Query Details

Back to Main Page

1. In the sentence beginning 'The meQTL are enriched for functionally relevant characteristics', please provide a definition for Hi-C.

2. In the author list, there are two authors with the name Panos Deloukas with different affiliations. Please confirm whether this is the same individual and whether the two authors can be merged.

3. Please check your article carefully, coordinate with any co-authors and enter all final edits clearly in the eproof, remembering to save frequently. Once corrections are submitted, we cannot routinely make further changes to the article.

4. Note that the eproof should be amended in only one browser window at any one time; otherwise changes will be overwritten.

5. Author surnames have been highlighted. Please check these carefully and adjust if the first name or surname is marked up incorrectly. Note that changes here will affect indexing of your article in public repositories such as PubMed. Also, carefully check the spelling and numbering of all author names and affiliations, and the corresponding email address(es).

6. You cannot alter accepted Supplementary Information files except for critical changes to scientific content. If you do resupply any files, please also provide a brief (but complete) list of changes. If these are not considered scientific changes, any altered Supplementary files will not be used, only the originally accepted version will be published.

7. If applicable, please ensure that any accession codes and datasets whose DOIs or other identifiers are mentioned in the paper are scheduled for public release as soon as possible, we recommend within a few days of submitting your proof, and update the database record with publication details from this article once available.

8. Your paper has been copy edited. Please review every sentence to ensure that it conveys your intended meaning; if changes are required, please provide further clarification rather than reverting to the original text. Please note that formatting (including hyphenation, Latin words,and any reference citations that might be mistaken for exponents) has been made consistent with our house style.

9. Please confirm or correct the city name inserted in affiliations 5, 10, 12, 13, 21, 22, 39 and 44.

10. In the sentence beginning 'Our meQTL replicate in data generated by the Illumina', please provide a definition for EPIC and confirm whether the definition provided for MeDIP-seq is correct.

11. In the legend for Fig. 2g, please provide a definition for SAT.

12. In the sentence beginning 'We used summary data-based Mendelian', please confirm whether it is correct to say '0.05 ÷ <Emphasis Type="Italic">n</Emphasis> tests'.

13. In the sentence beginning 'We then tested the 4,811 eQTL genes', please provide a definition for coloc PP4.

14. Please note, we reserve ‘significant' and its derivatives for statistical significance. This has been reworded where this was not the intended meaning (for example to important, notable, substantial) in the sentence beginning 'We used our meQTL as genetic instruments'. Please check and confirm.

15. In the sentence beginning 'At four of the 45 genetic loci with trans CpG', please include a definition for FET.

16. In the sentence beginning 'The <Emphasis Type="Italic">trans</Emphasis> CpG sites localize', please include the official gene symbol for <Emphasis Type="Italic">IKBE</Emphasis>.

17. In the legend for Fig. 4d, please include the official gene symbols for <Emphasis Type="Italic">ELE1</Emphasis> and <Emphasis Type="Italic">HKR1</Emphasis>.

18. In the legend for Fig. 4a, please indicate what is designated by the asterisk *.

19. In the sentence beginning 'To further test the hypothesis that ZNF333', please confirm whether it is correct to say 'FLAG- and/or Myc-tagged'.

20. In the legend for Fig. 6a, please provide definitions for KORA and LOLIPOP.

21. For Fig. 6f, please confirm whether the position is in units of bp.

22. The sentences beginning ‘As proof of principle’ and 'Our study provides new insights into the genetic regulation' have been edited to remove the priority claim, according to style. Please check and confirm.

23. In the sentence beginning 'We carried out similar analyses using', please confirm whether MethylEpic can be changed to MethylationEPIC.

24. For the sentence beginning 'Finally, we used the topology of the locus graph', please include complete reference details for 111 and 112.

25. In the sentence beginning 'We defined the symmetric transition matrix', please confirm whether the equation is represented correctly.

26. In the sentence beginning 'The final ranking of candidate gene' and the following sentence, please confirm whether the equations are represented correctly.

27. In the sentence beginning 'To avoid confounding by <Emphasis Type="Italic">cis</Emphasis> effects', please confirm whether the equation is represented correctly.

28. In the sentence beginning 'For the NFBC studies', please provide a definition for NFBC and clarify which author is meant by M.W. by including the full last name.

29. Please check that all funders have been appropriately acknowledged and that all grant numbers are correct.

30. In the Author contributions, two individuals appear who are not in the author list: Eva Reischl and James Scott. Please clarify.

31. In the sentence beginning 'The web links for the publicly available datasets', please include full reference details for Bonder et al. 2015, as no study by Bonder et al. in 2015 appears in the reference list.

32. Please check that the Competing Interests declaration is correct as stated. If you declare competing interests, please check the full text of the declaration for accuracy and completeness.

33. Please confirm or correct details for ref. 82 (Hawe et al., Zenodo).

Article

Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function

Johann S. Hawe Affiliationids : Aff1 Aff2
Rory Wilson Affiliationids : Aff3 Aff4
Katharina Schmid Affiliationids : Aff1 Aff2
Li Zhou Affiliationids : Aff5
Lakshmi Narayanan Lakshmanan Affiliationids : Aff5
Benjamin C. Lehne Affiliationids : Aff6
Brigitte Kühnel Affiliationids : Aff3 Aff4
William R. Scott Affiliationids : Aff6
Matthias Wielscher Affiliationids : Aff6
Yik Weng Yew Affiliationids : Aff5

Dominic P. Lee Affiliationids : Aff7
Eirini Marouli Affiliationids : Aff8 Aff9
Manon Bernard Affiliationids : Aff10 Aff11
Liliane Pfeiffer Affiliationids : Aff3 Aff4
Pamela R. Matías-García Affiliationids : Aff3 Aff4
Matias I. Autio Affiliationids : Aff7 Aff12
Stephane Bourgeois Affiliationids : Aff9
Christian Herder Affiliationids : Aff13 Aff14 Aff15
Ville Karhunen Affiliationids : Aff6 Aff16
Thomas Meitinger Affiliationids : Aff17 Aff18
Holger Prokisch Affiliationids : Aff18 Aff19
Wolfgang Rathmann Affiliationids : Aff13 Aff20
Michael Roden Affiliationids : Aff13 Aff14 Aff15
Sylvain Sebert Affiliationids : Aff16 Aff21 Aff22
Jean Shin Affiliationids : Aff10 Aff11
Konstantin Strauch Affiliationids : Aff23 Aff24 Aff25
Weihua Zhang Affiliationids : Aff6 Aff26
Wilson L. W. Tan Affiliationids : Aff7
Stefanie M. Hauck Affiliationids : Aff27
Juliane Merl-Pham Affiliationids : Aff27
Harald Grallert Affiliationids : Aff3 Aff4 Aff28
Eudes G. V. Barbosa Affiliationids : Aff1

MuTHER Consortium

Kourosh R. Ahmadi Affiliationids : Aff35
Chrysanthi Ainali Affiliationids : Aff36
Amy Barrett Affiliationids : Aff37
Veronique Bataille Affiliationids : Aff35
Jordana T. Bell Affiliationids : Aff35
Alfonso Buil Affiliationids : Aff38
Panos Deloukas Affiliationids : Aff39
Emmanouil T. Dermitzakis Affiliationids : Aff38
Antigone S. Dimas Affiliationids : Aff38
Richard Durbin Affiliationids : Aff40
Daniel Glass Affiliationids : Aff35
Elin Grundberg Affiliationids : Aff41
Neelam Hassanali Affiliationids : Aff37
Åsa K. Hedman Affiliationids : Aff42
Catherine Ingle Affiliationids : Aff40
David Knowles Affiliationids : Aff43
Maria Krestyaninova Affiliationids : Aff44
Cecilia M. Lindgren Affiliationids : Aff42
Christopher E. Lowe Affiliationids : Aff45 Aff46
Mark I. McCarthy Affiliationids : Aff37 Aff42
Eshwar Meduri Affiliationids : Aff40
Paola di Meglio Affiliationids : Aff47
Josine L. Min Affiliationids : Aff39
Stephen B. Montgomery Affiliationids : Aff38
Frank O. Nestle Affiliationids : Aff47
Alexandra C. Nica Affiliationids : Aff38
James Nisbet Affiliationids : Aff40
Stephen O'Rahilly Affiliationids : Aff45 Aff46

Simon Potter Affiliationids : Aff40

Johanna Sandling Affiliationids : Aff40

Magdalena Sekowska Affiliationids : Aff40

So-Youn Shin Affiliationids : Aff40

Kerrin S. Small Affiliationids : Aff35

Nicole Soranzo Affiliationids : Aff40

Tim D. Spector Affiliationids : Aff35

Gabriela Surdulescu Affiliationids : Aff35

Mary E. Travers Affiliationids : Aff37

Loukia Tsaprouni Affiliationids : Aff40

Sophia Tsoka Affiliationids : Aff36

Alicja Wilk Affiliationids : Aff40

Tsun-Po Yang Affiliationids : Aff40

Krina T. Zondervan Affiliationids : Aff42

Thomas Illig Affiliationids : Aff29 Aff30

Annette Peters Affiliationids : Aff3 Aff4 Aff31

Tomas Paus Affiliationids : Aff32

Zdenka Pausova Affiliationids : Aff10 Aff11

Panos Deloukas Affiliationids : Aff8 Aff9

Roger S. Y. Foo Affiliationids : Aff7 Aff12

Marjo-Riitta Jarvelin Affiliationids : Aff6 Aff16 Aff21 Aff33

Jaspal S. Kooner ✉

Email : j.kooner@ic.ac.uk

Affiliationids : Aff34, Correspondingaffiliationid : Aff34

Marie Loh ✉

Email : marie.loh@ntu.edu.sg

Affiliationids : Aff5 Aff6, Correspondingaffiliationid : Aff5

Matthias Heinig ✉

Email : matthias.heinig@helmholtz-muenchen.de

Affiliationids : Aff1 Aff2, Correspondingaffiliationid : Aff1

Christian Gieger ✉

Email : christian.gieger@helmholtz-muenchen.de

Affiliationids : Aff3 Aff4, Correspondingaffiliationid : Aff3

Melanie Waldenberger ✉

Email : waldenberger@helmholtz-muenchen.de

Affiliationids : Aff3 Aff4 Aff31, Correspondingaffiliationid : Aff3

John C. Chambers ✉

Email : john.chambers@ic.ac.uk

Affiliationids : Aff5 Aff6, Correspondingaffiliationid : Aff5

Aff1 Institute of Computational Biology, Deutsches Forschungszentrum für Gesundheit und Umwelt, Helmholtz Zentrum München, Neuherberg, Germany

Aff2 Department of Informatics, Technical University of Munich, Garching bei München, Germany

Aff3 Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

Aff4 Research Unit Molecular Epidemiology, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, Germany

Aff5 Lee Kong Chian School of Medicine, Singapore, Singapore

Aff6 Department of Epidemiology and Biostatistics, Imperial College London, London, UK

Aff7 Genome Institute of Singapore, Singapore, Singapore

Aff8 Centre for Genomic Health, Queen Mary University of London, London, UK

Aff9 William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK

Aff10 Departments of Physiology and Nutritional Sciences, University of Toronto, Toronto, Ontario, Canada

Aff11 The Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada

Aff12 Cardiovascular Research Institute, National University Health Systems, National University of Singapore, Singapore, Singapore

Aff13 German Center for Diabetes Research (DZD), partner site Düsseldorf, Düsseldorf, Germany

Aff14 Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University, Düsseldorf, Germany

Aff15 Division of Endocrinology and Diabetology, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany

Aff16 Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu, Finland

Aff17 Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

Aff18 Institute of Human Genetics, Technical University Munich, Munich, Germany

Aff19 Institute of Neurogenomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

Aff20 Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Aff21 Biocenter Oulu, University of Oulu, Oulu, Finland

Aff22 Department for Genomics of Common Diseases, School of Public Health, Imperial College London, London, UK

Aff23 Chair of Genetic Epidemiology, IBE, Faculty of Medicine, LMU Munich, Munich, Germany

Aff24 Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

Aff25 Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Mainz, Germany

Aff26 Department of Cardiology, Ealing Hospital, London North West Healthcare NHS Trust, Southall, UK

Aff27 Research Unit Protein Science, Helmholtz Zentrum München, German Research Centre for Environmental Health, Munich, Germany

Aff28 German Center for Diabetes Research (DZD), Munich-Neuherberg, Germany

Aff29 Hannover Unified Biobank, Hannover Medical School, Hannover, Germany

Aff30 Institute for Human Genetics, Hannover Medical School, Hannover, Germany

Aff31 German Research Center for Cardiovascular Disease (DZHK), partner site Munich Heart Alliance, Hannover, Germany

Aff32 Centre Hospitalier Universitaire Sainte-Justine, University of Montreal, Montreal, Canada

Aff33 Unit of Primary Care, Oulu University Hospital, Oulu, Finland

Aff34 National Heart and Lung Institute, Imperial College London, London, UK

Aff35 Department of Twin Research and Genetic Epidemiology, King’s College London, London, UK

Aff36 Department of Informatics, School of Natural and Mathematical Sciences, King’s College London, Strand, London, UK

Aff37 Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Churchill Hospital, Headington, Oxford, UK

Aff38 Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland

Aff39 William Harvey Research Institute, Queen Mary University of London, London, UK

Aff40 Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Aff41 Children’s Mercy Hospitals and Clinics, Kansas City, MO, USA

Aff42 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

Aff43 University of Cambridge, Cambridge, UK

Aff44 European Bioinformatics Institute, Hinxton, UK

Aff45 University of Cambridge Metabolic Research Labs, Institute of Metabolic Science, Addenbrooke’s Hospital Cambridge, Cambridge, UK

Aff46 Cambridge NIHR Biomedical Research Centre, Addenbrooke's Hospital, Cambridge, UK

Aff47 St. John's Institute of Dermatology, King's College London, London, UK

Received: 22 August 2019 / Accepted: 18 October 2021

Abstract

We determined the relationships between DNA sequence variation and DNA methylation using blood samples from 3,799 Europeans and 3,195 South Asians. We identify 11,165,559 SNP–CpG associations (methylation quantitative trait loci (meQTL), $P < 10^{-14}$), including 467,915 meQTL that operate in *trans*. The meQTL are enriched for functionally relevant characteristics, including shared chromatin state, Hi-C interaction and association with gene expression and metabolic and clinical traits. We use molecular interaction and colocalization analyses to identify multiple nuclear regulatory pathways linking meQTL loci to phenotypic variation, including *UBASH3B* (body mass index), *NFKBIE* (rheumatoid arthritis), *MGA* (blood pressure) and *COMMD7* (white cell counts). For rs6511961, chromatin immunoprecipitation followed by sequencing (ChIP–seq) validates zinc finger protein (ZNF)333 as the likely *trans* acting effector protein. Finally, we used interaction analyses to identify population- and lineage-specific meQTL, including rs174548 in *FADS1*, with the strongest effect in $CD8^+$ T cells, thus linking fatty acid metabolism with immune dysregulation and asthma. Our study advances understanding of the potential pathways linking genetic variation to human phenotype.

Editor's Summary

Genome-wide association analyses of DNA methylation in peripheral blood from 3,799 Europeans and 3,195 South Asians identify unique SNP–CpG associations (meQTL), providing insights into molecular mechanisms and the potential links to phenotypic variation.

These authors contributed equally: Johann S. Hawe, Rory Wilson, Katharina Schmid.

These authors jointly supervised this work: Jaspal S. Kooner, Marie Loh, Matthias Heinig, Christian Gieger, Melanie Waldenberger, John C. Chambers.

A list of authors and their affiliations appears at the end of the paper.

Main

Methylation of DNA plays a key role in determining genomic structure and function, including regulation of cellular differentiation and coordination of gene expression[1, 2, 3, 4]. Disturbances in DNA methylation have AQ1 been implicated in the development of atherosclerosis, cancer, obesity, type 2 diabetes and neuropsychiatric illness and other AQ2 complex multifactorial diseases and predict all-cause mortality[5, 6, 7, 8, 9, 10, 11, 12]. Improved understanding of the mechanisms influencing DNA methylation is therefore anticipated to provide new insights into the biological pathways that determine genome regulation, molecular phenotypes and development of disease. AQ3 AQ4 AQ5 AQ6 AQ7 AQ8

DNA methylation is strongly influenced by underlying genetic variation, both in *cis* (same chromosome) and in *trans* (across chromosomes)[9, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. Genetic variants that influence DNA methylation in *trans* are of particular interest and identify nuclear regulatory pathways that play a critical role in the coordination of genomic function and impact multiple biological processes[14, 18, 19, 20, 23]. We AQ9 aimed to build on this previous work and to advance understanding of the molecular mechanisms linking regulatory genetic variation to gene expression, molecular interactions, phenotypic variation and disease susceptibility.

Results

Genome-wide association and replication testing

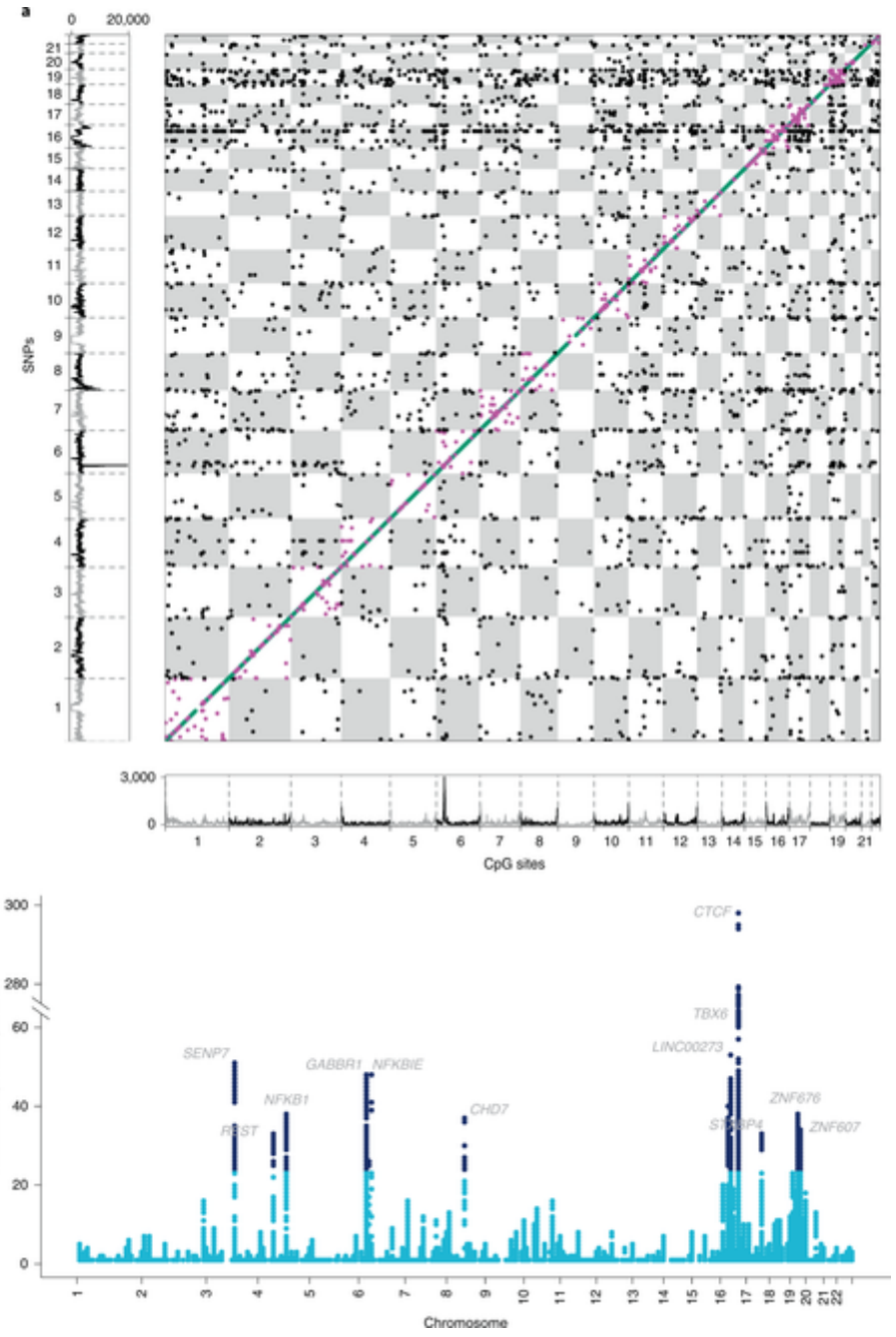
Our study design is summarized in Extended Data Fig. 1. We first carried out a genome-wide association study (GWAS) of DNA methylation in peripheral blood, with replication testing, of 3,799 Europeans ($n = 1{,}731$, discovery; $n = 2{,}068$, replication) and 3,195 South Asians ($n = 1{,}841$, discovery; $n = 1{,}354$, replication). DNA methylation was quantified using the Illumina Infinium HumanMethylation450 BeadChip. Genome-wide association was performed on Europeans and South Asians separately[24]. meQTL reaching genome-wide significance ($P < 10^{-14}$) were selected for replication testing. This stringent statistical threshold for genome-wide significance provides complete Bonferroni correction for the ~4.3 trillion statistical tests carried out and reduces the risk of false positive results (Methods). Replication testing was first performed using an ancestry-specific approach; this was followed by a final trans-ancestry analysis (Extended Data Fig. 1). At each stage of replication, we required that meQTL reach (1) $P < 0.05$ with consistent direction of effect and (2) $P < 10^{-14}$ in the combined analysis of discovery and replication results. The meQTL identified by genome-wide association showed a high rate of replication (>90%) in both ancestry-specific and cross-ancestry replication testing (Supplementary Table 1 and Extended Data Fig. 2). Replication rates were comparable or higher than those for meQTL reported in published studies (Supplementary Tables 2 and 3). AQ10 Our meQTL replicate in data generated by the Illumina MethylationEPIC array (>96% at $P <$ 0.05 with the same direction of effect, $n = 1{,}848$ samples, Methods) and by methylated DNA immunoprecipitation coupled with next-generation sequencing (MeDIP–seq) peripheral blood DNA methylomes (47% of testable meQTL at $P < 0.05$; Supplementary Table 4 and Extended Data Fig. 2)[25], demonstrating that our findings are generalizable across platforms

