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

Yang Wu[1,7], Ting Qi[1,7], Huanwei Wang[1], Futao Zhang[1], Zhili Zheng[1,2], Jennifer E. Phillips-Cremins[3], Ian J. Deary[4,5], Allan F. McRae[1], Naomi R. Wray[1,6], Jian Zeng[1], Jian Yang[1,2,6,*]

[1] Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia

[2] Institute for Advanced Research, Wenzhou Medical University, Wenzhou, Zhejiang 325027, China

[3] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104, USA

[4] Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh EH8 9JZ, UK

[5] Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, UK

[6] Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072, Australia

[7] These authors contributed equally to this work.

[*] Correspondence: Jian Yang (jian.yang.qt@gmail.com)

Abstract

Promoter-anchored chromatin interactions (PAIs) play a pivotal role in transcriptional regulation. Current high-throughput technologies for detecting PAIs, such as promoter capture Hi-C, are not scalable to large cohorts. Here, we present an analytical approach that uses summary-level data from cohort-based DNA methylation (DNAm) quantitative trait locus (mQTL) studies to predict PAIs. Using mQTL data from human peripheral blood (*n*=1,980), we predicted 34,797 PAIs which showed strong overlap with the chromatin contacts identified by previous experimental assays. The promoter-interacting DNAm sites were enriched in enhancers or near expression QTLs. Genes whose promoters were involved in PAIs were more actively expressed, and gene pairs with promoter-promoter interactions were enriched for co-expression. Integration of the predicted PAIs with GWAS data highlighted interactions among 601 DNAm sites associated with 15 complex traits. This study demonstrates the use of mQTL data to predict PAIs and provides insights into the role of PAIs in complex trait variation.

Introduction

Genome-wide association studies (GWASs) in the past decade have identified tens of thousands of genetic variants associated with human complex traits (including common diseases) at a stringent genome-wide significance level[1,2]. However, most of the trait-associated variants are located in non-coding regions[3,4], and the causal variants as well as their functional roles in trait etiology are largely unknown. One hypothesis is that the genetic variants affect the trait through genetic regulation of gene expression[4]. Promoter-anchored chromatin interaction (PAI)[5,6] is a key regulatory mechanism whereby non-coding genetic variants alter the activity of cis-regulatory elements and subsequently regulate the expression levels of the target genes. Therefore, a genome-wide map of PAIs is essential to understand transcriptional regulation and the genetic regulatory mechanisms underpinning complex trait variation.

High-throughput experiments, such as Hi-C[7] and ChIA-PET (chromatin interaction analysis by paired-end tag sequencing)[8], have been developed to detect chromatin interactions by a massively parallelized assay of ligated DNA fragments. Hi-C is a technique based on chromosome conformation capture (3C)[9] to quantify genome-wide interactions between genomic loci that are close in three-dimensional (3D) space, and ChIA-PET is a method that combines ChIP-based methods[10] and 3C. However, these high-throughput assays are currently not scalable to population-based cohorts with large sample sizes because of the complexity of generating a DNA library for each individual (tissue or cell line) and the extremely high sequencing depth needed to achieve high detection resolution[11]. On the other hand, recent technological advances have facilitated the use of epigenomic marks to infer the chromatin state of a specific genomic locus and further to predict the transcriptional activity of a particular gene[12,13]. There have been increasing interests in the use of epigenomic data (e.g., DNA methylation (DNAm) and/or histone modification) to infer chromatin interactions[14-17]. These analyses, however, rely on individual-level chromatin accessibility data often only available in small samples[14,16], and it is not straightforward to use the predicted chromatin interactions to interpret the variant-trait associations identified by GWAS.

In this study, we proposed an analytical approach to predict chromatin interaction by detecting the association between DNAm levels of two CpG sites due to the same set of genetic variants (i.e., pleiotropic association between DNAm sites). This can be achieved because if the methylation levels (unmethylated, partly methylated or fully methylated) of a pair of relatively distal CpG sites covary across individuals and such covariation is not (or at least not completely) caused by environmental or experimental factors (evidenced by the sharing of a common set of causal genetic variants in cis) (**Fig. 1b**), it is very likely that the two genomic regions interact (having contacts or functional links because of their close physical proximity in 3D space). Our analytical approach was based on two recently developed methods, i.e., the summary-data–based Mendelian randomization (SMR) test and the test for heterogeneity in dependent instruments (HEIDI)[18], which are often used in combination to detect pleiotropic association between a molecular phenotype (e.g. gene expression or DNA methylation) and a complex trait[18] or between two molecular phenotypes[19]. The SMR & HEIDI approach only requires summary-level data from DNA methylation quantitative trait locus (mQTL) studies, providing the flexibility of using mQTL data from studies with large sample sizes to ensure efficient power. Since the proposed method is based on cohort-based genetic data, it also allows us to integrate the predicted chromatin interactions with GWAS results to understand the genetic regulatory mechanisms for complex traits. In this study, we analyzed mQTL summary data from a meta-analysis of two cohort-based studies on 1,980 individuals with DNAm levels measured by Illumina 450K methylation arrays and SNP data from SNP-array-based genotyping followed by imputation to the 1000 Genome Project (1KGP) reference panels[19,20].

Results

Predicting promoter-anchored chromatin interactions using mQTL data

As described above, our underlying hypothesis was that if the variation between people in DNAm levels of two relatively distal CpG sites are associated due to the same set of causal genetic variants (**Fig. 1b**), then it is very likely that these two chromatin regions have contacts or functional links because of their close physical proximity in 3D space. Hence, we set out to predict promoter-anchored chromatin interactions (PAIs) from mQTL data. We applied the SMR & HEIDI approach[18] to test for pleiotropic associations of a DNAm site in the promoter region of a gene with all the other DNAm sites within 2 Mb of the focal promoter in either direction (excluding those in the focal promoter) using mQTL summary data from peripheral blood samples (**Fig. 1**, **Fig. S1** and **Methods**). Therefore, our analysis was a scan for genomic regions that are functionally associated with promoter regions likely because of chromatin contacts or close physical proximity in 3D space. Note that we limited the analysis to a 2 Mb window because chromatin interactions between genomic sites more than 2 Mb apart are rare[21], because summary data from epigenetic QTL studies are often only available for genetic variants in cis-regions, and because it reduces the computational and multiple testing burdens. The mQTL summary data were generated from a meta-analysis of two mQTL data sets from McRae et al. (n = 1,980)[19,20]. The mQTL effects were in standard deviation (SD) units of DNAm levels. In the SMR analysis, the promoter DNAm site was used as the "exposure" and each of the other DNAm sites in the region was used as the "outcome" (**Fig. 1**). For exposure probes, we included in the SMR analysis only the DNAm sites with at least one cis-mQTL (SNPs within 2 Mb of the CpG site associated with variation in DNAm level) at P_{mQTL}

$< 5 \times 10^{-8}$. We used such a stringent significance level because a basic assumption of Mendelian randomization is that the SNP instrument needs to be strongly associated with the exposure[22,23]. There were 90,749 DNAm probes with at least one cis-mQTL at $P_{\text{mQTL}} < 5 \times 10^{-8}$, 28,732 of which were located in promoters annotated based on data from blood samples of the Roadmap Epigenomics Mapping Consortium (REMC)[13]. We used the 1KGP-imputed Health and Retirement Study (HRS)[24] data as a reference sample for linkage disequilibrium (LD) estimation to perform the HEIDI test, which was used to reject SMR associations between DNAm sites not driven by the same set of causal variants (called linkage model in Zhu et al.[18]). In total, we identified 34,797 PAIs between pairwise DNAm sites that passed the SMR test ($P_{\text{SMR}} < 1.76 \times 10^{-9}$ based on a Bonferroni correction for multiple tests) and were not rejected by the HEIDI test ($P_{\text{HEIDI}} > 0.01$; see Wu et al.[19] for the justification of the use of this HEIDI threshold *P* value). The significant PAIs comprised of 21,787 unique DNAm sites, among which 10,249 were the exposure probes in promoter regions of 4,617 annotated genes. Most of the DNAm sites in promoters showed pleiotropic associations with multiple DNAm sites (mean = 4) (**Fig. S2a**). The distances between 95% of the pairwise interacting DNAm sites were less than 500 Kb (mean = 79 Kb and median = 23 Kb). Only ~0.7% of the predicted PAIs were between DNAm sites greater than 1 Mb apart (**Fig. S2b**). The summary statistics of the predicted PAIs are publicly available through the M2Mdb Shiny online application (**URLs**).

