

## **Supplementary Online Material**

### **Promoter-anchored chromatin interactions predicted from genetic analysis of epigenomic data**

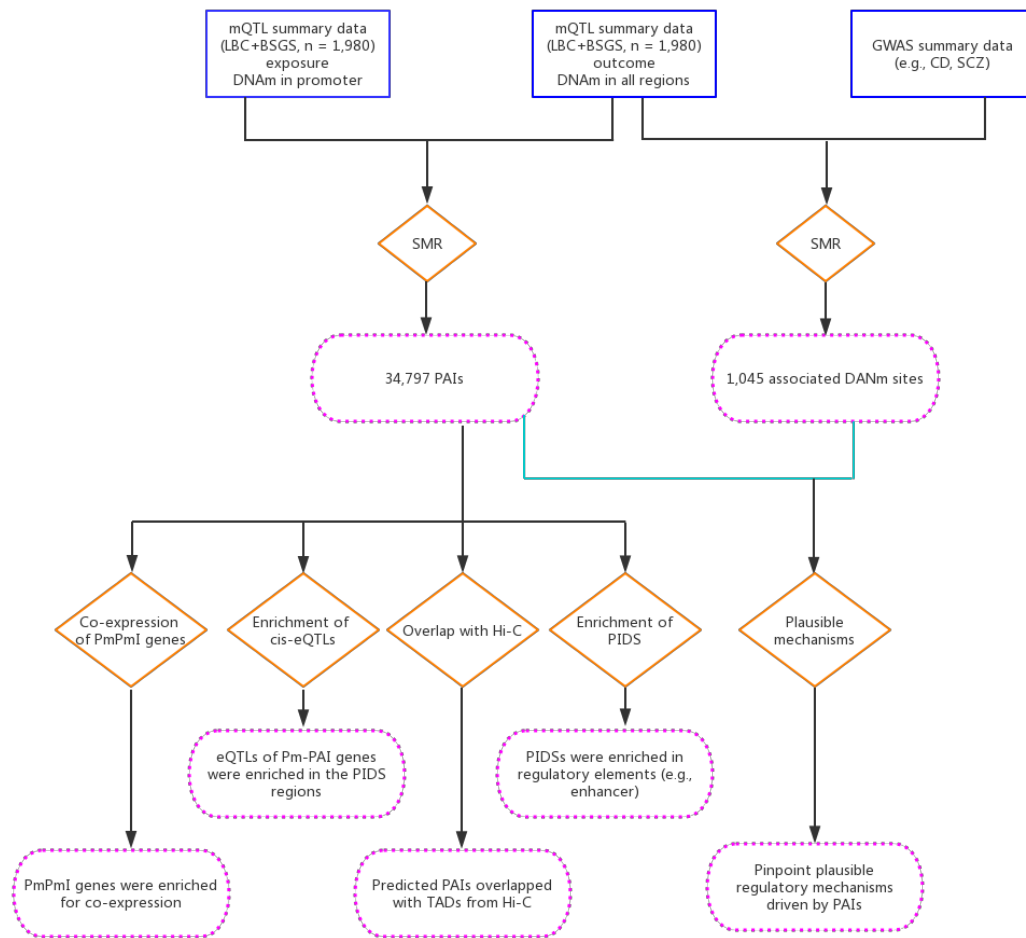
*Wu et al.*

#### **Contents**

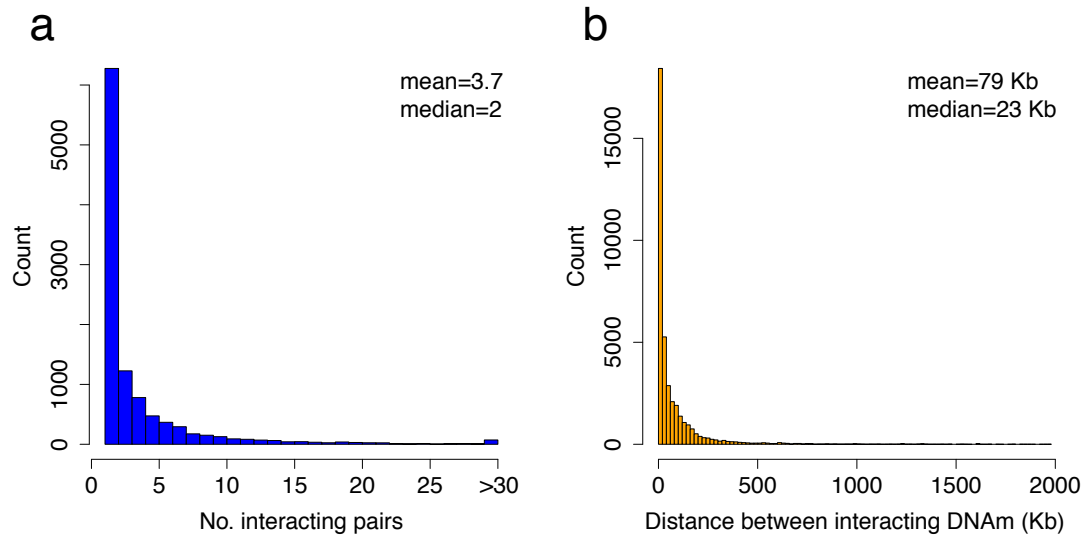