The output is a high-confidence cosmopolitan set of 11,165,559 meQTL (comprising 2,709,428 SNPs and 70,709 CpG sites) that are experimentally stringent and highly reproducible and operate across human populations (Fig. 1 and Supplementary Table 5). The median effect size for the 11.2 million meQTL was 2.0% (interquartile range, 1.2–3.5%) absolute change in methylation per allele copy. On average, the SNPs explain 10.3% (interquartile range, 4.4–11.5%) of variation in methylation at the respective CpG sites (Supplementary Table 6 and Extended Data Fig. 3).

Fig. 1

Summary of results for genome-wide association and replication testing.

a, Chessboard plot. Each dot represents a unique SNP–CpG pair reaching genome-wide significance in discovery ($P < 10^{-14}$) and showing both ancestry-specific and cross-ancestry replication. CpG position and background CpG density (450K array) are annotated on the *x* axis, and SNP position and background SNP density are annotated on the *y* axis. SNP–CpG pairs are color coded according to proximity of the SNP and the CpG site: *cis*, within 1 Mb ($n = 10{,}346{,}172$, green markers appearing as a diagonal line); long-range *cis*, distance >1 Mb but on the same chromosome ($n = 351{,}472$, purple markers); *trans*, SNP and CpG sites are on different chromosomes ($n = 467{,}915$, black markers). **b**, Manhattan plot of *trans* acting SNP–CpG associations. Each marker represents the number of CpG sites associated in *trans* with the identified *trans* acting SNPs. Results are for the cosmopolitan set of SNP–CpG pairs showing both ancestry-specific and cross-ancestry replication. SNPs with the highest number of CpG sites in *trans* (top 1%) are highlighted in dark blue, and the gene nearest the sentinel SNP is displayed.



The identified meQTL operate across diverse cell types

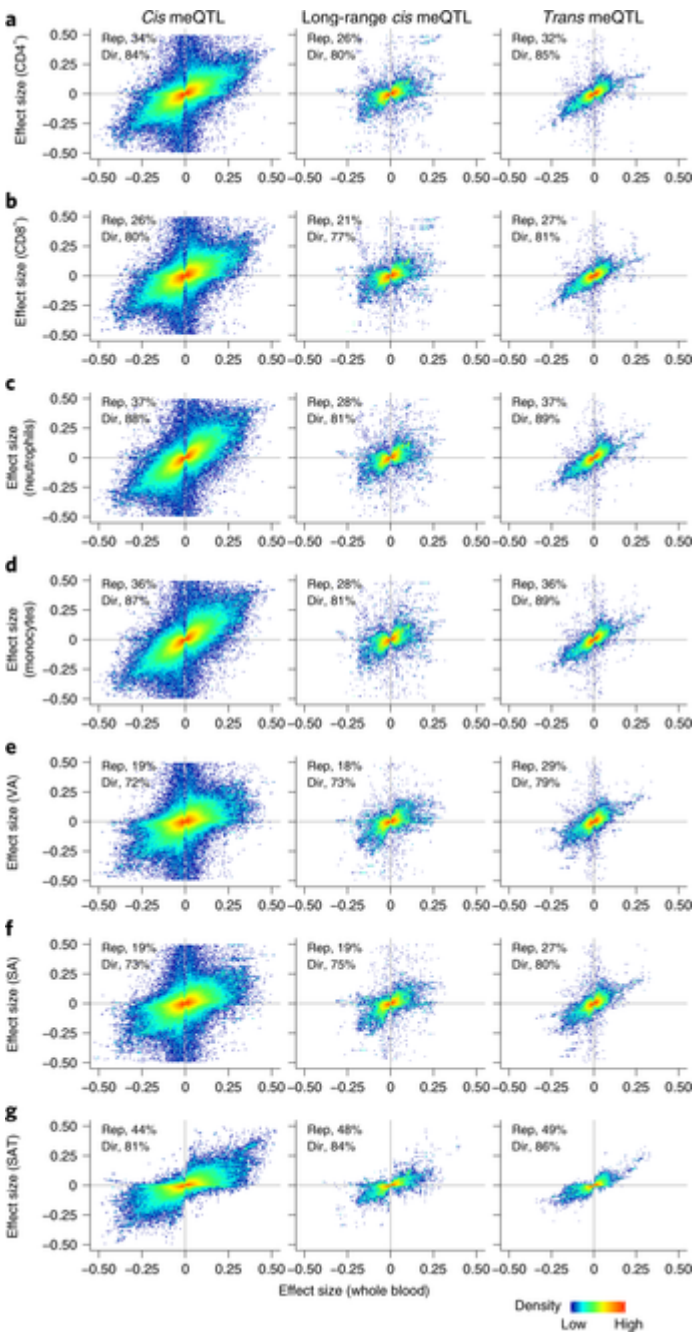
We show that 80–87% of the 11.2 million meQTL have a consistent direction of effect and 26–37% replicate at $P < 0.05$ in isolated white cell subsets ($n = 57$ samples; Fig. 2 and Supplementary Table 7). We also show that 72–86% of our meQTL have a consistent direction of effect in isolated adipocytes (subcutaneous and visceral, $n = 47$ samples) and in adipose tissue ($n = 603$ samples; $P < 1 \times 10^{-324}$ for each comparison, binomial test). A further 19.2% replicate in isolated visceral adipocytes, 19.4% replicate in subcutaneous adipocytes, and 44.2% replicate in subcutaneous adipose tissue ($P < 0.05$ and same direction of effect; Fig. 2 and Supplementary Table 7). These proportions are consistent with expectations based on sample size. Our results demonstrate that many of the meQTL operate across diverse cell lineages and are thus likely to be relevant to tissues and biological systems other than blood.

Fig. 2

Replication in isolated white cells, isolated adipocytes and adipose tissue.

Density plots summarizing replication of the SNP–CpG pairs identified by genome-wide association. **a**–**d**, Four isolated white cell subsets ($CD4^+$ lymphocytes, $CD8^+$ lymphocytes, neutrophils and monocytes). **e**,**f**, Isolated visceral and subcutaneous adipocytes (VA and SA). **g**, Whole adipose tissue. Results are presented as the effect size (change in AQ11 methylation on a scale of 0–1 in which 1 represents 100% methylation) per allele copy of the identified SNP in whole blood (*x* axis) and in the respective isolated cell type (*y* axis), stratified by SNP–

CpG proximity (*cis*, long-range *cis* and *trans* associations). Plotting area is limited to effect sizes between −0.5 and 0.5. Results show highly concordant effect sizes between whole blood and each cell type. The inset in each panel shows replication rates in the respective cell type (rep, $P < 0.05$ and same direction of effect) as well as the percent of directional consistency between effect sizes (dir).



Annotation of the meQTL identified

SNPs are enriched for association with DNA methylation on their *cis* chromosome, even beyond the conventional 1-Mb interval (Extended Data Fig. 4 and Methods). As underlying genomic mechanisms may differ according to proximity, we separated our findings into (1) *cis* meQTL (SNP–CpG distance <1 Mb, $n = 10{,}346{,}172$ pairs; 2,650,691 SNPs and 67,694 CpG sites), (2) long-range *cis* meQTL (>1 Mb apart but on the same chromosome, $n = 351{,}472$ pairs; 120,593 SNPs and 1,846 CpG sites) and (3) *trans* meQTL (associations between SNPs and CpG sites on different chromosomes, $n = 467{,}915$ pairs; 200,761 SNPs and 3,592 CpG sites). We used conditional analyses, correlation structure and genomic distance to estimate the total number of independent loci in our cosmopolitan SNP–CpG associations (Supplementary Fig. 1 and Methods). This identified 34,001 independent genetic loci associated with 46,664 independent methylation loci in *cis*, 467 independent genetic loci associated with 499 independent methylation loci in long-range *cis* and 1,847 independent genetic loci associated with 3,020 independent methylation loci in *trans*. For each of these, we selected a single sentinel SNP and a single CpG site (lowest *P* value in any pairwise association, Supplementary Tables 8 and 9 and Methods) to represent the individual loci in downstream analyses.

Functional genomic evaluation of the meQTL SNPs and CpG sites

Sentinel meQTL SNPs are enriched for location in multiple active chromatin regions, supporting a role in genome regulation (Extended Data Fig. 5)[14]. Expression array data for our cohort participants (Europeans, $n = 853$; South Asians, $n = 693$; Methods) identified 2,696 sentinel SNPs to be expression quantitative trait loci (eQTL; total eQTL pairs, 3,131; *cis*, 3,018; long-range *cis*, 50; *trans*, 63) at $P < 7.98 \times 10^{-11}$ ($P < 0.05$ after Bonferroni correction for all possible SNP–transcript tests, Supplementary Table 10) and showed that sentinel SNPs were enriched for eQTL both in *cis* and in *trans* (range, 4.1–22.1-fold compared to expectations under the null hypothesis, $P = 8.10 \times 10^{-18}$–2.45×10^{-66}; Extended Data Fig. 6). We separately showed that sentinel meQTL SNPs were strongly enriched for protein quantitative trait loci (QTL) (1.6–2.1-fold; $P < 0.001$) and metabolite QTL in *cis* (1.4-fold, $P < 0.001$) and for association with phenotypic traits and diseases (1.9–3.4-fold, $P < 0.001$). Results are summarized in Extended Data Fig. 7 and Supplementary Fig. 2.

Sentinel CpG sites influenced by genetic variants in *cis* are enriched in flanking regions of active transcription start sites and enhancers and depleted in heterochromatin regions, while SNP–CpG pairs in *trans* are additionally enriched at active transcription start sites (Extended Data Fig. 5)[14]. Using the extensive baseline phenotypic data for our participants, we show that meQTL CpG sites are enriched for association with metabolic, physiologic and clinical traits (252 of 277 available traits at $P < 1.8 \times 10^{-4}$ (Bonferroni correction for 277 tests) compared to expectations under the null hypothesis (median enrichment, 1.10; interquartile range, 1.06–1.15; Extended Data Fig. 7 and Supplementary Table 11). These findings support a potential role for the identified CpG sites (or their correlated

Next, we defined both the *cis* and *trans* relationships between DNA methylation and gene expression (expression quantitative trait methylation loci, eQTM) in our participants. Using similar analytic approaches as those in published studies initially suggested 90,666 putative *cis* eQTM in our dataset at $P = 8.7 \times 10^{-12}$ ($P < 0.05$ after Bonferroni correction for the number of possible CpG–expression pairs)[14]. However, this result appeared strongly confounded by variation in white cell composition, and adjustment for estimated cell type proportions reduced the number of *cis* eQTM identified to 769, of which 155 overlapped with our sentinel CpG sites. AQ12 We used summary data-based Mendelian randomization (SMR)[26] to further confirm this interpretation; putative *cis* eQTM identified with correction for white cell subsets were strongly replicated by SMR, while uncorrected eQTM were not (SMR $P < 0.05 \div n$ tests: 73% versus 17%, respectively, $P = 2.0 \times 10^{-29}$; Supplementary Table 12). In parallel, we identified 97,281 *trans* eQTM, of which 11,562 overlap one of our sentinel CpG sites; 627 of these *trans* eQTM are supported by SMR (Supplementary Table 12), a proportion consistent with the statistical power of our analysis (Supplementary Table 13). Finally, we show that sentinel CpG sites that are part of *cis* meQTL pairs are strongly enriched for being *cis* eQTM (that is, associated with gene expression in *cis*, Extended Data Fig. 6). Our results confirm the potential for white cell subset composition to confound analyses of gene expression in whole blood and provide experimental approaches for resolving potential biases.

Physical and regulatory interactions between meQTL SNPs and CpG sites

We tested whether *cis* meQTL might represent a direct effect of the sequence variant on the interaction between chromatin-associated factors and *cis* regulatory elements harboring the CpG site[27, 28]. Using data from the Roadmap Epigenomics Consortium, we showed that 88% of CpG sites with *cis* acting meQTL were associated with SNPs localizing to the same chromatin state (empirical $P = 9.9 \times 10^{-3}$; Extended Data Fig. 5). We similarly hypothesized that long-range *cis* meQTL might reflect physical interactions between distal enhancers and promoters[14, 29, 30]. In support of this, we showed that long-range *cis* associations occurred more frequently within topologically associated domains (15.5-fold, empirical $P < 0.01$; Extended Data Fig. 5) and more frequently had a Hi-C contact between SNP and CpG sites at promoter regions in 17 primary blood cell types characterized in the BLUEPRINT project[31] (2.5-fold, empirical $P < 0.01$; Extended Data Fig. 5). Annotating these associated pairs with chromHMM epigenetic states revealed 145 promoter–promoter, 178 enhancer–promoter and 49 enhancer–enhancer interactions. We demonstrated that the *trans* acting SNP–CpG pairs were also enriched for location in regions of chromosomal interaction in primary blood cells (3.7-fold, empirical $P = 6.6 \times 10^{-3}$; Extended Data Fig. 5) and in lymphoblastoid cell lines (1.8-fold, empirical $P < 0.01$; Supplementary Table 14)[31, 32]. In sum, these results indicate that genetic variants associate with methylation levels of CpG sites localized in the same or in physically interacting regulatory elements, consistent with a coordinated role in genomic regulation.

Intersection of DNA methylation and gene expression at meQTL

Few studies have explored *trans* acting relationships between DNA methylation and gene expression. *Trans* meQTL, in particular, provide new opportunities to understand the coordination of genomic function, including identification of the proximal candidate gene(s) underlying the *trans* acting effect of meQTL SNPs[23]. To address this systematically, we first used data from the eQTLGen Consortium ($n = 31{,}684$ samples) to identify 4,811 *cis* eQTL associated with the 1,847 *trans* acting sentinel meQTL SNPs ($P < 1 \times 10^{-6}$, Bonferroni correction for 48,237 eQTL tests). AQ13 We then tested the 4,811 eQTL genes for association with DNA methylation in our participants and found 1,607 *trans* eQTM at $P < 0.05$. SMR supported 929 of these *trans* eQTM (SMR $P < 3.1 \times 10^{-5}$; Bonferroni correction for 1,607 tests), while 34 *trans* eQTM were likely regulated by a common genetic mechanism (coloc PP4 > 0.6)[33] The 34 *cis* eQTL identified as likely to be mediating *trans* methylation signatures identified include *ZFP57* (associated with the *trans* meQTL SNP rs2747429), which encodes a DNA-binding protein critical for maintenance of epigenetic memory[33, 34], as well as other *ZNF* or *ZFP* genes anticipated to be involved in genome regulation (Supplementary Table 15).

Intersection of DNA methylation with clinical phenotypes at meQTL

We used our meQTL as genetic instruments to examine the potential causal relationships between DNA methylation and body mass index (BMI) as a model phenotype of global public health AQ14 importance. Our sentinel meQTL SNPs and CpG sites were both strongly enriched for association with BMI (Extended Data Fig. 7 and Supplementary Table 11, respectively), consistent with a role in the etiology of adiposity. Using the 941 SNPs independently associated with BMI at $P < 10^{-8}$ in the GWAS as genetic instruments[35], SMR suggested a potential causal relationship between DNA methylation and BMI at 374 loci ($P < 0.05$ after Bonferroni correction, Supplementary Table 16), of which 239 showed evidence for a shared underlying causal variant (coloc PP4 > 0.6). At the *UBASH3B* locus, we identified SNP rs7115089 as influencing both DNA methylation and BMI (SMR $P = 2.5 \times 10^{-10}$, coloc PP4 = 1.0). *UBASH3B* encodes a protein with tyrosine phosphatase activity, which has been previously linked to advanced neoplasia[36]. SNP rs7115089 is strongly associated with BMI[35] and is in linkage disequilibrium (LD) ($R^2 > 0.8$) with genetic variants linked to other cardiovascular and metabolic traits in GWASs[37, 38, 39, 40]. SNP rs7115089 was associated with differential methylation at our sentinel CpG site (cg26684673), which we previously showed to be associated with BMI in adults[8]. SNP rs7115089 was associated with expression of *UBASH3B* ($P = 1.7 \times 10^{-17}$). Animal models show that expression of *Ubash3b* is an early transcriptomic-based biomarker of gestational calorie restriction that may drive programmed susceptibility to obesity and other chronic diseases in later life[41], and expression of *UBASH3B* in peripheral blood is also strongly associated with BMI and other measures of adiposity in humans (Supplementary Table 17) [42]. Our results thus identify *UBASH3B* as a potential mediator of both genetic and environmental exposures underlying adiposity and cardiometabolic disease.