Overlap of the predicted PAIs with chromatin contacts identified from experimental assays

We first examined whether the predicted PAIs are consistent with chromatin contacts identified by experimental assays, such as Hi-C[21] and promoter captured Hi-C (PCHi-C)[5]. While the majority of experimental assays are measured in primary cell lines, topological associated domains (TADs) annotated from Hi-C are relatively conserved across cell types[25]. We therefore tested the overlap of our predicted PAIs with the TADs identified from recent Hi-C and PCHi-C studies[5,21,26] (see **Supplementary Table 1** for a full list of data sets from experimental assays used in this study). We found that 22,024 (63.3%) of the predicted PAIs were between DNAm sites located in the TADs identified by Rao et al. using Hi-C in the GM12878 cell lines[21], 27,200 (78.2%) in those by Dixon et al. using Hi-C in embryonic stem cells[26], and 27,716 (79.7%) in those by Javierre et al. using PCHi-C in primary hematopoietic cells[5]. These overlaps with Hi-C and PCHi-C data were significantly higher than expected by chance ($P < 0.001$ for all the three Hi-C/PCHi-C data sets; **Fig. 2a-c**). Note that the *P* value was computed by comparing the observed number to a null distribution generated by resampling the same number of DNAm pairs at random from distance-matched DNAm pairs included in the SMR analysis (**Methods**); the *P* value was truncated at 0.001 due to the finite number of resampling. One example was the *MAD1L1* locus (a ~450 Kb region) on chromosome 7 (**Fig. 2d** and **Fig. 2e**) where there were a large number of predicted PAIs highly consistent with TADs identified by Hi-C from the Rao et al. study[21]. There were also scenarios where the predicted PAIs were not aligned well with the TAD data. For example, 107 of the 183 predicted PAIs at the *RPS6KA2* gene locus did not overlap with the TADs identified by Hi-C from the Rao et al. study[21] (**Fig. S3a**). These predicted interactions, however, are very likely to be functional as indicated by our subsequent analysis with GWAS and omics data (see below). Additionally, the predicted PAIs were slightly enriched for the Hi-C loops identified from Rao *et al.*[21] (1.49-fold, $P < 0.001$, m = 130; **Fig. 3a**) and the *POLR2A* ChIA-PET loops from the ENCODE[27] project (1.44-fold, $P < 0.001$, m = 2,315; **Fig. 3b**), although the numbers of overlaps were small. One notable example was the *GNB1* locus where the predicted PAI between the promoter region of *GNB1* and an enhancer nearby is consistent with the enhancer-promoter interaction identified by both Hi-C from Rao *et al.*[21] and PCHi-C from Jung *et al.*[28] in the GM12878 cell lines (**Fig. S4**).

Comparison with other prediction methods

To assess the performance of our PAI prediction method, we compared it with two state-of-the-art approaches of this kind, i.e., the correlation-based method used in Gate *et al.*[29] and the pairwise hierarchical model (PHM) method developed by Kumasaka *et al.*[17], using the DNAm data described above or the chromatin accessibility data (measured by Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)) from Kumasaka *et al.*[17]. We used a recently released chromatin interaction data (PCHi-C loops) generated by Jung *et al.*[28] in GM12878 cell lines for validation, and quantified the enrichment of the predicted interactions in the PCHi-C loops defined based on a range of PCHi-C *P* value thresholds. We chose the PCHi-C data from Jung *et al.* because the *P* values of all the tested loops are available and because compared to other Hi-C data sets, chromatin interactions identified in GM12878 cell lines may be more relevant to the predicted PAIs in whole blood. We computed the fold enrichment of the predicted interactions by the three methods in the PCHi-C loops by a 2 × 2 contingency table and used the Fisher's exact test to assess the statistical significance of the enrichment (**Methods**). The results showed that our predicted PAIs using either DNAm or chromatin accessibility data were highly enriched in the PCHi-C loops and that the fold enrichment increased with the increase of the significance level used to claim the PCHi-C loops (**Fig. 3c**), consistent with the observation from previous work that Hi-C loops with lower *P* values are more reproducible between biological replicates[30]. Our SMR & HEIDI method outperformed the correlation-based method using either DNAm or chromatin accessibility data, as evidenced by the larger fold enrichment of our method compared to the correlation-based method at all the PCHi-C significance levels (**Fig. 3c**). We also compared the predicted PAIs with the interactions identified from the PHM approach[17] using the chromatin accessibility data. Of the 15,487 interactions identified by the PHM approach, 10,416 were tested in our SMR & HEIDI analysis; 98.4% were replicated at a nominal significance level ($P_{\text{SMR}} < 0.05$

and $P_{\text{HEIDI}} > 0.01$), and 36% were significant after multiple testing correction ($P_{\text{SMR}} < 4.8 \times 10^{-6}$ (0.05/10,416) and $P_{\text{HEIDI}} > 0.01$). While the PHM approach requires individual-level genotype and chromatin accessibility data and is less computationally efficient due to the use of Bayesian hierarchical model, our SMR & HEIDI method that requires only summary-level data is more flexible and can be potentially applied to all epigenetic QTL data.

Enrichment of the predicted PAIs in functional annotations

To investigate the functional role of the DNAm sites that showed significant interactions with the DNAm sites in promoter regions (called promoter-interacting DNAm sites or PIDSs hereafter), we conducted an enrichment analysis of the PIDSs (m = 14,361) in 14 main functional annotation categories derived from the REMC blood samples (**Methods**). The fold-enrichment was computed as the proportion of PIDSs in a functional category divided by the mean of a null distribution generated by resampling variance-matched "control" probes at random from all the outcome probes used in the SMR analysis. We found a significant enrichment of PIDSs in enhancers (fold-enrichment=2.17 and $P_{\text{enrichment}} < 0.001$), repressed Polycomb regions (fold-enrichment=1.56 and $P_{\text{enrichment}} < 0.001$), primary DNase (fold-enrichment=1.43 and $P_{\text{enrichment}} < 0.001$) and bivalent promoters (fold-enrichment=1.12 and $P_{\text{enrichment}} < 0.001$) and a significant underrepresentation in transcription starting sites (fold-enrichment=0.21 and $P_{\text{enrichment}} < 0.001$), quiescent regions (fold-enrichment=0.74 and $P_{\text{enrichment}} < 0.001$), promoters around transcription starting sites (fold-enrichment=0.77 and $P_{\text{enrichment}} < 0.001$), and transcribed regions (fold-enrichment=0.90 and $P_{\text{enrichment}} < 0.001$) in comparison with the control probes (**Fig. 4a** and **Fig. 4b**). On one hand, the enrichment test is not biased by the fact that the Illumina 450K methylation array probes are preferentially distributed towards certain genomic regions (e.g., promoters; **Fig. 4a**) because it tests against control probes sampled from probes on the array rather than random genomic positions. On the other hand, however, this test is over conservative because the control probes are enriched in certain functional genomic regions (**Fig. S5a**) and can possible contain some of the PIDSs, which may explain the relatively small fold enrichments observed above. The depletion of PIDSs in promoters was due to the exclusion of outcome probes from the focal promoters (**Methods**; **Fig. S5b**). In addition, a large proportion (~18%) of the predicted PAIs were promoter-promoter interactions (PmPmI), consistent with the results from previous studies[5,31] that PmPmI were widespread.

Relevance of the predicted PAIs with gene expression

We then turned to test whether pairwise genes with significant PmPmI were enriched for co-expression. We used gene expression data (measured by Transcript Per Kilobase Million mapped reads or TPM) from the blood samples of the Genotype-Tissue Expression (GTEx) project[32] and computed the Pearson correlation of expression levels across individuals between pairwise genes (r_P). To assess the statistical significance of the enrichment, we compared the observed mean Pearson correlation of all the significant PmPmI gene pairs (m = 2,236) to a null distribution of mean Pearson correlation values, generated by resampling a set of distance-matched control gene pairs either from the genes whose promoters were involved in the SMR analysis or from all genes. The mean correlation for the significant PmPmI gene pairs ($\bar{r}_P$) was 0.367, significantly ($P < 0.001$) higher than that for the control gene pairs sampled either from the genes whose promoters were involved in SMR (mean $\bar{r}_P = 0.292$; **Fig. 4c**) or from all genes (mean $\bar{r}_P = 0.156$; **Fig. 4c**), suggesting that pairwise genes with PmPmI are more likely to be co-expressed.

We also tested whether genes whose promoters were involved in significant PAI (called Pm-PAI genes hereafter, **Fig. 1**) were expressed more actively than the same number of control genes randomly sampled from the genes whose promoters were involved in SMR or from all genes. Similar to the analysis above, we used the gene expression data (measured by TPM) from the blood samples of the GTEx project and tested the enrichment of Pm-PAI genes in different expression level groups (**Methods**). In comparison to the control sets sampled from the genes whose promoters were involved in SMR, Pm-PAI genes were significantly overrepresented ($P <$ 0.001) among the group of genes with the highest expression levels and significantly underrepresented ($P < 0.001$) among genes that were not actively expressed (median TPM < 0.1) (**Fig. 4d**). These results implicate the regulatory role of the PIDSs in transcription and their asymmetric effects on gene expression. The enrichment was much stronger if the control sets were sampled from all genes (**Fig. S6a**). We also performed a similar enrichment analysis (testing against the control sets sampled from all genes) for the predicted target genes from the PCHi-C data from Jung *et al.*[28]. There was a significant enrichment of the PCHi-C target genes in the active gene groups, but the fold enrichment was slightly smaller than that of the Pm-PAI genes (**Fig. S6**), suggesting that PAIs could be more functionally relevant than PCHi-C loops.

