Unit for Lifelong Health and Ageing, Institute of Cardiovascular Science, University College London, London, UK
[4] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK
[5] Population, Policy and Practice, UCL Great Ormond Street Institute of Child Health, University College London, London, UK
[6] Centre for Longitudinal Studies, Institute of Education, University College London, London, UK
[7] School of Sport, Exercise & Health Sciences, Loughborough University, Loughborough, UK
[8] UCL Social Research Institute, University College London, London, UK
[9] MRC Epidemiology Unit and Department of Paediatrics, Wellcome Trust-MRC Institute of Metabolic Science, University of Cambridge School of Clinical Medicine, Cambridge, UK

Abstract

Background: Pinpointing genetic impacts on DNA methylation can improve our understanding of pathways that underlie gene regulation and disease risk.

Results: We report heritability and methylation quantitative trait locus (meQTL) analysis at 724,499 CpGs profiled with the Illumina Infinium MethylationEPIC array in 2358 blood samples from three UK cohorts. Methylation levels at 34.2% of CpGs are affected by SNPs, and 98% of effects are *cis*-acting or within 1 Mbp of the tested CpG. Our results are consistent with meQTL analyses based on the former Illumina Infinium HumanMethylation450 array. Both SNPs and CpGs with meQTLs are overrepresented in enhancers, which have improved coverage on this platform compared to previous approaches. Co-localisation analyses across genetic effects on DNA methylation and 56 human traits identify 1520 co-localisations across 1325 unique CpGs and 34 phenotypes, including in disease-relevant genes, such as *USP1* and *DOCK7* (total cholesterol levels), and *ICOSLG* (inflammatory bowel disease). Enrichment analysis of meQTLs and integration with expression QTLs give insights into mechanisms underlying *cis*-meQTLs (e.g. through disruption of transcription factor binding sites for CTCF and SMC3) and *trans*-meQTLs (e.g. through regulating the expression of *ACD* and *SENP7* which can modulate DNA methylation at distal sites).

Conclusions: Our findings improve the characterisation of the mechanisms underlying DNA methylation variability and are informative for prioritisation of GWAS variants for functional follow-ups. The MeQTL EPIC Database and viewer are available online at https://epicmeqtl.kcl.ac.uk.

Keywords: DNA methylation, Heritability, GWAS, Methylation quantitative trait loci, meQTL

Background

DNA methylation is a major regulator of gene function, with important roles in development and over the life course [1–3]. In humans, DNA methylation and demethylation occur predominantly at cytosine-guanine dinucleotides (CpG sites) through the action of DNA methyltransferases and TET enzymes, respectively [4, 5]. The human methylome consists of a mosaic of regions exhibiting variable stability over time, including both longitudinally stable regions, as well as dynamic regions where changes can relate to ageing or reflect environmental exposures, such as smoking [6, 7].

Multiple studies have shown that genetic effects have considerable impacts on DNA methylation levels at specific CpGs. Family and twin-based estimates of narrow-sense heritability in DNA methylation levels in blood [8, 9] report a wide range from 0 to 1 heritability at individual CpGs profiled by the Illumina Infinium HumanMethylation450 array (450K). Most studies typically associate genetic variation at single nucleotide polymorphisms (SNPs), or methylation quantitative trait loci (meQTLs), to DNA methylation levels at a specific CpG. A large proportion of reported meQTLs are in close proximity to the tested CpG (usually within 1 Mbp, in *cis*), while long-range and inter-chromosomal associations (*trans*) only represent a small fraction of meQTL associations. A recent large-scale study in 27,750 European samples estimated that DNA methylation levels at up to 45% of CpGs in the blood 450K methylome are associated with meQTL SNPs [10], which are in turn more likely to be GWAS signals than expected by chance. Another recent analysis of 3799 European and 3195 South Asian samples further explored trans-ancestry effects, and confirmed multiple links between meQTLs and phenotype variation [11]. In addition to analyses based on blood, a variety of studies have also identified meQTL SNPs in different tissues, for instance in brain [12, 13], adipose [14] and buccal tissue samples [15].

The most extensively used profiling technology for human methylome analyses to date has been the Illumina 450K array, comprising approximately 480,000 probes [16], and the vast majority of meQTL reports are based on this platform. However, the 450K array has limited coverage outside of CpG islands (CGIs) and genic regions. The most recent Illumina methylation array, the Infinium MethylationEPIC BeadChip (EPIC), improves genomic coverage of enhancers which are key regulatory regions. The EPIC array assays 853,307 sites, adding 333,265 novel CpGs in enhancers to the near entire set of 450K CpGs [17]. Accordingly, there is a need for follow-up analyses to identify new genetic influences on DNA methylation levels profiled by the EPIC array. To date, only two studies have explored meQTLs on the EPIC array in blood, but both included relatively modest sample sizes ($n = 156–1111$) [18, 19] and did not consider both genome-wide *cis*- and *trans*-meQTL effects.

Here, we report novel genome-wide meQTL analyses of DNA methylation profiles on the Illumina EPIC array, applying a meta-analysis across 2358 samples from three UK population cohorts: TwinsUK [20], the MRC National Survey of Health and Development or 1946 British Birth Cohort (1946BC) [21, 22], and the National Child Development Study or 1958 British Birth Cohort (1958BC) [23, 24]. We characterised novel meQTLs specific to the Illumina EPIC array, and carried out genome annotation enrichments and meQTL integration with summary statistics from 54 GWASs and previously reported eQTLs. A database and viewer of results is available online at https://epicmeqtl.kcl.ac.uk.

Results

We explored genetic impacts on all CpGs profiled by the Illumina EPIC array by initially estimating twin-based narrow-sense heritability, and subsequently identifying common genetic variants associated with DNA methylation levels in *cis* and *trans* genome-wide. We independently analysed samples from each of the three UK cohorts (TwinsUK, 1946BC and 1958BC cohort) separately, and meta-analysed the results. Results are presented at a permutation-based false discovery rate (FDR). We validated a subset of our findings in an external meQTL catalogue from the GoDMC study [10], and replicated selected meQTLs in target regions using methylated DNA immunoprecipitation sequencing (MeDIP-seq) in an independent sample. Follow-up analyses

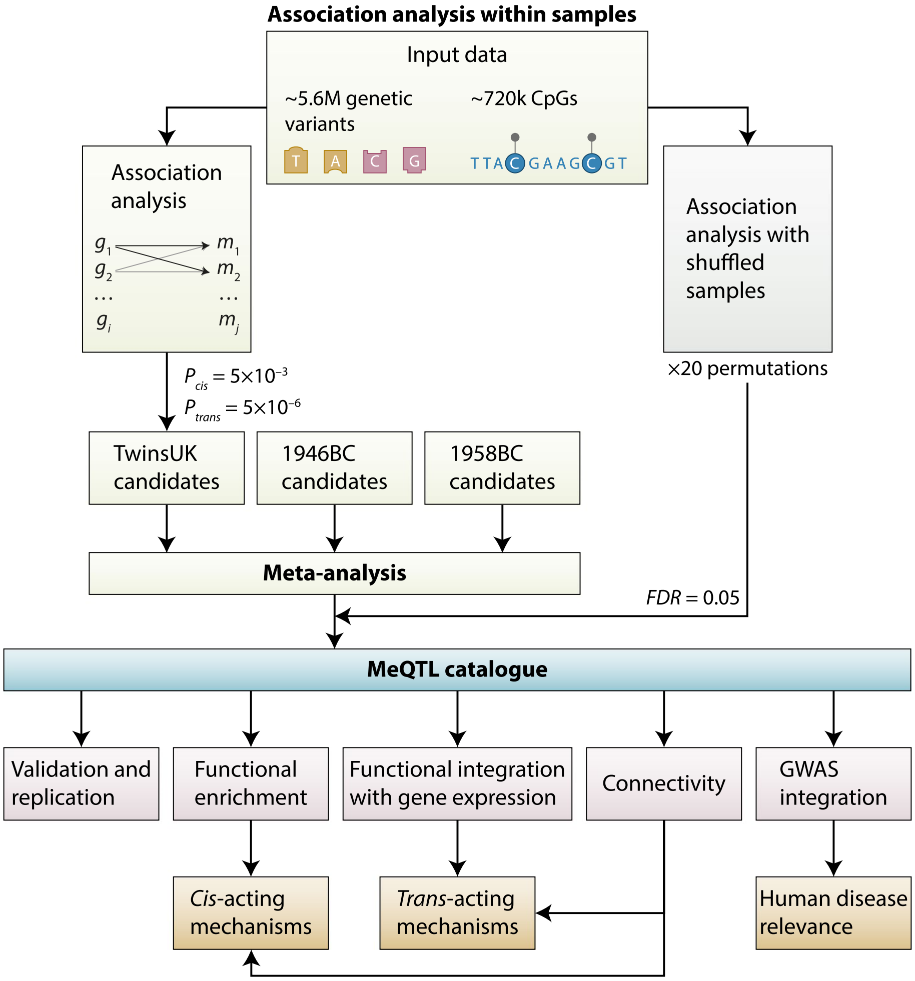

Fig. 1 Study design. Genome-wide association analyses compared genotypes and DNA methylation levels profiled by the EPIC array. Each cohort sample was independently tested, and results were meta-analysed. Results are presented at a permutation-based false discovery rate (FDR). Follow-up analyses aimed to find evidence of underlying mechanisms and their relevance to human disease

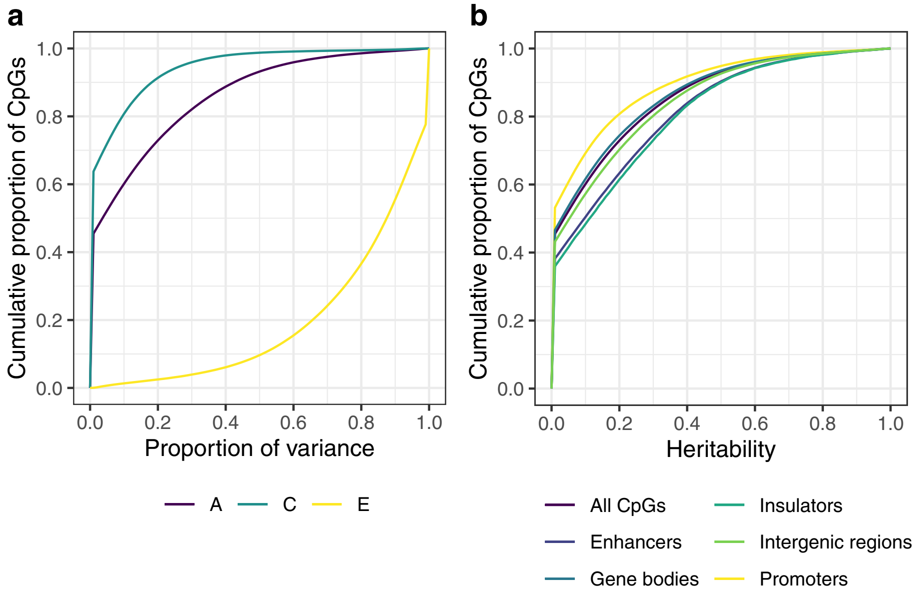

Fig. 2 Proportion of variance of genome-wide DNA methylation levels attributed to genetic variation. Estimates for the 723,814 CpGs sites covered by the EPIC array after a classical twin study of 88 MZ and 70 DZ twin pairs from the TwinsUK cohort. **a** Cumulative proportion of variance components of the *ACE* model: variance explained by additive genetic effects, or heritability (*A*), common environmental effects (*C*) and nonshared environmental effects (*E*). **b** Cumulative proportion of heritability estimates by genomic annotations

included enrichment analyses within genomic annotations and ontologies, and co-localisation integrating meQTLs with previously reported eQTLs and summary statistics from 56 GWASs, as well as clumping SNPs based on linkage-disequilibrium (LD) (Fig. 1).