**Figure S1 to S8**

**Supplementary Note 1-2**

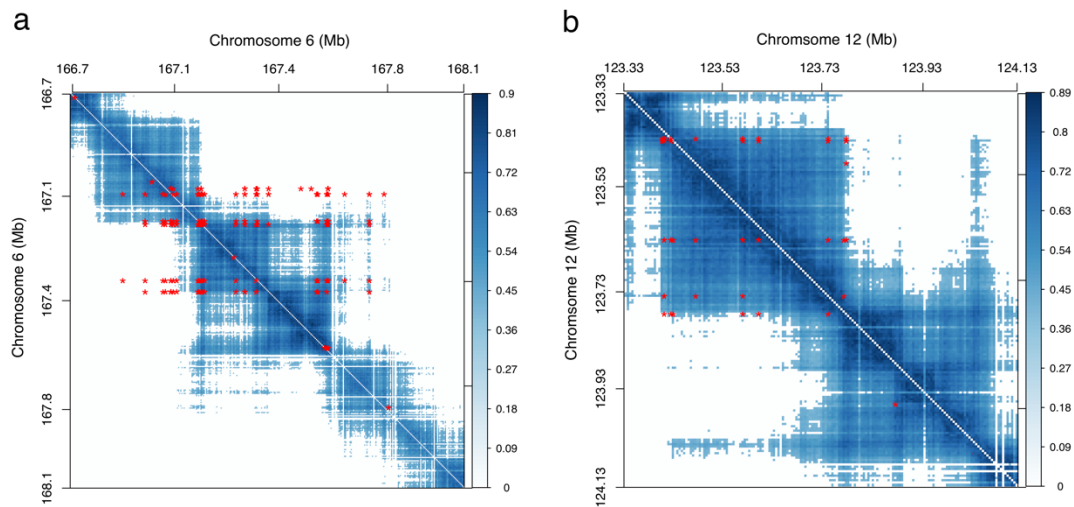
**References**



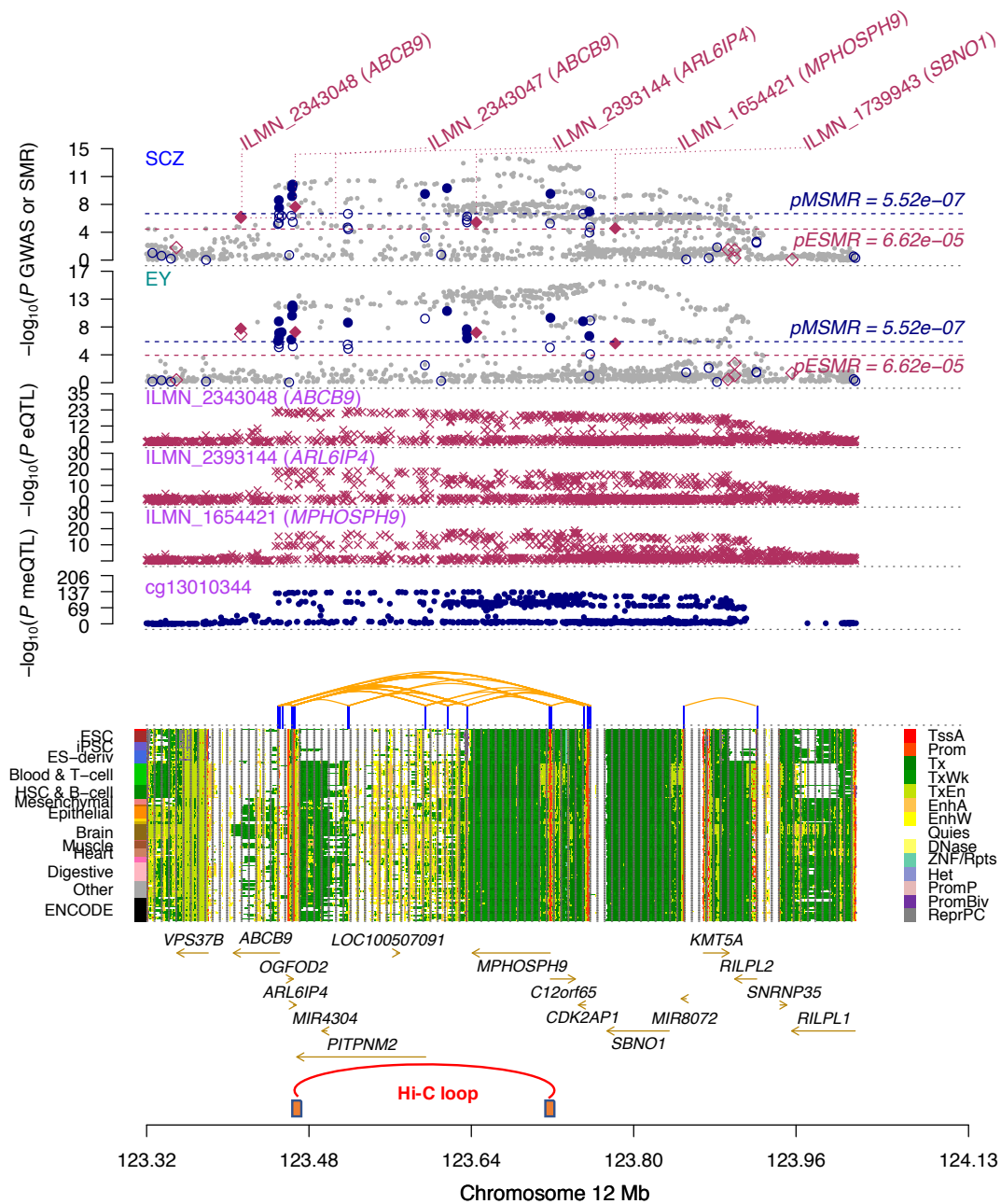
**Figure S1** Schematic overview of this study.



**Figure S2** Summary of the predicted PAIs. Panel a): distribution of the number of PIDSs (promoter interacting DNAm sites) for each bait probe (located in the promoter of a gene). Panel b): distribution of physical distances between pairwise interacting DNAm sites of the significant PAIs.