Integrating molecular information at *trans* acting loci

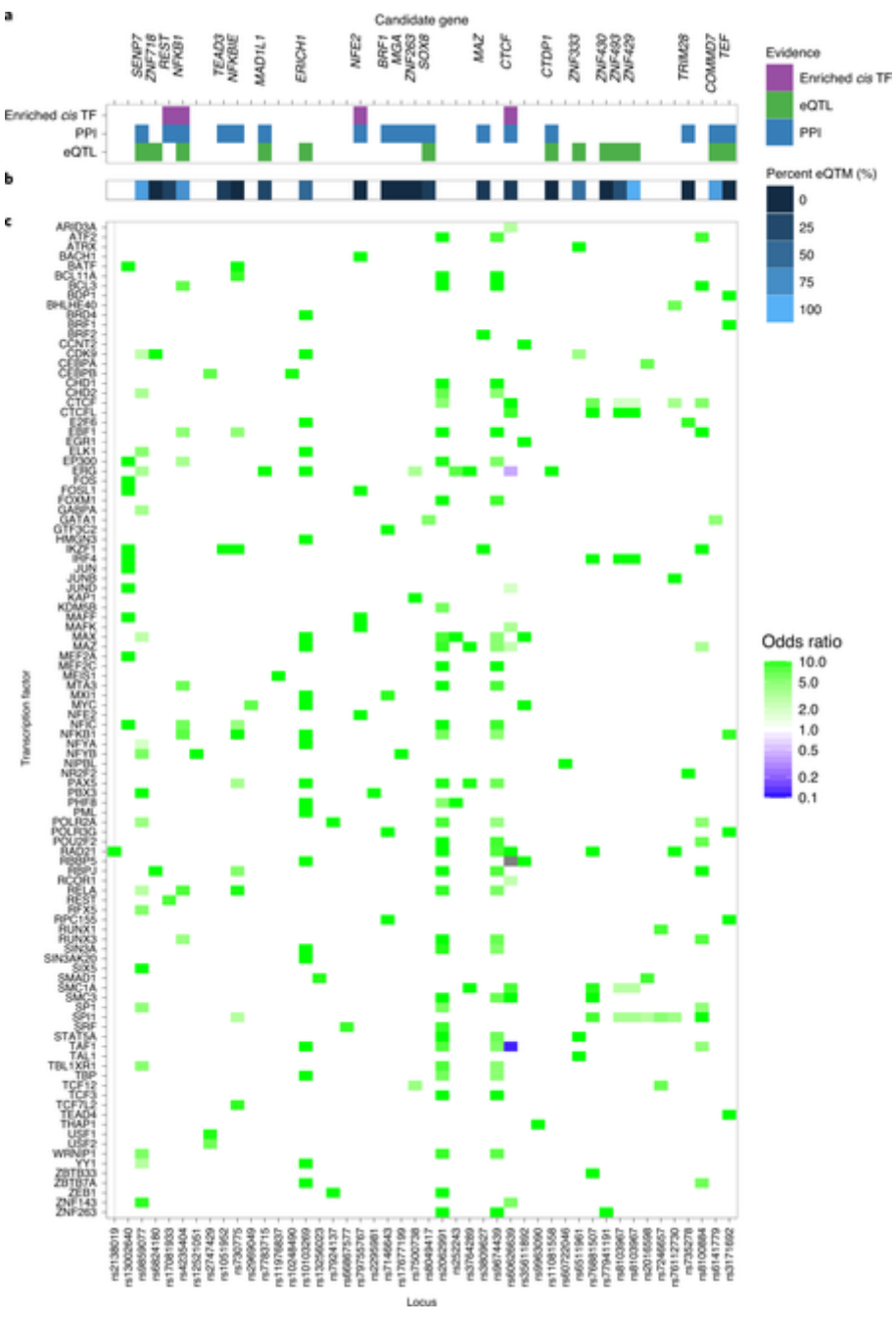
We identified 467,915 *trans* acting SNP–CpG pairs, comprising pairwise relationships between 200,761 unique SNPs and 3,592 unique CpG sites. Based on conditional analysis, these represent 1,847 distinct loci with genetic variants that influence DNA methylation in *trans* (range, 1–298 *trans* CpG sites per genetic locus, Fig. 1). The genes in *cis* to the sentinel *trans* acting SNPs were enriched for genes with known regulatory function (mean enrichment, 1.64-fold, empirical $P = 5.99 \times 10^{-3}$; gene list and pathway analysis in Supplementary Tables 18 and 19, respectively), including documented transcription factors such as those encoded by *CTCF*, *NFKB1*, *REST* and *TBX6*. Our results support the view that the *trans* meQTL identify genetic loci encoding factors with key roles as master regulators of genome structure and function and that the effects of these *trans* acting loci may be mediated through their remote effects on DNA methylation.

To generate new knowledge of the nuclear proteins involved in mediating *trans* SNP–CpG relationships, we next identified known transcription factors with binding sites that overlap the *trans* CpG signatures of the *trans* acting genetic loci. Based on power calculations, we limited the analysis to the 115 sentinel *trans* meQTL with $n \geq 5$ associated CpG sites (Methods). At 45 genetic loci (39%), the *trans* CpG sites of the respective sentinel SNPs overlapped binding sites of one or more known transcription factors (Fig. 3, Extended Data Fig. 8 and Supplementary Table 20; false discovery rate (FDR) < 0.05). This represents a 1.8-fold enrichment compared to expectation under the null hypothesis ($P = 7.4 \times 10^{-6}$, binomial test, Methods). As a sensitivity analysis, we repeated the experiment using data generated on a MethylationEPIC array to test the impact of increased coverage of methylation markers on identification of overlapping transcription factors (Methods and Supplementary Table 21). There was no evidence for false positive findings, but the higher-density marker set of the EPIC array did increase the number of overlapping transcription factors identified by 14% (Supplementary Table 21).

Fig. 3

Candidate genes for sentinel SNPs that are associated with *trans* CpG sites that overlap transcription factor-binding sites.

a, Evidence for each candidate. Genes that are transcription factors (TFs) in *cis* and that overlap the *trans* CpG signatures ('enriched *cis* TF', purple); genes selected by the random walk analysis including PPIs (blue); and genes that are *cis* eQTL for the sentinel SNPs (green). The heatmap in **b** shows the percentage of associated CpG sites with *trans* eQTM at each locus (*x* axis). The heatmap in **c** shows the enrichment or depletion of binding of transcription factors (*y* axis) at the associated CpG sites of each locus (*x* axis). Odds ratios comparing the frequency of state annotations at associated CpG sites with background CpG sites are color coded. Odds ratios greater than 10 or less than 0.1 have been set to 10 or 0.1 for improved readability of the color scale. Odds ratios greater than 1 indicate enrichment, while odds ratios less than 1 indicate depletion.



At four of the 45 genetic loci with *trans* CpG AQ15 signatures overlapping a transaction factor, the genes in *cis* to the sentinel SNP encoded the respective nuclear transcription factor (*REST*, *NFE2*, *CTCF* and *NFKB1*; FET $P = 1.7 \times 10^{-5}$–3.4×10^{-89}; Extended Data Fig. 9 and Supplementary Table 22). For this subset of loci, the identified *cis*-encoded transcription factor is likely to be directly responsible for the respective *trans* methylation signature. By contrast, at the remaining 41 loci, the genes in *cis* to the sentinel SNP did not encode the transcription factor overlapping the *trans* CpG sites (Supplementary Table 23). We hypothesized that the causal gene in *cis* at these

regulatory pathways. To identify the most likely candidate gene and accompanying molecular pathway for these loci, we integrated the comprehensive SNP–methylation (meQTL), SNP–expression (eQTL) and methylation–expression (eQTM) data generated in our study with publicly available protein–protein interaction (PPI) networks and transcription factor-binding maps using an approach based on random walks (Methods). Our approach identified strong candidate genes and their corresponding molecular networks at 19 loci (Fig. 3, Supplementary Tables 24 and 25, Extended Data Fig. 9 and Supplementary Fig. 3). In addition, we prioritized six candidate genes for the remaining loci, which were unambiguous *cis* eQTL for only a single gene (Methods). To corroborate the candidate genes identified in *cis* at these 25 genetic loci, we quantified the number of *trans* eQTM associated with expression for each of the candidate genes. We observed significantly more *trans* eQTM compared to the remaining genes encoded at the *trans* acting loci ($P = 4.5 \times 10^{-6}$, Wilcoxon test; Supplementary Fig. 4).

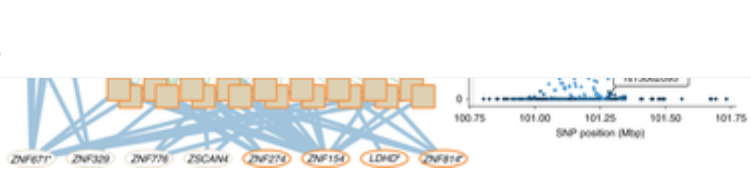
The *NFKBIE* locus

To illustrate the results of our approach, we highlight SNP rs730775, which is associated with 49 CpG sites in *trans* (Fig. 4). *NFKBIE* (empirical $P < 0.01$; Supplementary Table 24) is the most likely *trans* acting gene for this locus. The SNP is located in the first intron of *NFKBIE* and is a *cis* eQTL for *NFKBIE* in whole blood (eQTLGen, $P = 1.2 \times 10^{-23}$). Nuclear factor (NF)-κB inhibitor ε (NFKBIE) directly inhibits NF-κB1 activity and is significantly coexpressed ($P = 2.2 \times 10^{-4}$) with NF-κB1, which directly binds at 31 of the 49 *trans*-associated CpG sites (odds ratio = 7.8, $P = 9.1 \times 10^{-7}$; Supplementary Table 23). AQ16 The *trans* CpG sites localize to genes of the NF-κB pathway such as *IKBE* and *TRAF6* and are enriched for the gene ontology (GO) term ‘regulation of interleukin (IL)-6 biosynthetic process’ (GO:0045408; $P = 3.75 \times 10^{-5}$; hypergeometric test). The *NFKBIE* locus is associated with rheumatoid arthritis[43], which is characterized by IL-6-mediated autoimmunity and can be treated with IL-6-targeting drugs[44,45]. We performed a colocalization analysis of molecular QTL and genome-wide associations using enloc[46]. On average, the posterior colocalization probability was 70% at the sentinel SNP rs730775 (Fig. 4a), supporting a shared causal variant for the majority of the CpG sites. Our results suggest that genetic variation at the *NFKBIE* locus is linked to rheumatoid arthritis through *trans* acting regulation of DNA methylation by NF-κB.

Fig. 4

Regulatory networks and locus colocalization analyses.

a–**d**, Identified random walk networks and results for the individual colocalization analyses for the *NFKBIE*, *MGA*, *COMMD7* and *SENP7* loci, respectively. The networks illustrate connections between the genotype at SNPs (yellow rectangles), the identified candidate genes (yellow ellipses), which AQ17 are connected through a network of protein–protein and protein–DNA interactions to methylation at the *trans*-associated CpG sites (beige squares) and the expression of genes encoded at CpG sites. Ellipses represent genes (1) encoded at the genetic locus identified by the sentinel and prioritized by the random walk (yellow filled symbol), (2) encoded at CpG loci (beige border) or (3) part of the PPI network (black border). For genes in the PPI network, the filled color of ellipses represents the random walk score as indicated in the colored bar legend. Edges connecting genes, SNPs and CpG sites represent (1) PPIs, (2) protein–DNA interactions identified by transcription factor-binding site (TFBS) overlap and (3) genomic proximity (<1 Mb). Bold edges indicate significant correlation with gene expression. Other plots show (1) the genome-wide association signal AQ18 ($-\log_{10}(P)$) and (2) the colocalization signal (mean per-SNP colocalization probability (mean SCP) over all *trans* CpG sites) on the *y* axis for available SNPs in the genomic region around the respective genetic loci (*x* axis). Coloring of individual SNPs indicates LD (R^2) with the lead SNP in the locus.



The *MGA*, *COMMD7* and *SENP7* loci

The *trans* CpG sites linked to rs17677199 overlap the binding sites of three transcription factors encoded at other loci: MYC-associated factor X (MAX), E2F transcription factor 6 (E2F6) and nuclear transcription factor Y subunit β (NFYβ) (Fig. 4b). SNP rs17677199 lies in *cis* to *MGA*, encoding a known interacting protein for MAX, and *MGA*, *MAX* and *E2F6* expression shows strong covariation. *MGA* is thus a strong candidate linking rs17677199 with disturbances in MAX and E2F6 binding. SNP rs17677199 is associated with heightened blood pressure, aortic aneurysms and subarachnoid hemorrhage. Both *MAX* and *E2F6* are compelling candidates for mediating the effects of rs17677199 on DNA methylation and vascular disease. Mutations in *MAX* are associated with abnormalities of blood pressure regulation, including development of pheochomocytoma, a catecholamine-secreting tumor[47]. In addition, the E2F family of transcription factors is implicated in vascular function and blood pressure regulation[48]. E2F transcription factors regulate synthesis of dihydrofolate reductase (DHFR), the rate-limiting salvage enzyme for tetrahydrobiopterin, an essential cofactor for endothelial nitric oxide synthase. Colocalization analysis with fastenloc supports a shared causal variant underlying DNA methylation of *trans* meQTL CpG sites and diastolic blood pressure (Fig. 4b).

SNP rs6141779 is associated with ten *trans* CpG sites. The only gene at this locus is *COMMD7* (COMM domain-containing 7), which is also an eQTL for the sentinel SNP and thus a highly plausible *cis* candidate gene. Our pathway analysis linked *COMMD7* to *NFKB1* through covariation in expression (Fig. 4c). COMMD7 interacts with the NF-κB complex and suppresses its transcriptional activity[49]. Sentinel SNP rs6141779 is strongly associated with white cell subset composition[50]. Colocalization analysis supports multiple shared causal variants for basophil counts and DNA methylation with average posterior probabilities over CpG sites ranging from 7% to 66% (Fig. 4c).

We also replicated and extended results for the known *trans* acting locus *SENP7* (refs. [18, 23]) identified by SNP rs9859077 (Fig. 4d). Our pathway and colocalization analyses provide new insights into the molecular mechanism linking *SENP7* with *trans* regulation of both DNA methylation and gene expression on chromosome 19 and into the body composition, leukocyte traits and inflammatory diseases linked to this locus[51].

Experimental validation at the *ZNF333* locus

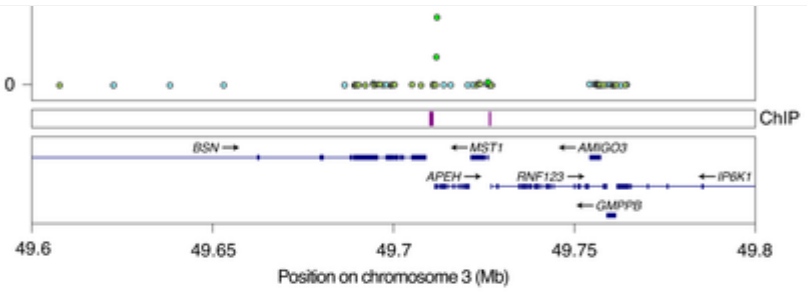
At the genetic locus identified by rs6511961, the putative candidate gene is *ZNF333* (Supplementary Table 24)[52]. Expression of *ZNF333* in our participants is associated with rs6511961 and covaries with expression of *TAL1* and *CDK9*, genes known to encode nuclear transcription factors (Extended Data Fig. 10). SMR supports a causal relationship between *cis* expression of *ZNF333* and *trans* methylation, with colocalization analyses providing some evidence for rs6511961 as a common underlying genetic driver (coloc PP4, 0.27).

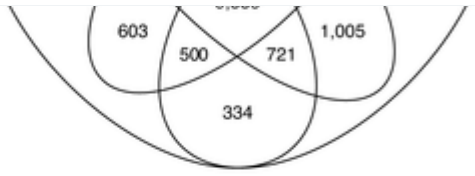
To further test the hypothesis that ZNF333 AQ19 contributes to the relationship of rs6511961 with its *trans* CpG signature, we carried out ChIP–seq using FLAG- and/or Myc-tagged ZNF333 constructs. ChIP–seq confirmed site-specific DNA binding (Fig. 5 and Extended Data Fig. 10). The putative binding motif for ZNF333 is TG(AG)*TCA. The binding sites for ZNF333 were enriched for motifs of known transcription factors ($P < 10^{-700}$), supporting the view that ZNF333 binds sites involved in genome regulation. Furthermore, we found that 35% of the CpG sites associated with rs6511961 in *trans* were in or near (<500 bp) ZNF333 DNA-binding sites (FET $P < 0.05$, Fig. 5). Immunoprecipitation mass spectrometry (IP–MS; Supplementary Note and Supplementary Tables 26–28) experiments provided further experimental evidence to support the hypothesis that *ZNF333* encodes a DNA-binding protein that determines, at least in part, the *trans* CpG signature of rs6511961.

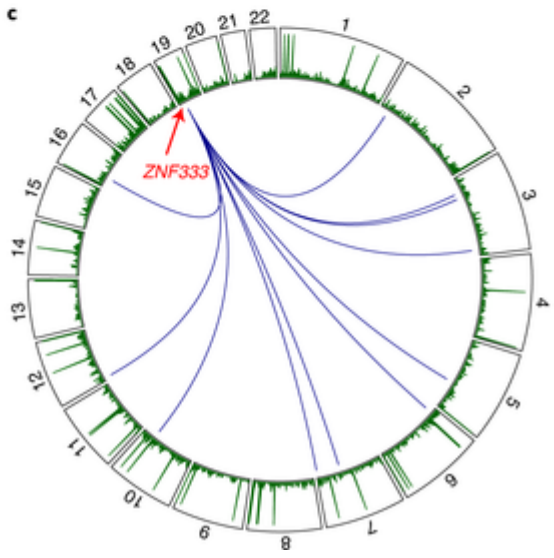
Fig. 5

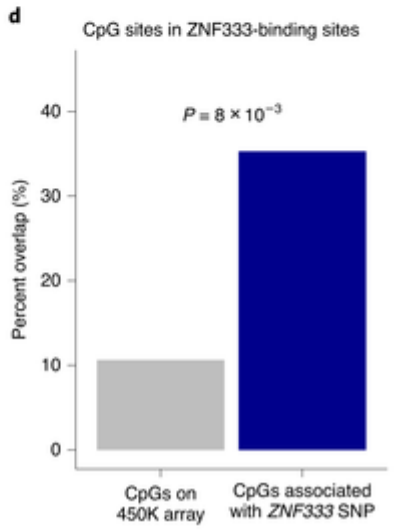
Experimental evaluation of ZNF333 by ChIP–seq.

a, Regional plot illustrating the overlap of the *trans* CpG signature for SNP rs6511961, with the ChIP–seq signature for ZNF333. Top, $-\log_{10}$ (*P* values) (*y* axis) of the association of each CpG site in the region (genomic position on the *x* axis) with the *trans* acting SNP rs6511961. The lead CpG associated with rs6511961 is identified by a diamond; color coding of other CpG sites at loci (circles) describes their correlation (*r*) with the lead CpG site. Middle, genomic coordinates of binding sites of ZNF333 identified by ChIP–seq are shown as purple boxes. Bottom, gene annotation (exons, blue boxes; introns, blue lines). **b**, Venn diagram showing the overlap between binding sites from biological replicates (rep) of ZNF333 ChIP–seq using either anti-FLAG or anti-Myc antibodies. **c**, Circos plot summarizing (1) the genomic distribution of CpG sites associated in *trans* (inner connections) with rs6511961 at the *ZNF333* locus and (2) the DNA-binding sites of ZNF333 identified by ChIP–seq studies (green bars). **d**, The observed and expected proportions of CpG sites that overlap ZNF333 DNA-binding sites (interval size around the peak of 500 bp) compared to the background frequency of all tested CpG sites. Significant enrichment is shown by permutation testing with matched background (Methods). Enrichment is robust to selection of interval size around the peak, from 100 bp (2.7-fold) to 1,000 bp (4.5-fold).









Population-specific effects at meQTL

Among our 11.2 million meQTL, 1,354,623 (12%) showed evidence for an interaction with ancestry at $P < 4.5 \times 10^{-9}$ (that is, $P < 0.05$ after Bonferroni correction for 11.2 million tests). Identified SNPs were enriched for blood composition and immune and cardiometabolic traits compared to background expectations (Supplementary Tables 29 and 30 and Extended Data Fig. 7). Our results are in line with findings that genetic loci associated with blood cell counts display substantial heterogeneity between populations and that gene regulatory programs in immune cells are subject to recent population-specific adaptation[53,54].