Heritability of the Illumina EPIC DNA methylome

We initially applied a classical twin study of 88 monozygotic (MZ) and 70 dizygotic (DZ) twin pairs from the TwinsUK cohort, to decompose the DNA methylation variance at each of 723,814 CpGs into additive genetic effects (A), common environmental effects (C) and nonshared environmental effects (E) (full summary statistics available at [25]). The heritability distribution was zero-inflated (45.5% of sites have $A < 0.01$), and the maximum individual CpG heritability was 0.998 (cg21906335 in the promoter region of *ZNF155*). Across all tested CpGs, the mean genome-wide narrow-sense heritability was $A = 0.138$ ($sd = 0.198$; median $A = 0.037$, IQR $= 0.220$) (Fig. 2a). When stratifying by genomic annotations, CpGs in enhancers tend to have overall greater heritability estimates (mean $A = 0.179$, $sd = 0.217$, 95% CI [0.178, 0.181]), for example, compared to promoters, which have one of the lowest heritability estimates (mean $A = 0.106$, $sd = 0.179$, 95% CI [0.105, 0.106]) (Fig. 2b). The improved representation of enhancer regions on the EPIC array may be reflected in a modestly greater mean heritability across novel EPIC-only sites ($A = 0.142$, $sd = 0.198$, $n = 348,091$) than that across 450K legacy probes ($A = 0.135$, $sd = 0.198$, $n = 375,336$; one-tailed t-test, $t_{(723,425)} = 15.2$, $P < 2.2 \times 10^{-16}$) (Additional file 1: Fig. S1a). Overall, the heritability patterns of the genomic annotations are consistent between EPIC-only probes and 450K legacy probes

Table 1 Summary characteristics for the five UK cohort sample sets

Cohort[a]	Sample size	Percentage females	Percentage smokers	Median age [range]
TwinsUK	394	100	5.6	64.5 [42.4, 86.6]
1946BC-99	1348	52.3	24.2	53.4 [53, 54]
1946BC-09	197	59.4	11.2	63.2 [60.3, 64.6]
1958BC-1	183	50.8	21.3	45.1 [44.5, 45.8]
1958BC-2	236	54.2	42.4	45.1 [44.3, 46.0]
Total	2358	60.9	21.5	53.5 [42.4, 86.6]

[a] 1946BC-99 and 1946BC-09 refer to independent samples from the 1946 birth cohort collected at two different time points and stratified to facilitate data handling. 1958BC-1 and 1958BC-2 refer to samples from the 1958 birth cohort processed in two different batches (see the 'Methods' section for further details)

(Additional file 1: Fig. S1b–c), and are broadly in line with previously reported 450K heritability estimates across genomic annotations [9]. We also found that variable CpG sites (with methylation β-values $sd > 0.025$, see Additional file 1: Section 1.2) tend to be the most heritable. For example, the average heritability of the most variable sites ($A = 0.278$) was double that estimated genome-wide ($A = 0.138$) and the zero-inflation rate was substantially lower (Additional file 1: Section 1.2).

Common genetic variation has major impacts on the blood methylome

To identify specific genetic variants that impact the methylome, a meQTL analysis was carried out with a total of 2358 whole blood samples across five datasets from three non-overlapping human cohort studies: TwinsUK, 1946BC and 1958BC (Table 1). Initially all SNP-CpG pairs were tested for association within each dataset. CpG and SNP associations within 1 Mbp (upstream and downstream) were considered to be in *cis*, and all others were considered to be in *trans* (Additional file 1: Table S1). SNP-CpG associations that surpassed relaxed significance thresholds within each dataset were retained for meta-analysis, with a total of 189,202,234 unique candidate *cis* meQTL-CpG pairs ($P \leq 5 \times 10^{-3}$) and 100,814,822 *trans* pairs ($P \leq 5 \times 10^{-6}$). After meta-analysis we retained meQTL-CpG pairs where the strength of association surpassed FDR 5% ($P_{cis} \leq 2.21 \times 10^{-4}$, $P_{trans} \leq 3.35 \times 10^{-9}$), and where pairs were identified as candidates in more than one dataset with a consistent direction of effect.

We identified 244,491 CpGs (33.7% of tested probes) to be under the influence of *cis*-meQTL SNPs, and 5219 CpGs (0.7% of tested probes) to be influenced by *trans*-meQTL SNPs (full summary statistics available at [25]). Of these, 2281 CpGs (0.9% of CpGs with *cis*-meQTL; 43.7% of CpGs with *trans*-meQTL) were influenced by at least one *cis* and one *trans* meQTL SNP simultaneously. There were 4,609,875 unique genetic variants identified as *cis*-meQTLs, and 240,866 identified as *trans*-meQTLs. Of these, 229,908 meQTLs were both *cis*- and *trans*-meQTLs for CpGs at different sites. The meQTL SNPs and CpGs under genetic control altogether formed 39,110,128 *cis* and 805,319 *trans* SNP-CpGs pairs (Fig. 3a). We carried out sensitivity analysis by splitting the CpGs into two sets, 450K legacy probes and EPIC-specific probes, and repeating the meQTL discovery process, and found that the resulting proportions of meQTLs reported remained very similar (Additional file 1: Section 1.12). The strength of the associations was stronger for *trans* SNP-CpG pairs than for *cis* SNP-CpG pairs, although

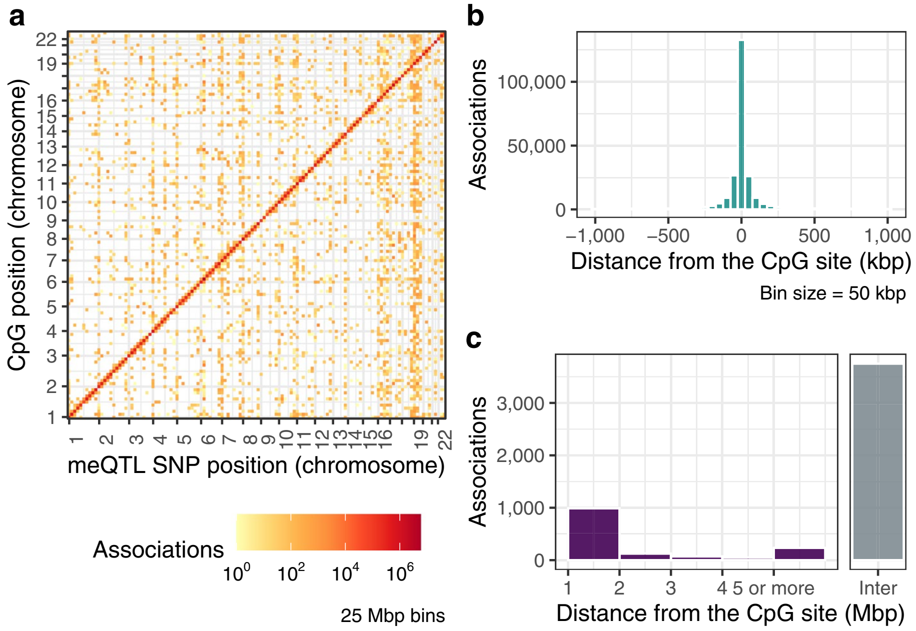

Fig. 3 DNA methylation quantitative trait loci (meQTLs) for CpG sites genome-wide. Association analysis carried out between 724,499 CpGs vs. 6,361,063 SNPs. **a** Genomic distribution of meQTL associations at a significance level of FDR < 0.05. The *x*-axis corresponds to the position of the SNPs within the 22 chromosomes and the *y*-axis to the position of the CpGs, with each pixel binning a range of 25 Mbps. The colour scale indicates the number of associations between specific CpGs/SNPs locations, on a logarithmic scale. **b** Histogram of distances between the CpGs and their most significant *cis*-meQTL SNPs. **c** Bar plot of absolute distances between the CpGs and their most significant *trans*-meQTL SNPs. Intra-chromosomal associations are shown in purple, and inter-chromosomal in grey

this is an expected result due to the difference in *P*-value thresholds for *cis* and *trans* associations (Additional file 1: Section 1.3). On the other hand, *trans* effects were more heterogeneous across samples compared to *cis* effects, and therefore the reported *trans* effects should be interpreted with caution. We estimated that, on average, a *cis*-meQTL explains 7.6% of the methylation variance of its associated CpG, while a *trans*-meQTL explains 11.5%, which is a significant difference (Additional file 1: Section 1.4). CpGs with both *cis* and *trans* associations, and SNPs that act as both *cis*- and *trans*-meQTLs, have associations with higher R^2 estimates (Additional file 1: Section 1.4). CpG sites with meQTLs were evenly distributed across chromosomes according to number of genes per chromosome. However, this pattern was not observed for meQTL SNPs (Additional file 1: Section 1.5).

Cis-meQTLs exhibited relatively short-range effects as expected [10, 18]. The median distance between each SNP *cis*-meQTL and its target CpG was 20.5 kbp (interquartile range (IQR) = 65.5 kbp) if considering the most significant association (Fig. 3b), and 75.5 kbp (IQR = 165.5 kbp) if considering all significant associations (Additional file 1: Fig. S6a). CpGs with *trans* associations have almost exclusively intra-chromosomal or inter-chromosomal meQTLs, and cases in which both types occur are rare (only 25 CpGs). For *trans* associations, 71.8% of the most significant associated SNPs per CpG are inter-chromosomal (Fig. 3c). When considering all *trans* associations the number of inter-chromosomal SNP-CpG pairs decreases to 45.4% (Additional file 1: Fig. S6b).

We also explored evidence for cell type-specific meQTL effects. These analyses considered only *cis*-meQTLs effects specific to CD4$^+$ T cells (mean ratio $= 0.199$, $sd = 0.073$ across samples) and monocytes (mean ratio $= 0.05$, $sd = 0.026$ across samples). We focused on these cell types due to their relatively greater degree of homogeneity compared to other blood cell types (e.g. granulocytes), and previous results from large scale meQTL studies [19, 26]. We observed that 8.9% of all CpGs had *cis*-meQTL effects specific for either CD4$^+$ T cells or monocytes ($P \leq 2.21 \times 10^{-4}$; see Additional file 1: Section 1.6 and Additional file 1: Table S2). Our cell-specific results replicate a proportion of previously reported cell-specific meQTLs in CD4$^+$ T cells (17.7%) and in monocytes (17.3%) [26]. Of the CpGs that had *cis*-meQTL SNPs in whole blood, 1.1% also showed evidence for cell-specific meQTL effects ($P \leq 2.21 \times 10^{-4}$), suggesting that the majority of genetic effects that we detect on CpGs in whole blood are stable across different blood cell types (Additional file 1: Section 1.6).

Overall, meQTLs explain 14.2% of the variance in the DNA methylation heritability ($F_{(2,\,723{,}424)} = 5.99 \times 10^4$, $P < 2 \times 10^{-16}$) (Additional file 1: Fig. S8). CpGs without detected meQTLs have lower mean heritabilities ($A = 0.085$, $sd = 0.146$, $n = 476,192$) compared to CpGs that have meQTLs. CpGs that have meQTLs can be split into three groups showing increasing mean heritabilities, from CpGs with only *cis*-meQTLs ($A = 0.238$, $sd = 0.239$, $n = 242,021$), to CpGs with only *trans*-meQTLs ($A = 0.295$, $sd = 0.251$, $n = 2933$), and to CpGs with both *cis*- and *trans*-meQTLs simultaneously ($A = 0.435$, $sd = 0.259$, $n = 2281$). Overall, CpGs with both *cis*- and *trans*-meQTLs have the largest evidence for DNA methylation heritability.

Replication of novel EPIC-specific and 450K legacy CpGs with meQTLs

We pursued replication of meQTL effects at selected CpGs using previously published MeDIP-seq data in an independent sample of 2319 individuals from the TwinsUK cohort. CpGs selected for replication included a subset of ten CpGs, which had the largest effect sizes (cg07143125, cg13904258, cg00918944, cg05808124), or with the largest number of meQTL SNP associations (cg25014118, cg00128506, cg18111489, cg16423305), or with meQTL SNPs that co-localised with GWAS signals (cg11024963, cg06162668). We replicated *cis*-meQTLs for 80% of the selected CpGs, and *trans*-meQTLs for 30% (Additional file 1: Table S3 and Additional file 2), after multiple testing correction and with a consistent direction of effect.

The EPIC array doubles the coverage of the 450K array. We observed that 51.2% of CpGs with *cis*-meQTLs (125,251 CpGs) and 37% of CpGs with *trans*-meQTLs (1933 CpGs) are specific to the EPIC array. For the remaining 450K legacy CpGs with meQTLs, we validated our results by comparing them to the GoDMC database based on 32,851 blood samples [10]. Altogether, 97.0% of our 450K specific CpGs with meQTLs (in *cis* or *trans*) were also under genetic influence in the GoDMC dataset.

