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

Josine L Min[1,2*], Gibran Hemani[1,2*], Eilis Hannon[3], Koen F Dekkers[4], Juan Castillo-Fernandez[5], René Luijk[4], Elena Carnero-Montoro[5,6], Daniel J Lawson[1,2], Kimberley Burrows[1,2], Matthew Suderman[1,2], Andrew D Bretherick[7], Tom G Richardson[1,2], Johanna Klughammer[8], Valentina Iotchkova[9], Gemma Sharp[1,2], Ahmad Al Khleifat[10], Aleksey Shatunov[10], Alfredo Iacoangeli[10,11], Wendy L McArdle[2], Karen M Ho[2], Ashish Kumar[12,13,14], Cilla Söderhäll[15], Carolina Soriano-Tárraga[16], Eva Giralt-Steinhauer[16], Nabila Kazmi[1,2], Dan Mason[17], Allan F McRae[18], David L Corcoran[19], Karen Sugden[19,20], Silva Kasela[21], Alexia Cardona[22,23], Felix R Day[22], Giovanni Cugliari[24,25], Clara Viberti[24,25], Simonetta Guarrera[24,25], Michael Lerro[26], Richa Gupta[27,28], Sailalitha Bollepalli[27,28], Pooja Mandaviya[29], Yanni Zeng[7,30,31], Toni-Kim Clarke[32], Rosie M Walker[33,34], Vanessa Schmoll[35], Darina Czamara[35], Carlos Ruiz-Arenas[36,37,38], Faisal I Rezwan[39], Riccardo E Marioni[33,34], Tian Lin[18], Yvonne Awaloff[35], Marine Germain[40], Dylan Aïssi[41], Ramona Zwamborn[42], Kristel van Eijk[42], Annelot Dekker[42], Jenny van Dongen[43], Jouke-Jan Hottenga[43], Gonneke Willemsen[43], Cheng-Jian Xu[44,45], Guillermo Barturen[6], Francesc Català-Moll[46], Martin Kerick[47], Carol Wang[48], Phillip Melton[49], Hannah R Elliott[1,2], Jean Shin[50], Manon Bernard[50], Idil Yet[5,51], Melissa Smart[52], Tyler Gorrie-Stone[52], BIOS Consortium[53], Chris Shaw[10,54], Ammar Al Chalabi[10,54,55], Susan M Ring[1,2], Göran Pershagen[12], Erik Melén[12,56], Jordi Jiménez-Conde[16], Jaume Roquer[16], Deborah A Lawlor[1,2], John Wright[17], Nicholas G Martin[57], Grant W Montgomery[18], Terrie E Moffitt[19,20,58,61], Richie Poulton[59], Tõnu Esko[21,60], Lili Milani[21], Andres Metspalu[21], John RB Perry[22], Ken K Ong[22], Nicholas J Wareham[22], Giuseppe Matullo[24,25], Carlotta Sacerdote[25,62], Salvatore Panico[63], Avshalom Caspi[19,20,58,61], Louise Arseneault[61], France Gagnon[26], Miina Ollikainen[27,28], Jaakko Kaprio[27,28], Janine F Felix[64,65], Fernando Rivadeneira[29], Henning Tiemeier[66,67], Marinus H van IJzendoorn[68,69], André G Uitterlinden[29], Vincent WV Jaddoe[64,65], Chris Haley[7], Andrew M McIntosh[32,34], Kathryn L Evans[33,34], Alison Murray[70], Katri Räikkönen[71], Jari Lahti[71], Ellen A Nohr[72,73], Thorkild IA Sørensen[1,2,74,75], Torben Hansen[74], Camilla Schmidt Morgen[76], Elisabeth B Binder[35,77], Susanne Lucae[35], Juan Ramon Gonzalez[36,37,38], Mariona Bustamante[36,37,38,78], Jordi Sunyer[36,37,38,79], John W Holloway[80,81], Wilfried Karmaus[82], Hongmei Zhang[82], Ian J Deary[34], Naomi R Wray[18,83], John M Starr[34,84], Marian Beekman[4], Diana van Heemst[85], P Eline Slagboom[4], Pierre-Emmanuel Morange[86], David-Alexandre Trégouët[40], Jan H Veldink[42], Gareth E Davies[87], Eco JC de Geus[43], Dorret I Boomsma[43], Judith M Vonk[88], Bert Brunekreef[89,90], Gerard H Koppelman[44], Marta E Alarcón-Riquelme[6,12], Rae-Chi Huang[91], Craig Pennell[48], Joyce van Meurs[29], M Arfan Ikram[92], Alun D Hughes[93], Therese Tillin[93], Nish Chaturvedi[93], Zdenka Pausova[50], Tomas Paus[94], Timothy D Spector[5], Meena Kumari[52], Leonard C Schalkwyk[52], Peter M Visscher[18,83], George Davey Smith[1,2], Christoph Bock[8,95], Tom R Gaunt[1,2], Jordana T Bell[5‡], Bastiaan T Heijmans[4‡], Jonathan Mill[3‡], Caroline L Relton[1,2‡]

* These authors contributed equally to this research.
‡These authors jointly supervised this work.