Enrichment of eQTLs in the PIDS regions

We have shown that the PIDSs are located in regions enriched with regulatory elements (e.g., enhancers) (**Fig. 4b**) and that the Pm-PAI genes tend to have higher expression levels (**Fig. 4d**). We next investigated if genomic regions near PIDS are enriched for genetic variants associated with expression levels of the Pm-PAI genes using data from an expression QTL (eQTL) study in blood[33]. There were 11,204 independent cis-eQTLs at $P_{\mathrm{eQTL}} < 5 \times 10^{-8}$ for 9,967 genes, among which 2,019 were Pm-PAI genes (**Methods**). We mapped cis-eQTLs to a 10 Kb region centered around each PIDS (5 Kb on either side) and counted the number of cis-eQTLs associated with expression levels of the corresponding Pm-PAI gene for each PIDS. There were 548 independent eQTLs located in the PIDS regions of the Pm-PAI genes, significantly higher than (P < 0.001) the mean of a null distribution (mean = 415) generated by randomly resampling distance-matched pairs of DNAm sites used in the SMR analysis (**Fig. 5a**). These results again imply the regulatory role of the PIDSs in transcription through eQTLs and provide evidence supporting the functional role of the predicted PAIs.

There were examples where a cis-eQTL was located in a PIDS region predicted to interact with the promoters of multiple genes. For instance, a cis-eQTL was located in an enhancer predicted to interact with the promoters of three genes (i.e., *ABCB9*, *ARL6IP4*, and *MPHOSPH9*) (**Fig. S7**), and the predicted interactions were consistent with the TADs identified by Hi-C from Rao *et al.*[21] (**Fig. S3b**). Furthermore, the predicted interactions between promoters of *ARL6IP4* and *MPHOSPH9* are consistent with the chromatin contact loops identified by Hi-C in the GM12878 cells[21] (**Fig. S7**). The eQTL association signals were highly consistent for the three genes, and the pattern was also consistent with the SNP association signals for schizophrenia (SCZ) and years of education (EY) as shown in our previous work[19], suggesting a plausible mechanism whereby the SNP effects on SCZ and EY are mediated by the expression levels of at least one of the three co-regulated genes through the interactions of the enhancer and three promoters (**Fig. S7**).

We have shown previously that the functional association between a DNAm site and a gene nearby can be inferred by the pleiotropic association analysis using SMR & HEIDI considering the DNAm level of a CpG site as the exposure and gene expression level as the outcome[19]. We further tested if the PIDSs are enriched among the DNAm sites showing pleiotropic associations with the expression levels of the neighboring Pm-PAI genes. We found that approximately 15% of the PIDSs were the gene-associated DNAm sites identified in our previous study[19], significantly higher (P < 0.001) than that computed from the distance-matched control probe pairs (1.3%) described above (**Fig. 5b**).

Replication of the predicted PAIs across tissues

To investigate the robustness of the predicted PAIs across tissues, we performed the PAI analysis using brain mQTL data from the Religious Orders Study and Memory and Aging Project (ROSMAP)[34] (n = 468). Of the 11,082 PAIs with $P_{\text{SMR}} < 1.76 \times 10^{-9}$ and $P_{\text{HEIDI}} > 0.01$ in blood and available in brain, 2,940 (26.5%) showed significant PAIs in brain after Bonferroni correction for multiple testing ($P_{\text{SMR}} < 4.51 \times 10^{-6}$ and $P_{\text{HEIDI}} > 0.01$). If we use a less stringent threshold for replication, e.g., the nominal P value of 0.05, 66.31% of PAIs predicted in blood were replicated in brain. Here, the replication rate is computed based on a P value threshold, which is dependent of the sample size of the replication data. Alternatively, we can estimate the correlation of PAI effects (i.e., the effect of the exposure DNAm site on the outcome site of a predicted PAI) between brain and blood using the r_b method[35]. This method does not rely on a P value threshold and accounts for estimation errors in the estimated effects, which is therefore not dependent of the replication sample size. The estimate of r_b was 0.527 (SE = 0.0051) for 11,082 PAIs between brain and blood, suggesting a relatively strong overlap in PAI between brain and blood.

It is of note that among the 2,940 blood PAIs replicated at $P_{\text{SMR}} < 4.51 \times 10^{-6}$ and $P_{\text{HEIDI}} > 0.01$ in brain, there were 268 PAIs for which the PAI effects in blood were in opposite directions to those in brain (**Supplementary Table 2**). For example, the estimated PAI effect between the *SORT1* and *SYPL2* loci was 0.49 in blood and -0.86 in brain. This tissue-specific effect is supported by the differences in gene expression correlation (correlation of expression levels between *SORT1* and *SYPL2* was -0.07 in whole blood and -0.37 in brain frontal cortex; $P_{\text{difference}} = 0.0018$) and the chromatin state of the promoter of *SYPL2* (bivalent promoter in blood and active promoter in brain; **Fig. S8**) between brain and blood. Taken together, while there are tissue-specific PAIs, a substantial proportion of the predicted PAIs in blood are consistent with those in brain.

Putative target genes of the disease-associated PIDSs

We have shown above the potential functional roles of the predicted PAIs in transcriptional regulation. We then turned to ask how the predicted PAIs can be used to infer the genetic and epigenetic regulatory mechanisms at the GWAS loci for complex traits and diseases. We have previously reported 1,203 pleiotropic associations between 1,045 DNAm sites and 15 complex traits and diseases by an integrative analysis of mQTL, eQTL and GWAS data using the SMR & HEIDI approach[19]. Of the 1,045 trait-associated DNAm sites, 601 (57.5%) sites were involved in the predicted PAIs related to 299 Pm-PAI genes (**Supplementary Table 3**). We first tested the enrichment of the Pm-PAI genes of the trait-associated PIDSs using FUMA[36]. For the 15 complex traits analysed in Wu *et al.*[19], our FUMA analyses identified enrichment in multiple GO and KEGG pathways relevant to the corresponding phenotypes such as the inflammatory response pathway for Crohn's disease (CD) and steroid metabolic process for body mass index (BMI) (**Supplementary Table 4**), demonstrating the regulatory role of the trait-associated PIDSs in biological processes and tissues relevant to the trait or disease.

There were a number of examples where the predicted PAIs provided important insights to the functional genes underlying the GWAS loci and the underlying mechanisms by which the DNA variants affect the trait through genetic regulation of gene expression. One notable example was a PIDS (cg00271210) in an enhancer region predicted to interact in 3D space with the promoter regions of two genes (i.e., *RNASET2* and *RPS6KA2*), the expression levels of both of which were associated with ulcerative colitis (UC) and CD as reported in our previous study[19] (**Fig. 6**). The SNP-association signals were consistent across CD GWAS, eQTL, and mQTL studies, suggesting that the genetic effect on CD is likely to be mediated through epigenetic regulation of gene expression. Our predicted PAIs further implicated a plausible mechanism whereby the expression levels of *RNASET2* and *RPS6KA2* are co-regulated through the interactions of their promoters with a shared enhancer (**Fig. 6**), although only 41.5% of the predicted PAIs in this region overlapped with the TADs identified by Hi-C from the Rao *et al.* study[21] (**Fig. S3a**) as mentioned above. According to the functional annotation data derived from the REMC samples, it appears that this shared enhancer is highly tissue-specific and present only in B cell and digestive system that are closely relevant to CD (**Fig. 6**). The over-expression of *RNASET2* in spleen (**Fig. S9**) is an additional piece of evidence supporting the functional relevance of this gene to CD. Another interesting example is the *ATG16L1* locus (**Fig. S10**). We have shown previously that five DNAm sites are in pleiotropic associations with CD and the expression level of *ATG16L1*[19]. Of these five DNAm sites, three were in an enhancer region and predicted to interact in 3D space with two DNAm sites in the promoter region of *ATG16L1* (**Fig. S10**), suggesting a plausible mechanism that the genetic effect on CD at this locus is mediated by genetic and epigenetic regulation of the expression level of *ATG16L1* through promoter–enhancer interactions.