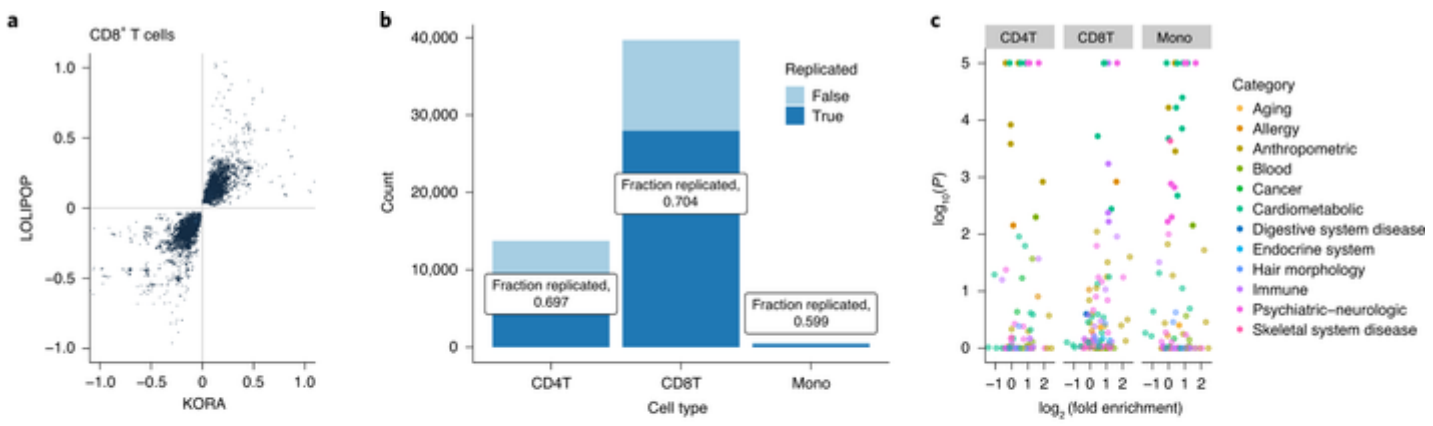
Interaction analysis of meQTL with environmental context

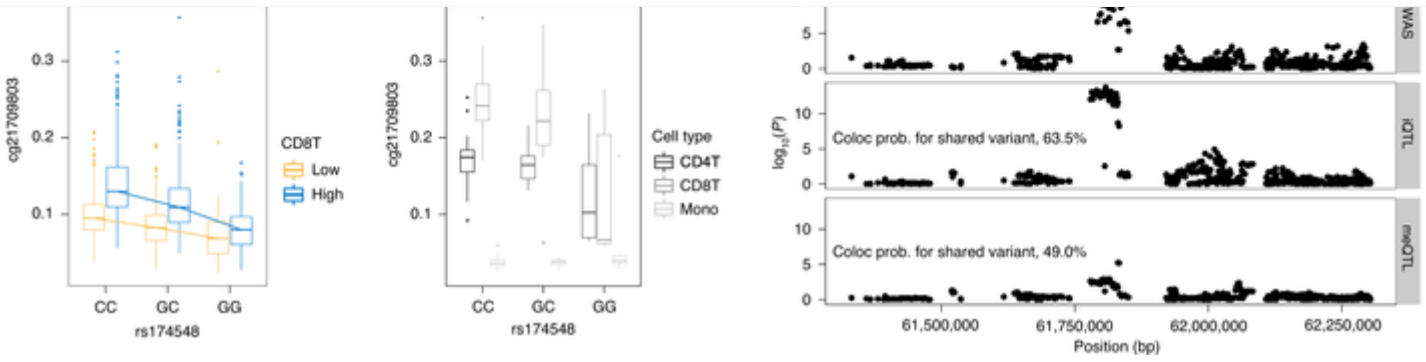
As a final experiment, we re-examined the relationship of SNPs with CpG sites in the cosmopolitan set of meQTL, seeking evidence for an interaction with white blood cell composition, BMI or cigarette smoking (Methods) as examples of biological traits that are anticipated or previously reported to have a strong relationship with DNA methylation[8,55,56,57,58,59]. We found that 130,016 (~1.1%) of our 11.2 million meQTL showed evidence for an interaction with one or more of the phenotypes tested (at a Bonferroni-corrected threshold of $P < 4.5 \times 10^{-9}$; Supplementary Table 31). White cell subsets generated the highest number of interaction meQTL ('iQTL'), and these showed evidence for replication between Europeans and South Asians (Fig. 6a). By contrast, there was little evidence for an effect of BMI or smoking on the genetic regulation of methylation in blood cells.

Fig. 6

White cell iQTL.

a, AQ20 Plot shows replication of effect sizes of significant iQTL ($CD8^+$ T cells) between KORA and LOLIPOP cohorts. Axes indicate genotype–cell type interaction effect sizes; points show individual associations. **b**, Bar plot indicates replication of iQTL in isolated cells. The *y* axis shows the total number of associations, and the *x* axis shows the respective cell types. Dark blue areas indicate the proportion of replicating associations; light blue areas indicate the proportion of non-replicating associations. CD4T, $CD4^+$ T cells; CD8T, $CD8^+$ T cells; mono, monocytes. **c**, Volcano plots highlighting the enrichment of iQTL SNPs with genome-wide association information in diverse traits. The *y* axis shows $-\log_{10}$ values of QTLEnrich *P* values; the *x* axis shows the $\log_2$ fold enrichment of observed GWAS SNPs among iQTL compared to expected values. Plots are split by analyzed cell types. Points reflect individual genome-wide associations; their colors represent the respective phenotype category. **d**, An example association plot for the rs174548–cg21709803 iQTL in KORA data, separated into individuals with 'high' and 'low' abundance (above and below the median, respectively) of $CD8^+$ T cells. The *y* axis indicates methylation residuals; the *x* axis shows genotypes. Box plots indicate medians (center lines) and first and third quartiles (lower and upper box limits, respectively; whisker extents, 1.5-fold of interquartile ranges). Points indicate outliers. **e**, The same association plot as in **d** but using data from isolated cells (indicated by different shades of gray). **f**, Manhattan plot of meQTL, asthma GWAS and iQTL results for the selected AQ21 iQTL example show colocalization of association signals. The *x* axis indicates the genomic region around the rs174548 SNP; the *y* axis indicates the $-\log_{10}$ of association *P* values. Individual points represent SNPs in the locus. Coloc prob., colocalization probability.





Significant interactions with blood cell proportions can be indicative of meQTL with stronger or weaker effects in specific cell types[60]. Cell type specificity of iQTL is supported by the high replication rates of iQTL in isolated $CD4^+$ and $CD8^+$ T cells (Fig. 6b). We expanded our iQTL analysis from cosmopolitan meQTL to a genome-wide *cis* iQTL analysis and discovered a total of 16,135 iQTL ($P < 8.8 \times 10^{-11}$; Supplementary Table 31), of which 64% were independent of cosmopolitan meQTL (LD $R^2 < 0.2$). The presence of an iQTL indicates that the relationship between methylation levels and genotype varies depending on the abundance of a specific cell type. SNPs that are part of white cell iQTL were enriched for association with phenotypic variation in GWASs (number of phenotypes enriched at FDR < 0.05 in the QTLEnrich analysis: $CD4^+$ T cells, $n = 18$; $CD4^+$ T cells, $n = 11$; monocytes, $n = 23$; Supplementary Table 32), including blood cell traits, immune traits and allergies (Fig. 6c). We showed that rs174548 in the *FADS1* gene shows increased correlation with DNA methylation in participants with high abundance of $CD8^+$ T cells (Fig. 6d,e). Fatty acid desaturase 1 (FADS1) is a key enzyme in the metabolism of fatty acids. SNP rs174548 is strongly associated with concentrations of arachidonic acid and other metabolites in fatty acid metabolism[61, 62], blood eosinophil counts[50] and inflammatory diseases such as asthma (GWAS, $P = 2.5 \times 10^{-10}$)[63]. Colocalization analysis indicated a shared causal variant for rs174548 and asthma (coloc PP4 = 0.63, Fig. 6f), providing a pathway linking fatty acid metabolism in $CD8^+$ T cells with immune phenotypes. This SNP is not detected as a cosmopolitan meQTL, highlighting the potential for iQTL analysis to improve annotation of functional genetic variants and to generate hypotheses about the cellular specificity of traits.

Discussion

We identify 11.2 million unique SNP–CpG associations in peripheral blood, including 467,915 meQTL associations that operate in *trans* and that comprise pairwise relationships between 1,847 genetic loci and 3,020 methylation loci. Key strengths of our study design include use of stringent statistical thresholds and replication testing across population groups and tissues to enable identification of high-confidence generalizable meQTL. Both the SNPs and CpG sites that form meQTL pairs are enriched for multiple functionally relevant characteristics, including shared chromatin state, Hi-C interaction, association with *cis* and *trans* gene expression and links to multiple metabolic and clinical traits. Candidate genes at *trans* acting genetic loci are enriched for nuclear transcription factors and their interacting proteins. Molecular interaction data, supported by colocalization analyses, identify multiple nuclear regulatory pathways, linking sequence variation to disturbances in DNA methylation and molecular and phenotypic variation. This includes the *UBASH3B* (BMI), *NFKBIE* (rheumatoid arthritis), *MGA* (blood pressure) and *COMMD7* (white cell counts) loci. As proof of principle, we use ChIP–seq to provide experimental support for ZNF333 as a AQ22 new *trans* acting genomic regulator. Finally, we use interaction analyses to identify both population- and cell lineage-specific meQTL effects that are biologically relevant. This includes meQTL SNP rs174548 in *FADS1*, with the strongest effect in $CD8^+$ T cells, linking fatty acid metabolism with immune dysregulation and asthma. Our study thus advances understanding of the relationships between DNA sequence variation and DNA methylation, thereby providing new insights into the molecular networks involved in nuclear regulation and the potential pathways linking genetic variation with human phenotype.

To move beyond investigation of cosmopolitan regulatory effects in mixed cellular populations, we extended our analyses to identify cell lineage- and population-specific processes. White cell subset interaction analyses revealed meQTL with stronger or weaker effects in specific cell types. We identified many thousands of white cell-specific iQTL, which were strongly supported by high replication rates in isolated $CD4^+$ and $CD8^+$ T cells. SNPs that are part of white cell iQTL are enriched for association with phenotypic variation in GWASs, notably blood cell traits, immune traits and allergies. We highlight the iQTL SNP rs174548 in the *FADS1* gene, which shows increased correlation with methylation in $CD8^+$ T cells. FADS1 plays a key role in fatty acid metabolism, and genetic variation at this locus is well known to be a determinant of concentrations of arachidonic acid, eicosanoids and blood lipid levels[61, 62]. Our iQTL analysis suggests that genetic variation at *FADS1* has a specific impact on regulation of *FADS1* in $CD8^+$ T cells and may help explain the relationship of this locus with inflammatory diseases such as asthma[63]. $CD8^+$ T cells contribute to the development of asthma, including recruitment to pulmonary sites and secretion of the pro-inflammatory cytokines IL-13 and IL-4 (ref. [64]). People with asthma have increased cytokine release by $CD8^+$ T cells, and cytokine activity is related to asthma severity[65]. Our interaction analyses of meQTL data thus shed new light on the mechanisms impacting DNA methylation in white blood cells, an approach that may enable identification of cell-specific patterns of DNA regulation in other studies of tissue samples with mixed cellular composition[60].

Our study provides new insights into the genetic regulation of DNA methylation and reveals multiple new nuclear regulatory networks. Our findings advance understanding of the biological pathways underpinning phenotypic variation and will inform hypothesis-driven experimental studies to define the specific molecular mechanisms involved.

Methods

Further details of experimental methods and data analyses are provided in the Supplementary Note.

Discovery and replication of genetic variants influencing DNA methylation

A summary of the participating population cohorts is provided in Supplementary Tables 33 and 34. Genome-wide association was carried out in Europeans and South Asians separately[24]. First, methylation residuals were derived from a linear regression of the percentage methylation (outcome) with technical and clinical predictors: age, sex, estimates of white blood cell subpopulations and principal components of control probe intensities (Supplementary Table 34). Association testing of methylation residuals with genotypes was carried out using Quicktest. Genome-wide significance was set to $P < 10^{-14}$, which corresponds to $P < 0.05$ after Bonferroni correction for the ~4.3 trillion statistical tests performed, a choice consistent with other recent publications[19, 20]. Replication testing was performed using linear regression in R, and combined analysis of discovery and replication data was performed with inverse-variance meta-analysis (R package meta). Associations were considered replicated when the association showed consistent direction of effect between discovery and replication, replication $P < 0.05$ and combined $P < 10^{-14}$. We assessed our meQTL for enrichment with SNPs known to influence white blood cell count to test for confounding by variation in white cell subsets (Supplementary Table 35).

Replication across platforms and cell types

We used DNA methylome data to carry out cross-platform replication of meQTL, with permutation testing to establish whether the overlaps observed were more than those expected by chance[25]. Replication across tissues was initially tested using genomic DNA from (1) isolated white cell subsets ($n = 60$ individuals), (2) isolated visceral adipocytes ($n = 48$ individuals) and (3) isolated subcutaneous adipocytes ($n = 48$ individuals). Genome-wide genotyping (Illumina OmniExpress) and quantification of DNA methylation (Illumina EPIC array) was performed according to the manufacturer's recommended protocols. Imputation of unmeasured genotypes was performed using the reference panel from the 1000 Genomes project phase 3. We tested associations between SNPs and CpG sites using linear regression. We additionally carried out replication testing in 603 subcutaneous adipose tissue samples collected in the MuTHER study. Methylation profiling was performed using the Illumina Infinium HumanMethylation450 BeadChip. Genotyping was performed with a combination of Illumina arrays (HumanHap300, Human Hap610Q, 1M-Duo and 1.2MDuo 1M). Associations between SNPs and DNA methylation levels were tested in samples of related individuals using GEMMA software[66].

Conditional analysis and linkage disequilibrium pruning

Local correlations between SNPs (LD) and between neighboring CpG sites lead to redundant pairs of SNPs and CpG sites representing the same meQTL. We used a two-stage approach to identify independent associations among all identified SNP–CpG pairs (Supplementary Fig. 1). We first performed iterative conditional analysis using individual-level data from the European and South Asian discovery datasets. For each CpG, the most strongly associated SNP (lowest P value) was selected. Association testing was then repeated for all SNPs that had previously been associated at $P < 10^{-14}$ with that CpG but including the most strongly associated SNP as a predictor in the regression model. Analysis was carried out in Europeans and South Asians separately, followed by meta-analysis. From the SNPs that remained significantly associated ($P < 10^{-14}$), the most strongly associated SNP was selected, and the process was repeated until no SNPs remained. Independently associated SNPs for the respective CpG were then carried forward. This yielded a parsimonious set of 84,456 SNPs independently associated with one or more CpG sites (Supplementary Table 8).

While this step reduces redundancy introduced by LD between SNPs, it creates a scenario in which the same genetic locus can be represented by different SNPs. This is caused by the fact that the most strongly associated SNP for each genetic locus (that is, the SNP conditioned on) will vary from one CpG to another. To further reduce the impact of local correlation (Supplementary Fig. 1), we combined highly correlated SNPs into SNP loci and highly correlated CpG sites into methylation loci. To achieve this, the most strongly associated marker (lowest P value) was selected, and all markers with $R^2 > 0.2$ and distance <1 Mb were then assigned to a corresponding locus. Of the remaining markers, the most strongly associated marker was again chosen, and the process was repeated until no markers remained. This approach was applied to SNPs and CpG sites within each category (*cis*, long-range *cis* and *trans*) separately. Supplementary Fig. 5 shows a sensitivity analysis on the number of independent loci for varying R^2 thresholds.

Enrichment of meQTL within chromatin states

We obtained chromatin state annotations (15-state model) defined by chromHMM segmentation of histone modification ChIP–seq data[67] from the Roadmap Epigenomics Project for primary blood cells[68]. As we were working with whole blood, we combined these primary epigenomes into a weighted epigenome annotation based on estimated cell fractions in whole blood (Supplementary Note and Supplementary Table 36). We used permutation testing to assess for enrichment compared to expectations under the null hypothesis.

Genetic variants influencing gene expression in Europeans and South Asians

Transcriptome-wide measurements of gene expression in blood along with measurements of DNA methylation from the same blood sample are available for European ($n = 853$) and South Asian ($n = 693$) participants of the KORA and LOLIPOP studies (Illumina HumanHT-12 version 3 and 450K methylation arrays, respectively). These data enable evaluation of relationships between SNPs, methylation and gene expression using individual-level data in relevant populations and with a range of statistical models to allow for sensitivity analyses and investigation of potential confounding effects. Expression values were summarized to gene-level estimates by averaging the $\log_2$ transformed expression levels of probes mapping to the same gene. To quantify the relationship between genetic variation and gene expression, we first derived residuals for gene expression using linear regression of gene expression levels against sex, age, RNA integrity number (RIN), RNA-amplification plate (KORA), RNA-conversion batch (LOLIPOP), sample storage time (KORA) and RNA-extraction batch (LOLIPOP). Expression residuals were then used as outcome variables in a linear regression model with SNP dosage as the independent variable, corresponding to the following linear model formulae: (1) gene ≈ SNP + sex + age + RIN + RNA-amplification plate + storage time (KORA) and (2) gene ≈ SNP + sex + age + RIN + RNA-conversion batch + RNA-extraction batch (LOLIPOP). Data analysis was performed using Matrix eQTL[69], and results were analyzed separately for Europeans and South Asians. We then combined results between Europeans and South Asians using inverse-variance meta-analysis. Statistical significance was inferred

at $P = 7.98 \times 10^{-11}$ ($P < 0.05$ after Bonferroni correction for the number of SNP–expression pairs tested). We supplemented results from our participants ('KORA–LOLIPOP eQTL dataset') with eQTL results from publicly available resources (GTEx and eQTLGen)[70, 71]. The specific datasets used for each experiment are documented.

SNPs influencing DNA methylation are enriched for association with gene expression

To confirm that SNPs influencing methylation are more likely to affect gene expression, we randomly selected 100 sets comprising 1,000 SNPs 'observed' to be associated with DNA methylation from the list of SNP–CpG associations after pruning. For each 'observed' set, we generated a 'background' set of SNPs to quantify expectations under the null hypothesis. Each set of 'background' SNPs comprised 1,000 SNPs that were (1) not part of a significantly associated SNP–CpG pair and (2) matched with the 'observed' SNPs for minor allele frequency (±2%) and distance to the nearest gene (±10 kb) but were selected otherwise at random. We then determined the proportion of SNPs associated with gene expression in 100 'observed' sets and 100 'background' sets. Association of observed and background SNPs with gene expression was tested in our KORA–LOLIPOP eQTL dataset (statistical significance was inferred at $P = 5.06 \times 10^{-11}$ as above). The probability of enrichment was calculated by comparison of 'observed' sets with 'background' sets using a t-test.

Association of DNA methylation with gene expression

We quantified associations of DNA methylation with gene expression using our KORA–LOLIPOP gene expression dataset (Europeans, $n = 853$; South Asians, $n = 693$). To test for and exclude CpG–gene pairs that arose due to confounding by the underlying genetic background, we derived methylation residuals by correcting methylation (β) values for the sentinel SNP(s) associated with the corresponding CpG (formula, $\mathrm{CpG} \approx \Sigma \mathrm{SNP}_{\mathrm{associated}}$). Gene expression residuals were used as outcome variables in a regression model with methylation residuals as the independent variable (formula, $\mathrm{gene}_{\mathrm{residuals}} \approx \mathrm{CpG}_{\mathrm{residuals}}$). Data analysis was performed using Matrix eQTL[69], and results were analyzed in Europeans and South Asians separately. At Bonferroni-corrected P-value thresholds, there was a high degree of reproducibility for eQTM results between the populations (Supplementary Table 37). We therefore combined results between Europeans and South Asians using inverse-variance meta-analysis (R package meta). Statistical significance was inferred at $P = 8.7 \times 10^{-12}$ ($P < 0.05$ after Bonferroni correction for all possible CpG–expression pairs). We carried out association tests with or without adjustment of the methylation residuals for white cell subsets (that is, with or without Houseman white cell subset estimates using the formula $\mathrm{CpG}_{\mathrm{residuals}} \approx \mathrm{CD8}^+$ T cells + $\mathrm{CD4}^+$ T cells + NK cells + B cells + monocytes) to test for confounding by cell subset composition (Supplementary Table 38).