Genomic annotations of local and distal genetic effects show consistent enrichment in enhancers

We found overall contrasting patterns of genomic annotations for CpGs with *cis*- and *trans*-meQTLs (Fig. 4a and Additional file 3: Table S6). CpGs located in CpG islands (CGIs), promoters and transcription factor binding sites (TFBSs) are less likely to

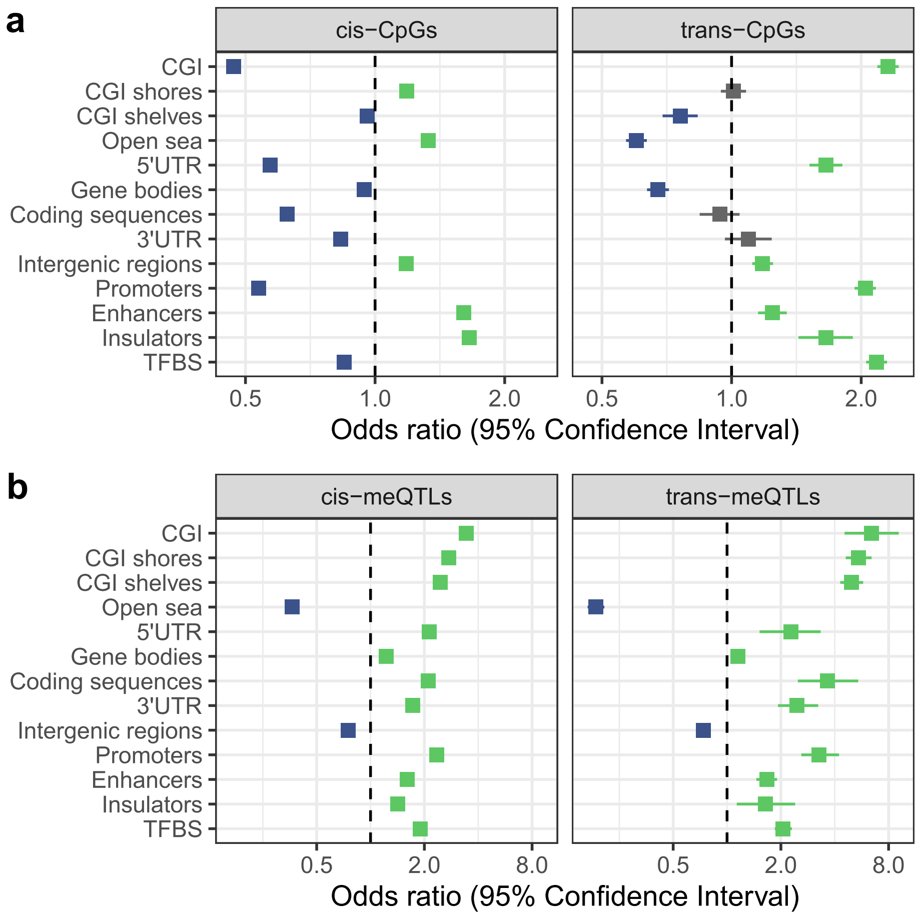

Fig. 4 Enrichment in genomic annotations of meQTL SNPs and their CpGs. The *x*-axis indicates the odds ratio and its 95% confidence interval (in logarithmic scale) for **a** CpGs with meQTLs or **b** meQTL SNPs, located within a specific genomic annotation. Significant enrichment is marked in green, depletion in blue, and non-significant genomic annotations in grey

harbour *cis*-meQTLs (odds ratio (OR) $<$ 1, FDR $\leq$ 0.05, two-tailed Fisher's exact test), but are more likely to have *trans*-meQTLs (OR $>$ 1, FDR $\leq$ 0.05). CpGs located in intergenic regions, enhancers and insulators are more likely to have both *cis*- and *trans*-meQTLs. We also explored the overrepresentation of CpGs with meQTLs near to genes, and with respect to gene ontology (GO) terms related to molecular functions and biological processes (Additional file 3: Table S22).

We next explored enrichment or depletion of meQTL SNPs in different genomic annotations. To this end, we compared the proportions of the most significantly associated meQTL SNPs per CpG site to the full panel of tested genetic variants in different annotations. Contrary to results observed for CpGs, we found a consistent pattern in the distribution of *cis* and the *trans*-meQTLs according to genomic category (Additional file 1: Fig. S9 and Additional file 3: Table S7). Overall, coding regions, promoters, enhancers, insulators and TFBSs are over-represented for genetic variants that are meQTLs (OR $>$ 1, FDR $\leq$ 0.05), either in *cis* or in *trans*. On the other hand, intergenic regions are under-represented for genetic variants that are meQTLs (OR $<$ 1,

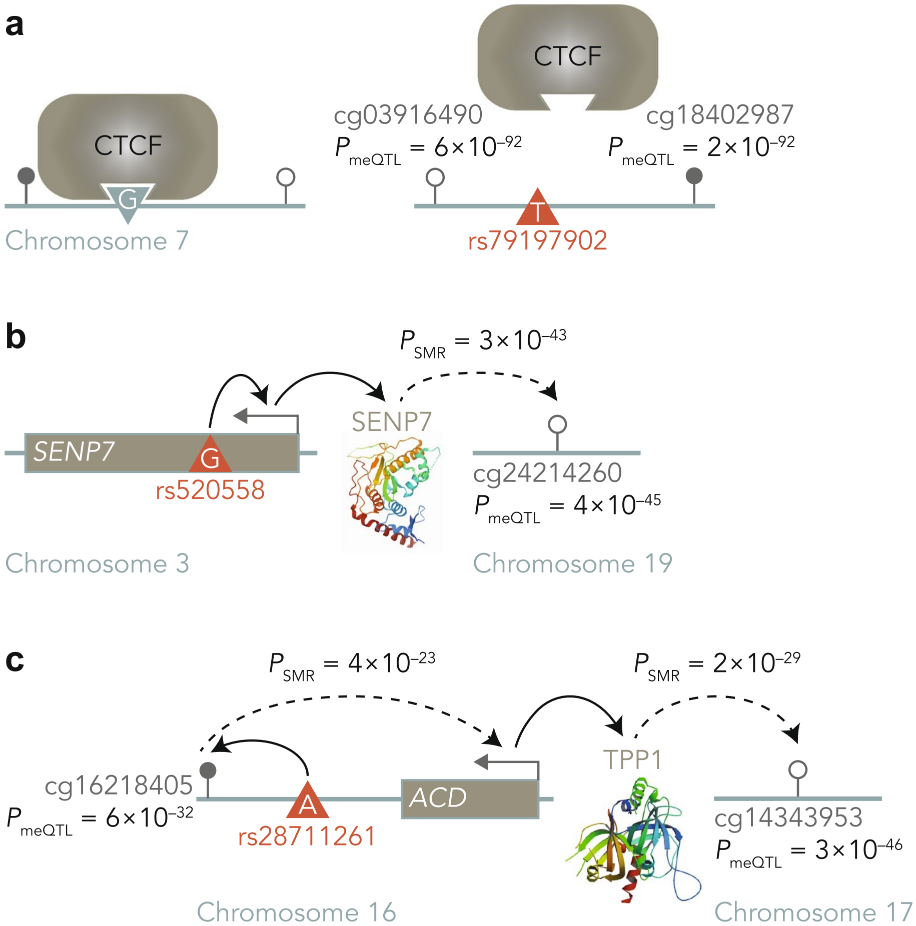

Fig. 5 Underlying mechanisms of meQTL SNPs. **a** Example of a *cis*-meQTL mechanism. The disruption of a TFBS (e.g. CTCF binding site) by a genetic variant (rs79197902), leads to reduced protein binding affinity, which changes local methylation (cg03916490, cg18402987). **b** Example of an 'eQTL-mediation mechanism' for *trans*-meQTLs. SNP rs520558 that is an eQTL for a gene involved in DNA methylation regulation (*SENP7*) indirectly affects distal CpG sites (cg24214260). Dashed lines represent associations for which there is suggestive, but not conclusive, evidence of directionality. **c** Example of a '*cis*-meQTL-mediation mechanism' for *trans*-meQTLs. SNP rs28711261 is associated with a nearby CpG (cg16218405), which in turn is associated with a gene involved in DNA methylation regulation (*ACD* gene of TPP1), and indirectly affects distal CpG sites (cg14343953)

FDR ≤ 0.05). The results remain consistent in sensitivity analyses taking as background reference subsamples of SNPs with a similar distribution of minor allele frequencies (MAF) and distance to target CpGs (Fig. 4b and Additional file 3: Table S8). Therefore, unlike the genomic patterns observed for CpGs under genetic control, the genetic variants driving meQTL effects show similar genomic distributions for local and distal genetic effects.

The location of meQTL SNPs and CpGs helps to elucidate genetic mechanisms of methylome regulation. As previously proposed, TFBSs may play a critical role in *cis* associations, as a genetic variant could prevent protein binding and alter methylation of the surrounding loci [27–29] (Fig. 5a). The observed over-representation of meQTL

SNPs in TFBSs, both for *cis* and *trans* results, supports this hypothesis ($\mathrm{OR}_{\mathrm{cis}} = 2.47$, 95% $\mathrm{CI}_{\mathrm{cis}}$ [2.44, 2.50], $\mathrm{OR}_{\mathrm{trans}} = 1.92$, 95% $\mathrm{CI}_{\mathrm{trans}}$ [1.74, 2.10]). We further explored this observation through enrichment analyses considering TFBS for sixteen specific transcription factors (TFs) of interest, previously identified as relevant for chromatin interactions and in the modulation of DNA methylation. These TFs included CTCF [30], ZNF143 [31] and EBF1 [32] (Additional file 1: Fig. S10). The direction of the effect in all cases was consistent with the results from the overall TFBS enrichment analysis.

We inspected CpGs in enhancers in more detail, motivated by their targeted coverage on the EPIC array. A set of 39,450 CpGs with *cis*-meQTLs and 789 CpGs with *tran*-meQTLs were annotated to enhancers (strong and weak/poised enhancers, see the 'Methods' section), based on ChromHMM annotations [33]. We find that the corresponding *cis*-meQTL genetic variants of CpGs in enhancers also tend to be in enhancer regions and TFBSs, when compared to the total set of meQTL SNPs (Additional file 1: Fig. S11 and Additional file 3: Tables S18–S19). This observation is not simply attributable to the genomic location of the associated CpGs in enhancers (Additional file 1: Section 1.7). In short, we observe a clear enrichment of CpGs with meQTLs, and of meQTL SNPs, in enhancers.

Functional integration gives insights into long range genetic impacts on the methylome

To explore potential mechanisms underlying meQTL associations, we carried out several functional integration analyses.

First, we combined our meQTL findings with data from the eQTLGen Consortium [34], the most extensive eQTL resource to date, conducted on 31,684 blood samples from individuals from 37 cohorts of predominantly European ancestry. We used *cis*-eQTLs results for 19,250 genes, and applied Summary-based Mendelian Randomization (SMR) [35] to co-localise signals and infer putative pleiotropic or causal effects on DNA methylation and gene expression.

Overall, we observe robust evidence for co-localisation between *cis*-meQTLs and eQTLs, which is in line with previous findings [18, 36]. Analysis of *cis*-meQTLs identified 19,267 unique SNPs that co-localise with *cis*-eQTLs of 8511 genes and with *cis*-meQTLs of 21,663 CpGs, resulting in 31,395 unique gene-CpG associations ($P_{\mathrm{SMR}} \leq 9.82 \times 10^{-9}$, $P_{\mathrm{HEIDI}} > 0.05$) (Additional file 4). Altogether, 44.2% of expressed genes shared a genetic basis with DNA methylation, which is greater than previously reported [18, 37]. CpGs typically have shared genetic effects with a single gene (median = 1, IQR = 1). Site cg11024963 had the highest number of co-localisation events (13 genes, including *DUS2*, *ZDHHC1*, *TPPP3* and *ECD4*) through the *cis*-meQTL rs8054034, successfully replicated in the MeDIP-seq dataset. Correspondingly, genes have shared genetic effects with a median of two CpGs (IQR = 2), and at most 88 CpGs, in the case of the *MSRA* gene. We observed an enrichment of CpGs with shared meQTL/eQTLs in genic and regulatory regions (Additional file 1: Fig. S12 and Additional file 3: Table S11). The resulting SMR genes were related to immunological processes in GO analyses (Additional file 3: Table S23). If we consider CpG-gene pairs with co-localised QTLs, we observe that methylation levels at the CpGs tend to be negatively correlated with the corresponding gene expression levels, regardless of the location of the CpG within the gene (Additional file 1: Fig. S13 and Additional file 1: Section 1.8). In summary, we observe the largest to date shared genetic basis between local genetic impacts on DNA methylation and gene expression, suggesting presence of joint regulatory mechanisms.