Discussion

We have presented an analytical approach on the basis of the recently developed SMR & HEIDI method to predict promoter-anchored chromatin interactions using mQTL summary data. The proposed approach uses DNAm level of a CpG site in the promoter region of a gene as the bait to detect its pleiotropic associations with DNAm levels of the other CpG sites (**Fig. 1**) within 2 Mb distance of the focal promoter in either direction. In contrast to experimental assays, such as Hi-C and PCHi-C, our approach is cost-effective (because of the reuse of data available from experiments not originally designed for this purpose) and scalable to large sample sizes. Our method utilises a genetic model to perform a Mendelian randomization analysis so that the detected associations are not confounded by non-genetic factors, which is also distinct from the methods that predict chromatin interactions from the correlations of chromatin accessibility measures[14,16].

Using mQTL summary-level data from human peripheral blood (n = 1,980), we predicted 34,797 PAIs for the promoter regions of 4,617 genes. We showed that the predicted PAIs were enriched in TADs detected by published Hi-C and PCHi-C assays and that the PIDS regions were enriched with eQTLs of target genes. We also showed that the PIDSs were enriched in enhancers and that the Pm-PAI genes tended to be more actively expressed than matched control genes. These results demonstrate the functional relevance of the predicted PAIs to transcriptional regulation and the feasibility of using data from genetic studies of chromatin status to infer three-dimensional chromatin interactions. The proposed approach is applicable to data from genetic studies of other chromatin features such as histone modification (i.e., hQTL)[37] or chromatin accessibility (caQTL)[29]. The flexibility of the method also allowed us to analyse data from different tissues or cell types. Using summary data from a brain mQTL study (n = 468), we replicated 26.5% of blood PAIs in brain at a very stringent threshold (P_{SMR} < 0.05 / m with m being the number of tests in the replication set and P_{HEIDI} > 0.01) and 66.31% at a less stringent threshold (P_{SMR} < 0.05). Together with an estimate of r_{b} of 0.527 for the correlation of PAI effects between brain and blood, we demonstrated a substantial overlap of the predicted PAIs between blood and brain, in line with the finding from a recent study that cis-mQTLs are largely shared between brain and blood[35].

The use of a genetic model to detect PAIs also facilitated the integration of the predicted PAIs with GWAS data. In a previous study, Wu *et al.*[19] mapped DNAm sites to genes and then to a trait by checking the consistency of pleiotropic association signals across all the three layers. In this study, we have shown examples of how to integrate the predicted PAIs with GWAS, eQTL and functional annotation data to better understand the genetic and epigenetic regulatory mechanisms underlying the GWAS loci for complex traits (**Figs. 6**, **S7**, and **S10**). The pleiotropic associations between DNAm sites involved in PAIs and a complex trait are also helpful to link genes to the trait at GWAS loci even in the absence of eQTL data. If both DNAm sites of a PAI show pleiotropic association with the trait, the corresponding Pm-PAI gene is likely to be a functionally relevant gene of the trait. Of the 1,045 DNAm sites that showed pleiotropic associations with 15 complex traits as reported in Wu *et al.*[19], 601 sites were involved in the PAIs for 299 Pm-PAI genes identified in this study. In this case, these Pm-PAI genes are very likely to be the functionally relevant genes at the GWAS loci. In comparison with 66 gene targets identified in Wu *et al.*[19] (34/66 overlapped with 299 Pm-PAI genes), integration of PAIs with GWAS facilitates the discovery of more putative gene targets for complex traits.

There are several reasons why the overlaps between the predicted PAIs and Hi-C loops were limited. First, Hi-C loops were detected with errors. We observed that the concordances between different Hi-C data sets were very limited (**Fig. S11**), consistent with the conclusion from Forcato *et al.* that the reproducibility of Hi-C loops is low at all resolutions[38]. Second, most (65%) of our predicted PAIs are interactions between DNAm sites within 50 Kb (**Fig. S2b**), which are often not well captured by the 3C-based methods due to its low resolution[17]. Third, the chromatin interactions are cell type specific[5] so that differences between the Hi-C loops identified in cell lines and our PAIs identified in whole blood are expected. For the PAIs that were between DNAm sites not located in TADs or Hi-C loops, we have shown specific examples that these predicted PAIs are likely to be functionally interacted (**Fig. 2d and Fig. S3**), suggesting that these PAIs are likely to be interactions yet to be identified by experimental assays. On the other hand, compared to the loops identified based on 3C-based methods, our predicted PAIs are more likely to be functional interactions due to the use of genetic and regulatory epigenomic data, as evidenced by the observation that our predicted Pm-PAI genes showed stronger enrichment in active gene groups compared to the predicted target genes from the PCHi-C data (**Fig. S6**).

There are some limitations of this study. First, chromatin interactions are likely to be tissue- and temporal-specific whereas our PAI analyses were limited to mQTL data from blood and brain owing to data availability and thus were unable to detect PAIs in specific tissues or at different developmental stages. Second, although the sample size of our blood mQTL summary data is large (n = ~2,000), the PAI analysis could be underpowered if the proportion of variance in exposure or outcome explained by the top associated cis-mQTL is small. Third, the predicted PAIs are relatively sparse as illustrated in **Fig. 2d** because of the sparsity of the DNAm array used, the underlying hypothesis of the SMR method, and the stringent statistical significance level used to claim significant PAIs (**Supplementary Note 1**). Fourth, the functional annotation data derived from the REMC samples could potentially include noise due to the small sample sizes, leading to uncertainty in defining the bait promoter regions. Fifth, if the DNAm levels of two CpG sites are affected by two sets of causal variants in very high LD, these two DNAm sites will appear to be associated in the SMR analysis and the power of the HEIDI test to reject such an SMR association will be limited because of the high LD[18,19]. However, this phenomenon is likely to be rare given that most of the promoter-anchored DNAm sites were predicted to interact with multiple DNAm sites which are very unlikely to be all caused by distinct sets of causal variants in high LD. Sixth, the predicted PAIs including those falling in chromatin loops and TAD regions were not necessarily functional interactions and need to be validated by functional assays in the future. Despite these limitations, our study provides a novel computational paradigm to predict PAIs from genetic effects on epigenetic markers with high resolution. Integrating of the predicted PAIs with GWAS, gene expression, and functional annotation data provides novel insights into the regulatory mechanisms underlying GWAS loci for complex traits. The computational framework is general and applicable to other types of chromatin and histone modification data, to further decipher the functional organisation of the genome.

Methods

Predicting PAIs from mQTL data by the SMR and HEIDI analyses

We used summary-level mQTL data to test whether the variation between people in DNAm levels of two CpG sites are associated because of a set of shared causal variants. Mendelian Randomization (MR) is an approach developed to test for the causal effect of an exposure and an outcome using a genetic variant as the instrumental variable[22,23]. Summary-data–based Mendelian Randomization (SMR) is a variant of MR, originally designed to test for association between the expression level of a gene and a complex trait using summary-level data from GWAS and eQTL studies[18] and subsequently applied to test for associations between DNAm and gene expression and between DNAm and complex traits[19]. Here, we applied the SMR analysis to detect associations between DNAm sites. Let *x* be an exposure DNAm, *y* be an outcome DNAm, and *z* be an instrument SNP associated with exposure DNAm (e.g., $P_{\mathrm{mQTL}} < 5 \times 10^{-8}$). The SMR estimate of the effect of exposure DNAm on the outcome DNAm (i.e., $\hat{b}_{xy}$) is the ratio of the estimated effect of instrument on exposure ($\hat{b}_{zx}$) and that on outcome ($\hat{b}_{zy}$), $\hat{b}_{xy} = \hat{b}_{zy}/\hat{b}_{zx}$, where $\hat{b}_{zx}$ and $\hat{b}_{zy}$ are available from the summary-level mQTL data. We specified the DNAm level of a probe within the promoter region of a gene as the exposure and tested its associations with the DNAm levels of other probes (outcomes) within 2 Mb of the exposure probe (**Fig. 1** and **Fig. S1**). Probe pairs in the same promoter region were not included in the analysis. For a pair of probes in two different promoter regions, the one with higher variance explained by its top associated cis-mQTL was used as the exposure and the other one was used as the outcome. The associations passed the SMR test could possibly be due to linkage (i.e., distinct sets of causal variants in LD, one set affecting the exposure and the other set affecting the outcome), which is less of biological interest in comparison with pleiotropy (i.e., the same set of causal variants affecting both the exposure and the outcome). We then applied the HEIDI (heterogeneity in dependent instruments) test to distinguish pleiotropy from linkage. In brief, the HEIDI test was developed to test against the null hypothesis that the two DNAm sites are affected by the same set of causal variants. This is equivalent to testing whether there is a difference between the $\hat{b}_{xy}$ estimated from any mQTL *i* ($\hat{b}_{xy(i)}$) and that estimated from the top associated mQTL ($\hat{b}_{xy(top)}$). If we define the difference in estimate between $\hat{b}_{xy}$ at mQTL *i* and that at top associated mQTL as $\hat{d}_i = \hat{b}_{xy(i)} - \hat{b}_{xy(top)}$, then for multiple mQTLs (i.e., top 20 associated mQTLs after pruning out SNPs in very strong LD), we have $\hat{\mathbf{d}} \sim MVN(\mathbf{d}, \mathbf{V})$, where $\hat{\mathbf{d}} = \{\hat{d}_1, \cdots, \hat{d}_{20}\}$ and $\mathbf{V}$ is the covariance matrix that can be estimated using summary-level mQTL data and LD information from a reference panel[18] (e.g., the 1KGP-imputed HRS[24] data). Therefore, we can test the evidence for heterogeneity through evaluating whether $\mathbf{d} = \mathbf{0}$ using an approximate multivariate approach[39]. We rejected the SMR associations with $P_{\mathrm{HEIDI}} < 0.01$. All these analyses have been implemented in the SMR software tool (**URLs**).