In addition, we compared the proportion of putative *cis* eQTM from analyses with and without correction for white cell subsets that were supported by SMR. SMR tests for association of an exposure with an outcome using summary-level data from GWAS and other QTL studies and using a genetic variant as the instrumental variable to avoid non-genetic confounding[26]. Colocalization analysis was subsequently performed for loci with a potentially causal relationship between DNA-methylation levels and gene expression in *cis* (PP4 > 0.6)[46].

Enrichment of meQTL SNPs and CpG sites for association with phenotypes

We performed enrichment analyses of meQTL and iQTL SNPs for association with clinical traits using QTLEnrich[72], which includes uniformly processed summary statistics of 114 genome-wide associations. We tested meQTL SNPs for enrichment as protein QTL and metabolite QTL using the PhenoScanner version 2 database[63, 73]. To evaluate the biological relevance of our sentinel CpG sites, we first quantified the association of DNA methylation with 49 clinical traits (physical measures, health status, lifestyle behaviors and biochemical traits) and with the concentration of 228 metabolites measured by NMR metabolomics in the LOLIPOP cohort ($n = 2{,}866$ participants with DNA-methylation data available). We used permutation testing to determine expectations under the null hypothesis (Supplementary Note and Supplementary Table 39).

Identification of *cis* eQTL influencing CpG sites in *trans*

We used SMR analysis to assess whether the proximal candidate gene at a *trans* acting genetic locus showed covariation with the *trans* methylation signature (triangulation of *cis* eQTL, *trans* meQTL and *trans* eQTM data). Results for *cis* SNP–expression (*cis* eQTL) associations were obtained from eQTLGen[71], while *trans* SNP–methylation (*trans* meQTL) and SNP–expression (*trans* eQTM) associations were as reported in the current study. We started with *trans* sentinel meQTL SNPs reported in our current study and identified significant *cis* eQTL associations at a Bonferroni-corrected threshold. For loci for which SMR estimates suggested a potential causal relationship between *cis* gene expression and *trans* methylation levels ($P < 0.05$ after Bonferroni correction), this was followed up with a coloc analysis (PP4 > 0.6). In addition, we also evaluated the complementary model in which the causal inference analysis started with observed *trans* eQTM and assessed the proportion that was correctly inferred by SMR.

Enrichment of *trans* CpG sites in transcription factor-binding sites

We obtained TFBSs for 145 distinct DNA-binding proteins from 246 ChIP–seq experiments performed on blood-related cell lines (Supplementary Table 20). Data were uniformly processed by the ReMap resource[74]. We defined a CpG site to be bound if a binding site was located within a window of 100 bp (50 bp in each direction; Supplementary Fig. 6). To examine the relationship between the *trans* CpG signatures of the sentinel SNPs and the TFBS of DNA-binding proteins, we first determined the minimum number of *trans* CpG sites associated with a sentinel SNP needed for detection of enrichment in the TFBS. This number depends on whether the smallest achievable P value in the Fisher test is less than an adjusted significance level, P_{adj} (Supplementary Note). Based on this analysis, we tested each of the 115 sentinel SNPs with ≥5 associated *trans* CpG sites for over-representation or under-representation in the TFBS for each of the 246 ChIP–seq datasets for DNA-binding proteins. For each sentinel SNP, we resampled 10,000 sets of CpG sites of equal size

to compute empirical *P* values for the overlap of the observed *trans* CpG sites with the TFBS. AQ23 We carried out similar analyses using the MethylationEPIC array to validate our findings.

Random walk analysis

We set out to identify the most likely *trans* acting gene for each locus with at least five *trans* acting SNP–CpG pairs overlapping a TFBS by linking the genes in the locus to the associated CpG sites through a sequence of PPIs and protein–DNA interactions. We used PPIs that had experimental evidence or database information available in the STRING database[75]. The initial network comprised 12,769 proteins and 186,674 interactions. In addition, we restricted the network to 8,880 proteins that were expressed (median reads per kilobase per million sequenced (RPKM) > 0.1) in whole blood in the GTEx dataset[70] and further to the largest connected component of the network comprising 8,668 proteins and 99,143 interactions used for the analysis. Formally, we defined the PPI network $P = (V_P, E_P)$, where V_P is the set of nodes (or vertices) corresponding to proteins and E_P is the set of undirected edges corresponding to interactions between proteins. Similarly, we represent protein–DNA interactions as graph $D = (V_D, E_D)$, where V_D is the union of 145 proteins for which ChIP–seq data were available (see above) and the CpG sites that were within 50 bp of sites bound by these proteins.

For each locus, we identified the set of candidate genes *C* as all genes encoded at the SNP locus that are part of the PPI network. Locus regions were defined based on the results of the pruning analysis that identified sentinel SNPs. Specifically, we identified all *trans*-associated CpG sites that were assigned to the same sentinel SNP. For these *trans* CpG sites, we obtained all SNPs that were (1) associated with the CpG in the complete, cosmopolitan pairwise analysis of SNP–CpG associations and (2) located in *cis* (within 1 Mb) with the sentinel SNP. In this way, the *trans* acting loci are refined by patterns of LD and observed associations with methylation levels but are not larger than 1 Mb.

Next, we identified the set of CpG sites *S* that were associated with the respective sentinel SNP at the *trans* acting genetic locus. We added the CpG sites *S* and their protein–DNA interactions $E_D(S)$ to the PPI graph *P* to form the locus graph $G = (V_P + S, E_P + E_D(S))$. AQ24 Finally, we used the topology of the locus graph *G* to rank candidate genes *C*. The ranking is based on random walks and is conceptually similar to published studies[111,112]. We represent graphs (V, E) by their adjacency matrix $A = (a_{ij})$ with entries $a_{ij} = 1$ if (i, j) in *E* and 0 otherwise. AQ25 We defined the symmetric transition matrix $T = (t_{ij})$ with $t_{ij} = a_{ij} \times (d_i d_j)^{-0.5}$, where d_i is the degree of node *i*, specifying the probability to move from node *i* to node *j* in one step of the random walk[76]. Consequently, transition probability matrices for paths with *t* steps can be computed as T^t. We initiated random walks at the CpG sites *S* and computed the transition probability T^t_{sc} to start at CpG site *s* in *S* and reach candidate gene *c* in *C* in *t* steps. As the lengths of the paths *t* are not known a priori, we sum the transition probabilities over all possible path lengths $t = (0,\ldots, \infty)$. The random walk has a stationary state with a distribution that is defined by the degree distribution of the nodes, which corresponds to the first eigenvector ψ_0 of the transition matrix *T* with eigenvalue $\lambda_0 = 1$ (ref. [76]). We were not interested in this stationary state, so we removed the contribution of the first eigenvector from the transition matrix and computed the aggregated transition probability matrix $M = \Sigma^{\infty}_{t=0} (T - \psi^T_0 \psi_0)^t$. This infinite sum has a closed-form solution[77]; however, the resulting matrix *M* is not sparse, and therefore the computation is very memory intensive. Alternatively, the solution can be approximated using spectral decomposition of the transition matrix[77]:

$$M = \sum_{i=1}^{n-1} \left(\frac{\lambda_i}{1 - \lambda_i} - \psi_i^T \psi_i \right)^t .$$

To compute the ranking of candidate genes while saving memory, we approximated the aggregated transition matrix *M* using the first $n = 500$ eigenvectors and stored only the submatrix of *M* that holds the transitions from CpG sites *s* in *S* and candidate genes *c* in *C*. AQ26 The final ranking of candidate gene *c* was computed as the average aggregated transition probability over all CpG sites $p_c = 1 \times (|S| \Sigma_s M_{sc})^{-1}$. To assess whether the score p_c of a candidate gene was significantly higher than that expected by chance, we performed the same analysis on $B > 100$ randomized graphs and computed scores p^b_c for all genes in *C* to determine the empirical *P* values for the maximum score at each locus $P(p_c) = 1 \times (B \Sigma_b \delta(p_c > \max_C p^b_c))^{-1}$. Randomized graphs were constructed by randomly sampling the same number of $|S|$ CpG sites S_b with matched mean and standard deviation of methylation levels (TFBS analysis). The random CpG sites S_b were then added to the PPI graph *P* to form the background locus graph $G_b = (V_P + S_b, E_P + E_D(S_b))$. In this manner, we empirically assessed the probability of ranking scores as extreme as the one observed by transitioning from a random set of CpG sites through the original PPI and ChIP–seq graph to each of the candidate genes. For each locus, the set of significant candidate genes was defined as $C^* = (c \mid P(p_c) < 0.05)$.

To visualize the results of the random walk analysis, we first defined weights w_i for each node *i* of the locus graph *G* by the sum of the random walk score to transition from the CpG sites in *S* to node *i* and of the random walk score for transitioning from *i* to the selected candidate genes in *C** in the *trans* locus. These weights were normalized and inverted to $w^*_i = \max_i(w_i) - w_i$, such that the highest scoring nodes receive the lowest weights. These weights *w** were then used to determine the minimal weight paths from each of the CpG sites in *S* to the candidate genes in *C** in the *trans* locus, thus representing paths through nodes with high random walk scores. Nodes on these minimal weight paths were recorded in the set *Q*. For each locus, we defined the candidate pathway G_c as the subgraph of the locus graph *G* with the nodes defined by the union of *C**, *Q* and *S* and all edges of *G* between this subset of nodes.

Identification of candidate genes for sentinel SNPs at *trans* acting genetic loci

We combined all available information from transcription factor signatures, PPI random walks and eQTL results (nominal $P < 0.01$ in our data or in GTEx whole-blood data) to select candidate gene(s) responsible for the effect of the sentinel SNPs on DNA methylation in *trans*. We evaluated random walk-based candidate predictions using GO enrichment analysis and overlap with eQTL results (Supplementary Fig. 7). We observed that a definition of SNP locus based on the association results (LD regions) yielded a higher proportion of candidates annotated to GO terms for regulators such as ‘regulation of biological process’, ‘DNA binding’ and ‘regulation

only PPIs between genes expressed in whole blood were considered. Therefore, we used PPIs of genes expressed in whole blood and the LD-based definition of *trans* loci to identify candidates by random walk analysis. We established the following order of evidence for prioritization: (1) transcription factors encoded at the *trans* acting genetic locus that are enriched for binding at the associated CpG sites and that are a *cis* eQTL for the sentinel SNP, (2) transcription factors encoded at the *trans* acting locus that are enriched for binding at the associated CpG sites but that do not have an eQTL with the sentinel SNP, (3) candidates that were identified through the random walk analysis (empirical $P < 0.05$) and have a *cis* eQTL, (4) random walk candidates without a *cis* eQTL and (5) singular *cis* eQTL at the *trans* acting genetic locus without other evidence.

Integrated network analysis

To set the results of the random walk analysis into context, we integrated the candidate pathways defined above for each *trans* acting genetic locus with genotype, gene expression and methylation data for Europeans (KORA) and South Asians (LOLIPOP). Hence, we collected for both cohorts (1) genotype data for the sentinel SNP (2) methylation β residuals (methylation data) for all CpG sites associated in *trans* and (3) gene transcript expression residuals (gene expression data) for all genes within a 1-Mb window of the respective SNP and CpG sites as well as the genes used in the random walk analysis. Genetic variation in *cis* could also influence expression and methylation measurements. AQ27 To avoid confounding by *cis* effects, we therefore adjusted expression and methylation data for previously reported *cis* eQTL[78] and for *cis* acting SNPs identified in our study using a linear regression model (that is, getting residuals (1) for genes using $\text{gene}_A \approx \text{gene}_A + \text{eQTL}_{SNP1} + \text{eQTL}_{SNP2} + \ldots + \text{eQTL}_{SNPi}$) and (2) for CpG sites using $\text{CpG}_A \approx \text{CpG}_A + \text{meQTL}_{SNP1} + \text{meQTL}_{SNP2} + \ldots + \text{meQTL}_{SNPi}$). The residuals were used to test for association individually in each cohort and subsequently combined using fixed effects meta-analysis. Resulting *P* values were adjusted for multiple testing using the Benjamini–Hochberg method[79]. In the resulting network, vertices represent variables (genotype, gene expression and methylation), and edges represent significant correlation between these variables (FDR < 0.05). Correlation edges found between a CpG and a CpG gene (that is, a gene found within the 1-Mb window around the CpG) were added to the candidate pathway graph (Random walk analysis) for each locus.

Colocalization analysis of *trans* meQTL

Colocalization analysis of *trans* meQTL and GWAS was performed using fastenloc[46], a Bayesian method to determine the probability of a shared causal variant for a pair of molecular (meQTL) and physiological (genome-wide associated) traits. First, we used PhenoScanner version 2 (refs. [63, 73]) and the GWAS catalog[80] to select genome-wide associated traits and studies of interest for each locus. We obtained GWAS summary statistics for each trait of interest for the region (±500 kb) around the sentinel SNP (Supplementary Table 40). fastenloc was used to determine SNP-level posterior colocalization probabilities for molecular and physiological traits for all CpG sites associated with the same locus in *trans*. We summarized the colocalization probabilities across all *trans* CpG sites using the average SNP-level posterior colocalization probabilities.

ChIP–seq validation of ZNF333 binding at the identified DNA-methylation sites

The plasmid used to overexpress dual-tagged (Myc and FLAG) human ZNF333 (RC216457) was purchased from OriGene Technologies. The ZNF333 plasmid and the control GFP plasmid (pmax-GFP, Lonza) were transfected into HCT116 cells with jetPRIME transfection reagent (Polyplus) according to the manufacturer's instructions in 15-cm tissue culture dishes. Culture medium was refreshed after 24 h, and cells were maintained for another 24 h. At 48 h, cell lysates were used for ChIP–seq. Western blotting using anti-Myc and anti-FLAG antibodies was also performed to confirm high ZNF333 expression. Raw sequencing from ChIP–seq experiments was mapped using BWA. The overlap between ZNF333 ChIP–seq peaks (union of Myc and FLAG) and rs6511961 target CpG sites (in *trans*) was calculated using a window size of 500 bp. Statistical significance was calculated based on permutation testing.

Interaction analysis of meQTL with their environmental context

We ran interaction analyses for the cosmopolitan SNP–CpG pairs using linear regression models with the methylation β value as the dependent variable and an interaction between the SNP and phenotype of interest as the independent variable of interest. The phenotypes of interest examined were smoking (yes or no), BMI (kg m^{-2}) and estimated proportions of $CD8^+$ T cells, $CD4^+$ T cells and monocytes. The analyses were run for KORA F4 and LOLIPOP separately. Significant results in one cohort were examined for replication ($P < 0.05$, same direction of effect) in the other cohort. In a second step, we repeated the interaction analysis with the covariates age, sex, BMI and white blood cell count for all CpG–SNP pairs in *cis* using tensorQTL (version 1.0.3)[81]. Statistical significance was inferred at a Bonferroni-corrected *P* value of 0.05 per number of tested pairs. We used GOstats for pathway analysis of the iQTL (Supplementary Table 41).

Reporting Summary

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

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00969-x.

Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00969-x.

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

Peer review information *Nature Genetics* thanks Charles Danko and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Acknowledgements

The KORA study was initiated and financed by the Helmholtz Zentrum München (German Research Center for Environmental Health), which is funded by the German Federal Ministry of Education and Research (BMBF) and by the state of Bavaria. KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. The work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the EU Joint Programming Initiative 'a Healthy Diet for a Healthy Life' (DIMENSION grant number 01EA1902A). The work was further supported by the Bavarian State Ministry of Health and Care through the research project DigiMed Bayern (https://www.digimed-bayern.de/). The German Diabetes Center (DDZ) is supported by the Ministry of Culture and Science of the State of North Rhine–Westphalia and the German Federal Ministry of Health. This study was supported in part by a grant from the German Federal Ministry of Education and Research to the German Center for Diabetes Research (DZD). The LOLIPOP study is supported by the National Institute for Health Research (NIHR) Comprehensive Biomedical Research Centre Imperial College Healthcare NHS Trust, the British Heart Foundation (SP/04/002), the Medical Research Council (G0601966, G0700931), the Wellcome Trust (084723/Z/08/Z), the NIHR (RP-PG-0407-10371), European Union FP7 (EpiMigrant, 279143) and European Union Horizon 2020 (iHealth-T2D, 643774). B.C.L. is supported by the Imperial College Junior Research Fellowship scheme as well as an Academy of Medical Sciences Springboard award. J.C.C. is also supported by the Singapore NMRC (NMRC/STaR/0028/2017). We thank the participants and research staff who made the study possible. AQ28 For the NFBC studies, M.W. was supported by the European Union's Horizon 2020 research and innovation program (grant 633212). NFBC1966 received financial support from the Academy of Finland (grants 104781, 120315, 129269, 1114194 and 24300796, Center of Excellence in Complex Disease Genetics and SALVE), University Hospital Oulu, Biocenter, University of Oulu, Finland (75617), NHLBI grant 5R01HL087679-02 through the STAMPEED program (1RL1MH083268-01), the NIH–NIMH (5R01MH63706:02), the ENGAGE project and grant agreement HEALTH-F4-2007-201413, EU FP7 EurHEALTHAgeing (277849), the Medical Research Council, UK (G0500539, G0600705, G1002319, PrevMetSyn/SALVE) and an MRC Centenary Early Career Award. NFBC1986 received financial support from EU QLG1-CT-2000-01643 (EUROBLCS) grant E51560, NorFA grant nos. 731, 20056 and 30167 and USA/NIHH 2000 G DF682 grant 50945. The NFBC programs are also funded by the H2020-633595 DynaHEALTH action, the Academy of Finland Exposomic, Genomic and Epigenomic Approach to Prediction of Metabolic and Cardiorespiratory Function and Ill-Health project (285547) and the EU H2020 ALEC project (grant agreement 633212). The MuTHER study was funded by the WT (081917/Z/07/Z). TwinsUK was funded by the WT and the European Community's Seventh Framework Programme (FP7/2007-2013). The study also received support from the NIHR Clinical Research Facility at Guy's and St. Thomas' and King's College London. Analysis was funded by British Heart Foundation grant RG/14/5/30893 to P.D. and forms part of the research themes contributing to the translational research portfolio of the Barts Cardiovascular Biomedical Research Unit, which is funded by the NIHR. The Saguenay Youth Study has been funded by the Canadian Institutes of Health Research (T.P., Z.P.), the Heart and Stroke Foundation of Canada (Z.P.) and the Canadian Foundation for Innovation (Z.P.). We acknowledge G. Möller and J. Adamski (Helmholtz Center Munich) for their support in the IP–MS transfection experiment. We used data generated by the PCHI-C Consortium[31], funded by the UK NIHR, the Medical Research Council (MR/L007150/1) and the Biotechnology and Biological Research Council (BB/J004480/1). AQ29

Author contributions

Data collection and analysis in the contributing population studies: KORA, A.P., B.K., C.G., C.H., C.B., AQ30 Eva Reischl, H.P., K. Strauch, L. Pfeiffer, M. Waldenberger, M.R., R.W., T.I., T.M. and W.R.; LOLIPOP, B.C.L., James Scott, J.S.K., J.C.C., W.Z. and W.R.S.; MuTHER, E. Marouli; MuTHER Consortium, P.D. and S.B.; NFBC, M.-R.J., M. Wielscher, S.S. and V.K.; SYS, J. Shin, M.B., T.P. and Z.P. Data collection and molecular follow-up analyses: ChIP–seq, D.P.L., M.I.A., R.S.Y.F. and W.L.W.T.; ChIP–MS, S.M.H., J.M.-P. and P.R.M.-G. Data analysis and writing group (alphabetical order): J.C.C., J.S.H., M.H., C.G., B.C.L., M.L., K. Schmid, M. Waldenberger and R.W.