SMR analysis with *trans*-meQTLs also identified a number of meQTL and eQTL co-localisation events. Altogether, 642 unique *trans*-meQTL SNPs co-localised with *cis*-eQTLs (1520 co-localisation events), simultaneously affecting 709 CpGs and 782 genes ($P_{\mathrm{SMR}} \leq 3.71 \times 10^{-7}$, $P_{\mathrm{HEIDI}} > 0.05$) (Additional file 4). A median of one CpG is associated per gene (IQR = 1) and one gene per CpG (IQR = 2). The results could reflect a scenario where genetic variants that influence the expression of genes involved in direct or indirect global epigenetic regulation are also *trans*-meQTLs (i.e. 'eQTL-mediation mechanism' from Villicaña and Bell [29], also proposed by Huan et al. [38]). The gene with the most associations to CpGs through co-localised QTLs was *SENP7* (19 CpGs in chromosome 19, one on chromosome 5 and one on chromosome 10; Fig. 5b). Our findings are in line with recent studies indicating that *SENP7* interacts with epigenetic regulators in the context of DNA repair [11, 39]. Within these CpG-gene pairs with shared *trans*-meQTLs/*cis*-eQTLs, we identified an enrichment of CpGs in enhancers and TFBSs (Additional file 1: Fig. S12 and Additional file 3: Table S11). Furthermore, the genes are annotated to GO terms related to DNA-binding transcription repressor activity including predominantly zinc finger proteins, which are known to act as epigenetic regulators in different contexts [40–42] (Additional file 3: Table S23). In addition, the results of GO enrichments also replicate findings from previous studies [11, 38].

The *cis*-meQTL to *cis*-eQTL co-localisation results also allow us to make inferences into mechanisms of distal genetic impacts on DNA methylation levels. We observed a significant enrichment of *trans*-meQTLs in the co-localised *cis*-meQTL to *cis*-eQTL SNPs, compared to the non-co-localised *cis*-meQTLs (OR = 3.73, 95% CI [3.59, 3.88]). In light of this, we then used these *trans*-meQTL SNPs (that co-localised with *cis*-meQTL and *cis*-eQTL) as instrumental variables in SMR to test for associations between the corresponding eQTL gene expression levels and DNA methylation levels of the corresponding CpGs in *trans*. We identified a total of 511 *trans*-associations through 279 SNPs (hereafter 'multi-QTLs'), between 323 CpGs and 292 genes ($P_{\mathrm{SMR}} \leq 9.82 \times 10^{-9}$, $P_{\mathrm{HEIDI}} > 0.05$) (Additional file 4). These results could reflect a genetic mechanism of *trans*-meQTL effects, where a *cis*-meQTL impacts nearby CpG sites. These CpG sites in turn may affect the expression of genes involved in epigenetic regulatory processes, and whose products affect the methylation of multiple distal sites (i.e. '*cis*-meQTL-mediation mechanism' from [29]). The adrenocortical dysplasia homolog (*ACD)* gene fits this scenario (Fig. 5c)). *ACD* has four eQTLs (rs28711261, rs9936153, rs12935253 and rs2059850989) that co-localised with *cis*-meQTLs of six CpGs, and *trans*-meQTLs of 16 CpGs on different chromosomes. *ACD* produces the TPP1 protein, which is part of the shelterin complex that maintains telomere length [43]. A correlation between DNA methylation patterns and telomere length has been reported previously [44, 45], although multiple mechanisms likely underlie these links given that condensation of telomeric chromatin by the shelterin complex does not primarily occur through DNA methylation [46].

We carried out two additional functional exploration analyses of meQTLs. First, we searched for meQTL-CpG associations that overlapped three-dimensional (3D) conformations of the genome, such as topologically associated domains (TADs). The rationale behind this analysis was that some intra-chromosomal *trans*-meQTLs may act as 'long-range' *cis*-meQTLs [11, 38] that TADs bring into physical proximity [29, 47]. We integrated our meQTL results with TADs predicted from multiple-tissue Hi-C experiments [48–52]. We found that 36.5% of CpGs with intra-chromosomal *trans*-meQTLs share the same TAD with their most associated meQTL. In comparison, 17.1% of CpGs with *cis*-meQTLs share the same TAD with their most associated meQTL. Furthermore, TADs containing intra-chromosomal *trans*-meQTL associations are significantly larger than TADs with *cis*-meQTL associations (mean TAD size$_{cis}$ = 1.2 Mbp; mean TAD size$_{trans}$ = 3.4 Mbp; $P \leq 2.54 \times 10^{-13}$), which supports our hypothesis that TADs may bring *trans*-meQTLs into physical proximity with their target CpG (Additional file 1: Section 1.9). In summary, our results are consistent with the hypothesis that some intra-chromosomal *trans*-meQTLs may act as 'long-range' *cis*-meQTLs within TADs.

Second, we focused on GO analysis of *trans*-meQTLs that lie within coding regions to test for evidence that *trans*-meQTLs may alter the function of proteins such as TFs. Our motivation was that such SNPs may impact the binding affinity of the TFs, and therefore change DNA methylation levels of distal unoccupied binding sites. A total of 79 *trans*-meQTLs (1.8% of the 4398 top *trans*-meQTLs) were annotated in coding regions of 168 protein-coding genes. We found enrichment in 37 GO terms relative to molecular functions and 182 terms for biological processes (Additional file 3: Table S24). Of these, 11 corresponded to categories related to protein binding and 56 to regulation of biological processes. Therefore, these results support the hypothesis that *trans*-meQTLs may alter the function of proteins such as TFs that then impact DNA methylation levels at multiple genomic regions.

Highly connected CpGs and meQTLs

We calculated the effective number of meQTL SNP associations per CpG, discarding redundant SNPs due to LD. To this end, we merged all *cis*-meQTL SNPs and following LD clumping generated '*cis*-meQTL regions', and repeated the process for *trans*-meQTLs (see Additional file 1: Section 1.10). Overall, CpGs under genetic control tend to have few associations after LD clumping, with a median of two meQTLs in *cis* (IQR = 3) and one meQTL in *trans* (IQR = 1) per CpG. However, a subset of CpGs have a high number of clumped meQTLs, or are 'highly regulated' or connected. Specifically, such highly regulated CpGs include 1.4% of CpGs with *cis*-meQTLs that have over 13 clumped meQTLs, and 2.9% with *trans*-meQTLs with over 5 meQTL associations (thresholds correspond to $Q_3 + 3\text{IQR}$). From the CpGs with both *cis*- and *trans*-meQTLs (2281 sites), 627 CpGs are highly regulated by either *cis*-meQTLs, *trans*-meQTLs or both.

Highly regulated CpGs with *cis*-meQTLs are overrepresented in genic and regulatory regions, such as enhancers, compared to other CpGs with *cis*-meQTLs (Additional file 1: Fig. S14, Additional file 3: Table S12 and S20). In the case of highly regulated CpGs with *trans*-meQTLs, coding sequences are enriched, while promoters, TFBSs and intergenic regions are depleted. Moreover, 32 immune-related GO annotations are enriched for highly regulated CpGs in *cis* but not in *trans* (Additional file 3: Table S25). The CpGs that have the most associations overall both *cis* and *trans* are the novel EPIC probes cg16423305 (42 *cis* and 21 *trans*-meQTLs), cg00128506 (48 *cis* and 13 *trans*-meQTLs) and cg25014118 (50 *cis* and 6 *trans*-meQTLs). All these CpGs are located on chromosomal region 8p23.1 near to or in genes *PRAG1*, *MFHAS1* and *XKR6*. Additionally, CpG cg00128506 is in an enhancer region, and in the binding site of transcription factors ELF1, USF2, IKZF2 and RAD51. We replicated 81% of cg00128506 *cis*- and *trans*-meQTL associations in the MeDIP-seq dataset (Additional file 1: Table S3).

We next considered the connectivity of meQTL SNPs. We observed a median of five unique *cis*-CpGs (IQR $= 10$) associated with each region of clumped meQTLs, and a median of one *trans*-CpG (IQR $= 1$). Highly connected meQTL clumped regions, or 'key regulatory regions', were defined as 4.4% (*cis*) and 7.8% (*trans*) of genetic regions associated with more than 42 *cis*-CpGs and 5 *trans*-CpGs, respectively (thresholds correspond to $Q_3 + 3\text{IQR}$). A relatively large proportion of 71.9% of the meQTLs that act simultaneously in *cis* and *trans* (165,290 SNPs) are located in key regulatory regions in either *cis*, *trans* or both. MeQTLs located within key regulatory regions are enriched/depleted in the same genomic regions as the top meQTLs previously described (Additional file 1: Fig. S15, Additional file 3: Table S13 and S21).

Particularly noticeable among highly connected CpGs and genetic regions is the major histocompatibility complex (MHC) region, which is overrepresented with both CpGs with genetic effects and SNPs that are meQTL. This locus contains multiple highly regulated CpGs and key regulatory meQTL regions for *cis* and especially *trans* associations (Additional file 1: Section 1.11). However, the very high genetic diversity and complexity of this genomic region necessitates further follow-ups with higher resolution genetic and epigenetic sequence datasets. Apart from the MHC region, other genomic regions with high level of CpG connectivity include the above-mentioned region on chromosome 8p23.1 (6,200,001–12,700,000 region spanning many genes). For meQTL SNP-level connectivity other genomic region hotspots included chromosomes 17q25.3 (in *B3GNTL1*) and 21q22.3 (in multiple genes) for *cis* associations, and 19p13.2 (*ZNF* gene family) and 7p22.3 (*MAD1L1*) for *trans* associations.

We also compared the number of clumped meQTLs per CpGs in enhancer regions. We detected a small but significant increase in the mean number of *cis*-meQTL associations for CpGs in enhancers (3.14 meQTLs per CpG outside of enhancers, compared to 3.58 in enhancers; two-tailed *t*-test, unequal variances, $t_{(52,211)} = 23.7$, $P < 2.2 \times 10^{-16}$), but no difference in the median number of associations (two *cis*-meQTLs for both categories). For *trans*-meQTLs, neither the mean (two-tailed *t*-test, unequal variance, $t_{(1,042.7)} = 0.4$, $P = 0.69$) or the median differed across these categories. Overall, the CpG sites with most genetic associations are found in enhancers, as confirmed by the enrichment observed of highly regulated CpGs.

The interplay between genetic variation, DNA methylation and complex traits

We used our meQTL findings to identify co-localisations between our *cis*-meQTLs SNPs and GWAS SNPs from 56 common human complex traits grouped in seven phenotypic categories (Additional file 5: Table S26).

After Bonferroni correction for 186,817 tested CpG sites and seven phenotypic classes, we identified 1520 associations through co-localisation between 1325 unique CpGs and 34 traits, involving 1180 unique *cis*-meQTLs ($P_{\text{SMR}} \leq 3.82 \times 10^{-8}$, $P_{\text{HEIDI}} > 0.05$) (Additional file 5: Tables S26–S27 and Additional file 6). Height was the phenotype with the most GWAS signals co-localised with meQTLs (501 CpG sites). 'Growth and ageing' (which includes height) was the phenotypic class with most co-localisations and 598 unique CpG sites. The CpGs with most associations were cg06162668 and cg27288595, with six traits each. CpG cg06162668 is in an intergenic region in chromosome 2, and was associated with obesity and metabolic disease phenotypes through SNP rs7561317. The association between cg06162668 and rs7561317 was replicated in the MeDIP-seq dataset. Site cg27288595 was also associated with obesity and metabolic disease, along with growth and ageing, and is located in the *ZBTB38*. This gene encodes a zinc finger that binds to DNA methylation sites and acts a transcriptional repressor [40]. Overall, CpGs with GWAS co-localisations are enriched in CGIs, coding and regulatory regions, compared with other CpGs with meQTLs (Additional file 1: Fig. S18 and Additional file 3: Table S17).

The strongest associations were between total cholesterol levels with cg17260184 and cg27123834, annotated upstream of the transcription starting site of *USP1* and *DOCK7*, respectively. *USP1* encodes a deubiquitinase which regulates the cellular response to DNA damage [53]. *DOCK7*—primarily involved in axon formation and neurogenesis—also overlaps the gene encoding angiopoietin-like protein (*ANGPTL3)* that regulates plasma lipid levels [54, 55]. The associated genetic variants rs2131925 and rs12136083, respectively, are in non-coding regions. To our knowledge, the function of these two variants has not been characterised.

Another example of note is the observed association between inflammatory bowel disease (IBD) and cg19297788 ($\beta_{\mathrm{SMR}} = -0.41$, $P_{\mathrm{SMR}} = 1.44 \times 10^{-12}$, $P_{\mathrm{HEIDI}} = 0.06$), a CpG in a weak enhancer region of chromosome 21 (Fig. 6). The CpG also falls within three TFBSs for TCF12, EBF1 and RUNX3 and was not previously covered by the 450K array. We found evidence of association with both conditions comprised by IBD, ulcerative colitis ($\beta_{\mathrm{SMR}} = -0.39$, $P_{\mathrm{SMR}} = 9.27 \times 10^{-9}$, $P_{\mathrm{HEIDI}} = 0.11$) and Crohn's disease ($\beta_{\mathrm{SMR}} = -0.47$, $P_{\mathrm{SMR}} = 3.11 \times 10^{-10}$, $P_{\mathrm{HEIDI}} = 0.08$). This locus is surrounded by five genes, including *ICOSLG*, a coding gene for a ligand of the T-cell surface receptor ICOS. This gene has been identified in previous studies as a risk locus for IBD [56–58], where the interaction between ICOS/ICOSLG in IBD and decreased expression of *ICOSL* can affect IBD risk [57]. However, it was unclear how genetic variants in the locus lead to the change in gene expression of *ICOSL*. According to our results, IBD and CpG site cg19297788 share the common genetic variant rs2876932 (chr21:45,618,536).