Corresponding author: Josine L Min, josine.min@bristol.ac.uk

Affiliations

[1] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK
[2] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
[3] University of Exeter Medical School, College of Medicine and Health, University of Exeter, Exeter, UK
[4] Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
[5] Department of Twin Research and Genetic Epidemiology, King's College London, London, UK
[6] Pfizer - University of Granada - Andalusian Government Center for Genomics and Oncological Research (GENYO), Spain
[7] MRC Human Genetic Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
[8] CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria
[9] MRC Weatherall Institute of Molecular Medicine, Oxford, UK
[10] Department of Basic and Clinical Neuroscience, Maurice Wohl Clinical Neuroscience Institute, London, UK
[11] Department of Biostatistics and Health Informatics, King's College London, London, UK
[12] Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Solna, Sweden
[13] Chronic Disease Epidemiology unit, Swiss Tropical and Public Health Institute, Basel, Switzerland
[14] University of Basel, Basel, Switzerland
[15] Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden
[16] Neurology Department, Hospital del Mar - IMIM (Institut Hospital del Mar d'Investigacions Mèdiques), Barcelona, Spain
[17] Bradford Institute for Health Research, Bradford, UK
[18] Institute for Molecular Bioscience, University of Queensland, Australia
[19] Center for Genomic and Computational Biology, Duke University, Durham, NC, USA
[20] Department of Psychology and Neuroscience, Duke University, Durham, NC, USA
[21] Estonian Genome Center, Institute of Genomics, University of Tartu, Estonia
[22] MRC Epidemiology Unit, University of Cambridge, School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge CB2 0QQ, United Kingdom

[23] Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, United Kingdom
[24] Department of Medical Sciences, University of Turin, Turin, Italy
[25] Italian Institute for Genomic Medicine (IIGM), Turin, Italy
[26] Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
[27] Institute for Molecular Medicine, University of Helsinki, Helsinki, Finland
[28] Department of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland
[29] Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands
[30] Faculty of Forensic Medicine, Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China
[31] Guangdong Province Key Laboratory of Brain Function and Disease, Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China
[32] Division of Psychiatry, Royal Edinburgh Hospital, University of Edinburgh, Edinburgh EH10 5HF, UK
[33] Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, Western General Hospital, University of Edinburgh, Crewe Road, Edinburgh EH4 2XU, UK
[34] Centre for Cognitive Ageing and Cognitive Epidemiology, Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK
[35] Department of Translational Research in Psychiatry, Max-Planck-Institute of Psychiatry, Munich, Germany
[36] ISGlobal, Barcelona Global Health Institute, Barcelona, Spain
[37] Universitat Pompeu Fabra (UPF), Barcelona, Spain
[38] CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[39] School of Water, Energy and Environment, Cranfield University, Cranfield, Bedfordshire MK43 0AL, UK
[40] INSERM UMR_S 1219, Bordeaux Population Health Center, University of Bordeaux, 33076 Bordeaux Cedex, France
[41] Department of General and Interventional Cardiology, University Heart Center Hamburg, Hamburg, Germany
[42] Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, 3584 CG, The Netherlands
[43] Department of Biological Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Van Der Boechorststraat 7-9, 1081 BT, Amsterdam, The Netherlands
[44] University of Groningen, University Medical Center Groningen, Department of Pediatric Pulmonology and Pediatric Allergology, Beatrix Children’s Hospital, GRIAC Research Institute Groningen, The Netherlands
[45] CiiM and TWINCORE, joint ventures between the Hannover Medical School and the Helmholtz Centre for Infection Research, Hannover, Germany

[46] Chromatin and Disease Group, Cancer Epigenetics and Biology Programme (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), 08908 L'Hospitalet de Llobregat, Barcelona, Spain
[47] Instituto de Parasitología y Biomedicina López Neyra, CSIC, Granada, Spain
[48] School of Medicine and Public Health, Faculty of Medicine and Health, The University of Newcastle, Newcastle, NSW, Australia
[49] Menzies Institute for Medical Research, College of Health and Medicine, University of Tasmania, Hobart, Australia, School of Global Population Health, Faculty of Health and Medical Sciences, The University of Western Australia, Perth, WA, Australia; School of Pharmacy and Biomedical Sciences, Faculty of Health Sciences, Curtin University, Perth, WA, Australia
[50] The Hospital for Sick Children, University of Toronto, Toronto, Canada
[51] Department of Bioinformatics, Institute of Health Sciences, Hacettepe University, 06100, Ankara, Turkey
[52] School of Life Sciences, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ
[53] A list of members and affiliations appears at the end of the paper.
[54] Department of Neurology, King's College Hospital, London, UK
[55] United Kingdom Dementia Research Institute, King's College London, London, UK
[56] Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden
[57] QIMR Berghofer Medical Research Institute, Brisbane, Australia
[58] Department of Psychiatry and Behavioral Sciences, Duke University Medical School, Durham, NC, USA
[59] Dunedin Multidisciplinary Health and Development Research Unit, Department of Psychology, University of Otago, Dunedin, New Zealand
[60] Program in Medical and Population Genetics, Broad Institute, Broad Institute, Cambridge, MA, USA
[61] Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King’s College London, London, UK
[62] Piemonte Centre for Cancer Prevention, Turin, Italy
[63] Dipartimento Di Medicina Clinica E Chirurgia, Federico II University, Naples, Italy
[64] The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands
[65] Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands
[66] Department of Child and Adolescent Psychiatry, Erasmus Medical Center, Rotterdam, Netherlands
[67] Department of Social and Behavioral Science, Harvard TH Chan School of Public Health, Boston, USA
[68] School of Clinical Medicine, University of Cambridge, UK
[69] Department of Clinical, Educational and Health Psychology, Division on Psychology and Language Sciences, Faculty of Brain Sciences, UCL, London, UK
[70] Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK