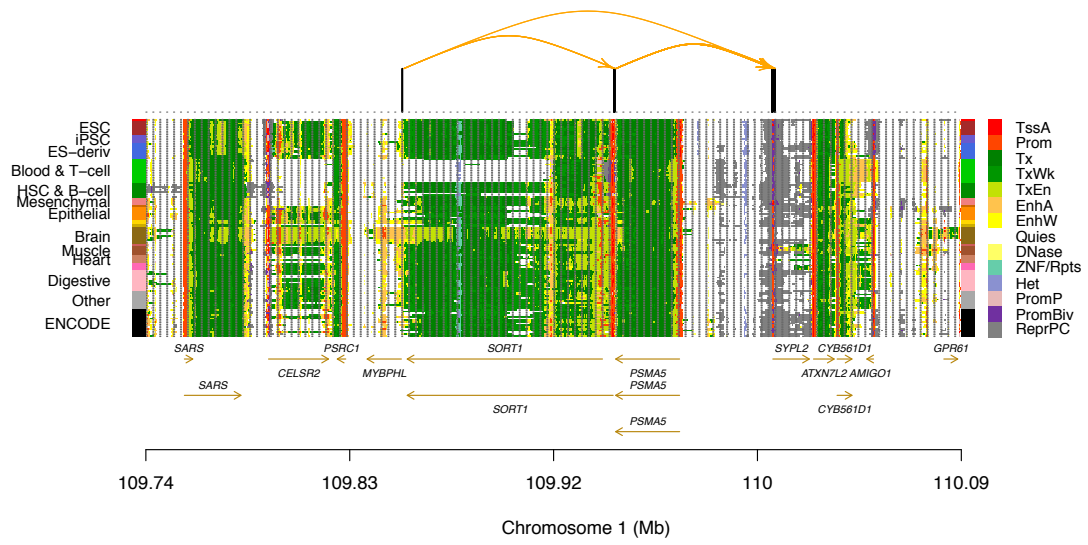


**Figure S3** Overlap of the predicted PAIs with TADs annotated from the Rao et al. <sup>1</sup> Hi-C data. Panel a): a heatmap of the predicted PAIs (red asterisks) and chromatin interactions with correlation score > 0.4 (blue dots) identified by Hi-C in a 1.38 Mb region on chromosome 6. Only 41.5% of the predicted PAIs in this region showed overlap with the TADs. This region harbours the *RPS6KA2* locus as shown in **Fig. 5**. Panel b): a heatmap of the predicted PAIs (red asterisks) and chromatin interactions with correlation score > 0.4 (blue dots) identified by Hi-C in a 0.81 Mb region on chromosome 12. The predicted PAIs were highly consistent with the chromatin interactions identified by Hi-C. This region harbours the *ABCB9* locus as shown in **Fig. S4**. The heatmap is asymmetric for the PAIs with the x- and y-axes representing the physical positions of “outcome” and “exposure” probes respectively.

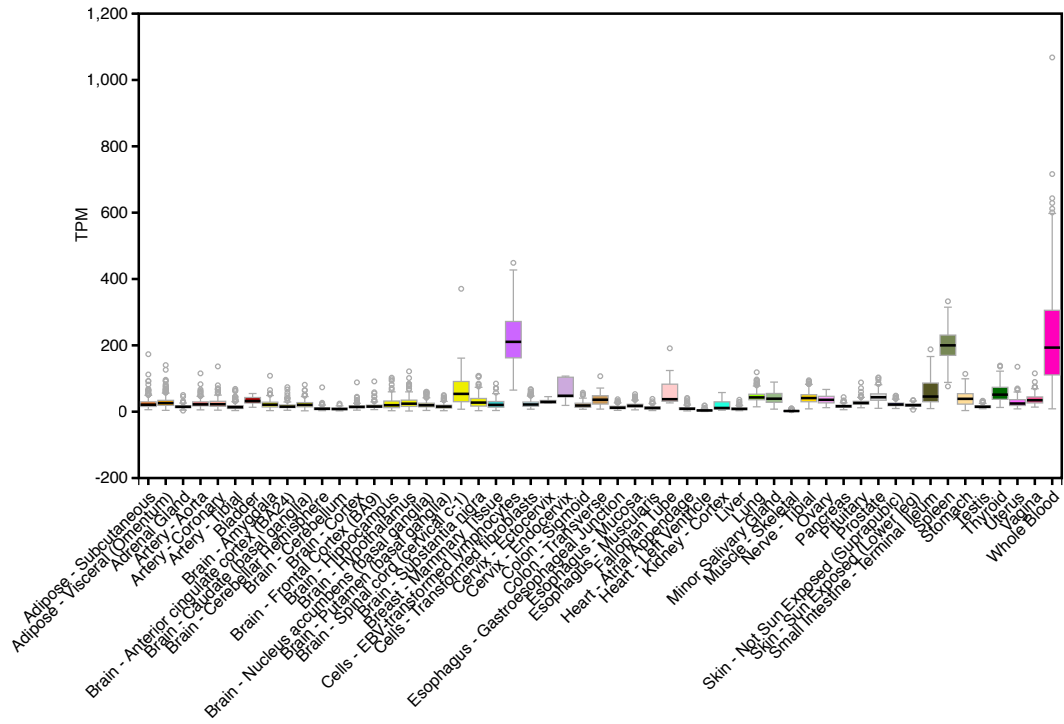


**Figure S4** A shared PIDS region with eQTLs predicted to interact with the promoters of multiple genes (i.e., *ABCB9*, *ARL6IP4*, *MPHOSPH9*). The top two plots show  $-\log_{10}(P \text{ values})$  of SNPs from the GWAS meta-analyses (grey dots) for schizophrenia (SCZ) and educational years (EY). Red diamonds and blue circles represent  $-\log_{10}(P \text{ values})$  from SMR tests for associations of gene expression and DNAm with SCZ and EY, respectively. Solid diamonds and circles are the probes not rejected by the HEIDI test. The following three plots show  $-\log_{10}(P \text{ values})$  of SNP associations for the gene expression probes ILMN\_2343048 (tagging *ABCB9*), ILMN\_2393144 (tagging *ARL6IP4*), and ILMN\_1654421 (tagging *MPHOSPH9*) from the CAGE study. The sixth plot shows  $-\log_{10}(P \text{ values})$  of SNP associations for the DNAm probe cg13010344 from the mQTL meta-analysis. The