Another example of note includes the associations observed for CpG cg17459721 and phenotypes for waist ($\beta_{\mathrm{SMR}} = -0.11$, $P_{\mathrm{SMR}} = 1.73 \times 10^{-12}$, $P_{\mathrm{HEIDI}} = 0.98$) and hip circumference ($\beta_{\mathrm{SMR}} = -0.11$, $P_{\mathrm{SMR}} = 3.59 \times 10^{-12}$, $P_{\mathrm{HEIDI}} = 0.98$), through rs7187776 on chromosome 16p11.2. Previous GWAS have described this region in the context of

(See figure on next page.)

Fig. 6 Association between IBD and DNA methylation at site cg19297788. **a** Locus association plot. The grey dots represent the *P*-values of the SNPs from the IBD GWAS [56], the violet diamond the *P*-value of the SMR test, and the violet crosses the *P*-values of the meQTLs of cg19297788. **b** Effect sizes of IBD GWAS SNPs vs. effect sizes of meQTLs of cg19297788, for SNPs used in the HEIDI test. The slope of the dashed line represents the β_{SMR} estimate at the co-localised SNP. Error bars represent standard errors of estimated SNP effects. SNPs in LD with the top co-localised meQTL are expected to have a consistent effect under the causality/pleiotropy scenario

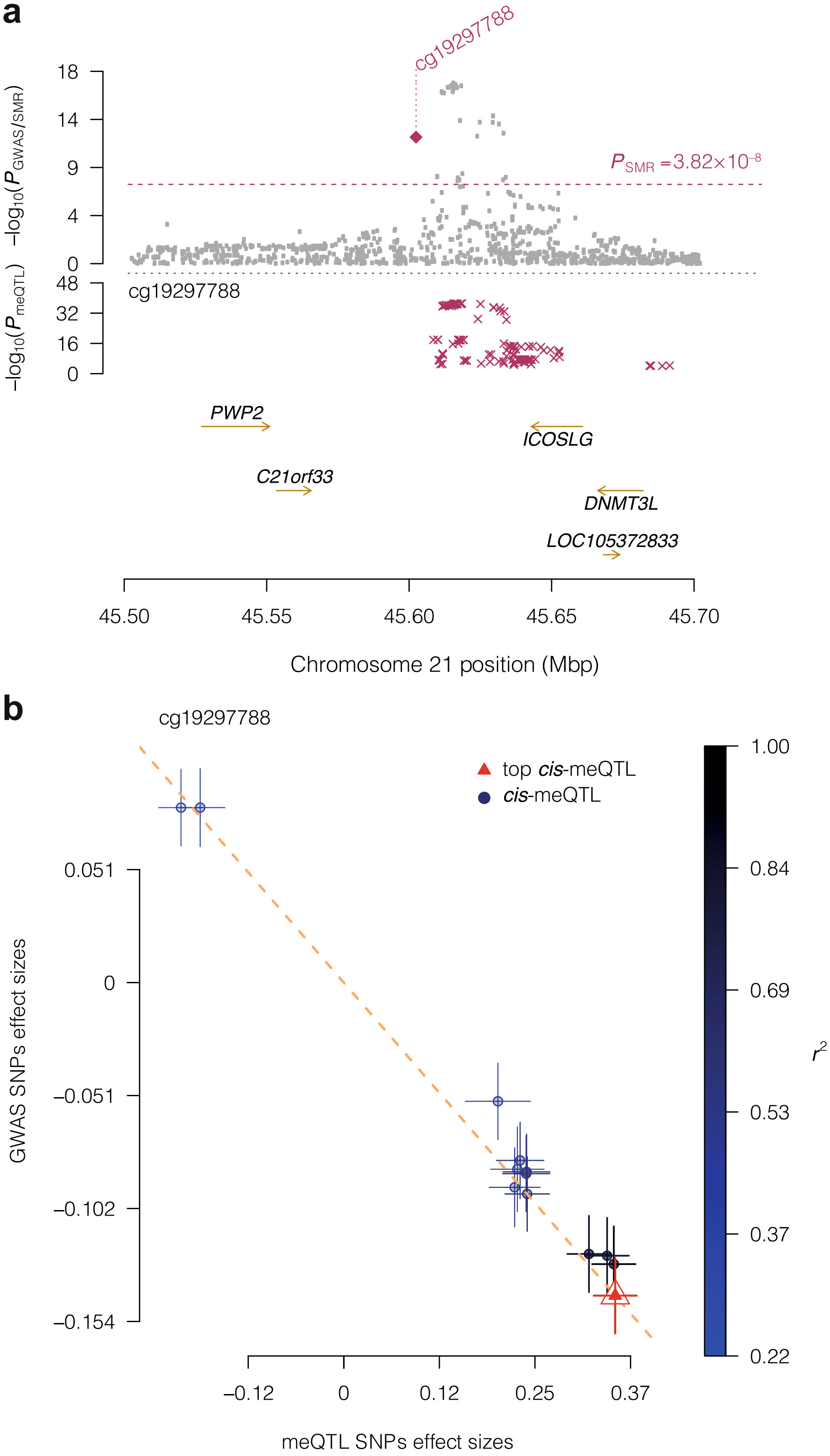

Fig. 6 (See legend on previous page.)

body mass index and body fat distribution [59, 60], but the mechanisms of action remain unclear. Here we identified from the SMR with gene expression that this meQTL SNP co-localises with eQTLs of the *TUFM* and *SPNS1* genes, and *trans*-meQTL for cg03969070 (chromosome 1). The latter CpG is in the promoter of *STK40*, involved in the glycogen metabolism (GO:0005977), among other biological processes. Therefore, we hypothesise that the action of rs7187776 is through a *cis*-meQTL-mediation mechanism.

Altogether these examples of integrative analyses highlight connections between target genetic variants and DNA methylation at multiple CpGs, gene expression at several genes, and a number of complex metabolic traits and diseases. These novel links provide functional insights into mechanisms of action for specific GWAS variants in selected human phenotypes.

Discussion

We investigated the impact of genetic variation on DNA methylation levels at genomic regions profiled by the Illumina Infinium MethylationEPIC BeadChip in three UK cohort populations. To our knowledge, previous meQTL studies have not yet explored both *cis*- and *trans*-meQTLs across the genome on the EPIC array in a large number of samples in blood. The increased coverage of the array, especially in intergenic regions such as enhancers, provides novel insights into the genetic regulation of DNA methylation, with downstream impacts into the regulation of gene expression and human complex traits.

We estimated that more than 33% of the EPIC methylome is under genetic control, the majority of which is in *cis*. Our *cis* results are in line with previous studies on the Illumina EPIC and 450K, in terms of proportion of sites, distance to target, allele frequency, and genomic annotations [10, 11, 19, 38]. The proportion of *trans* signals that we detected is somewhat lower than previous studies [10, 18, 38], although this likely in part reflects power as our two-stage meta-analysis approach may reduce power to detect *trans* associations. Specifically, before filtering associations in at least two cohorts, the detected *trans* associations were 10-fold greater compared to the final list (compared to *cis*, only 2-fold higher). This likely represents lower reproducibility of *trans* signals, which may be more likely subject to cohort specific differences, batch effects, or may potentially represent biological factors. Therefore, the reported *trans*-meQTL results should be interpreted with caution and validated in future studies. Furthermore and consistent with previous studies [11, 38], we also observe evidence that intra-chromosomal *trans*-meQTLs are likely to be 'long-range' *cis*-meQTLs, as the vast majority are located within 5 Mbp from the target CpG and a proportion fall within TADs. Lastly, in line with recent large-scale findings from the blood 450K meQTLs [10], our results confirm that SNPs and CpGs that exhibit both *cis* and *trans* associations, are highly reproducible, appear to have large effects, and exhibit high connectivity with other genetic variants and CpGs. Our results also highlight multiple highly connected genomic regions of interest, both putative key regulatory regions of SNPs, and regions containing highly-regulated CpGs. These connectivity results improve our understanding of specific mechanisms of genetic regulation on the epigenome, transcriptome, and human phenotypes.

We estimated the proportion of variance explained by meQTLs, both in relation to variability of DNA methylation at each CpG and to methylation heritability. Although the mean values genome-wide appear relatively low, there are cases of CpGs where genetic factors explain close to 100% of CpG DNA methylation variance. A similar trend is observed in terms of the number of meQTLs per CpG, where there are few CpGs with a large number of associations after LD clumping. These extreme cases, instead of being seen as exceptions, can be further explored in future to better understand the underlying mechanisms, evolutionary selection, and epistatic and environmental interactions of meQTL.

We integrated our meQTL results with large-scale blood eQTL results, as well as with GWAS findings from 56 human phenotypes. Altogether, these integrative analyses highlighted sets of shared genetic impacts that allow us to make two key inferences. The first one through eQTL integration gives insights into specific mechanisms of long-range genetic impacts on DNA methylation, highlighting multiple examples consistent with two hypothesised mechanisms. Second, through integration of GWAS findings with meQTLs, and in cases with eQTLs, we highlight multiple examples of specific putative mechanisms underlying GWAS genetic impacts on human phenotypes. Our work is consistent with and extends previous efforts, both disease-specific [61, 62] and multi-trait [18, 37], that integrate different molecular data at the genome-wide level to provide new insights into disease processes and biological pathways.

One of the main strengths of our study is that the sample used is representative and age-homogenous of a well-characterised nationwide population. Limitations include, first, analyses were restricted to whole blood samples. Although blood cell heterogeneity was taken into account, the estimated cell proportions are relatively low resolution. We undertook cell-specific analysis and observed that the majority of whole blood meQTLs do not show evidence for cell-specific effects. However, we did not comprehensively explore cell-population specific meQTL effects and restricted our analysis to two cell types with modest to moderate proportions in our whole blood data. Second, we did not include conditional analyses and therefore the number of independent meQTLs per CpGs remains unknown. Third, we carried out validation of all legacy 450K signals in the GoDMC dataset, and pursued replication at targeted novel EPIC-specific sites. The resolution of MeDIP-seq methylation data (500 bp) is lower compared to EPIC data and therefore presents a more qualitative replication approach. Fourth, limitations to eQTL and meQTL integration include the assumption of shared genetic impacts, although the effects may be coincidental. Fifth, several limitations of the SMR approach include no explicit test for causal impacts, a limited selection of 56 phenotypes considered, and differences in power across phenotypes because different GWASs have differing samples sizes and therefore power. Sixth, we cannot rule out that our sample has a selection bias, with an overrepresentation of healthy participants able to give blood samples and information on health. Finally, our findings relate to middle-aged and older adults. Although there is evidence to suggest that the meQTL effects are stable across the life course [63], further studies should confirm whether the associations described here are valid in other age groups. In this same line, we cannot extrapolate our observations to other ethnic groups.

Conclusions

In summary, we present a novel large-scale DNA methylation quantitative trait locus analysis in blood samples from three UK cohorts profiled on the Illumina EPIC array. The results identify novel genetic impacts on DNA methylation levels across the genome, and integrative analyses with gene expression and GWAS findings give insights into mechanisms underlying genetic regulation of human functional and phenotypic variability.

Methods

Study cohorts

We analysed data collected from 2478 samples across three different UK population cohorts, of which 2358 samples passed quality control assessment and are included in the analyses in this manuscript. TwinsUK [20] (post-QC $n = 394$, from 236 unique families) is the UK's most comprehensive and detailed registry of adult monozygotic and dizygotic twins. The MRC National Survey of Health and Development (NSHD), or 1946 British birth cohort (1946BC) [21, 22] (post-QC $n = 1,545$), is the longest-running birth cohort in the UK, with data about individuals born during one week in March 1946. The National Child Development Study (NCDS), or 1958 British birth cohort (1958BC) [23, 24] (post-QC $n = 419$), surveys individuals born during the same week in March 1958. The 1946BC data contained samples of individuals at either age ≈ 53 or ≈ 63, and therefore, we stratified the cohort in two age-based groups to facilitate data handling, referred to as 1946BC-99 ($n = 1,348$) and 1946BC-09 ($n = 197$). The 1958BC samples were processed in two different batches and also stratified into 1958BC-1 ($n = 183$) and 1958BC-2 ($n = 236$). Local research ethics committees granted ethical approval of the study, and all participants provided written informed consent.