[71] Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Finland
[72] Research Unit for Gynaecology and Obstetrics, Institute of Clinical research, University of Southern Denmark, Odense, Denmark
[73] Centre of Women's, Family and Child Health, University of South-Eastern Norway, Kongsberg, Norway
[74] The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
[75] Department of Public Health (Section of Epidemiology), Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
[76] The National Institute of Public Health, University of Southern Denmark, Copenhagen
[77] Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA, USA
[78] Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain
[79] IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain
[80] Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK
[81] Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, UK
[82] Division of Epidemiology, Biostatistics, and Environmental Health Sciences, School of Public Health, University of Memphis, Memphis, USA
[83] Queensland Brain Institute, University of Queensland, Australia
[84] Alzheimer Scotland Dementia Research Centre, University of Edinburgh, University of Edinburgh, UK
[85] Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands
[86] C2VN, Aix-Marseille University, INSERM, INRAE, Marseille, France
[87] Avera Institute for Human Genetics, Sioux Falls, USA
[88] University of Groningen, University Medical Center Groningen, Department of Epidemiology, GRIAC Research Institute Groningen, Groningen, The Netherlands
[89] Institute for Risk Assessment Sciences, Universiteit Utrecht, Utrecht, The Netherlands
[90] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
[91] Telethon Kids Institute, University of Western Australia, Perth, WA, Australia
[92] Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands
[93] UCL Institute of Cardiovascular Science, London, UK
[94] Departments of Psychology and Psychiatry, University of Toronto, Toronto, Canada
[95] Institute of Artificial Intelligence and Decision Support, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria

Abstract – 146 words-max 150

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

Main – 4996 words - max 5000

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

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

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

Results

Genetic variants influence 45% of tested DNAm sites

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

Using linkage disequilibrium (LD) clumping, we identified 248,607 independent *cis*-mQTL associations ($p < 1e-8$, < 1Mb from the DNAm site, **Supplementary Figure 3**) with a median distance between single nucleotide polymorphisms (SNP) and DNAm sites of 36kb (IQR=118 kb, **Extended data 2**). We found 23,117 independent *trans* mQTL associations (using a conservative threshold of $p < 1e-14$[7], **Supplementary Figure 3**, **Supplementary Information**). These mQTL involved 190,102 DNAm sites, representing 45.2% of all those tested (**Figure 1b)** which is a 1.9x increase of sites with a *cis* association ($p<1e-8$) and 10x increase of sites with a *trans* association ($p<1e-14$) over a previous study whose sample size was 7x smaller[8]. As expected, mQTL effect sizes for each DNAm site (the maximum absolute additive change in DNAm level measured in standard deviation (SD) per allele) were lower for sites with a *trans* association (as compared to sites with a *cis* association (per allele SD change = -0.02 (s.e.=0.002, p=2.1e-14, **Extended data 3**). The differential improvement in yield between *cis* and *trans* associations is revealing in terms of the genetic architecture – relatively small sample sizes are sufficient to uncover the majority of large *cis* effects, whereas much larger sample sizes are required to identify the polygenic *trans* component.

The majority of *trans* associations (80%) were inter-chromosomal. Of the intra-chromosomal *trans* associations, 34% were >5 Mb from the DNAm site, **Extended data 2a**). We found a substantially lower number of inter-chromosomal *trans* associations per 5 Mb region (1.59) than intra-chromosomal associations (>1 Mb: 7.95; >6 Mb 4.81, excluding chromosome 6).

Next, using conditional analysis[14] we explored the potential for multiple independent SNPs operating within the locus of each mQTL, identifying 758,130 putative independent variants. Each DNAm site, for which a mQTL in *cis* had been detected, had a median of 2 independent variants (IQR=4 variants, **Supplementary Figure 5**). For all subsequent analyses, we used index SNPs from clumping procedures to be conservative and unbiased due to the non-independence of genetic variants.

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

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

We then investigated how much of the heritability of variable DNAm can be explained by our mQTL associations using family-based heritability studies of DNAm[2,15]. We found a strong positive relationship between variance explained by replication mQTL estimates (127,680 sites in GS) and heritability for both studies (family: Pearson r=0.41 across, 121,582 available sites; twin: Pearson r=0.37 across 118,955 available sites)

(**Figure 1d, Supplementary Data 2**). The mQTL that we identified explain 15%-17% of the additive genetic variance of DNAm (**Supplementary Figure 7**). Finally, there were strong positive relationships between the heritability of DNAm levels at a DNAm site and the number of independent mQTL (**Supplementary Figure 8**), heritability and effect size **(Supplementary Figure 9),** variance explained and the number of independent mQTL **(Supplementary Figure 10)** and variance explained and distribution of DNAm levels (**Supplementary Figure 11**)**.** Overall, our results support a mixed genetic architecture of polygenic genome-wide effects and larger *cis* effects.

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

Cis and *trans* mQTL operate through distinct mechanisms

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

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

OR=2.03). These enrichment patterns were consistent if we restricted to only inter-chromosomal associations (**Supplementary Information**, **Supplementary Figure 17**).

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

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

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

Trans mQTL SNPs and DNAm exhibit patterned TF binding

Recent studies have uncovered multiple types of transcription factor (TFs)/DNA interactions influenced by DNAm including the binding of DNAm-sensitive TFs[26-28] and cooperativity between TFs[27,29]. To gain insights into how SNPs induce long-range DNAm changes, we mapped enrichments for DNAm sites and SNPs across binding sites for 171 TFs in 27 cell types[30,31]. We found strong enrichments for the majority of TFs and cell types amongst DNAm sites with a *trans* association (*cis+trans:* 55%; *trans only*: 80%; *cis only:* 18%) and amongst *cis-acting* SNPs (*cis only*: 96%, *cis+trans*: 91%, *trans only*: 1%) **(Figures 2b, Supplementary Tables 7-8, Supplementary Figures 22-23).** Consistent with the observation that *trans only* DNAm sites are enriched for CpG islands (**Supplementary Figure 13**), DNAm sites that overlap TFBS were relatively hypomethylated (weighted mean DNAm levels = 21% vs 52%, p<2.2e-16) (**Supplementary Figure 24**).