heatmap-like panel on the bottom shows the 14 REMC annotations with the significant PAIs annotated by orange curved lines on the top (see **Fig. S3b** for the overlap of the predicted PAIs with Hi-C data) and the Hi-C loop identified by Rao et al.<sup>1</sup> annotated on the x-axis (two orange bars connected by a red curved line).

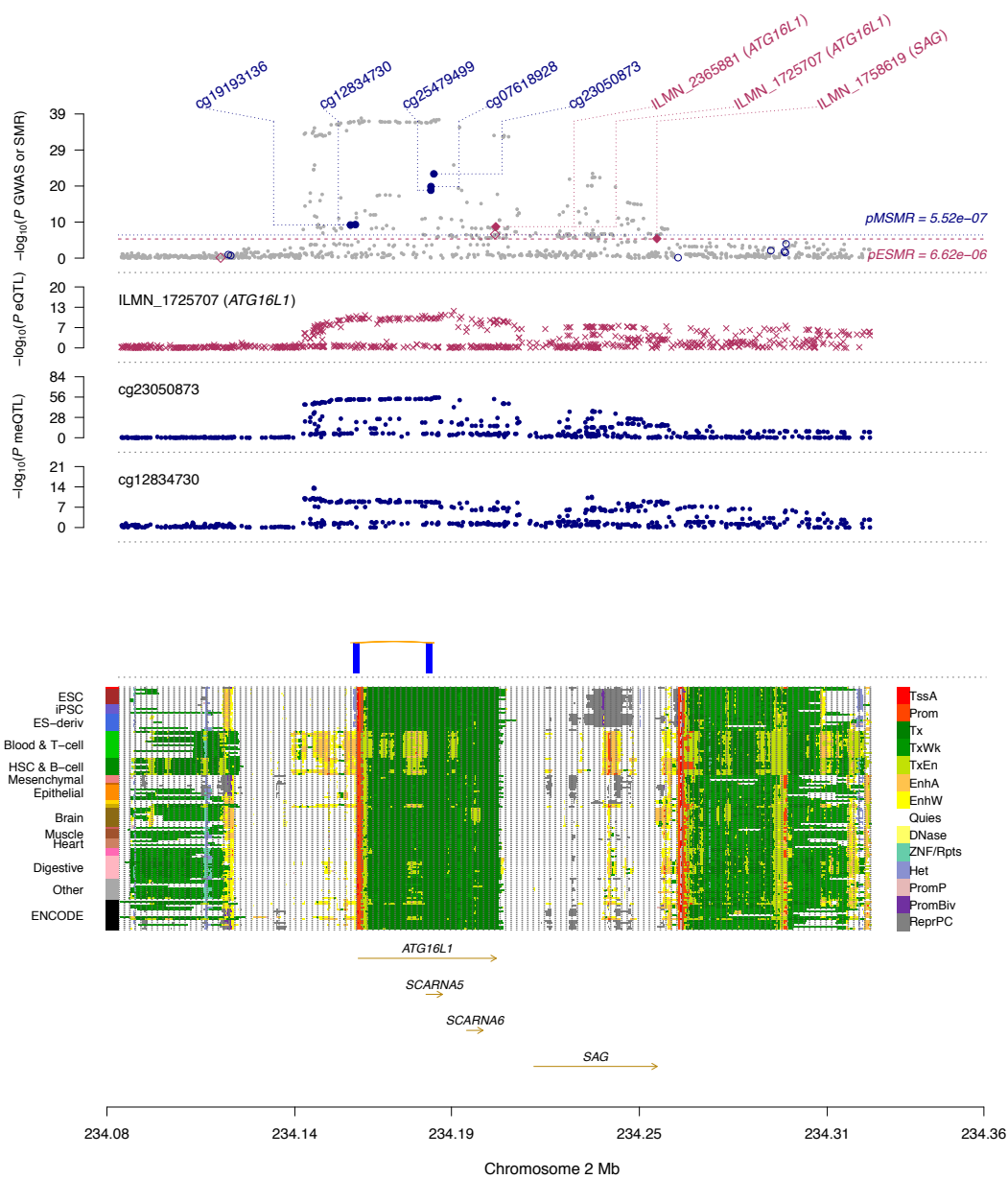


**Figure S5** Predicted PAIs at the *SORT1* locus. Shown are the 14 REMC chromatin state annotations with the significant PAIs labelled on the top.