Because the mQTL data for the exposure and the outcome were obtained from the same sample, we investigated whether the SMR and HEIDI test-statistics were biased by the sample overlap. To this end, we computed the phenotypic correlation between each pair of exposure and outcome probes as well as the variance explained by the top associated cis-mQTL of each exposure probe, and performed the simulation based on these observed distributions (**Supplementary Note 2**). The simulation results showed that *P* values from both SMR and HEIDI tests were evenly distributed under the null model without inflation or deflation (**Fig. S12**). We have made all the PAIs analysis scripts publicly available at GitHub https://github.com/wuyangf7/PAI.

Data used for the PAI analysis

The peripheral blood mQTL summary data were from the Brisbane Systems Genetics Study (BSGS)[40] (*n*=614) and Lothian Birth Cohorts (LBC) of 1921 and 1936[41] (*n*=1,366). We performed a meta-analysis of the two cohorts and identified 90,749 DNAm probes with at least a cis-mQTL at $P_{\mathrm{mQTL}} < 5\times10^{-8}$ (excluding the probes in the major histocompatibility complex (MHC) region because of the complexity of this region), of which 28,732 DNAm probes were in the promoter regions defined by the annotation data derived from 23 REMC blood samples (T-cell, B-cell, and Hematopoietic stem cells). The prefrontal cortex mQTL summary data were from the Religious Orders Study and Memory and Aging Project (ROSMAP)[34] (*n*=468), comprising 419,253 probes and approximate 6.5 million genetic variants. In the ROSMAP data, there were 67,995 DNAm probes with at least a cis-mQTL at $P_{\mathrm{mQTL}} < 5\times10^{-8}$ (not including the probes in the MHC region), of which 22,285 DNAm probes were in the promoter regions defined by the annotation data derived from 10 REMC brain samples. For all the DNAm probes, enhanced annotation data from Price *et al.*[42] (**URLs**) were used to annotate the closest gene of each DNAm probe.

We included in the analysis 15 complex traits (including disease) as analysed in Wu *et al.*[19]. They are height[43], body mass index (BMI)[44], waist-hip-ratio adjusted by BMI (WHRadjBMI)[45], high-density lipoprotein (HDL)[46], low-density lipoprotein (LDL)[46], thyroglobulin (TG)[46], educational years (EY)[47], rheumatoid arthritis (RA)[48], schizophrenia (SCZ)[49], coronary artery disease (CAD)[50], type 2 diabetes (T2D)[51], Crohn's disease (CD)[52], ulcerative colitis (UC)[52], Alzheimer's disease (AD)[53] and inflammatory bowel disease (IBD)[52]. The GWAS summary data were from the large GWAS meta-analyses (predominantly in samples of European ancestry) with sample sizes of up to 339,224. The number of SNPs varied from 2.5 to 9.4 million across traits.

Annotations of the chromatin state

The epigenomic annotation data used in this study were from the Roadmap Epigenomics Mapping Consortium (REMC), publicly available at http://compbio.mit.edu/roadmap/. We used these data to annotate the functional relevance of the DNAm sites and their cell type or tissue specificity. The chromatin state annotations from the Roadmap Epigenomics Project[13] were predicted by ChromHMM[12] based on the imputed data of 12 histone-modification marks. It contains 25 functional categories for 127 epigenomes in a wide range of primary tissue and cell types (**URLs**). The 25 chromatin states were further combined into 14 main functional annotations (as shown in **Fig. 4b** and Wu *et al.*[19]).

Overlap of the predicted PAIs with Hi-C, PCHi-C and ChIA-PET data

To test the overlap between our predicted PAIs and chromatin contacts detected by Hi-C, PCHi-C or ChIA-PET, we used chromatin contact loops and topological associated domains (TADs) data from the Rao *et al.* study called in the GM12812 cells[21] and the Dixon *et al.* study in embryonic stem cells[26], PCHi-C interaction data generated from human primary hematopoietic cells[5], and the *POLR2A* ChIA-PET chromatin loops from the ENCODE project[27] (**Supplementary Table 1**). To assess the statistical significance of the enrichment, we generated a null distribution by randomly sampling 1,000 sets of control probe pairs (with the same number as that of the predicted PAIs) from the distance-matched probe pairs tested in the SMR analysis. We mapped both the predicted PAIs and the control probe pairs to the TAD regions or chromatin contact loops detected by previous experimental assays and quantified the number of overlapping pairs. We estimated the fold enrichment by the ratio of the overlapping number for the predicted PAIs to the mean of the null distribution and computed the empirical *P* value by comparing the overlapping number for the predicted PAIs with the null distribution.

We used the chromatin interaction data generated by Jung *et al.*[28] in GM12878 cell lines as a validation set to evaluate the performance of different interaction prediction methods. We quantified the enrichment of the predicted interactions by different methods in the significant PCHi-C loops defined based on a range of PCHi-C *P* value thresholds and used the Fisher's exact test to assess the statistical significance of the enrichment.

Enrichment of the PIDSs in functional annotations

To conduct an enrichment test of the promoter interacting DNAm sites (PIDSs) in different functional annotation categories, we first extracted chromatin state data of 23 blood samples from the REMC samples. We then mapped the PIDSs to 14 main functional categories based on the physical positions, and counted the number of PIDSs in each functional category. Again, we generated a null distribution by randomly sampling the same number of control probes (with variance in DNAm level matched with the PIDSs) from all the probes tested in the PAI analysis and repeated the random sampling 1,000 times. The fold enrichment was calculated by the ratio of the observed value to the mean of the null distribution, and an empirical *P* value was computed by comparing the observed value with the null distribution.

Quantifying the expression levels of Pm-PAI genes

To quantify the expression levels of genes whose promoters were involved in the predicted PAIs (Pm-PAI genes), we used gene expression data (measured by Transcript Per Kilobase Million mapped reads (TPM)) from blood samples of the Genotype-Tissue Expression (GTEx) project[32]. We classified all the genes into two groups based on their expression levels in GTEx blood, i.e., active and inactive (TPM < 0.1). For the active genes, we further divided them into four quartiles based on their expression levels in GTEx blood, and counted the number of Pm-PAI genes in each of the five groups. To generate the null distribution, we randomly sampled the same number of control genes whose promoter DNAm sites were included in the SMR analysis, and repeated the random sampling 1,000 times. We computed the number of Pm-PAI genes and control genes in each group and assessed the significance by comparing the number of Pm-PAI genes with the null distribution in each group. We further tested the enrichment of the Pm-PAI genes against a null distribution sampled from all genes.

Enrichment of eQTLs and gene-associated DNAm in the PIDS regions

The eQTL enrichment analysis was conducted using all the independent cis-eQTLs (m=11,204) from the CAGE[33] study. The independent cis-eQTLs were from SNP-probe associations ($P < 5\times10^{-8}$) after clumping analysis in PLINK[54] followed by a conditional and joint (COJO) analysis in GCTA[55]. We only retained the cis-eQTLs whose target genes had at least a PIDS and mapped the cis-eQTL to a 10 Kb region centred around each corresponding PIDS of a Pm-PAI gene. To assess the significance of the enrichment, we generated a null distribution by mapping the cis-eQTLs to the same number of control gene-DNAm pairs (strictly speaking, it is the bait DNAm probe in the promoter of a gene together with another non-promoter DNAm probe) randomly sampled (with 1,000 repeats) from those included in the PAI analysis with the distance between a control pair matched with that between a Pm-PAI gene and the corresponding PIDS. In addition, we have identified a set of DNAm sites that showed pleiotropic associations with gene expressions in a previous study[19]. We used the same approach as described above to test the significance of enrichment of the gene-associated DNAm sites in the PIDSs.

Supplemental information

Supplemental data include 12 supplemental figures and 4 supplemental tables.