Next we hypothesized that if a *trans* mQTL is driven by TF activity[8,10] then particular TF-TF pairs may exhibit preferential enrichment[32]. A mQTL has a pair of TFBS annotations[31], one for the SNP and one for the DNAm site. We evaluated if the annotation pairs amongst 18,584 inter-chromosomal *trans*-mQTL were associated to TF binding in a non-random pattern (**Supplementary Information, Extended Data 6a-b).** We found that 6.1% (22,962 of 378,225) of possible pairwise combinations of SNP-DNAm site annotations were more over- or under-represented than expected by chance after strict multiple testing correction (**Supplementary Information**, **Supplementary Table 9, Extended Data 6c**).

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

Next, we compared the locations of inter-chromosomal *trans* mQTL (n=18,584) to known regions of chromatin interactions[36] as alternative mechanism for *trans* coordination[8,37]. We found 1175 overlaps for 637 SNP-DNAm site pairs (3.4%) where the LD region of the mQTL SNP and the corresponding site overlapped with any interacting regions (525 SNPs, 602 sites) as compared to a mean of 473 SNP-DNAm site pairs in 1000 permuted datasets (OR=1.36, p_{Fisher}=6.5e-7, $p_{empirical}$<1e-3) (**Supplementary Figure 25).** To summarise, our results show that *trans* mQTL are in part driven by long-range cooperative TF interactions and, that for a small proportion of interchromosomal *trans* mQTL the spatial distance *in vivo* is likely to be small.

Trans-mQTL effects form DNAm communities

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

We hypothesised that *cis* sites were causally influencing multiple *trans* sites within their communities. We evaluated whether the estimated causal effect (obtained from the *trans*-mQTL effect divided by the *cis*-mQTL effect i.e. the Wald ratio) of the *cis* site on the *trans* site was consistent with the observational correlation between the *cis*- and *trans*-site. While there was an association, the relationship was weak (Pearson r=0.096, p=1.73e-6, **Supplementary Figure 26**), indicating that changes in *cis* sites causing changes in *trans* sites is likely not the predominant mechanism. We did observe that the *cis-trans* DNAm levels were more strongly correlated than we would expect by chance (**Supplementary Figure 27**), suggesting that they are jointly regulated without generally being causally related.

Next, we evaluated if DNAm sites within each community were enriched for regulatory annotations and/or gene ontologies **(Supplementary Tables 11-14, Supplementary Figures 28-29)**. Multiple communities showed enrichments (FDR <0.001); community 9 DNAm sites were strongly enriched for TFBS annotations relating to the cohesin complex in multiple cell types, community 22 DNAm sites were enriched for NFKB and EBF1 in B lymphocytes and community 76 DNAm sites were enriched for EZH2 and SUZ12 and bivalent promotor and repressed polycomb states (**Figure 3b**). Community 2 (comprising 253 sites) was enriched for active enhancer state in 3 cell types and for lymphocyte activation (GO:0046649 FDR = 0.016) and multiple KEGG pathways including the JAK-STAT signalling pathway (I04630: FDR =8.53e-7) (**Supplementary Tables 13 and 14**).

Regulatory features within a network may share a set of biological features that are related to complex traits. We performed enrichment analysis to evaluate if the loci tagged by DNAm sites in a community were related to each of 133 complex traits (**Supplementary Table 15**), accounting for non-random genomic properties of the selected loci. Restricting the analysis to only the 56 communities with ten or more sites, we found eleven communities that tagged genomic loci that were enriched for small p-values with 22 complex traits (FDR < 0.05) (**Figure 3c, Supplementary Table 16**). Blood related phenotypes were overrepresented (11 out of 23 enrichments being related to metal levels or haematological measures, binomial test p-value = 4.2e-5). Amongst the communities enriched for GWAS signals, community 16 was highly associated with iron and haemoglobin traits. Community 9 was associated to plasma cortisol (p = 8.27e-5). Finally, we performed enrichment analysis on 36 blood cell count traits[39]. We found that community 16 was enriched for hematocrit (p=4.34e-10) and hemoglobin concentration (p=1.99e-8) and community 5 was enriched for reticulocyte traits (p=1.67e-6) **(Supplementary Figure 30).** The enrichments found for these DNAm communities indicate that a potentially valuable utility of mapping *trans*-mQTL is to indicate how distal regions of the genome are functionally related.

DNAm and complex traits share genetic factors

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

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

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

It has previously been difficult to conclude whether genetic colocalisation between DNAm and complex traits indicates a) a causal relationship where the DNAm level is on the pathway from genetic variant to trait (vertical pleiotropy) or b) a non-causal relationship where the variant influences the trait and DNAm independently through different pathways (horizontal pleiotropy)[51]. In Mendelian randomisation (MR) it is reasoned that under a causal model, multiple independent genetic variants influencing DNAm should exhibit consistent causal effects on the complex trait[52]. Amongst the putative colocalising signals, 440 (32%) involved a DNAm site that had at least one other independent mQTL. We cannot determine with certainty the causal relationship of any specific site with a trait. To test if there was a general trend of DNAm sites causally influencing a trait we evaluated if the MR effect estimate based on the colocalising signals were consistent with those obtained based on the secondary signals. There were substantially more large genetic effects of the secondary mQTL on respective traits than expected by chance (70 with $p < 0.05$, binomial test p = 2.4e-16). However only 41 (59%) of these had effect estimates in the same direction as the primary colocalising variant, which is not substantially better than chance (binomial test p = 0.19). Twelve of the 41 mQTL were located in the *HLA* region. Of the remaining mQTL, 27 were associated with anthropometric (*ESR1* and birth weight), immune response (*IRF5* and systemic lupus erythematosus) and lipid traits (*TBL2* and triglycerides). We then performed systematic colocalization analysis of all mQTL against 36 blood cell traits[39]. Here we discovered 94,738 instances of a DNAm site and a blood cell trait sharing a causal variant. In 28,138 instances the colocalising DNAm site had an independent secondary mQTL, and with these associations we again tested for a general trend of DNAm sites causally influencing the blood trait. The association between independent signals was very weak ($R^2 = 0.008$). Together, across the sites that were analysable in this manner, these results indicate that those blood measured DNAm sites that have shared genetic factors with traits cannot be typically thought of as mediating the genetic association to the trait (**Extended Figure 7, Supplementary Table 19**). Instead, if DNAm is a coregulatory phenomenon then the colocalising signals between DNAm sites and complex traits may be due to a common cause, for example genetic variants primarily acting on TF binding.[8,10]