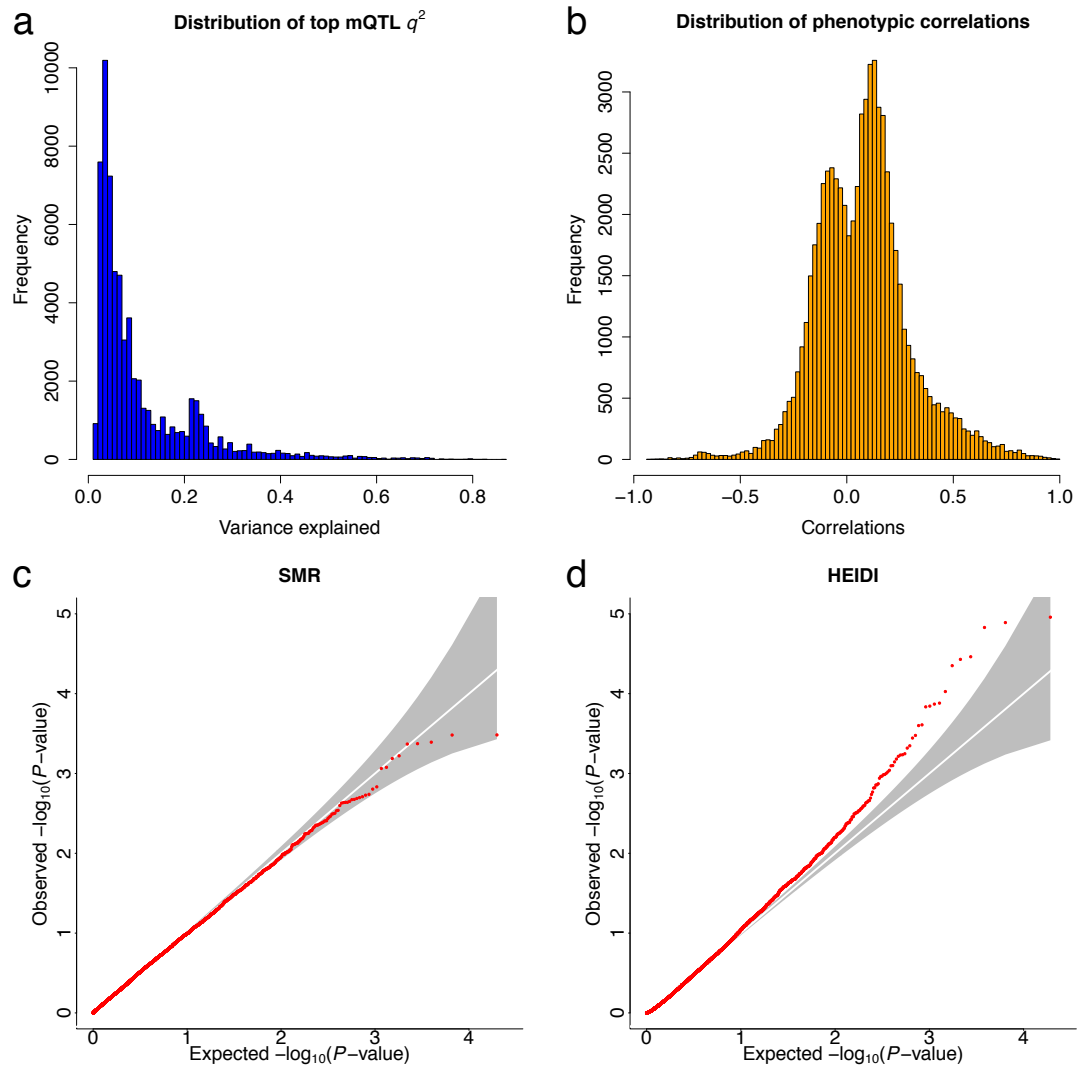


**Figure S6** Expression (measured by TPM) of *RNASET2* (ENSG00000026297.11) in 53 tissues of the GTEx (V7 data).





**Figure S7** Prioritizing gene and functional regions at the *ATG16L1* locus for Crohn's disease (CD). The top plot shows  $-\log_{10}(P \text{ values})$  of SNPs from the GWAS meta-analysis (grey dots) for CD<sup>2</sup>. Red diamonds and blue circles represent  $-\log_{10}(P \text{ values})$  from the SMR tests for associations of gene expression and DNAm with CD respectively. Solid diamonds are the probes not rejected by the HEIDI test. The second plot shows  $-\log_{10}(P \text{ values})$  of SNP associations for the gene expression probe ILMN\_1725707 (tagging *ATG16L1*) from the CAGE study. The third plot shows  $-\log_{10}(P \text{ values})$  of SNP associations for the DNAm probes cg23050873 and cg12834730. The plot in the middle shows the significant PAls between pairwise DNAm sites. The bottom plot shows the 14 REMC chromatin state annotations.



**Figure S8** Calibrating the test-statistics from SMR and HEIDI by simulation in overlapping samples. Details of the simulation can be found in **Supplementary Note 1**. Panel a): distribution of the variance in DNAm explained by the top-associated mQTL for each exposure probe. Panel b): distribution of correlations of DNAm levels between pairwise CpG sites (included in the PAI analysis) computed from LBC data<sup>3</sup> (**Methods**). Panel c): QQ plot of  $P$  values from the SMR test under the null model (i.e., there is no association between the two DNAm sites). Panel d): QQ plot of  $P$  values from the HEIDI test under the pleiotropic model (i.e., the two DNAm sites are associated due to the same causal variant).

### Supplementary Note 1. Simulations in overlapping samples

The SMR and HEIDI tests assume that the estimate of SNP effect on exposure is independent of that on outcome. This assumption can be violated if the sample from which the SNP effect on exposure was estimated overlapped with that from which the SNP effect on outcome was estimated, and there is phenotypic correlation between exposure and outcome in the overlapping