URLs

M2Mdb, http://cnsgenomics.com/shiny/M2Mdb/

SMR, http://cnsgenomics.com/software/smr

GTEx, http://www.gtexportal.org/home/

Annotation file for the Illumina HumanMethylation450 BeadChip, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL16304

Acknowledgements

We thank Peter Visscher for helpful discussion. This research was supported by the Australian Research Council (DP160101343, DP160101056 and FT180100186), the Australian National Health and Medical Research Council (1107258, 1083656, 1078901 and 1113400), and the Sylvia & Charles Viertel Charitable Foundation. The Lothian Birth Cohorts (LBC) are supported by Age UK (Disconnected Mind programme). Methylation typing was supported by Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, The Wellcome Trust Institutional Strategic Support Fund, The University of Edinburgh, and The University of Queensland. The LBC resource is prepared in the Centre for Cognitive Ageing and Cognitive Epidemiology, which is supported by the Medical Research Council and Biotechnology and Biological Sciences Research Council (MR/K026992/1), and which supports I.J.D.. This study makes use of data from dbGaP (accessions: phs000428.v1.p1 and phs000424.v1.p1) and EGA (accession: EGAS00001000108). A full list of acknowledgements to these data sets can be found in the **Supplementary Note 3**.

Author Contributions

J.Y. conceived and supervised the study. Y.W., T.Q. and J.Y. designed the experiment. Y.W. and T.Q. performed simulations and statistical analyses under the assistance or guidance from J.Y., J.Z., H.W., F.Z., and Z.Z.. I.J.D., N.R.W. and A.F.M. contributed the blood DNA methylation data. J.E.P.C. provided critical advice that significantly improved the interpretation of the results. N.R.W. and J.Y. contributed funding and resources. Y.W., T.Q., J.Z. and J.Y. wrote the manuscript with the participation of all authors.

Declaration of Interests

We declare that all authors have no competing interests.

References

1. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* **45**, D896-D901 (2017).
2. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017).

3. Farh, K.K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-343 (2015).

4. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *New England Journal of Medicine* **373**, 895-907 (2015).

5. Javierre, B.M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384 e19 (2016).

6. Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**, 93-108 (2016).

7. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289-293 (2009).

8. Fullwood, M.J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64 (2009).

9. Wit, E.d. & Laat, W.d. A decade of 3C technologies: insights into nuclear organization. *Genes & Development* **26**, 11-24 (2012).

10. Kuo, M.-H. & Allis, C.D. In Vivo Cross-Linking and Immunoprecipitation for Studying Dynamic Protein:DNA Associations in a Chromatin Environment. *Methods* **19**, 425-433 (1999).

11. Belton, J.M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268-76 (2012).

12. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-6 (2012).

13. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).

14. Zhu, Y. *et al.* Constructing 3D interaction maps from 1D epigenomes. *Nat Commun* **7**, 10812 (2016).

15. Huang, J., Marco, E., Pinello, L. & Yuan, G.-C. Predicting chromatin organization using histone marks. *Genome Biology* **16**, 162 (2015).

16. Fortin, J.-P. & Hansen, K.D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology* **16**, 180 (2015).

17. Kumasaka, N., Knights, A.J. & Gaffney, D.J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nature Genetics* **51**, 128-137 (2019).

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

19. Wu, Y. *et al.* Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nature Communications* **9**, 918 (2018).

20. McRae, A.F. *et al.* Identification of 55,000 Replicated DNA Methylation QTL. *Scientific Reports* **8**, 17605 (2018).

21. 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).

22. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics* **23**, R89-R98 (2014).

23. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32**, 1-22 (2003).

24. Sonnega, A. *et al.* Cohort Profile: the Health and Retirement Study (HRS). *Int J Epidemiol* **43**, 576-85 (2014).

25. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat Rev Mol Cell Biol* **16**, 245-57 (2015).

26. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-80 (2012).

27. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

28. Jung, I. *et al.* A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat Genet* **51**, 1442-1449 (2019).

29. Gate, R.E. *et al.* Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat Genet* **50**, 1140-1150 (2018).

30. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290 (2013).

31. Li, G. *et al.* Extensive Promoter-centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell* **148**, 84-98 (2012).

32. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648-660 (2015).

33. Lloyd-Jones, L.R. *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. *Am J Hum Genet* **100**, 371 (2017).

34. Ng, B. *et al.* An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nature Neuroscience* **20**, 1418 (2017).

35. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nature Communications* **9**, 2282 (2018).

36. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications* **8**, 1826 (2017).

37. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414 e24 (2016).

38. Forcato, M. *et al.* Comparison of computational methods for Hi-C data analysis. *Nat Methods* **14**, 679-685 (2017).

39. Kuonen, D. Saddlepoint Approximations for Distributions of Quadratic Forms in Normal Variables. *Biometrika* **86**, 929-935 (1999).

40. Powell, J.E. *et al.* The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. *PLoS One* **7**, e35430 (2012).

41. Chen, B.H. *et al.* DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging (Albany NY)* **8**, 1844-1865 (2016).

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

43. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-86 (2014).

44. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).

45. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187-96 (2015).

46. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274-83 (2013).

47. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539-42 (2016).

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

49. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-7 (2014).

50. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* **47**, 1121-30 (2015).

51. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics* **44**, 981 (2012).

52. Liu, J.Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979-986 (2015).

53. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452-8 (2013).

54. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

55. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44**, 369 (2012).

56. Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162**, 1051-65 (2015).

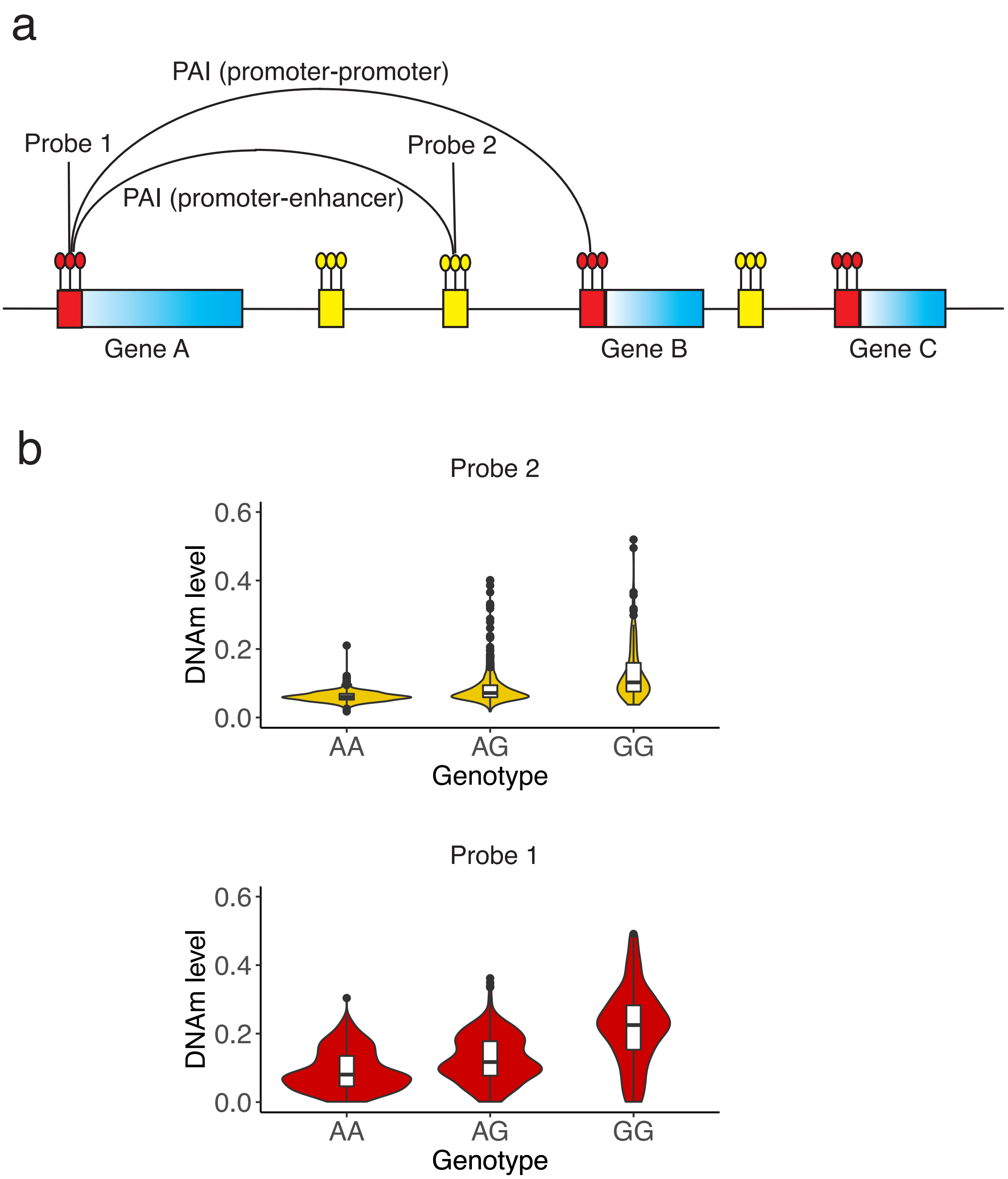

Figure 1 Schematic of the promoter-anchored interaction (PAI) analysis. Panel a): a schematic of the PAI analysis. The blue rectangles represent genes with their promoter regions color coded in red. The small yellow bars represent other functional regions (e.g., enhancers). In this toy example, the promoter region of Gene A is used as the bait for the PAI analysis. Genes (e.g., genes A and B) whose promoters are involved in significant PAIs are defined as Pm-PAI genes. DNAm sites (e.g., DNAm probe 2) that showed significant interactions with the DNAm sites in promoter regions are defined as promoter-interacting DNAm sites or PIDS. Panel b): variation between people in DNAm levels of two CpG sites are associated because of a shared causal variant. The DNAm level ranges

from 0 to 1 (with 0 being unmethylated and 1 being fully methylated). It is the ratio of the methylated probe intensity to the overall intensity (sum of methylated and unmethylated probe intensities).