The influence of traits on DNAm variation

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

We initially identified 4785 associations where risk factors or genetic liability to disease influences DNAm levels (multiple testing threshold $p < 1.4e-7$). However, causal inference on omic variables can lead to false positives due to violations in the MR assumptions. We developed a filtering process involving a novel causal inference method to help protect against these invalid associations (**Supplementary Information, Supplementary Figure 33**). This left 85 associations (involving 84 DNAm sites) in which DNAm sites were putatively influenced by 13 traits (nine risk factors or four diseases) (**Supplementary Table 20**). Further filtering that would exclude traits that were predominantly instrumented by variants in the *HLA* region or driven by one SNP would reduce the total number of associations substantially from 84 to 19. We replicated five associations for triglycerides influencing DNAm sites near *CPTA1* and *ABCG1*[56] and found associations for transferrin saturation/iron influencing DNAm sites near *HFE.*

We next evaluated if there was evidence for small, widespread changes in DNAm levels in response to complex trait variation, by calculating the genomic control inflation factor (GC_{in}) for the p-values obtained from the MR analyses of each trait against all DNAm sites. Five traits (fasting glucose, age at menarche, cigarettes smoked per day, immunoglobulin G index levels, serum creatinine), showed GC_{in} values above 1.05 (**Extended data 8**). GC_{in} calculations were performed at each chromosome singly for each trait (**Supplementary Figure 34**) and in a leave-one-chromosome-out analysis (**Supplementary Figure 35**). The GC_{in} remained consistent (except for immunoglobulin G index levels), indicating that the traits have small but widespread influences on DNAm levels across the genome.

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

Discussion

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

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

Overall our data and results have resulted in the most comprehensive atlas of genetic effects to date. We expect that this atlas will be of use to the scientific community for studies of genome regulation, contribute to the control of confounding in EWAS and to perform causality analysis.

Acknowledgements

C.L.R., G.D.S., G.S., J.L.M., K.B., M.Su., T.G.R. and T.R.G. are supported by the UK Medical Research Council Integrative Epidemiology Unit at the University of Bristol (MC_UU_00011/1, MC_UU_00011/4, MC_UU_00011/5). C.L.R. receives support from a Cancer Research UK Programme Grant (C18281/A191169). G.H. is funded by the Wellcome Trust and the Royal Society (208806/Z/17/Z). E.H. and J.M. were supported by Medical Research Council (MRC) project grants (MR/K013807/1 and MR/R005176/1 awarded to J.M.) and a MRC Clinical Infrastructure award (MR/M008924/1 awarded to J.M.). The study was also supported by JPI HDHL funded DIMENSION project (administered by the BBSRC UK, BB/S020845/1 to J.T.B., and by ZonMW The Netherlands, grant 529051021 to B.T.H). B.T.H. is supported by the Netherlands CardioVascular Research Initiative (the Dutch Heart Foundation, Dutch Federation of University Medical Centres, the Netherlands Organisation for Health Research and Development, and the Royal Netherlands Academy of Sciences) for the GENIUS project Generating the best evidence-based pharmaceutical targets for atherosclerosis (CVON2011-19, CVON2017-20). A.D.B. has been supported by a Wellcome Trust PhD Training Fellowship for Clinicians, the Edinburgh Clinical Academic Track (ECAT) programme (204979/Z/16/Z). J.Kl. was supported by a DOC fellowship of the Austrian Academy of Sciences. Cohort specific acknowledgements and funding are presented in **Supplementary Information**.

Author Contributions

Project management: G.H., G.S., J.L.M

Designed individual studies and contributed data:
A.A.C., A.Cas., A.D.H., A.G.U, A.Me., A.Mu., A.M.M., B.B., B.T.H.,C.H., C.L.R., C.P., C.Sa., C.Sh., C.Sö., D.A.L., D.v.H., D.I.B., D.T., E.A.N., E.B.B., E.J.C.d.G, E.M., F.G., F.R., G.E.D, G.H.K., G.P., G.W.M., H.R.E., H.T., H.Z., I.J.D., J.F.F., J.H.V., J.J.C., J.Ka., J.L., J.M., J.M.S., J.M.V., J.v.M., J.R., J.R.B.P., J.R.G., J.Sh., J.T.B., J.W., J.W.H., K.K.O., K.L.E., K.R., L.A., L.C.S., L.M., M.A.I., M.Bee., M.Bu., M.E.A.R., M.H.v.IJ., M.Ke., M.O., N.C., N.G.M., N.J.W., N.R.W., P.E.S., P.Mo., P.M.V., R.H., R.P., S.L., S.P., T.D.S., T.E., T.E.M., T.I.A.S, T.P., T.T., V.W.V.J., W.K., Z.P.