samples<sup>4</sup>. To investigate whether the SMR and HEIDI test-statistics were biased by sample overlap, we performed simulations under the null and pleiotropic models based on the whole-genome sequencing (WGS) data from the UK10K project<sup>5</sup>. We included only unrelated individuals ( $n = 3,642$ ) and  $\sim 8.3$  million SNPs with minor allele frequency (MAF)  $> 0.01$  and Hardy-Weinberg Equilibrium (HWE)  $P$  value  $> 1 \times 10^{-6}$ .

#### Simulation under the null model

We first simulated DNAm levels at two probes in the same sample (i.e., complete sample overlap) under the null model (i.e., there is no association between the DNAm levels of the two probes) and investigated whether the SMR statistics were biased or not. To do so, we randomly sampled SNPs within a 1 Mb region of the genome and chose one SNP at random as the causal SNP from the sampled SNPs. We then simulated the DNAm levels of 3,642 individuals at one probe ( $\mathbf{m}_1$ ) based on model  $\mathbf{m}_1 = \mathbf{z}b_{zm1} + \mathbf{e}_{m1}$ , where  $\mathbf{z}$  is a vector of genotype of the causal SNP,  $b_{zm1}$  is the effect of the causal SNP on DNAm level  $b_{zm1} \sim N(0, R_{zm1}^2)$  with  $R_{zm1}^2$  being the proportion of variance in  $\mathbf{m}_1$  explained by the causal SNP, and  $\mathbf{e}_{m1} \sim N(0, \sigma_{e_{m1}}^2)$  with  $\sigma_{e_{m1}}^2 = \text{var}(\mathbf{Z}b_{zm1})(1/R_{zm1}^2 - 1)$ . To generate data under the null model, we generated the DNAm levels of the other probe ( $\mathbf{m}_2$ ) based on model  $\mathbf{m}_2 = \mathbf{e}_{m2}$ , where  $\mathbf{e}_{m2} \sim N(0, \sigma_{e_{m2}}^2)$  with  $\sigma_{e_{m2}}^2 = 1$ . Correlation of errors in estimating the SNP effects ( $r_e$ ) may occur due to sample overlap ( $\rho$ ) and phenotypic correlation ( $r_p$ ). To mimic this, we generated residuals ( $\mathbf{e}$ ) of the two probes from a

multivariate normal distribution,  $\mathbf{e} = \begin{pmatrix} \mathbf{e}_{m1} \\ \mathbf{e}_{m2} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{e_{m1}}^2 & r_p \sigma_{e_{m1}} \sigma_{e_{m2}} \\ r_p \sigma_{e_{m1}} \sigma_{e_{m2}} & \sigma_{e_{m2}}^2 \end{pmatrix} \right)$ , where