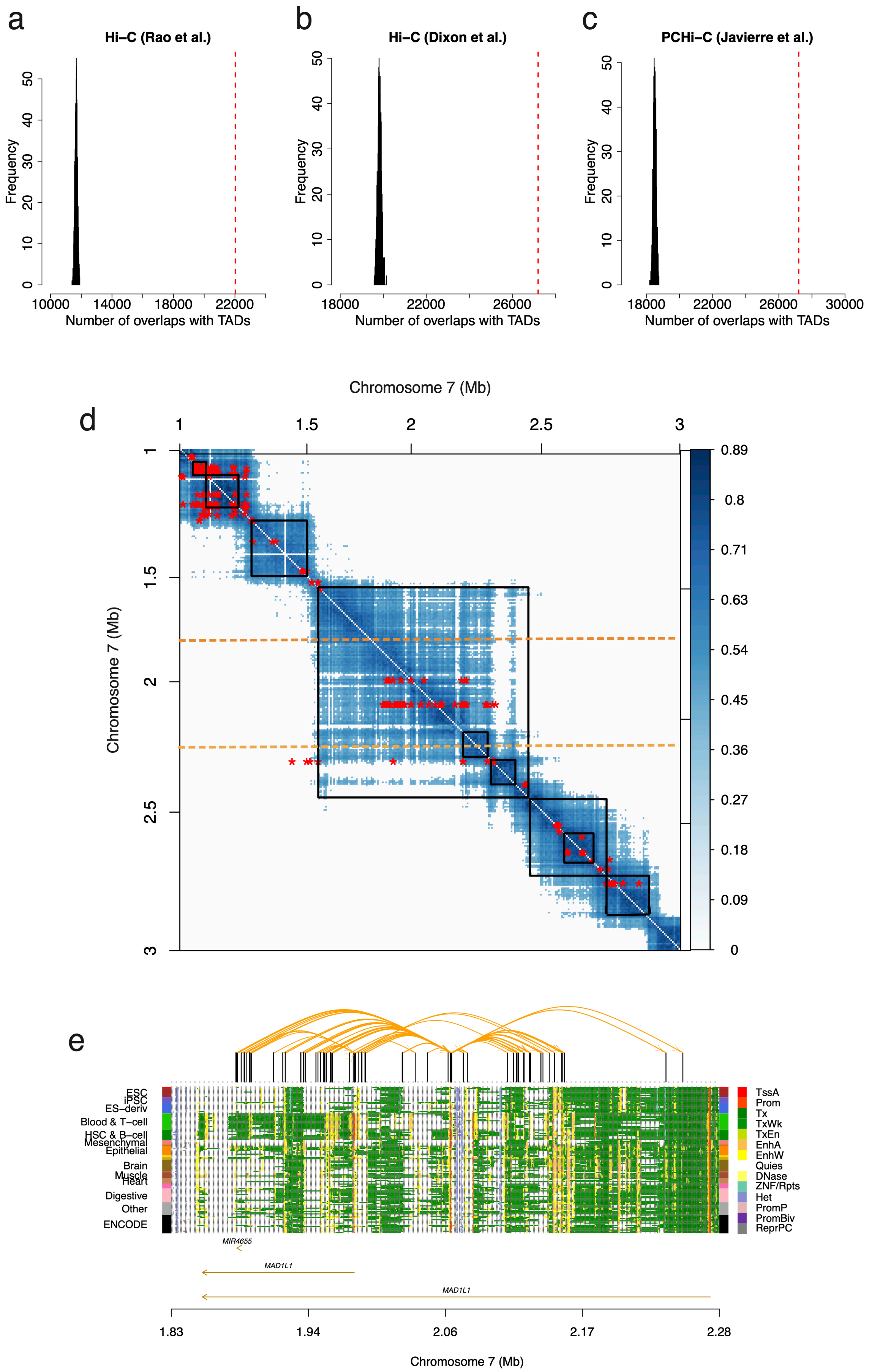
a
Hi–C (Rao et al.)
Frequency
0
10
20
30
40
50
10000
14000
18000
22000
Number of overlaps with TADs
b
Hi–C (Dixon et al.)
Frequency
0
10
20
30
40
50
18000
22000
26000
Number of overlaps with TADs
c
PCHi–C (Javierre et al.)
Frequency
0
10
20
30
40
50
18000
22000
26000
30000
Number of overlaps with TADs
d
Chromosome 7 (Mb)
1
1.5
2
2.5
3
Chromosome 7 (Mb)
0.89
0.8
0.71
0.63
0.54
0.45
0.36
0.27
0.18
0.09
0
e
ESC
iPSC
ES–deriv
Blood & T–cell
HSC & B–cell
Mesenchymal
Epithelial
Brain
Muscle
Heart
Digestive
Other
ENCODE
TssA
Prom
Tx
TxWk
TxEn
EnhA
EnhW
Quies
DNase
ZNF/Rpts
Het
PromP
PromBiv
ReprPC
MIR4655
MAD1L1
MAD1L1
1.83
1.94
2.06
2.17
2.28
Chromosome 7 (Mb)

Figure 2 Overlap of the predicted PAIs with TADs identified by Hi-C and PCHi-C. Panels a), b) and c): overlaps of the predicted PAIs with TADs identified by a) Rao *et al.*[21] and b) Dixon *et al.*[26] using Hi-C and by c) Javierre *et al.*[5] using PCHi-C. The red dash lines represent the observed number and histograms represent the distribution of control sets. Panel d): a heatmap of the predicted PAIs (red asterisks) and chromatin interactions with correlation scores > 0.4 (blue dots) identified by Grubert *et al.*[56] using Hi-C in a 2 Mb region on chromosome 7. Black squares represent the TADs identified by Rao *et al.*[21]. The heatmap is asymmetric for the PAIs (red asterisks) with the *x*- and *y*-axes representing the physical positions of outcome and exposure DNAm probes, respectively. Panel e): the predicted PAIs at the *MAD1L1* locus, a 450-Kb sub-region of that shown between two orange dashed lines in panel d). The orange curved lines on the top represent the significant PAIs between 14 DNAm sites in the promoter regions of *MAD1L1* (multiple transcripts) and other DNAm sites nearby. The panel on the bottom represents 14 chromatin state annotations (indicated by different colours) inferred from data of 127 REMC samples (one row per sample). Note that the predicted PAIs appear to be much sparser than the Hi-C loops largely because the PAIs were predicted from analyses with very stringent significance levels (**Supplementary Note 1**).