Generated and/or quality-controlled data: A.A.K., A.I., A.S., C.S.M., H.R.E., J.L.M., K.B., K.M.H., N.K., S.M.R., T.H., R.M.W., W.L.M.

Designed new statistical or bioinformatics tools: G.H., J.L.M., M.Su., T.R.G., V.I.

Analysed the data and/or provided critical interpretation of results:
A.D.B., A.Car., A.D., A.F.M., A.K., B.T.H., C.B., C.H., C.L.R., C.R.A., C.Sor., C.V., C.X., C.W., D.A., D.C., D.J.L., D.L.C., D.M., E.C.M., E.G., E.H., E.M., F.C.M., F.I.R., F.R.D., G.B., G.C., G.D.S., G.H., G.H.K., G.M., G.W., I.Y., J.C.F., J.v.D., J.J.H., J.Ka., J.Kl., J.L.M., J.M., J.Su., J.T.B., K.B., K.v.E., K.F.D., K.S., L.C.S., M.Ber., M.Bu., M.H.v.IJ., M.G., M.Ku., M.L., M.Sm., M.Su., N.K., P.Me., P.Ma., P.M.V., R.E.M., R.G., R.L., R.Z., S.B., S.G., S.K., T.C., T.G., T.G.R., T.I.A.S., T.L., T.R.G., Y.A., Y.Z., V.I., V.S.

Designed and/or managed the study: B.T.H., C.B., C.L.R., J.M., J.T.B., T.R.G.
Wrote the manuscript: A.D.B., B.T.H., C.B., C.L.R., D.J.L., E.C.M, E.H., G.D.S., G.H., J.C.F., J.Kl., J.L.M., J.M., J.T.B., K.B., K.F.D., M.Su., P.M.V., R.L., T.G.R., T.R.G., V.I.

Competing interests

The authors declare no competing interests.

Financial disclosures

T.R.G receives funding from GlaxoSmithKline and Biogen for unrelated research.

References

1. Petronis, A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* **465**, 721-7 (2010).
2. van Dongen, J. *et al.* Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat Commun* **7**, 11115 (2016).
3. Hannon, E. *et al.* Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS Genet* **14**, e1007544 (2018).
4. Kerkel, K. *et al.* Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat Genet* **40**, 904-8 (2008).
5. Schadt, E.E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297-302 (2003).
6. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* **23**, R89-98 (2014).
7. Gaunt, T.R. *et al.* Systematic identification of genetic influences on methylation across the human life course. *Genome Biol* **17**, 61 (2016).
8. Bonder, M.J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet* **49**, 131-138 (2017).
9. Hannon, E. *et al.* Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat Neurosci* **19**, 48-54 (2016).
10. Hop, P.J. *et al.* Genome-wide identification of genes regulating DNA methylation using genetic anchors for causal inference. *Genome Biol* **21**, 220 (2020).
11. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
12. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* **17**, 208 (2016).
13. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288-95 (2011).
14. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75, S1-3 (2012).
15. Shah, S. *et al.* Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res* **24**, 1725-33 (2014).
16. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2**, e00523 (2013).
17. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).
18. McRae, A.F. *et al.* Identification of 55,000 Replicated DNA Methylation QTL. *Sci Rep* **8**, 17605 (2018).

19. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
20. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81-86 (2017).
21. Elliott, G. *et al.* Intermediate DNA methylation is a conserved signature of genome regulation. *Nat Commun* **6**, 6363 (2015).
22. Feldmann, A. *et al.* Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genet* **9**, e1003994 (2013).
23. Grundberg, E. *et al.* Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am J Hum Genet* **93**, 876-90 (2013).
24. Kim-Hellmuth, S. *et al.* Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**(2020).
25. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat Commun* **9**, 2282 (2018).
26. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**(2017).
27. Domcke, S. *et al.* Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575-9 (2015).
28. Baubec, T. *et al.* Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature* **520**, 243-7 (2015).
29. Ginno, P.A. *et al.* A genome-scale map of DNA methylation turnover identifies site-specific dependencies of DNMT and TET activity. *Nat Commun* **11**, 2680 (2020).
30. Sánchez-Castillo, M. *et al.* CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res* **43**, D1117-23 (2015).
31. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
32. Waszak, S.M. *et al.* Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell* **162**, 1039-50 (2015).
33. Viny, A.D. *et al.* Dose-dependent role of the cohesin complex in normal and malignant hematopoiesis. *J Exp Med* **212**, 1819-32 (2015).
34. Battle, A. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
35. Kumasaka, N., Knights, A.J. & Gaffney, D.J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat Genet* **51**, 128-137 (2019).
36. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).
37. Delaneau, O. *et al.* Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* **364**(2019).
38. Vosa, U. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRXiv* (2018).
39. Astle, W.J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).
40. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-5 (2012).
41. Tachmazidou, I. *et al.* Whole-Genome Sequencing Coupled to Imputation Discovers Genetic Signals for Anthropometric Traits. *Am J Hum Genet* **100**, 865-884 (2017).
42. Kato, N. *et al.* Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat Genet* **47**, 1282-1293 (2015).
43. Iotchkova, V. *et al.* GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat Genet* **51**, 343-353 (2019).

44. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).

45. Reinius, L.E. *et al.* Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One* **7**, e41361 (2012).

46. Houseman, E.A. *et al.* Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics* **9**, 365 (2008).

47. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet* **13**, e1007081 (2017).

48. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).

49. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-7 (2016).

50. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat Genet* (2020).

51. Richardson, T.G. *et al.* Systematic Mendelian randomization framework elucidates hundreds of CpG sites which may mediate the influence of genetic variants on disease. *Hum Mol Genet* **27**, 3293-3304 (2018).

52. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum Mol Genet* **27**, R195-R208 (2018).

53. Brion, M.J., Shakhbazov, K. & Visscher, P.M. Calculating statistical power in Mendelian randomization studies. *Int J Epidemiol* **42**, 1497-501 (2013).

54. Pierce, B.L. & Burgess, S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am J Epidemiol* **178**, 1177-84 (2013).

55. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**(2018).

56. Dekkers, K.F. *et al.* Blood lipids influence DNA methylation in circulating cells. *Genome Biol* **17**, 138 (2016).

57. Braun, K.V.E. *et al.* Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. *Clin Epigenetics* **9**, 15 (2017).

58. Simpson, J.T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**, 407-410 (2017).

Figure Legends

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

Figure 2: *Cis* and *trans* mQTL operate through distinct mechanisms
a. Distributions of enrichments for chromatin states and gene annotations among mQTL sites and SNPs. Enrichment analyses were performed using 25 combinatorial chromatin states from 127 cell types (including 27 blood cell types) and gene annotations. The heatmap represents the distribution of odds ratios for *cis only*, *trans only*, or *cis+trans* sites and SNPs. For the enrichment of chromatin states, ORs were averaged across cell types. Significance has been categorised as: *=FDR<0.001;**=FDR<1e-10;***=FDR<1e-50 **b. Distributions of enrichment for occupancy of TFBS among mQTL sites and SNPs.** Each density curve represents the distribution of odds ratios for *cis only*, *trans only*, or *cis+trans* sites (left) and SNPs (right). **c. Distributions of enrichment of mQTL among 41 complex traits and diseases.** Each density curve represents the distribution of odds ratios for *cis only*, *trans only*, or *cis+trans* SNPs.

Figure 3: Communities constructed from *trans*-mQTL. a. A network depicting all communities in which there were twenty or more sites. Random walks were used to generate communities (colours), so occasionally a DNA site connects different communities. **b. The relationship between genomic annotations, mQTL and communities**. Communities 9 and 22 are comprised of DNAm sites that are related

through shared genetic factors. The sankey plots show the genomic annotations for the genetic variants (left) and for the DNAm sites (right). The DNAm sites comprising these communities are enriched for TFBS related to the cohesin complex and NFkB, respectively. **c. Enrichment of GWA traits among community SNPs**. The genomic loci for each of the 56 largest communities were tested for enrichment of low p-values in 133 complex trait GWAS (y-axis) against a null background of community SNPs. The x-axis depicts the two-sided -log10 p-value for enrichment, with the 5% FDR shown by the vertical dotted line. Colours represent log odds ratios. Enrichments were particularly strong for blood related phenotypes (including circulating metal levels).

Figure 4: Identifying putative causal relationships between sites and traits using bi-directional MR. Aggregated results from a systematic bi-directional MR analysis between DNAm sites and 116 complex traits. The y-axis represents the two-sided p-value from MR analysis. The top plot depicts results from tests of DNAm sites colocalising with complex traits. The light grey points represent MR estimates that either did not surpass multiple testing, or shared small p-values at both the DNAm site and complex trait but had weak evidence of colocalisation. Bold, coloured points are those that showed strong evidence for colocalisation (H4 > 0.8). The bottom plot shows the two-sided -log10 p-values from MR analysis of risk factor or genetic liability of disease on DNAm levels. Extensive follow up was performed on DNAm site-trait pairs with putative associations, and those that pass filters are plotted in bold and colored according to the trait category. A substantial number of MR results in both directions exhibited very strong effects but failed to withstand sensitivity analyses.

Online Methods

Study design overview

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

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

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

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

Data preparation

Participants

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

The Genetics of DNA Methylation Consortium (GoDMC) pipeline

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

Genotype data

Each study performed quality control on genotype data for all autosomes and chromosome X (if available) and imputed to 1000G phase 1 or above using hg19/build37. Dosages were converted to bestguess data without a probability cut-off. SNPs that failed Hardy Weinberg equilibrium (p<1e-6), had a MAF <0.01, an info score <0.8 or missingness in more than 5% of the participants were removed. We recoded SNPs to CHR:POS[11] format and removed duplicate SNPs. We then harmonized the recoded SNPs to the 1000G reference using easyQC_v9.2[60]. This harmonization script removed SNPs with mismatched alleles and recoded INDEL alleles to I and D.

We performed a gender check to remove participants with discordant gender to the covariate file. We extracted and pruned a set of common HapMap3 SNPs (MAF>0.2, without long-range LD regions before we calculated the first 20 genetic principal components (PCs) on LD pruned SNPs and excluding regions of high LD from the analysis. We used PLINK.2.0[61] for unrelated participants and GENESIS[62] for related participants to identify ethnic outliers. Ethnic outliers that deviated 7 SDs from the mean were removed. After outlier removal we recalculated genetic PCs for use in subsequent analyses. To identify relatedness in unrelated datasets, we pruned the genotype data to a set of independent HapMap 3 SNPs with MAF>0.01 and calculated genome-wide average identity by state (IBS) using PLINK2.0. Participants with IBS > 0.125 were removed.