Data availability

Summary statistics for the 11.2 million SNP–CpG pairs reaching genome-wide significance are available at https://zenodo.org/record/5196216#.YRZ3TfJxeUk. ChIP–seq data for ZNF333 are available through the NCBI SRA (accession code SRP284104). Raw genotype, methylation and expression data can be made available upon reasonable request by the authors. Controlled data access to data from the KORA cohort can be obtained through https://epi.helmholtz-muenchen.de. AQ31 The web links for the publicly available datasets used in the study are as follows: PhenoScanner version 2 (http://www.phenoscanner.medschl.cam.ac.uk), GWAS catalog (https://www.ebi.ac.uk/gwas/docs/file-downloads), meQTL and eQTM data from Bonder et al. (https://molgenis26.gcc.rug.nl/downloads/biosqtlbrowser/2015_09_02_Primary_cis_meQTLsFDR0.05-ProbeLevel.zip, https://molgenis26.gcc.rug.nl/downloads/biosqtlbrowser/2015_09_02_trans_meQTLsFDR0.05-CpGLevel.txt, https://molgenis26.gcc.rug.nl/downloads/biosqtlbrowser/2015_09_02_cis_eQTMsFDR0.05-CpGLevel.txt), GTEx version 6 eQTL results (https://storage.googleapis.com/gtex_analysis_v6/single_tissue_eqtl_data/GTEx_Analysis_V6_eQTLs.tar.gz), eQTLGen *cis* eQTL results (https://molgenis26.gcc.rug.nl/downloads/eqtlgen/cis-eqtl/cis-eQTLs_full_20180905.txt.gz), TWAS hub (http://twas-hub.org/genes/UBASH3B/), GWAS summary statistics of 114 traits for colocalization analysis (https://zenodo.org/record/3629742), ChIP–seq binding sites (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredWithCellsV3.bed.gz, http://tagc.univ-mrs.fr/remap/download/All/filPeaks_public.bed.gz), chromHMM states (http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/all.mnemonics.bedFiles.tgz), Hi-C data (EGAD00001003106), PPIs (http://string90.embl.de/newstring_download/protein.links.detailed.v9.0.txt.gz). Source data are provided with this paper.

Code availability

Code for the analysis is available at GitHub (https://github.com/heiniglab/hawe2021_meQTL_analyses) and also through Zenodo

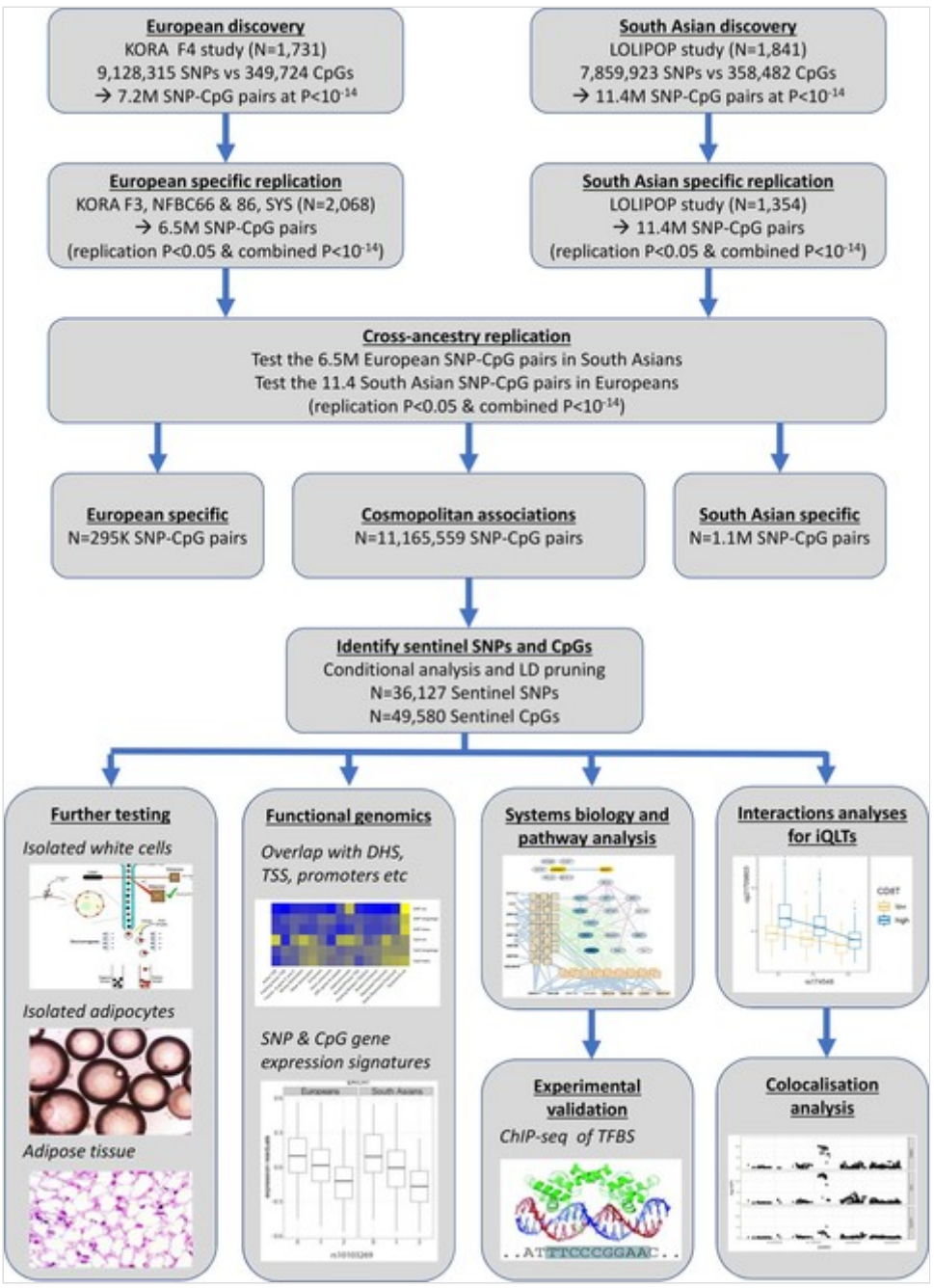
Competing interests The authors declare no competing interests. AQ32

Extended data

Extended Data Fig. 1

Study design.

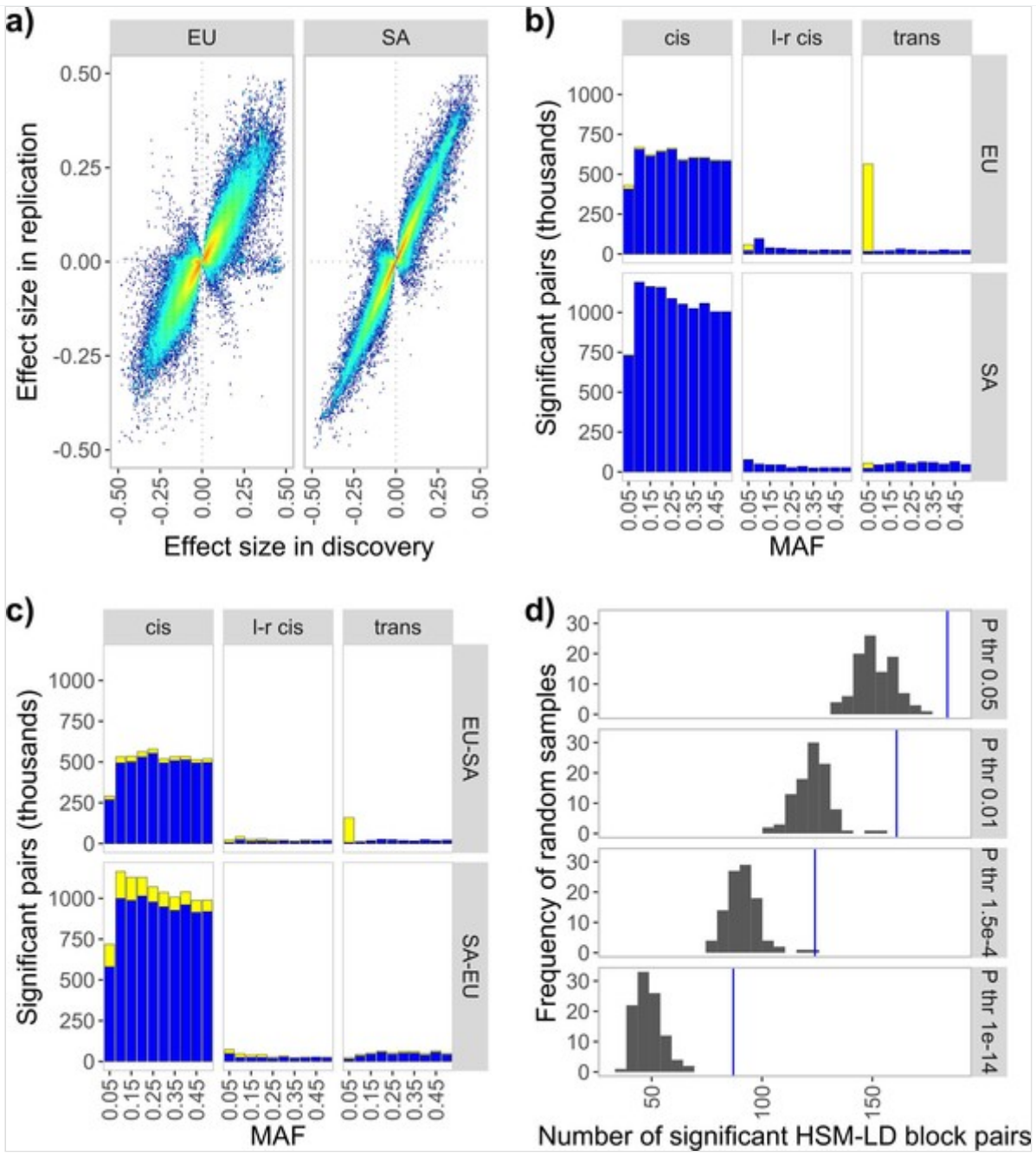
Overview of study design.



Extended Data Fig. 2

Replication testing of meQTLs within and across ancestries.

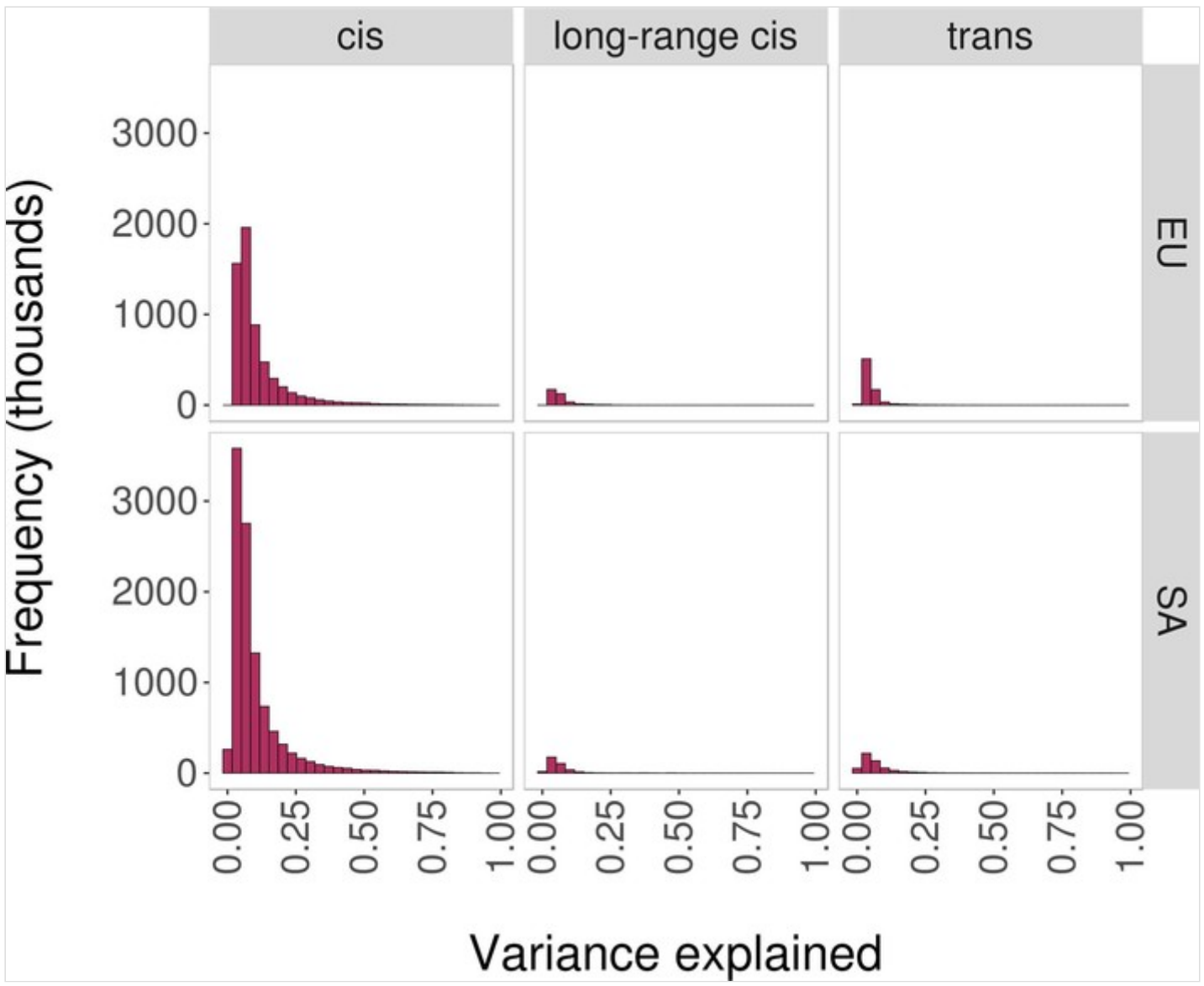
a: Ancestry-specific replication of SNP-CpG pairs identified by genome-wide association. Effect size: change in methylation (0-1 scale) per allele copy of the SNP. Axes set to [-0.5,0.5]. **b:** Ancestry-specific replication, by pair proximity and MAF. Bars: no. of pairs identified in discovery in given category. Blue: replicated; yellow: not replicated. **c:** Cross-ancestry replication, by pair proximity and MAF. Top: discovery in EU, replication in SA; bottom: discovery in SA, replication in EU. Bars: no. of pairs identified in discovery in given category. Blue: replicated; yellow: not replicated. **d:** Cross-platform: replication in KORA F4 (N<=1731) of published MeDIP-seq meQTLs, by significance threshold. Blue lines: no. of replicated results (of 328); histograms: no. of replicated results over 100 randomly selected matched datasets. P-values: one-sided, no adjustment for multiple testing. See Methods for test description. EU: European; SA: South Asian; MAF: Minor allele frequency.



Extended Data Fig. 3

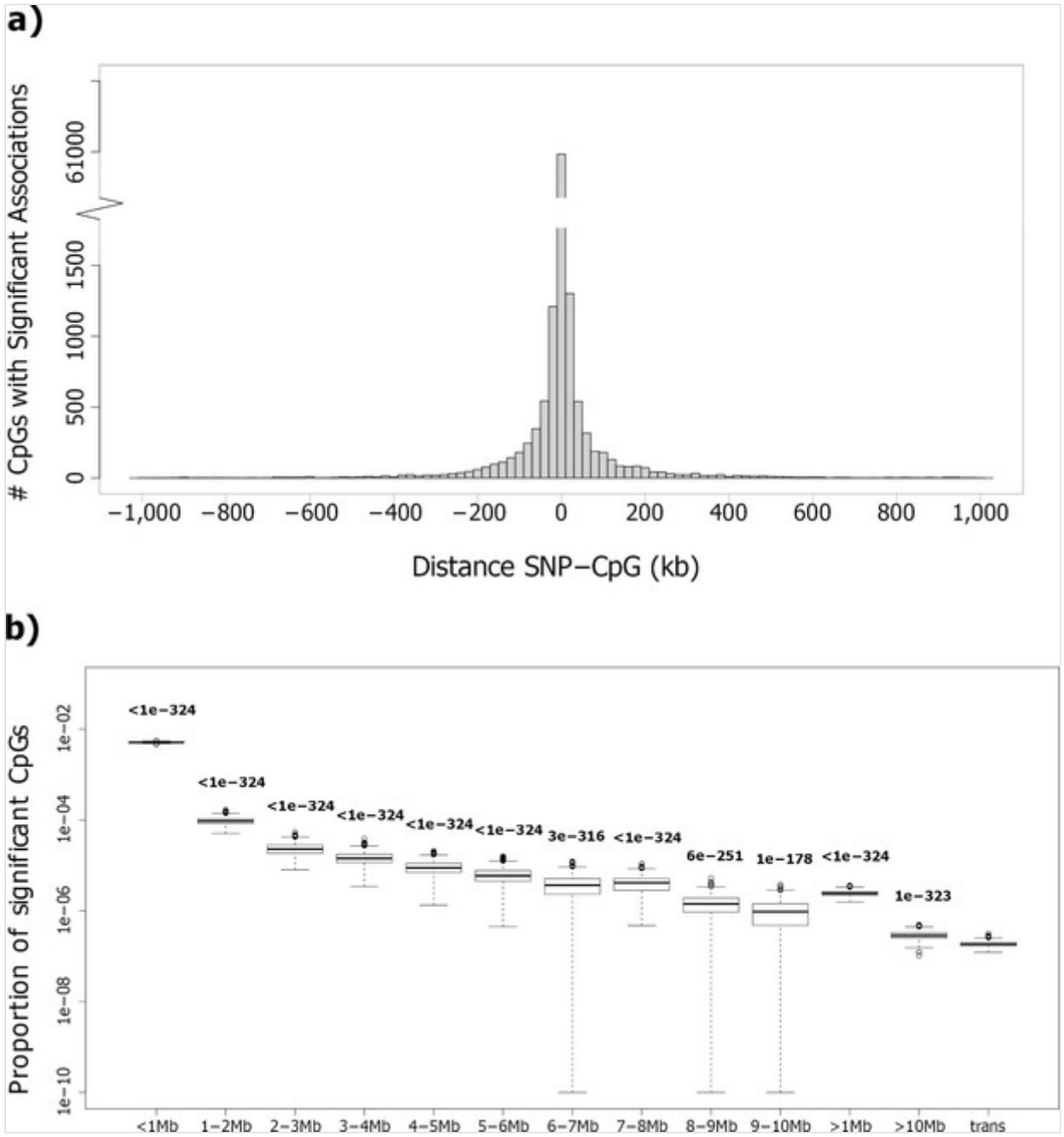
Variance in DNA methylation explained by meQTL SNPs.

Histograms showing the proportions of variance of DNA methylation explained by genetic variants in both populations when variants are located in *cis* (left), long-range *cis* (middle) or *trans* (right) of the associated CpG site. EU: European; SA: South Asian.