Genotyping and imputation

DNA was extracted from whole blood samples and genotyping was carried out with a combination of platforms across studies (Additional file 1: Section 1.1). Quality control of raw genotype data from each of the five samples was carried out separately in PLINK [64], and steps included filtering out low-frequency and rare variants (minor allele frequency, MAF < 0.01), with a Hardy-Weinberg equilibrium $P < 1 \times 10^{-6}$ or missingness rate $> 3\%$. We also removed samples with more than 5% of missing data. We imputed genotypes with the 1000 Genomes reference panel phase 3 version 5 [65] in the Michigan Imputation Server [66] and again filtered the resulting variants using a threshold for MAF > 0.05 and $r^2 > 0.8$. For the present study we used genome assembly GRCh37/hg19 [67] for reporting genomic positions. The final set included 6,361,063 unique genetic variants in at least one of the sample sets (Additional file 1: Table S1 with CpG sites per cohort).

DNA methylation profiling and data processing

DNA was bisulfite-converted using the EZ DNA methylation kit (Zymo Research). DNA methylation levels were profiled with the Infinium MethylationEPIC BeadChip (EPIC) at site-specific resolution, and raw intensities signals were obtained. Altogether five cohort samples were profiled, and detailed description of profiling and DNA methylation data initial processing is provided in Additional file 1: Section 1.1.

Briefly, raw intensities signals were processed (separately for each sample set) with the ENmix package [68] in the R environment [69] and converted into Illumina β-values (ratio of methylation at each CpG site) for downstream analysis. Background correction was performed using the Exponential-Normal mixture distribution (ENmix) method, dye-bias correction was performed using the Regression on Logarithm of Internal Control probes (RELIC) method [70], and probe design bias adjustment was performed implementing the Regression on Correlated Probes (RCP) method [71]. Filtering included exclusion of probes with missingness rates > 5% (detection $P > 1 \times 10^{-6}$) and exclusion of samples with missing methylation data at > 5% CpG (detection $P > 1 \times 10^{-6}$) and with no genotyping data. Additionally, we filtered out probes with a polymorphism with MAF > 0.05 in the interrogated CpG or the extension base (in case of type II probes), using the UK10K haplotype reference panel, plus the recommended list of masked probes published by Zhou et al. [72]. After data normalisation, we retained 724,499 unique CpGs in the autosomes across the sample sets (Additional file 1: Table S1 with CpG sites per cohort). For the analysis, the number of samples with DNA methylation and genotyping data was 2358 (see Table 1 for the final sample size of each cohort).

DNA methylation data adjustment

DNA methylation profiles are cell type-dependent, and cell composition is a major confounding variable in methylation studies in tissues with cellular heterogeneity, such as whole blood [73]. We estimated the cell composition for monocytes, granulocytes, plasmablasts and immune cells (natural killer, naïve $CD8^{+}$, $CD4^{+}$, and joined $CD8^{+}$/ $CD28^{-}$/ CD45RA cells), using the regression calibration approach proposed by Houseman et al. [73] and implemented in the R package minfi [74].

To ensure normality and reduce the impact of confounders in the analyses, we applied a rank-based inverse normal transformation (INT) to the DNA methylation β-values and fitted a linear mixed-effects model (LMM) with covariates with the lme4 package [75]. We specified as fixed effects the variables sex (only for the birth cohorts), blood cell proportions, smoking and age (only for TwinsUK), and as random effects the technical covariates plate ID and position on the chip (as well as family ID and zygosity for TwinsUK). The residuals of this model were used for downstream analyses.

Heritability estimation

We used a classical twin design to estimate the narrow-sense heritability (h^2) of DNA methylation at CpG-level for TwinsUK data, with the OpenMx package [76] in R. After removing singletons, we kept 70 DZ twin pairs and 88 MZ twin pairs from the cohort. We used adjusted residuals of β-values (without the correction for family ID and zygosity) of the 723,814 CpG sites available in the cohort. We applied structural equations and maximum likelihood estimation to decompose the variance proportion at each CpG site in additive genetic (A), shared environment (C) and unique environment plus residual (E) components. The h^2 corresponds to the proportion of phenotypic variance attributed to additive genetic effects (A component). We discarded CpGs where the model had critical optimization failures, keeping estimations for 723,427 CpGs. We compared the mean heritability between novel EPIC-only sites and the 450K legacy probes using a one-tailed t-test assuming equal variance.

Genome-wide association of DNA methylation

MeQTL analysis was performed in two stages. In the discovery phase, we identified candidate associations per sample. We fitted a linear regression between all possible pairs of SNPs and CpG sites, with the genotype variant as the explanatory variable—coded as doses of the alternative allele (0, 1 or 2)—and adjusted β-values for the CpG as the response. In total, 3.4 billion of *cis* pairs and 4.7 trillion of *trans* pairs were tested across the five cohort samples. SNPs separated by no more than 1 Mbp from the tested CpG were considered *cis*, and the remaining *trans*. In the discovery step we applied a liberal P-value to keep the associations for further analysis, specifically, $P \leq 5 \times 10^{-3}$ for *cis* and $P \leq 5 \times 10^{-6}$ for *trans* associations. The discovery step was performed in Matrix eQTL [77] implemented in R.

The second stage was a meta-analysis with the summary statistics of the subset of candidate associations kept from the discovery phase. As some of the sample sizes of the cohorts are substantially different, which impacts the variance of the estimated coefficients, we accounted for this heterogeneity in a random-effects inverse-variance weighted meta-analysis, using the open-source software GWAMA [78].

To account for multiple testing, we estimated the false discovery rate (FDR) with a permutation approach. Briefly, for each of the cohorts, we shuffled the labels of the individual samples for the methylation profiles (maintaining the family structure in TwinsUK), and association tests on the permuted data were carried out as before in Matrix eQTL and meta-analysed in GWAMA. A total of twenty permutations were performed overall, and the resulted P-values formed our null distribution. Then, we calculated the FDR as described in Hastie et al. [79], with the proportion of associations in the null distribution over the associations in the observed real data. SNP-CpG pairs were reported as significant meQTLs if they had an FDR ≤ 0.05 ($P \leq 2.21 \times 10^{-4}$ for *cis*, $P \leq 3.35 \times 10^{-9}$ for *trans*). Lastly, we only report those associations that were observed in two or more of the five sample sets and with the same direction of effect. As a sensitivity analysis, we also estimated the threshold P-values by dividing the EPIC CpGs into two sets (legacy 450K probes only and the novel EPIC-only probes). All the details are available in the Additional file 1: Section 1.12.

We replicated our results with the GoDMC meQTL catalogue [10]. We selected from our list of CpGs probes, those that were included in the 450K array and that were in the GoDMC study. We considered CpGs to replicate if they were also reported to be under genetic influence in the GoDMC study, with the same or with different SNP as that identified in our study.

For the integration of meQTL and heritability results, we fit a linear regression with the A estimate for each CpG as the dependent variable, and the categorical variables indicating the presence or absence of *cis*- or *trans*-meQTLs as independent.

Cell type-specific meQTLs

In addition to identifying whole blood meQTLs, we also explored evidence for blood cell-specific meQTL effects. To this end, we considered DNA methylation-based estimates of blood cell proportion for each sample cohort (Additional file 1: Fig. S7a). We focused on CD4^{+} T cells and monocytes as they exhibit a relatively greater level of homogeneity when compared to other blood cell types, such as granulocytes. Furthermore, these cell types have been included in previous large-scale meQTL studies [19, 26]. In cell-type specific analyses for CD4^{+} T cells, we first adjusted DNA methylation levels for covariates as described in the whole blood meQTL analysis, but did not include estimated proportion of CD4$^+$ T cells as a covariate. We then fitted a linear model to estimate *cis*-meQTLs in Matrix eQTL. We considered the genetic variant, the proportion of CD4$^+$ T cells, and the interaction term between these as predictors. For each cohort sample, we kept all associations where the interaction term surpassed $P \leq 5 \times 10^{-3}$. We then meta-analysed the summary statistics of the interaction terms in a random-effects model using GWAMA, and filtered associations observed in two or more sample sets with the same direction of effect. We used a similar process to estimate cell-specific meQTL effects for monocytes.

Linkage disequilibrium (LD)-based clumping of meQTLs

To account for LD structure among the identified meQTLs, we carried our LD clumping of the meQTL SNPs, performed separately for *cis*- and *trans*-meQTLs. Here, we kept the genetic variant with more associated CpGs as representative for each LD block—to ensure that all clumps were consistent across all CpGs. LD clumping was performed using PLINK with LD threshold of $r^2 > 0.1$ (calculated using all the samples in this study) within a window of 2 Mbp. Finally, as the representative SNP of each clump may not be the one associated with a given CpG, we used the most significant meQTL per CpG and per clump.

MeDIP-seq data

For meQTL replication of novel EPIC probes we used previously published methylation data profiled with methylated DNA immunoprecipitation sequencing (MeDIP-seq) in TwinsUK blood samples [80, 81]. We excluded individuals from the current study, resulting in a final independent sample of 2319 participants (from 1632 unique families, 93.5% females, median age 55, age range 16–82) from the TwinsUK cohort, with whole blood methylomes profiled using MeDIP-seq.

MeDIP-seq of whole blood samples was performed as previously described [81]. Briefly, following DNA fragmentation through sonication, sequencing libraries were prepared using Illumina's DNA Sample Prep kit for single-end sequencing. The anti-5mC antibody (Diagenode) was used for immunoprecipitation and MeDIP was validated by quantitative polymerase chain reaction. Captured DNA was purified and amplified with adaptor-mediated PCR, and fragments of size 200–500 bp were selected by gel excision and QC assessed by Agilent BioAnalyzer. Sequencing was carried out on the Illumina platform. Sequencing data were aligned using BWA [82] using build GRCh37/hg19 and a mapping quality score of Q10, and QC steps included FastQC, removal of duplicate reads, and SAMTools [83] QC. MeDIP-seq data quantification into methylation levels was carried out using MEDIPS v1.0 [84] reads per million (RPM), and further QC was carried out in R including batch effects inspection. The average high quality BWA aligned reads were ≈ 16.8 million per sample. Processed MeDIP-seq methylation data for analysis were quantified in genomic windows (bins) of 500 bp (250 bp overlap) with RPM scores.

We selected ten novel EPIC CpGs to replicate based on the number of associations, the strength of association, effect sizes, and the co-localisation of their meQTLs resulting after SMR. MeDIP-seq methylation levels in each bin were transformed with the rank-based INT and adjusted for covariates (sex, age, family and zygosity) in an LMM.

We excluded bins with evidence of methylation association with smoking ($P < 0.05$). Finally, we excluded bins with methylation data in less than 1160 samples. The final set of CpGs and the respective bins is listed in Additional file 1: Table S3.

We performed the meQTL analysis in Matrix eQTL as described above. We considered associations to replicated if they exceeded a statistical threshold of $P < 0.005$ (Bonferroni correction for 10 CpGs at a significance level of 0.05) and with the same direction of effect as in the original EPIC meQTL analysis.

Functional annotations

We obtained different genomic annotations in BED file format through UCSC Table Browser as of August 31, 2020 [85]. We used annotations for CpG islands, RefSeq genes [86], chromatin state segmentation for the GM12878 cell line [33], and TFBSs for GM12878 cell line from ENCODE 3 [87]. Then we mapped the DNA methylation sites and genetic variants to the functional annotations.

Chromosome sizes and number of coding genes were retrieved from the reference genome GRCh37/hg19 and Ensembl 104 [88] databases. We considered the major histocompatibility complex (MHC) locus to span chromosome 6 base-pair positions 28,477,797 to 33,448,354 bp [67].

We incorporated 3D genomic annotations in the meQTL functional annotations by using previously published data from Hi-C experiments in lymphoblastoid cell line GM12878 [49, 50]. We also considered additional Hi-C data across multiple relevant cells and tissues, including in GM12878, in spleen [51] and thymus [52]. TADs from these datasets were the generated in and obtained from the 3D Genome Browser [48]. We estimated the percentage of CpGs where the target CpG and its most associated meQTL fell within the same TAD. This analysis was carried out for all CpGs with *cis*-meQTLs and for all CpGs with intra-chromosomal *trans*-meQTLs. The estimation was performed for the GM12878 Hi-C data alone, as well as for TADs estimated from multiple cells and tissues. Further details and results are provided in the Additional file 1: Section 1.9.

Enrichment analyses

Fisher's exact tests were performed to investigate enrichment or depletion of CpGs/SNPs across genomic regions and functional annotations. We used a modified version of the R package LOLA [89] extending the default one-tailed to a two-tailed test, and incorporating the estimation of confidence intervals. The results for each independent analysis were corrected for multiple testing with the Benjamini–Hochberg procedure [90] to control the FDR.