DNAm data normalisation and quality control

DNAm was measured in whole blood or cord blood using HumanMethylation450 or EPIC arrays in at least 100 European individuals. Each study performed normalization and quality control on the DNAm data independently, with most studies using functional normalisation through the R package meffil v0.1.0[63] (see **Supplementary Table 1**). Briefly, meffil has been designed to preprocess raw idat files to a normalization matrix for large sample sizes without large computational memory requirements and to perform quality control in an automated way where the analyst can adjust default parameters easily. Sample quality control included removal of participants where more than 10% of the DNAm sites failed the detection p-value of 0.1 and/or threshold of 3 beads. In addition, mismatched samples were identified by comparing the 65 SNPs on the DNAm array to the genotype array and a gender check. Additional DNAm quality was checked by the methylated versus unmethylated ratio, dye bias using the normalisation control probes and bisulphate control probes. Protocols can be found here: https://github.com/perishky/meffil/wiki. For each DNAm site, we replaced outliers that were 10 SDs from the mean (3 iterations) with the DNAm site mean.

Covariates

We used sex, age at measurement, batch variables (slide, plate, row if available), smoking (if available) and recorded cell counts to adjust for possible confounding and to reduce residual variation. Additional confounders (genetic PCs, non-genetic DNAm PCs, and where necessary predicted smoking and cell counts) were calculated using the GoDMC pipeline. After quality control and normalization of the DNAm data, we predicted smoking status by using previously reported DNAm associations with smoking[64]. In addition, we predicted cell counts using the Houseman algorithm[46] implemented in meffil v0.1.0[63]. We performed a PC analysis on the 20,000 most variable autosomal DNAm sites and kept all PCs that cumulatively explained 80% of the variance. We performed a genome wide association analysis on the DNAm PCs and retained the PCs that were not associated with a genotype ($p > 1e-7$). We kept a maximum of 20 non-genetic PCs for subsequent adjustment.

DNAm data adjustment

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

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

Positive and negative controls

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

Association analyses

Phase 1: creating the candidate list of associations

We performed a fast, comprehensive analysis of all *cis*- and *trans*-associations on 420,509 reliable[59] residualised DNAm sites separately in 22 studies (N=16,907) using the R package Matrix eQTL v2.1.0[69]. For each DNAm site j the residual value y_{ji} was regressed against each SNP k

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

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

Phase 2: obtaining summary data from all studies for meta-analysis

This candidate list was sent to 36 studies (N=27,750) where effect sizes for all putative associations were recalculated by fitting linear models. For putative *cis*-mQTL we performed linear regression as in phase 1. To improve statistical power to estimate the *trans*-mQTL effects we recorded the top *cis* SNP x_c, for each DNAm site (based on lowest p-value within that study) and fit this as a covariate in the *trans*-mQTL regressions

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

Evaluation of DNAm data adjustment

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

Impact of two-stage design on power of study

Though the multi-stage study design was performed out of practical necessity, we evaluated the impact it had on statistical power in comparison to the hypothetical situation of analysing all the data together in a standard one stage mQTL design.
For *cis* mQTL associations we calculated the power of detecting an association in at least one of 22 studies at p < 1e-5. To do this we calculate what is the probability of missing an association as being the product of the probability of missing it in study 1 AND in study 2 AND in study 3 etc.

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

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

For *trans* mQTL associations we calculated the power to detect an association in at least two of 22 studies at p< 1e-5. We calculate what is the probability of missing an association as being the product of the probability of missing it in both study 1 and study 2 AND in study 1 and study 3 AND in study 1 and study 4 etc.

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

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

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

Meta-analyses

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

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

To explore heterogeneity further, we meta-analysed our SNP-DNAm pairs using FE, RE and MRE models and found that associations that were dropped in MRE analyses showed higher I^2 and tau^2 and smaller effect sizes and DNAm site SDs (**Supplementary Figures 3-4**).

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

Further downstream analyses have been described in **Supplementary Information**.

Data Availability

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

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

Code Availability

Datasets were processed using https://github.com/perishky/meffil unless stated otherwise. Individual study analysts used a github pipeline https://github.com/MRCIEU/godmc to conduct the mQTL analysis. We used https://github.com/MRCIEU/godmc_phase1_analysis for the phase1 analysis, https://github.com/explodecomputer/random-metal for the meta analyses and https://github.com/MRCIEU/godmc_phase2_analysis for the follow-up analyses.

BIOS consortium

Marian Beekman[4], Dorret I Boomsma[43], Jenny van Dongen[43], Diana van Heemst[85], Bastiaan T Heijmans[4], Jouke-Jan Hottenga[43], René Luijk[4], Joyce van Meurs[29], P Eline Slagboom[4], André G Uitterlinden[29], Jan H Veldink[42]

References

11. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

46. Houseman, E.A. *et al.* Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics* **9**, 365 (2008).

59. Zhou, W., Laird, P.W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res* **45**, e22 (2017).

60. Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**, 1192-212 (2014).

61. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

62. Conomos, M.P., Reiner, A.P., Weir, B.S. & Thornton, T.A. Model-free Estimation of Recent Genetic Relatedness. *Am J Hum Genet* **98**, 127-48 (2016).

63. Min, J.L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* **34**, 3983-3989 (2018).

64. Zeilinger, S. *et al.* Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS One* **8**, e63812 (2013).

65. Aulchenko, Y.S., de Koning, D.J. & Haley, C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**, 577-85 (2007).

66. Chen, Y.A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203-9 (2013).

67. Naeem, H. *et al.* Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics* **15**, 51 (2014).

68. Price, M.E. *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* **6**, 4 (2013).

69. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-8 (2012).

70. Dahl, A., Guillemot, V., Mefford, J., Aschard, H. & Zaitlen, N. Adjusting for Principal Components of Molecular Phenotypes Induces Replicating False Positives. *Genetics* **211**, 1179-1189 (2019).

71. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).

72. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control Clin Trials* **7**, 177-88 (1986).

73. Hedges, L.V. & Olkin, I. CHAPTER 9 - Random Effects Models for Effect Sizes. in *Statistical Methods for Meta-Analysis* 189-203 (Academic Press, San Diego, 1985).