Extended Data Fig. 4

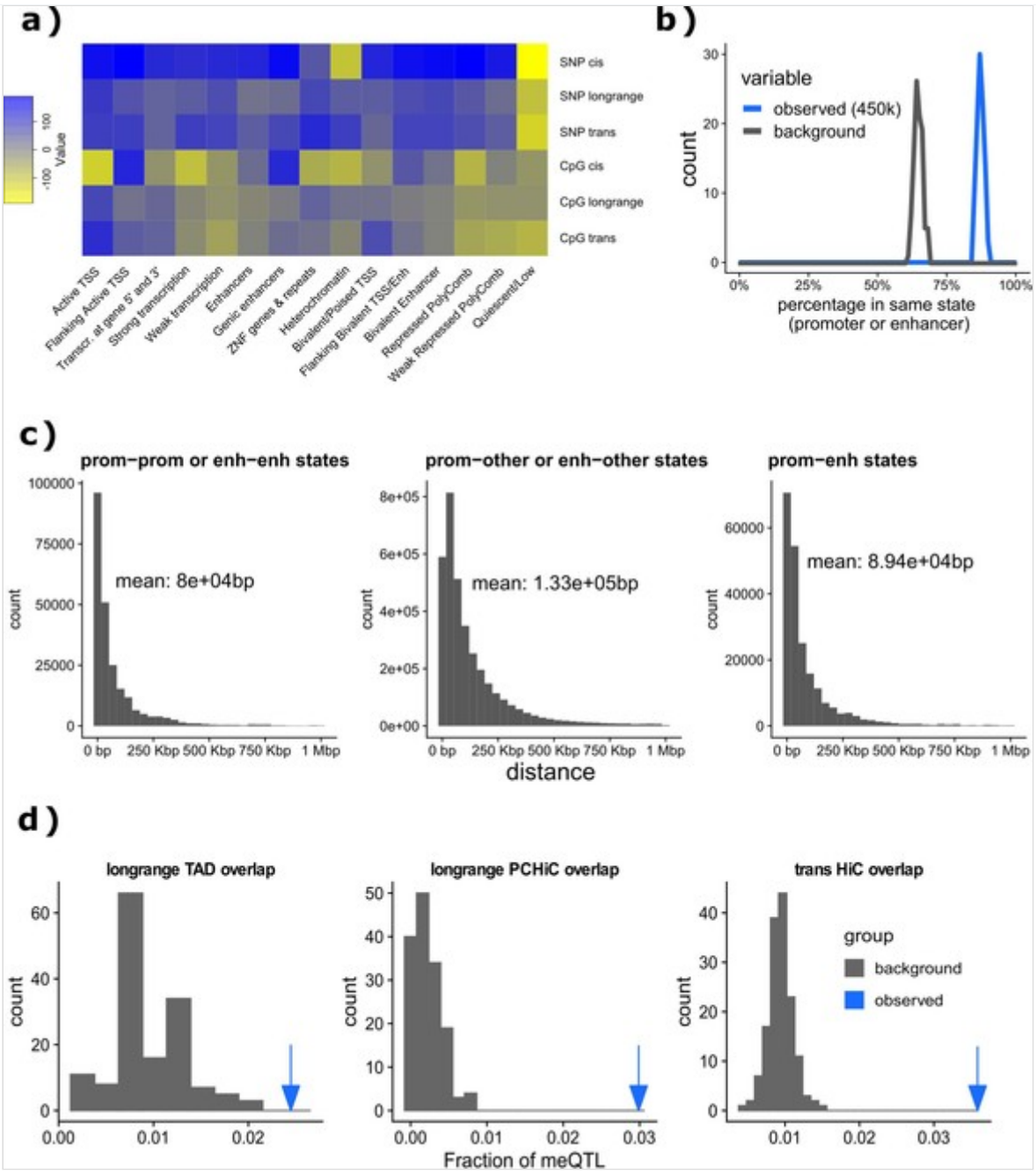
Panel a: Histogram showing for each CpG site the genomic distance between CpG and the closest associated SNP from the cosmopolitan set of 10,346,172 SNP-CpG pairs identified in cis (association confirmed in both Europeans and South Asians). **Panel b**: Boxplots showing the proportion of SNP-CpG pairs that reach genome-wide significance for different distance categories (x-axis), compared to SNP-CpG pairs on different chromosomes (trans).1,000 random samples of 10,000 SNPs were taken. P-values above each box are based on a comparison (one-sided t-test) between the proportion of SNP-CpG pairs in trans that reach significance, and the proportion that reach significance in the respective same-chromosome distance window. Boxplots show medians (center lines), first and third quartiles (lower and upper box limits, respectively), 1.5-fold interquartile ranges (whisker extents) and outliers (black circles).



Extended Data Fig. 5

Functional genomic context of meQTL SNPs and CpGs.

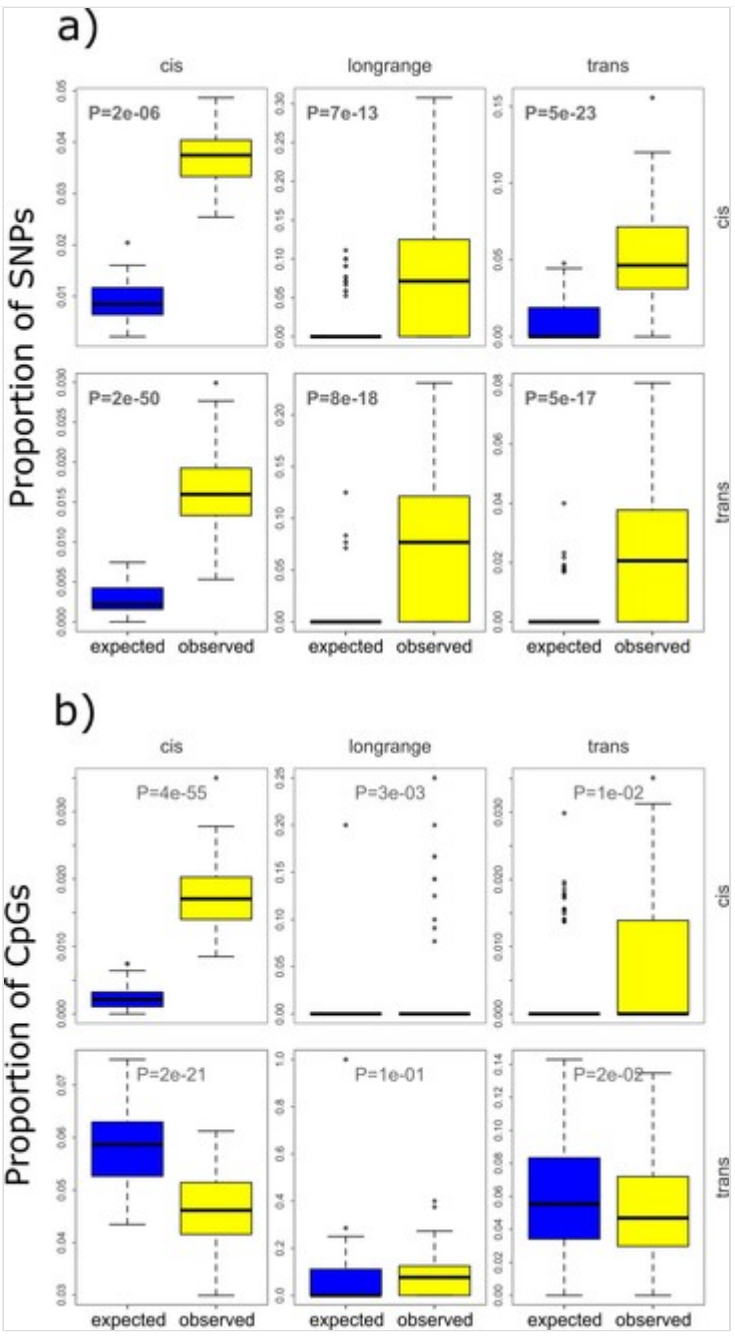
Panel a: Genomic overlap between chromatin state annotations (15- state model; Roadmap Epigenomics Project and SNPs/CpGs identified by genome-wide association and cross-ancestry replication testing. Results are presented as a heatmap showing the P-values for enrichment (blue) or depletion (yellow) in the respective chromatin state (two-sided t-test). P-values have been Bonferroni-adjusted for the total number of tests (see Methods for details). **Panel b:** Colocalisation of SNPs and CpG sites in promoter and enhancer chromatin states. The histograms show the frequency at which CpG sites that localise in promoter or enhancer chromatin states have at least one *cis*-meQTL SNP that localises to the same chromatin state. Observed (turquoise) *cis*-meQTL pairs colocalise to the same chromatin state more frequently than matched background SNP-CpG pairs (grey). **Panel c:** Distance distributions for *cis* SNP-CpG pairs 1) localising to the same state (left), 2) where one entity localises to a promoter/enhancer state and the other to neither promoter nor enhancer state (center) and 3) one entity localises to a promoter and the other to an enhancer state. **Panel d:** Overlap of SNP-CpG associations with chromatin contacts in primary cells. The x-axis shows the fraction of SNP-CpG pairs that localise within the same topologically associated domain (TAD, left panel) or that overlap with Hi-C contacts (center and right panels). The left panel shows localisation of long-range cis-meQTLs within the same TAD. The center panel shows the overlap of long range *cis*-meQTLs (same chromosome, distance SNP - CpG > 1Mb) with contacts from promoter capture Hi-C (PCHi-C). The right panel shows overlap of trans-meQTL with Hi-C contacts. The blue vertical arrows indicate the overlap observed in the data. The grey histograms show the distribution of the fraction of randomly sampled SNP-CpG pairs overlapping contact regions for each category.



Extended Data Fig. 6

Enrichment of meQTL SNPs and CPGs for association with gene expression.

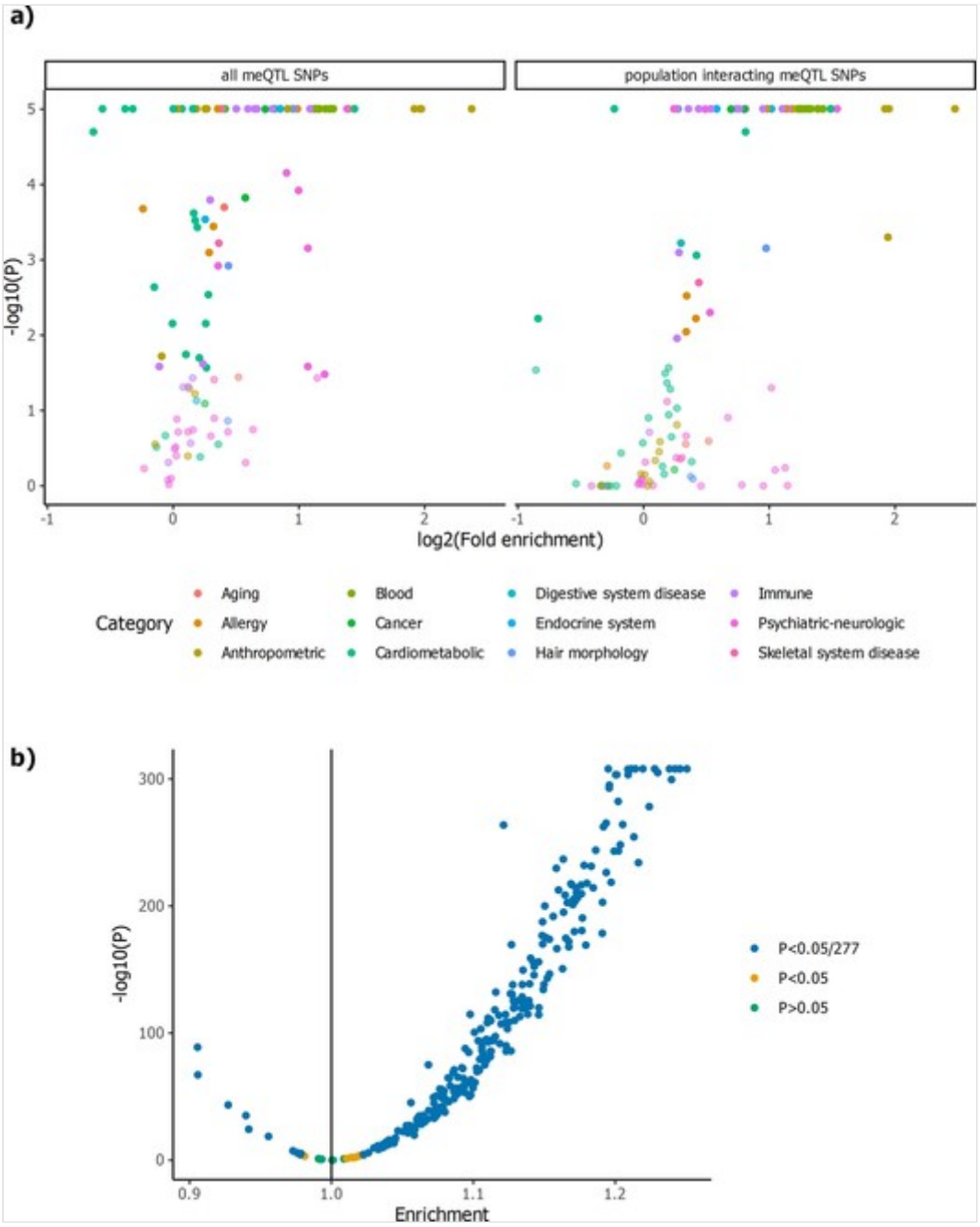
Sentinel meQTL SNPs and CpGs are enriched for association with gene expression in cis and trans (SNPs) and only in cis (CpGs). **Panel a:** Results are presented as the proportion of SNPs that are observed to be associated with gene expression in *cis* (top row) or in *trans* (bottom row), stratified by proximity between SNP and CpG for the respective SNP-CpG pair (*cis*, long-range *cis* and *trans* from left to right). **Panel b:** Similarly, results are presented as the proportion of CpGs that are observed to be associated with gene expression in *cis* (top row) or in *trans* (bottom row), stratified by proximity between SNP and CpG for the respective SNP-CpG pair (*cis*, long-range *cis* and *trans* from left to right). **Both panels:** In each plot, the observed proportion (yellow boxplots) is compared to the proportion expected under the null hypothesis based on permutation testing (blue boxplots, see Methods). Inset in each figure is the P-value for comparison between observed and expected proportions (t-test). Boxplots show medians (center lines), first and third quartiles (lower and upper box limits, respectively), 1.5-fold interquartile ranges (whisker extents) and outliers (black circles). Proportions were calculated based on 100 sets of permutations with 1,000 SNPs (Panel A) or 1,000 CpGs (Panel B) in each permutation.



Extended Data Fig. 7

Enrichment of meQTL SNPs and CpGs for associations with phenotypic traits.

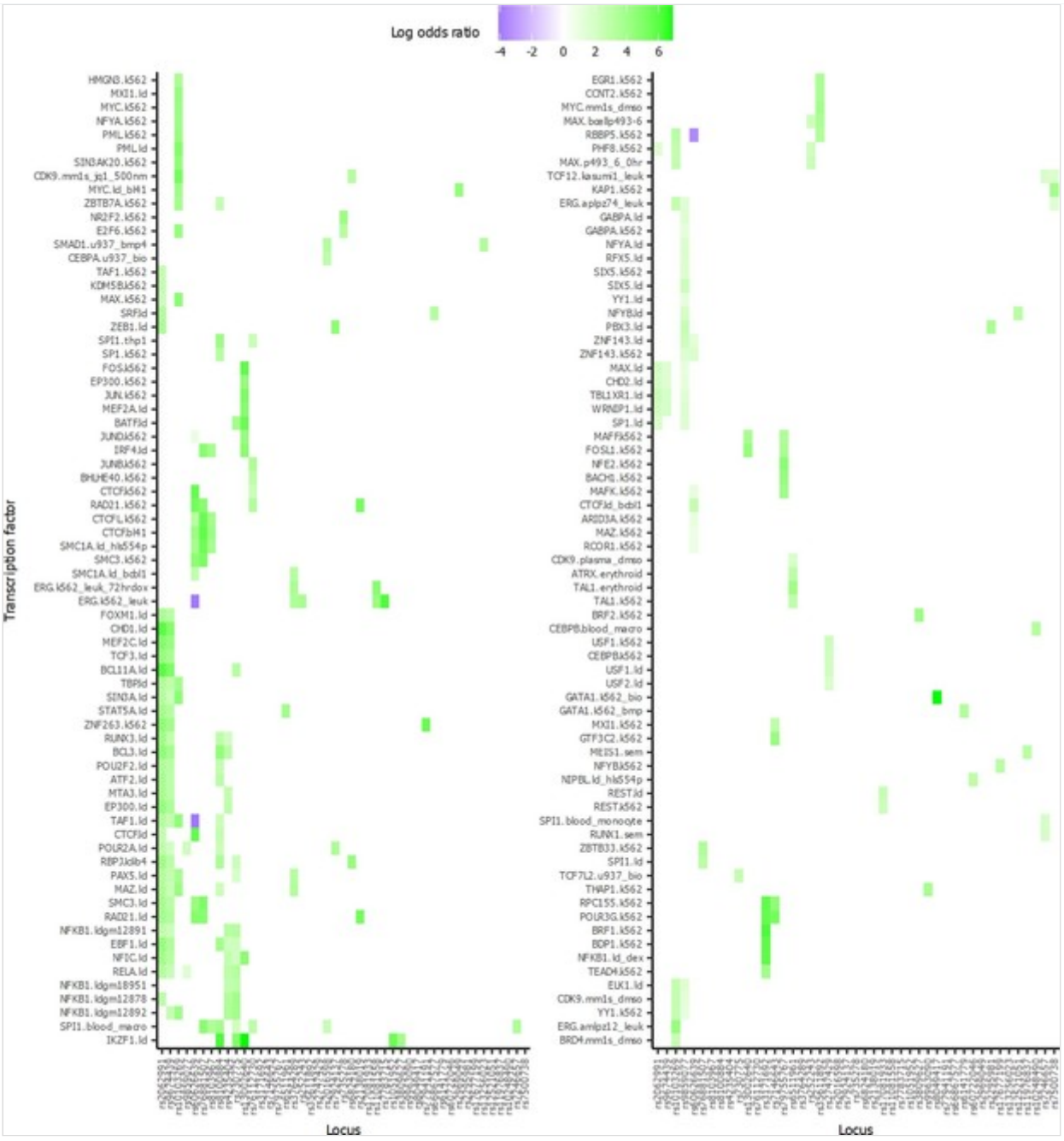
(A) SNPs influencing DNA methylation (left panel) and SNPs identified to be population interacting meQTL based on our cosmopolitan discovery analysis (right panel) are both enriched for association with phenotypic traits, Analysis carried out using using QTLEnrich and 114 uniformly processed GWAS summary statistics. The volcano plot shows the log2 fold enrichment of significant GWAS hits among iQTL on the x-axis and the -log10 of the P-value of the enrichment test on the y-axis. Each point represents one of 114 GWAS studies. The transparency of the fill colour indicates the false discovery rate (FDR < 5%: no transparency). (B) Sentinel CpGs are enriched for clinical and metabolic traits. We tested the Sentinel CpGs for association with 277 available clinical and metabolic traits (NMR metabolomics). We used permutation testing to generate expectations under the null hypothesis, and to determine both the magnitude and probability for enrichment. Results show strong evidence that our genetically regulated Sentinel CpGs are enriched for association with traits (enriched at P<0.05/277 for 252 phenotypes) with median enrichment 1.10 (IQR: 1.06-1.15).



Extended Data Fig. 8

CpG sites associated with *trans*-acting sentinel SNPs are enriched for location in transcription factor binding sites.

Heatmap showing the enrichment (or depletion) of CpG sites for *trans*-acting sentinel SNPs (x-axis) with the DNA binding sites of known transcription factors (y-axis). Log2 odds ratios compare the frequency of overlap for the CpGs associated with the respective SNP, compared to the background frequency of overlap for all tested CpG sites. Results are shown for the 45 sentinel SNPs that show evidence for overlap with known transcription factor binding sites (out of the 115 tested *trans*-acting sentinel SNPs with at least five associated CpG sites).