For all enrichment tests on SNPs, we used the most significant meQTL per CpG (referred to as top meQTL). For some of the analyses on SNPs, we first generated a background set from a resampling method in order to obtain a collection of SNPs with equivalent distributions of MAFs and distances to the CpGs. To do this, we categorised the available SNPs according to their distance from the EPIC CpGs and their MAF. We took a random sample of SNPs with the same size as that of the set of interest (sample without replacement within each category, and with replacement across all the categories). Then, we did the enrichment analysis as described before, using the random sample as the background set. We repeated the process one thousand times and saved for each iteration the OR estimates. Finally, we obtained the mean OR of the annotations for the point estimate, and for the confidence intervals at $\alpha = 0.05$, we used the 2.5% and 97.5% percentiles of resampling distribution of OR. All the enrichment tests results are presented specifying the set of CpGs/SNPs of interest and the background set used (Additional file 3: Tables S5–S25).

We carried out gene ontology (GO) [91] enrichment analyses with the R implementation of clusterProfiler [92]. This package uses a one-tailed Fisher's exact test to find the overrepresentation of genes in molecular functions, cellular components or biological processes. *P*-values were corrected using Benjamini–Hochberg procedure, and redundancy was reduced in the enriched GO terms with the Wang method [93] using a similarity cut-off value of 0.7. GO enrichment analysis at a CpG level was performed with the GOmeth method available in the missMethyl package, which corrects for probe number and multi-gene bias [94].

Summary-based Mendelian Randomisation with gene expression data

We applied Summary-based Mendelian Randomization (SMR) analysis [35] between DNA methylation and gene expression. The SMR approach consists of two stages. First, it employs the most significant genetic variant associated with a CpG site as an instrumental variable (IV) to test the association between the CpG and a phenotype through the two-step least-squares (2SLS) estimation. This association can be due to a causal relationship, horizontal pleiotropy, or LD. Under a causal/pleiotropic scenario, the estimations of the effect sizes of the DNA methylation on the expression levels are expected to be homogeneous when calculated with other SNPs in LD with the single causal variant. Therefore, for excluding spurious associations derived from LD, the second step is the heterogeneity in dependent instruments (HEIDI) test with up to twenty alternative SNPs for each CpG. A significant *P*-value in the HEIDI test is evidence of heterogeneity across the effects of the SNPs and, therefore, indicates that the phenotype and the CpG are associated with different causal variants in LD.

We used summary statistics from the eQTLGen Consortium [34], carried out on 31,684 blood samples from individuals from 37 cohorts (mostly of European ancestry). We stuck to the *cis*-eQTL results—pairs of SNPs and genes no more than 1 Mbp apart, considering the centre of the gene—available at 19,250 genes. We used SMR v1.03 with the default settings, with the European subset of 1000 Genomes phase 3 version 5 as reference panel.

In the first analysis, we only used the *cis*-meQTLs to find associations between genes and nearby CpGs. We tested a total of 5,092,588 pairs of CpGs–genes, setting a statistical threshold of $P_{\text{SMR}} \leq 9.82 \times 10^{-9}$ (Bonferroni correction for a significance level of 0.05) and $P_{\text{HEIDI}} > 0.05$ to filter out associations due to LD.

The second SMR analysis was with the *trans*-meQTLs as IVs to test long-range associations through co-localised QTLs between all the CpGs and genes. We compared 134,698 CpG–gene pairs and established a significant threshold of $P_{\text{SMR}} \leq 3.71 \times 10^{-7}$ (Bonferroni correction for a significance level of 0.05) and $P_{\text{HEIDI}} > 0.05$.

The final SMR eQTL analysis consisted of identifying associations between CpGs and distant genes through genetic variants that were also significant in the *cis* SMR analysis. For this, we made a list of targets SNP-CpG pairs (separated by more than 5 Mbp to exclude cases of long-range LD) to test for each gene and use the option *--extract-target-snp-probe* in SMR v1.03. We considered the same significance criterion as in the *cis*-meQTL–*cis*-eQTL co-localisation ($P_{\mathrm{SMR}} \leq 9.82 \times 10^{-9}$, $P_{\mathrm{HEIDI}} > 0.05$).

Summary-based Mendelian Randomisation with GWAS data

We repeated the SMR approach to test for co-localisation of our significant *cis*-meQTLs with GWAS signals from 56 phenotypic traits, using summary statistics from previously published studies (details of each study in Additional file 5: Table S26). We downloaded and prepared data, adding the chromosomal position of the variants using dbSNP 141 as reference [95] (where not annotated), and harmonising the ID format with that of the meQTLs. We used SMR v1.03 with the default settings, with the European subset of 1000 Genomes phase 3 version 5 as reference panel. We categorised post hoc the phenotypes into seven classes. For filtering co-localisations with sufficient statistical evidence, we set a threshold of $P_{\mathrm{SMR}} \leq 3.82 \times 10^{-8}$ after the Bonferroni adjustment of the significance level of 0.05 for the number of independent tests ($186,817 \times 7$ accounting the CpGs tested and the phenotypic classes), and a $P_{\mathrm{HEIDI}} > 0.05$.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-03011-x.

Additional file 1: Supplementary note. Supplementary methods and results, cohort specific acknowledgements, supplementary tables S1–S4, supplementary figures S1–S18 and supplementary references [97–101].

Additional file 2. Summary statistics of meQTL replication results for selected CpG sites using MeDIP-seq. TSV file with the following fields: (1) **SNP**: meQTL SNP ID in the format of chr:pos_A1_A2. (2) **bin**: MeDIP-seq bin ID in the format of chr.pos1.pos2. (3) **CpG**: CpG ID included within the bin. (4) **type**: type of association, can be cis or trans. (5) **beta**: association coefficient estimate. (6) **t-stat**: *t*-statistic value. (7) **p-value**: raw *P*-value of the association. (8) **R2**: coefficient of determination of the association.

Additional file 3: Tables S5–S25. Table S5. List of all the enrichment analyses, set of CpGs/SNPs of interest, and background set used in each. **Tables S6–S17**. Enrichment analyses on genomic annotations. **Tables S18–S21**. Enrichment analyses on specific transcription factor binding sites annotations. **Tables S22–S25**. Enrichment analyses on gene ontologies.

Additional file 4. Summary statistics of Summary-based Mendelian Randomisation with gene expression data. TSV file with the following fields: (1) **SMRtest**: type of SMR test, can be cis-meQTL, trans-meQTL or target_trans-meQTL. (2) **CpG**: CpG ID. (3) **Gene**: ENSEMBL gene ID. (4) **SNP**: SNP ID in the format of chr:pos, tested for co-localisation. (5–6) **chr_cpg**, **pos_cpg**: chromosome and position of CpG. (7) **gene_cpg**: gene annotated to the CpG. (8–9) **chr_snp**, **pos_snp**: chromosome and position of SNP. (10) **A1**: effect allele. (11) **A2**: alternative allele. (12) **Freq**: frequency of A1. (13–15) **b_eQTL**, **se_eQTL**, **p_eQTL**: coefficient estimate, standard error and *p*-value from the eQTL association. (16–18) **b_meQTL**, **se_meQTL**, **p_meQTL**: coefficient estimate, standard error and *p*-value from the meQTL association. (19–21) **b_SMR**, **se_SMR**, **p_SMR**: coefficient estimate, standard error and *p*-value from the SMR. (22) **p_HEIDI**: *p*-value from HEIDI. (23) **nsnp_HEIDI**: number of SNPs used in the HEIDI test.

Additional file 5: Tables S26–S27. Table S26. List of GWASs summary statistics used for the SMR. **Table S27**. Summary of co-localisations by phenotypic class.

Additional file 6. Summary statistics of Summary-based Mendelian Randomisation with GWAS data. TSV file with the following fields: (1) **CpG**: CpG ID. (2) **Trait**: phenotypic trait ID. (3) **Class**: phenotypic class of trait. (4) **SNP**: SNP ID in the format of chr:pos, tested for co-localisation. (5–6) **chr_cpg**, **pos_cpg**: chromosome and position of CpG. (7) **gene_cpg**: gene annotated to the CpG. (8–9) **chr_snp**, **pos_snp**: chromosome and position of SNP. (10) **A1**: effect allele. (11) **A2**: alternative allele. (12) **Freq**: frequency of A1. (13–15) **b_GWAS**, **se_GWAS**, **p_GWAS**: coefficient estimate, standard error and *p*-value from the GWAS association. (16–18) **b_meQTL**, **se_meQTL**, **p_meQTL**: coefficient estimate, standard error and *p*-value from the meQTL association. (19–21) **b_SMR**, **se_SMR**, **p_SMR**: coefficient estimate, standard error and *p*-value from the SMR. (22) **p_HEIDI**: *p*-value from HEIDI. (23) **nsnp_HEIDI**: number of SNPs used in the HEIDI test.

Additional file 7. Review history.

Acknowledgements
The study authors express their gratitude to the research volunteers who took part in the study. Further acknowledgments specific to each cohort can be found in the Additional file 1: Section 2.

Review history
The review history is available as Additional file 7.

Peer review information
Anahita Bishop and Wenjing She were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
SV, JC-F and JTB designed the study and outlined the main conceptual ideas. JTB, JC-F, P-CT, JM, DK, MS, CP, CR, GP, AW, RH, AG and KKO generated the primary datasets or contributed toward data generation. SV led the data analysis. JC-F, EH and CC contributed to data analysis. SV developed the web application. SV and JTB wrote the article, and all authors provided critical feedback and helped shape the research, analysis and manuscript. All authors read and approved the final manuscript.

Funding
The work was supported by the UK Economic and Social Research Council (ESRC ES/N000404/1 to JTB), JPI-HDHL DIMENSION project funded in the UK via the Biotechnology and Biological Sciences Research Council (BBSRC BB/S020845/1 and BB/T019980/1 to JTB), and Mexican National Council of Science and Technology (CONACYT) doctoral fellowship (2019-000021-01EXTF-00323 to SV). Cohort-specific funding acknowledgements are included in the Additional file 1: Section 2.

Availability of data and materials
An interactive web application with our results is available at https://epicmeqtl.kcl.ac.uk (built using the R Shiny framework [96]), with summary statistics of top meQTLs and associations after LD clumping. Full summary statistics of heritability models, and of significant meQTL associations are available at Zenodo DOI 10.5281/zenodo.8047777 [25]. Additional files 2, 3, 4, 5, and 6 present the results of the replication and co-localisation analyses.

The TwinsUK methylation data are uploaded on the ReShare UK Data Service, under Data collection id 853,526. Access to further individual-level data can be applied for through the cohort data access committee, see https://twinsuk.ac.uk/resources-for-researchers/access-our-data/. The 1946BC-09 methylation dataset is available in the public domain through https://doi.org/10.5522/NSHD/S202. Access to further individual-level data can be applied for through the cohort data access committee, see http://www.nshd.mrc.ac.uk/data. The 1958BC-1 and 1958BC-2 methylation datasets are available through https://doi.org/10.5255/UKDA-SN-5594-2. Access to further individual-level data can be applied for through the cohort data access committee, see https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000032.

Declarations

Ethics approval and consent to participate
Ethical approval for TwinsUK was granted by the National Research Ethics Service London-Westminster, the St Thomas' Research Ethics Committee (REC reference numbers: EC04/015 and 07/H0802/84). All research participants have signed informed consent prior to taking part in any research activities. Ethical approval for 1946BC was granted by the Central Manchester Research Ethics Committee (07/H1008/168 and 07/H1008/245) and the Scotland A Research Ethics Committee (08/MRE00/12). Ethical approval for 1958BC swas granted by the London Central REC (14/LO/0097, 12/LO/2010 and 08/H0718/29) and by South East MREC (01/1/44). This covered consent for the collection of blood samples for health research. Biosamples for the sweep are held at University of Bristol and this has ethical approval as a tissue bank under application 09/H1010/12 from North-West Haydock NRES committee.

Competing interests
The authors declare that they have no competing interests.

Received: 6 October 2022 Accepted: 10 July 2023
Published online: 31 July 2023

References
1. Pervjakova N, Kasela S, Morris AP, Kals M, Metspalu A, Lindgren CM, et al. Imprinted genes and imprinting control regions show predominant intermediate methylation in adult somatic tissues. Epigenomics. 2016;8(6):789–99.
2. Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, Dupre Y, et al. DNA methylation profiles of human active and inactive X chromosomes. Genome Res. 2011;21(10):1592–600.
3. Vilain A, Bernardino J, Gerbault-Seureau M, Vogt N, Niveleau A, Lefrancois D, et al. DNA methylation and chromosome instability in lymphoblastoid cell lines. Cytogenet Genome Res. 2000;90(1–2):93–101.
4. Bourc'his D, Xu GL, Lin CS, Bollman B, Bestor TH. Dnmt3L and the establishment of maternal genomic imprints. Science. 2001;294(5551):2536–9.
5. Song J, Rechkoblit O, Bestor TH, Patel DJ. Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. Science. 2011;331(6020):1036–40.

6. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet. 2009;5(8):e1000602.
7. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LTY, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500(7463):477–81.
8. McRae AF, Powell JE, Henders AK, Bowdler L, Hemani G, Shah S, et al. Contribution of genetic variation to transgenerational inheritance of DNA methylation. Genome Biol. 2014;15(5):1–10.
9. Van Dongen J, Nivard MG, Willemsen G, Hottenga JJ, Helmer Q, Dolan CV, et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. Nat Commun. 2016;7(1):1–13.
10. Min JL, Hemani G, Hannon E, Dekkers KF, Castillo-Fernandez J, Luijk R, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. Nat Genet. 2021;53(9):1311–21.
11. Hawe JS, Wilson R, Schmid KT, Zhou L, Lakshmanan LN, Lehne BC, et al. Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. Nat Genet. 2022;54(1):18–29.
12. Ng B, White CC, Klein HU, Sieberts SK, McCabe C, Patrick E, et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat Neurosci. 2017;20(10):1418–26.
13. Schulz H, Ruppert AK, Herms S, Wolf C, Mirza-Schreiber N, Stegle O, et al. Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus. Nat Commun. 2017;8(1):1–11.
14. Grundberg E, Meduri E, Sandling JK, Hedman ÅK, Keildson S, Buil A, 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. 2013;93(5):876–90.
15. Van Dongen J, Ehli EA, Jansen R, Van Beijsterveldt CE, Willemsen G, Hottenga JJ, et al. Genome-wide analysis of DNA methylation in buccal cells: a study of monozygotic twins and mQTLs. Epigenetics Chromatin. 2018;11(1):1–14.
16. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011;6(6):692–702.
17. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016;8(3):389–99.
18. Hannon E, Gorrie-Stone TJ, Smart MC, Burrage J, Hughes A, Bao Y, et al. Leveraging DNA-methylation quantitative-trait loci to characterize the relationship between methylomic variation, gene expression, and complex traits. Am J Hum Genet. 2018;103(5):654–65.
19. Husquin LT, Rotival M, Fagny M, Quach H, Zidane N, McEwen LM, et al. Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation. Genome Biol. 2018;19(1):1–17.
20. Verdi S, Abbasian G, Bowyer RC, Lachance G, Yarand D, Christofidou P, et al. TwinsUK: the UK adult twin registry update. Twin Res Hum Genet. 2019;22(6):523–9.
21. Kuh D, Pierce M, Adams J, Deanfield J, Ekelund U, Friberg P, et al. Cohort profile: updating the cohort profile for the MRC National Survey of Health and Development: a new clinic-based data collection for ageing research. Int J Epidemiol. 2011;40(1):e1–9.
22. Wadsworth M, Kuh D, Richards M, Hardy R. Cohort profile: the 1946 National Birth Cohort (MRC National Survey of Health and Development). Int J Epidemiol. 2006;35(1):49–54.
23. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol. 2006;35(1):34–41.
24. Fuller E, Power C, Shepherd P, Strachan D. Technical report on the National Child Development Study biomedical survey 2002–2004. National Centre for Social Research. 2006.
25. Villicaña S, Castillo-Fernandez J, Hannon E, Christiansen C, Tsai PC, Maddock J, et al. Genetic impacts on DNA methylation help elucidate regulatory genomic processes. Zenodo. 2023. https://doi.org/10.5281/zenodo.8047777.
26. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell. 2016;167(5):1398–414.
27. Banovich NE, Lan X, McVicker G, Van de Geijn B, Degner JF, Blischak JD, et al. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS Genet. 2014;10(9):e1004663.
28. Do C, Lang CF, Lin J, Darbary H, Krupska I, Gaba A, et al. Mechanisms and disease associations of haplotype-dependent allele-specific DNA methylation. Am J Hum Genet. 2016;98(5):934–55.
29. Villicaña S, Bell JT. Genetic impacts on DNA methylation: research findings and future perspectives. Genome Biol. 2021;22(1):1–35.
30. Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res. 2012;22(9):1680–8.
31. Michaud J, Praz V, Faresse NJ, JnBaptiste CK, Tyagi S, Schütz F, et al. HCFC1 is a common component of active human CpG-island promoters and coincides with ZNF143, THAP11, YY1, and GABP transcription factor occupancy. Genome Res. 2013;23(6):907–16.
32. Guilhamon P, Eskandarpour M, Halai D, Wilson GA, Feber A, Teschendorff AE, et al. Meta-analysis of IDH-mutant cancers identifies EBF1 as an interaction partner for TET2. Nat Commun. 2013;4(1):1–9.
33. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43–9.
34. Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis-and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53(9):1300–10.
35. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48(5):481–7.
36. Pierce BL, Tong L, Argos M, Demanelis K, Jasmine F, Rakibuz-Zaman M, et al. Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. Nat Commun. 2018;9(1):804.

37. Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z, et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nat Commun. 2018;9(1):1–14.
38. Huan T, Joehanes R, Song C, Peng F, Guo Y, Mendelson M, et al. Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. Nat Commun. 2019;10(1):4267.
39. Lemire M, Zaidi SH, Ban M, Ge B, Aïssi D, Germain M, et al. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. Nat Commun. 2015;6(1):1–12.
40. Filion GJ, Zhenilo S, Salozhin S, Yamada D, Prokhortchouk E, Defossez PA. A family of human zinc finger proteins that bind methylated DNA and repress transcription. Mol Cell Biol. 2006;26(1):169–81.
41. Monteagudo-Sánchez A, Hernandez Mora JR, Simon C, Burton A, Tenorio J, Lapunzina P, et al. The role of ZFP57 and additional KRAB-zinc finger proteins in the maintenance of human imprinted methylation and multi-locus imprinting disturbances. Nucleic Acids Res. 2020;48(20):11394–407.
42. Ying Y, Wang M, Chen Y, Li M, Ma C, Zhang J, et al. Zinc finger protein 280C contributes to colorectal tumorigenesis by maintaining epigenetic repression at H3K27me3-marked loci. Proc Natl Acad Sci. 2022;119(22):e2120633119.
43. Hockemeyer D, Palm W, Else T, Daniels JP, Takai KK, Ye JZ, et al. Telomere protection by mammalian Pot1 requires interaction with Tpp1. Nat Struct Mol Biol. 2007;14(8):754–61.
44. Buxton JL, Suderman M, Pappas JJ, Borghol N, McArdle W, Blakemore AI, et al. Human leukocyte telomere length is associated with DNA methylation levels in multiple subtelomeric and imprinted loci. Sci Rep. 2014;4(1):1–8.
45. Dong Y, Huang Y, Gutin B, Raed A, Dong Y, Zhu H. Associations between global DNA methylation and telomere length in healthy adolescents. Sci Rep. 2017;7(1):1–6.
46. Bandaria JN, Qin P, Berk V, Chu S, Yildiz A. Shelterin protects chromosome ends by compacting telomeric chromatin. Cell. 2016;164(4):735–46.
47. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49(1):131–8.
48. Wang Y, Song F, Zhang B, Zhang L, Xu J, Kuang D, et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. 2018;19(1):1–12.
49. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80.
50. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289–93.
51. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep. 2016;17(8):2042–59.
52. Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature. 2015;518(7539):350–4.
53. García-Santisteban I, Peters GJ, Giovannetti E, Rodríguez JA. USP1 deubiquitinase: cellular functions, regulatory mechanisms and emerging potential as target in cancer therapy. Mol Cancer. 2013;12(1):1–12.
54. Guo T, Yin RX, Huang F, Yao LM, Lin WX, Pan SL. Association between the DOCK7, PCSK9 and GALNT2 gene polymorphisms and serum lipid levels. Sci Rep. 2016;6(1):1–18.
55. Tikka A, Jauhiainen M. The role of ANGPTL3 in controlling lipoprotein metabolism. Endocrine. 2016;52(2):187–93.
56. Liu JZ, Van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47(9):979–86.
57. Hedl M, Lahiri A, Ning K, Cho JH, Abraham C. Pattern recognition receptor signaling in human dendritic cells is enhanced by ICOS ligand and modulated by the Crohn's disease ICOSLG risk allele. Immunity. 2014;40(5):734–46.
58. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491(7422):119–24.
59. van Dongen J, Willemsen G, Heijmans BT, Neuteboom J, Kluft C, Jansen R, et al. Longitudinal weight differences, gene expression and blood biomarkers in BMI-discordant identical twins. Int J Obes. 2015;39(6):899–909.
60. Volckmar AL, Song JY, Jarick I, Pütter C, Göbel M, Horn L, et al. Fine mapping of a GWAS-derived obesity candidate region on chromosome 16p11. 2. PLoS ONE. 2015;10(5):e0125660.
61. Zhao T, Hu Y, Zang T, Wang Y. Integrate GWAS, eQTL, and mQTL data to identify Alzheimer's disease-related genes. Front Genet. 2019;10:1021.
62. Kim S, Forno E, Yan Q, Jiang Y, Zhang R, Boutaoui N, et al. SNPs identified by GWAS affect asthma risk through DNA methylation and expression of cis-genes in airway epithelium. Eur Respir J. 2020;55(4):1–4.
63. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 2016;17(1):1–14.
64. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
65. 1000 Genomes Project Consortium, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68.
66. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284–7.
67. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. PLoS Biol. 2011;9(7):e1001091.
68. Xu Z, Niu L, Taylor JA. The ENmix DNA methylation analysis pipeline for Illumina BeadChip and comparisons with seven other preprocessing pipelines. Clin Epigenetics. 2021;13(1):1–8.
69. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2021.
70. Xu Z, Langie SA, De Boever P, Taylor JA, Niu L. RELIC: a novel dye-bias correction method for Illumina Methylation BeadChip. BMC Genomics. 2017;18(1):1–7.
71. Niu L, Xu Z, Taylor JA. RCP: a novel probe design bias correction method for Illumina Methylation BeadChip. Bioinformatics. 2016;32(17):2659–63.
72. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017;45(4):e22–e22.

73. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13(1):1–16.
74. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363–9.
75. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. 2015;67(1):1–48.
76. Neale MC, Hunter MD, Pritikin JN, Zahery M, Brick TR, Kirkpatrick RM, et al. OpenMx 2.0: Extended structural equation and statistical modeling. Psychometrika. 2016;81(2):535–549.
77. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28(10):1353–8.
78. Mägi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics. 2010;11(1):1–6.
79. Hastie T, Tibshirani R, Friedman J. High-dimensional problems: p N. In: The elements of statistical learning. New York: Springer; 2009. p. 649–98.
80. Bell CG, Gao F, Yuan W, Roos L, Acton RJ, Xia Y, et al. Obligatory and facilitative allelic variation in the DNA methylome within common disease-associated loci. Nat Commun. 2018;9(1):1–13.
81. Bell CG, Xia Y, Yuan W, Gao F, Ward K, Roos L, et al. Novel regional age-associated DNA methylation changes within human common disease-associated loci. Genome Biol. 2016;17(1):1–14.
82. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
83. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
84. Chavez L, Jozefczuk J, Grimm C, Dietrich J, Timmermann B, Lehrach H, et al. Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. Genome Res. 2010;20(10):1441–50.
85. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32(suppl_1):D493–6.
86. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.
87. ENCODE Project Consortium, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.
88. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Res. 2021;49(D1):D884–91.
89. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics. 2016;32(4):587–9.
90. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodological). 1995;57(1):289–300.
91. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021;49(D1):D325–34.
92. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation. 2021;2(3):100141.
93. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23(10):1274–81.
94. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analysing methylation data from Illuminas HumanMethylation450 platform. Bioinformatics. 2016;32(2):286–8.
95. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.
96. Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, et al. shiny: Web Application Framework for R. 2022. R package version 1.7.2.
97. Borghol N, Suderman M, McArdle W, Racine A, Hallett M, Pembrey M, et al. Associations with early-life socio-economic position in adult DNA methylation. Int J Epidemiol. 2012;41(1):62–74.
98. Suderman M, Borghol N, Pappas JJ, Pinto Pereira SM, Pembrey M, Hertzman C, et al. Childhood abuse is associated with methylation of multiple loci in adult DNA. BMC Med Genomics. 2014;7(1):1–12.
99. Josh Pasek and with some assistance from Alex Tahk and some code modified from R-core; Additional contributions by Gene Culter and Marcus Schwemmle. weights: Weighting and Weighted Statistics. 2021. R package version 1.0.4.
100. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4(8):1184–91.
101. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118.

Publisher's Note

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