$r_p$  and  $R_{zm1}^2$  were randomly sampled from the observed distributions computed from data used in the PAI analysis in the LBC cohorts. We then performed a regression analysis to detect the top associated mQTL for each simulated probe, and ran an SMR analysis for each pair of simulated probes. We repeated the simulation 5,000 times to quantify the inflation/deflation of SMR test-statistics under this simulation scenario (**Fig S8c**).

#### Simulation under a pleiotropic model

To examine the distribution of the HEIDI test-statistics under a pleiotropic model (i.e., the

DNAm levels of two CpG sites are associated due to the same causal variant), we sampled a region and a causal SNP to generate DNAm level of the first CpG site using the same strategy above. The DNAm levels of the second probe was simulated based on model  $\mathbf{m}_2 = \mathbf{z}b_{zm2} + \mathbf{e}_{m2}$ , where  $b_{zm2}$  is the effect of the causal SNP on DNAm level of the second probe with  $b_{zm2} \sim N(0, R_{zm2}^2)$ ,  $R_{zm2}^2$  being the proportion of variance in  $\mathbf{m}_2$  explained by the causal variant,  $\mathbf{e}_{m2} \sim N(0, \sigma_{e_{m2}}^2)$  with  $\sigma_{e_{m2}}^2 = \text{var}(\mathbf{Z}b_{zm2})(1/R_{zm2}^2 - 1)$ .

The residuals ( $\mathbf{e}$ ) of the two probes were sampled from a multivariate normal distribution

$$\mathbf{e} = \begin{pmatrix} \mathbf{e}_{m1} \\ \mathbf{e}_{m2} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{e_{m1}}^2 & r_p \sigma_{e_{m1}} \sigma_{e_{m2}} \\ r_p \sigma_{e_{m1}} \sigma_{e_{m2}} & \sigma_{e_{m2}}^2 \end{pmatrix} \right),$$

where  $r_p$ ,  $R_{zm1}^2$  and  $R_{zm2}^2$  were

randomly sampled from the observed distributions in the LBC cohorts mentioned above. We detected the mQTL for each probe using linear regression and performed an SMR analysis to test for association between the two probes. We then performed a HEIDI test if the SMR association p-value was  $< 0.05$ , and repeated the simulation 5,000 times to evaluate the inflation/deflation of the HEIDI test-statistics under this simulation scenario (**Fig. S8d**).

## **Supplementary Note 2. Acknowledgments**

**HRS (dbGaP accession: phs000428.v1.p1):** HRS is supported by the National Institute on Aging (NIA U01AG009740). The genotyping was funded separately by the National Institute on Aging (RC2 AG036495, RC4 AG039029). Genotyping was conducted by the NIH Center for Inherited Disease Research (CIDR) at Johns Hopkins University. Genotyping quality control and final preparation of the data were performed by the Genetics Coordinating Center at the University of Washington.

**UK10K (EGA accession: EGAS00001000108):** The UK10K project was funded by the Wellcome Trust award WT091310. Twins UK (TUK): TUK was funded by the Wellcome Trust and ENGAGE project grant agreement HEALTH-F4-2007-201413. The study also receives support from the Department of Health via the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St. Thomas' NHS Foundation Trust in partnership with King's College London. Dr Spector is an NIHR senior Investigator and ERC Senior Researcher. Funding for the project was also provided by the British Heart Foundation grant PG/12/38/29615 (Dr Jamshidi). A full list of the investigators who contributed to the UK10K sequencing is available from <http://www.UK10K.org>.

**GTEx (dbGaP accession: phs000424.v6.p1):** The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health ([commonfund.nih.gov/GTEx](http://commonfund.nih.gov/GTEx)). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc. (HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819), Harvard University

(MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822).

## References

1. Rao, Suhas S.P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665-1680 (2014).
2. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-381 (2014).
3. McRae, A.F. *et al.* Identification of 55,000 Replicated DNA Methylation QTL. *Scientific Reports* **8**, 17605 (2018).
4. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-7 (2016).
5. UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).