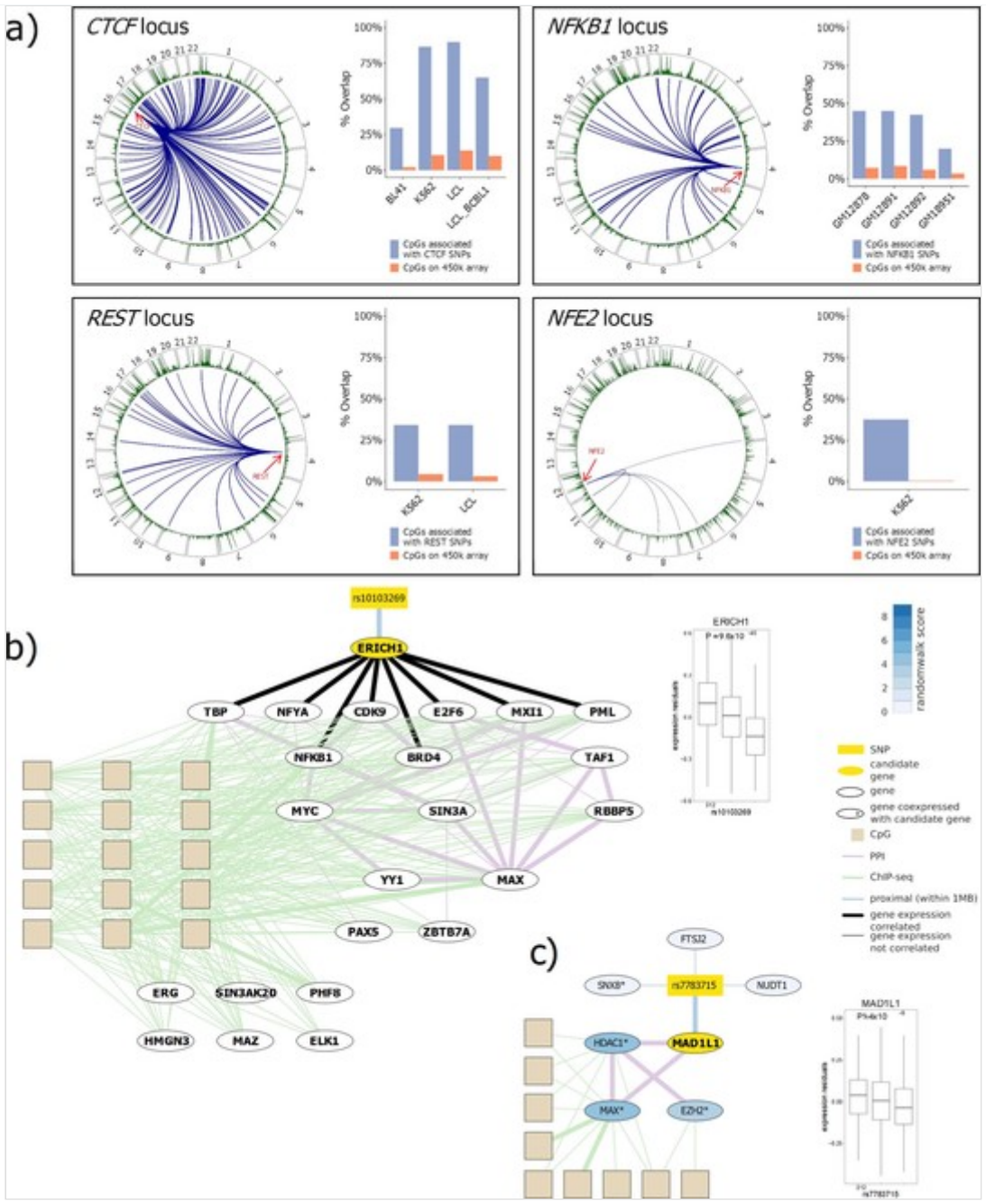


Extended Data Fig. 9

Trans-acting regulatory networks at the *CTCF*, *NFKB1*, *REST*, *NFE2*, *MAD1L1* and *ENRICH1* loci.

(a) Circos plots summarising i. genomic distribution of CpGs associated in *trans* [inner connections], and ii. known DNA binding sites of transcription factor encoded in *cis* [outer ring], for sentinel SNPs at *CTCF*, *NFKB1*, *REST* and *NFE2* loci. Inset are observed and expected proportions of CpG sites that overlap respective DNA binding sites as available for different cell lines (see Methods). FDR < 1.17×10^{-2} for all cell lines and transcription factors. (b) Regulatory network of *ERICH1* locus illustrating the connection between SNP rs10103269 (yellow rectangle) and expression of identified candidate gene *ERICH1* (yellow ellipse), which is connected through protein-protein and protein-DNA interactions to methylation at *trans*-associated CpG sites (beige rectangles). Ellipses represent genes encoded at the genetic locus identified by the sentinel or that are part of the protein-protein interaction network. Genes marked with an asterisk (*) show co-expression with the candidate gene. Bold gene names indicate a strong genetic effect of the sentinel on the expression of that gene (eQTL). Fill colour of ellipses represent the random walk score (colour bar legend). The colour of edges connecting genes and CpG sites represent: i. protein-protein interactions (purple), ii. protein-DNA interactions identified by TFBS overlap (green), and iii. proximity (distance < 1 Mb) between genes and SNPs or CpG sites (blue). The thickness of edges represents correlation with gene expression (thick) or no correlation of/with gene expression (thin). Boxplot shows the effect of sentinel SNP (rs10103269) in *cis* on expression of *ERICH1* with the p-value from linear regression of expression ~ genotype (n=1,546 biologically independent samples combined from both cohorts). Center line indicates median, lower and upper box limits correspond to the first and third quartiles, respectively; whisker extent indicates 1.5-fold interquartile range; outliers not shown. (c) *MAD1L1* locus pathway analysis. Annotations and symbols are as described in (b).

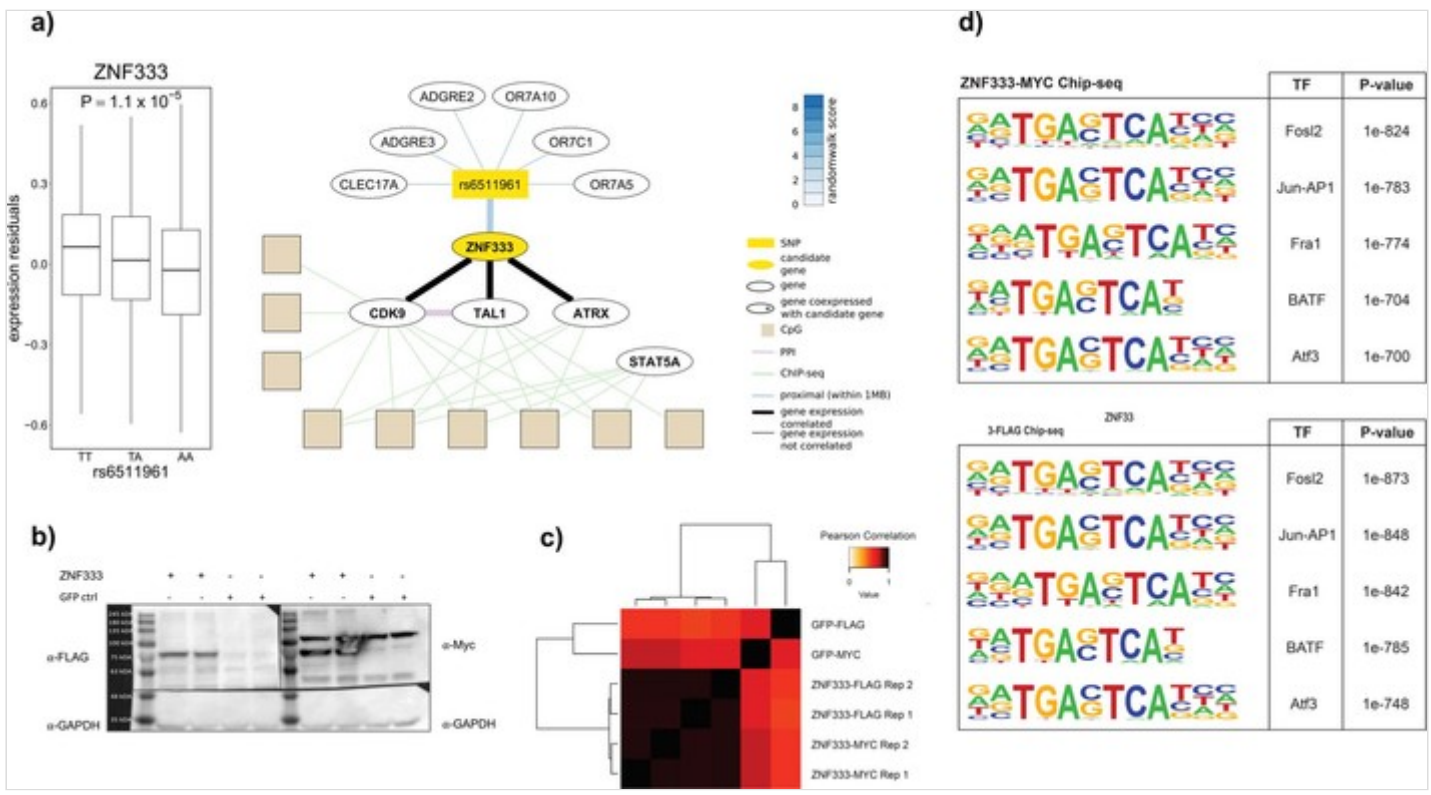
Source data



Extended Data Fig. 10

Experimental validation at the ZNF333 locus.

Panel a. Regulatory network of the *ZNF333* locus. Annotations and symbols are as described in Extended Figure 9. The boxplot shows the effect of sentinel SNP (rs6511961) in cis on expression of the candidate gene *ZNF333* with the p-value from the linear regression of expression ~ genotype (n=1,546 biologically independent samples combined from both cohorts). **Panels b-d**. HCT116 cells were transfected with ZNF333-FLAG/Myc tagged or GFP-control plasmids in biological replicates. **Panel b**. Protein lysates were Western blotted for ZNF333 expression using FLAG or MYC antibodies as validations. GAPDH was used as loading control (n=2). Source data: Membranes were cut into three pieces for optimisation of exposure. Top left panel: Original uncropped and unprocessed scans. Top right panel: Scan exposure optimized for molecular ladder. Bottom left panel: Scan exposure optimized for GAPDH. Bottom right panel: Final overlay figure. **Panel c**. Heatmap showing the Pearson correlation between ChIP-seq performed for ZNF333 using either FLAG or MYC antibodies. **Panel d**. Motifs of known TFs enriched in ZNF333 binding sites showing perfect overlap between ChIP with FLAG and MYC antibodies.



Supplementary information

Supplementary Information

Supplementary Note and Figs. 1–9

Reporting Summary

Supplementary Tables

Supplementary Tables 1–41.

Source data

Source Data Extended Data Fig. 9

Unprocessed data for ZNF333 ChIP–seq experiment.

References

1. Bird, A. Perceptions of epigenetics. *Nature* **447**, 396–398 (2007).

2. Schubeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).

3. Parry, A., Rulands, S. & Reik, W. Active turnover of DNA methylation during cell fate decisions. *Nat. Rev. Genet.* **22**, 59–66 (2021).

4. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33**, 245–254 (2003).

5. Chambers, J. C. et al. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case–control study. *Lancet Diabetes Endocrinol.* **3**, 526–534 (2015).

6. Marioni, R. E. et al. DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* **16**, 25 (2015).

7. van der Harst, P., de Windt, L. J. & Chambers, J. C. Translational perspective on epigenetics in cardiovascular disease. *J. Am. Coll. Cardiol.* **70**, 590–606 (2017).

8. Wahl, S. et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).

9. Zhang, Y. et al. DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat. Commun.* **8**, 14617 (2017).

10. Sugiura, M. et al. Epigenetic modifications in prostate cancer. *Int. J. Urol.* **28**, 140–149 (2020).

11. Blokhin, I. O., Khorkova, O., Saveanu, R. V. & Wahlestedt, C. Molecular mechanisms of psychiatric diseases. *Neurobiol. Dis.* **146**, 105136 (2020).

12. Darwiche, N. Epigenetic mechanisms and the hallmarks of cancer: an intimate affair. *Am. J. Cancer Res.* **10**, 1954–1978 (2020).

13. Bonder, M. J. et al. Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC Genomics* **15**, 860 (2014).

14. Bonder, M. J. et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).

15. Gibbs, J. R. et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6**, e1000952 (2010).

16. Grundberg, E. et al. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.* **93**, 876–890 (2013).

17. Gutierrez-Arcelus, M. et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2**, e00523 (2013).

18. Lemire, M. et al. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat. Commun.* **6**, 6326 (2015).

19. Huan, T. et al. Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. *Nat. Commun.* **10**, 4267 (2019).

20. Hannon, E. et al. Leveraging DNA-methylation quantitative-trait loci to characterize the relationship between methylomic variation, gene expression, and complex traits. *Am. J. Hum. Genet.* **103**, 654–665 (2018).

21. Gaunt, T. R. et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.* **17**, 61 (2016).

22. McRae, A. F. et al. Identification of 55,000 replicated DNA methylation QTL. *Sci. Rep.* **8**, 17605 (2018).

23. Hop, P. J. et al. Genome-wide identification of genes regulating DNA methylation using genetic anchors for causal inference. *Genome Biol.* **21**, 220 (2020).

24. Peterson, R. E. et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).

25. Bell, C. G. et al. Obligatory and facilitative allelic variation in the DNA methylome within common disease-associated loci. *Nat. Commun.* **9**, 8 (2018).

26. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

27. Brenner, C. et al. Myc represses transcription through recruitment of DNA methyltransferase corepressor. *EMBO J.* **24**, 336–346 (2005).

28. Esteve, P. O., Chin, H. G. & Pradhan, S. Human maintenance DNA (cytosine-5)-methyltransferase and p53 modulate expression of p53-repressed promoters. *Proc. Natl Acad. Sci. USA* **102**, 1000–1005 (2005).

29. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).

30. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).

31. Javierre, B. M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384 (2016).

32. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

33. Liu, Y., Toh, H., Sasaki, H., Zhang, X. & Cheng, X. An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence. *Genes Dev.* **26**, 2374–2379 (2012).

34. Shi, H. et al. ZFP57 regulation of transposable elements and gene expression within and beyond imprinted domains. *Epigenetics Chromatin* **12**, 49 (2019).

35. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).

36. Lee, S. T. et al. Protein tyrosine phosphatase UBASH3B is overexpressed in triple-negative breast cancer and promotes invasion and metastasis. *Proc. Natl Acad. Sci. USA* **110**, 11121–11126 (2013).

37. Pulit, S. L. et al. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).

38. Kichaev, G. et al. Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).

39. Zhu, Z. et al. Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. *J. Allergy Clin. Immunol.* **145**, 537–549 (2020).

40. Richardson, T. G. et al. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: a multivariable Mendelian randomisation analysis. *PLoS Med.* **17**, e1003062 (2020).

41. Konieczna, J., Sanchez, J., Palou, M., Pico, C. & Palou, A. Blood cell transcriptomic-based early biomarkers of adverse programming effects of gestational calorie restriction and their reversibility by leptin supplementation. *Sci. Rep.* **5**, 9088 (2015).

42. Mancuso, N. et al. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).

43. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

44. Emery, P. et al. IL-6 receptor inhibition with tocilizumab improves treatment outcomes in patients with rheumatoid arthritis refractory to anti-tumour necrosis factor biologicals: results from a 24-week multicentre randomised placebo-controlled trial. *Ann. Rheum. Dis.* **67**, 1516–1523 (2008).

45. Navarro-Millan, I., Singh, J. A. & Curtis, J. R. Systematic review of tocilizumab for rheumatoid arthritis: a new biologic agent targeting the interleukin-6 receptor. *Clin. Ther.* **34**, 788–802 (2012).

46. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLoS Genet.* **13**, e1006646 (2017).

47. Burnichon, N. et al. MAX mutations cause hereditary and sporadic pheochromocytoma and paraganglioma. *Clin. Cancer Res.* **18**, 2828–2837 (2012).

48. Li, H. et al. Novel treatment of hypertension by specifically targeting E2F for restoration of endothelial dihydrofolate reductase and eNOS function under oxidative stress. *Hypertension* **73**, 179–189 (2019).

49. Burstein, E. et al. COMMD proteins, a novel family of structural and functional homologs of MURR1. *J. Biol. Chem.* **280**, 22222–22232 (2005).

50. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).

51. Suhail, A. et al. DeSUMOylase SENP7-mediated epithelial signaling triggers intestinal inflammation via expansion of γδ T cells. *Cell Rep.* **29**, 3522–3538 (2019).

52. Jing, Z., Liu, Y., Dong, M., Hu, S. & Huang, S. Identification of the DNA binding element of the human ZNF333 protein. *J. Biochem. Mol. Biol.* **37**, 663–670 (2004).

53. Chen, M. H. et al. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell* **182**, 1198–1213 (2020).

54. Nedelec, Y. et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell* **167**, 657–669 (2016).

55. Joehanes, R. et al. Epigenetic signatures of cigarette smoking. *Circ. Cardiovasc. Genet.* **9**, 436–447 (2016).

56. Singmann, P. et al. Characterization of whole-genome autosomal differences of DNA methylation between men and women. *Epigenetics Chromatin* **8**, 43 (2015).

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

58. Giri, A. K. et al. DNA methylation profiling reveals the presence of population-specific signatures correlating with phenotypic characteristics. *Mol. Genet. Genomics* **292**, 655–662 (2017).

59. Breeze, C. E. et al. eFORGE: a tool for identifying cell type-specific signal in epigenomic data. *Cell Rep.* **17**, 2137–2150 (2016).

60. Westra, H. J. et al. Cell specific eQTL analysis without sorting cells. *PLoS Genet.* **11**, e1005223 (2015).

61. Guan, W. et al. Genome-wide association study of plasma N^6 polyunsaturated fatty acids within the cohorts for heart and aging research in genomic epidemiology consortium. *Circ. Cardiovasc. Genet.* **7**, 321–331 (2014).

62. Shin, S. Y. et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).

63. Kamat, M. A. et al. PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).

64. Gelfand, E. W. & Dakhama, A. $CD8^+$ T lymphocytes and leukotriene B4: novel interactions in the persistence and progression of asthma. *J. Allergy Clin. Immunol.* **117**, 577–582 (2006).

65. Cho, S. H., Stanciu, L. A., Holgate, S. T. & Johnston, S. L. Increased interleukin-4, interleukin-5, and interferon-γ in airway $CD4^+$ and $CD8^+$ T cells in atopic asthma. *Am. J. Respir. Crit. Care Med.* **171**, 224–230 (2005).

66. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).

67. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).

68. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

69. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).

70. GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

71. Kim, K. A. et al. Environmental risk factors and comorbidities of primary biliary cholangitis in Korea: a case–control study. *Korean J. Intern. Med.* **36**, 313–321 (2020).

72. GTEx Consortium The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

73. Staley, J. R. et al. PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).

74. Griffon, A. et al. Integrative analysis of public ChIP–seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.* **43**, e27 (2015).

75. Franceschini, A. et al. STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2013).

76. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).

77. Haghverdi, L., Buttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).

78. Schramm, K. et al. Mapping the genetic architecture of gene regulation in whole blood. *PLoS ONE* **9**, e93844 (2014).

79. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* **125**, 279–284 (2001).

80. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

81. Taylor-Weiner, A. et al. Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).

82. Hawe, J. S., Heinig, M. & Loh, M. Code for the analyses described in Hawe et al. Nature Genetics. *Zenodo* https://doi.org/10.5281/zenodo.5529828 (2021). **AQ33**