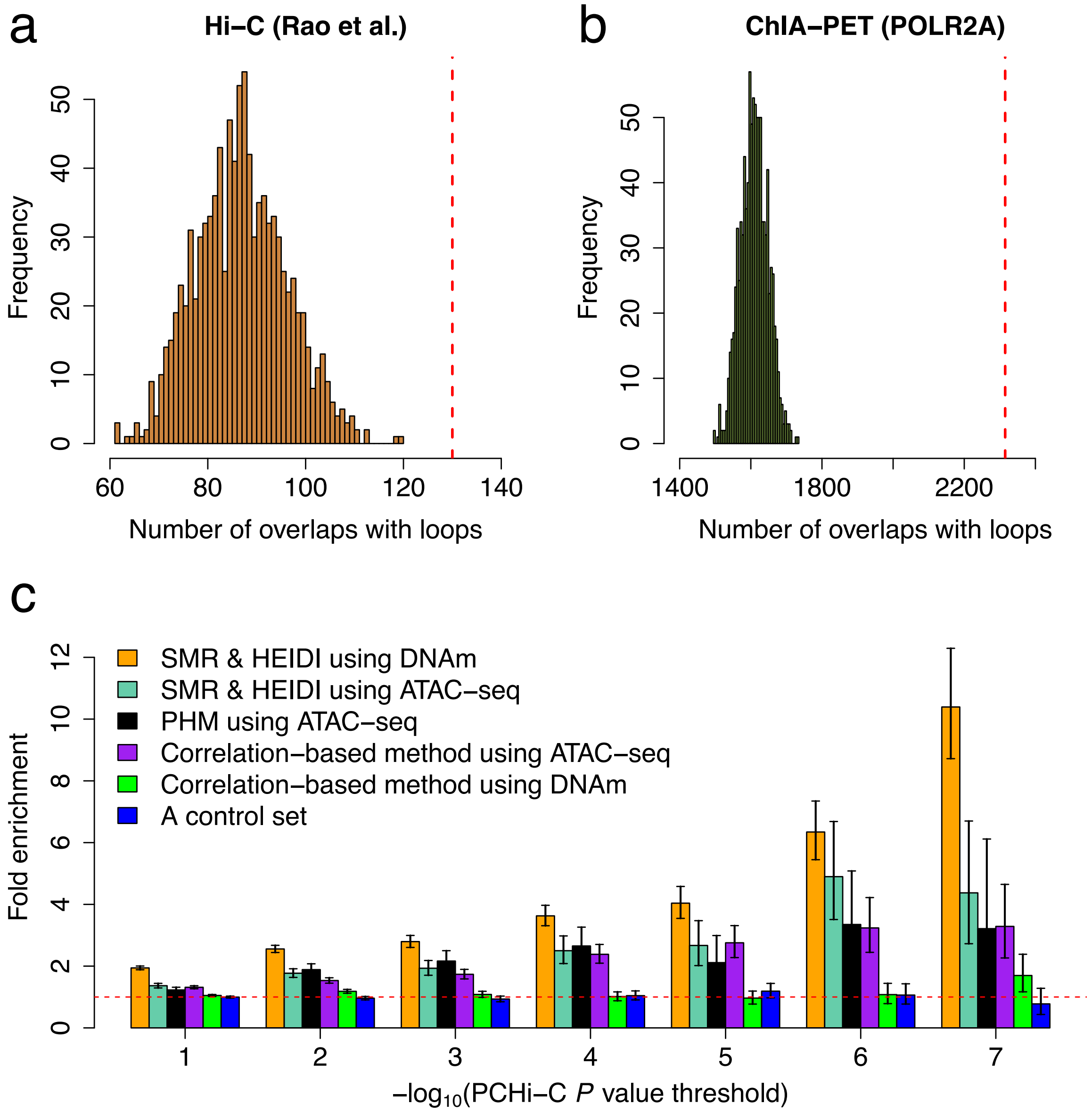

Figure 3 Enrichment of the predicted interactions in chromatin loops identified by experimental assays. Panels a) and b): overlaps of the predicted PAIs with the chromatin loops identified by a) Hi-C from Rao *et al.*[21] and b) *POLR2A* ChIA-PET from the ENCODE project[27]. The red dash lines represent the observed number and histograms represent the distribution of control sets. Panel c): enrichment of the predicted interactions in the significant PCHi-C loops defined based on a range of *P* value thresholds. We used the PCHi-C loops identified from Jung *et al.* in GM12878 cell lines[28]. PHM: the pairwise hierarchical model developed by Kumasaka *et al.*[17]. The error bar around each estimate represents the 95% confidence interval. The horizontal red dashed line indicates no enrichment.

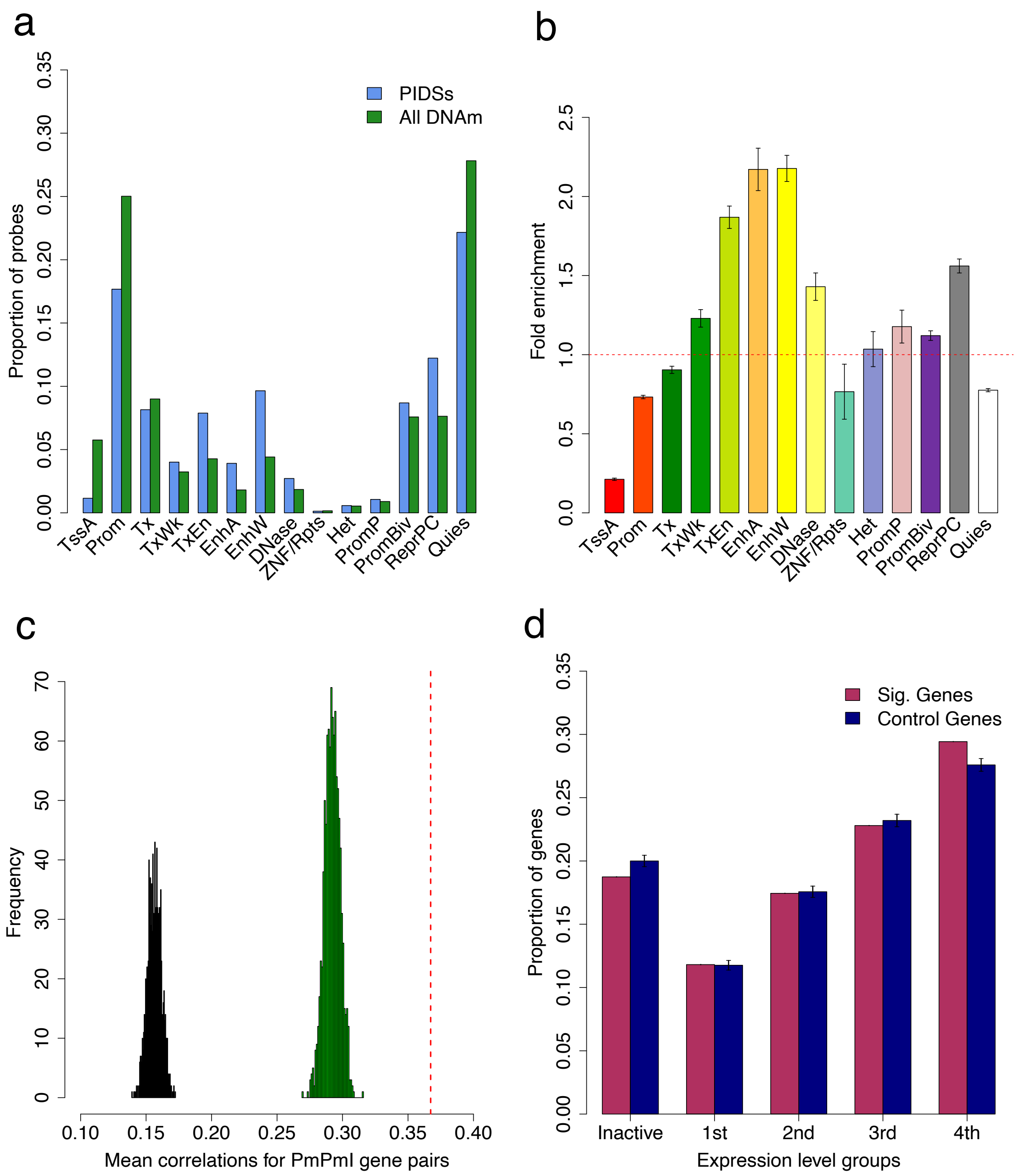

Figure 4 Enrichment of PIDSs and Pm-PAI genes. Panels a) and b): enrichment of PIDSs in 14 main functional annotation categories inferred from the 127 REMC samples. Fold enrichment: a ratio of the proportion of PIDSs in an annotation category to the mean of the control sets. The error bar in panel b) represents standard deviation of the estimate under the null obtained from the control sets. The 14 functional categories are: TssA, active transcription start site; Prom, upstream/downstream TSS promoter; Tx, actively transcribed state; TxWk, weak transcription; TxEn, transcribed and regulatory Prom/Enh; EnhA, active enhancer; EnhW, weak enhancer; DNase, primary DNase; ZNF/Rpts, state associated with zinc finger protein genes; Het, constitutive heterochromatin; PromP, Poised promoter; PromBiv, bivalent regulatory states;

ReprPC, repressed Polycomb states; and Quies, a quiescent state. Panel c): mean Pearson correlation of expression levels for gene pairs whose promoters were involved in PmPmI. The red dash line represents the observed mean Pearson correlation value of the significant PmPmI gene pairs and the histograms represent the null distributions of mean Pearson correlation values generated by repeated resampling of a set of distance-matched control gene pairs either from the genes whose promoters were involved in the SMR analysis (green) or from all genes (black). Panel d): proportion of Pm-PAI genes in five gene activity groups with the first group being the inactive group (TPM <0.1) together with four quartiles defined based on the expression levels of all genes in the GTEx blood samples. The error bar represents the standard deviation estimated from the 1,000 control sets.

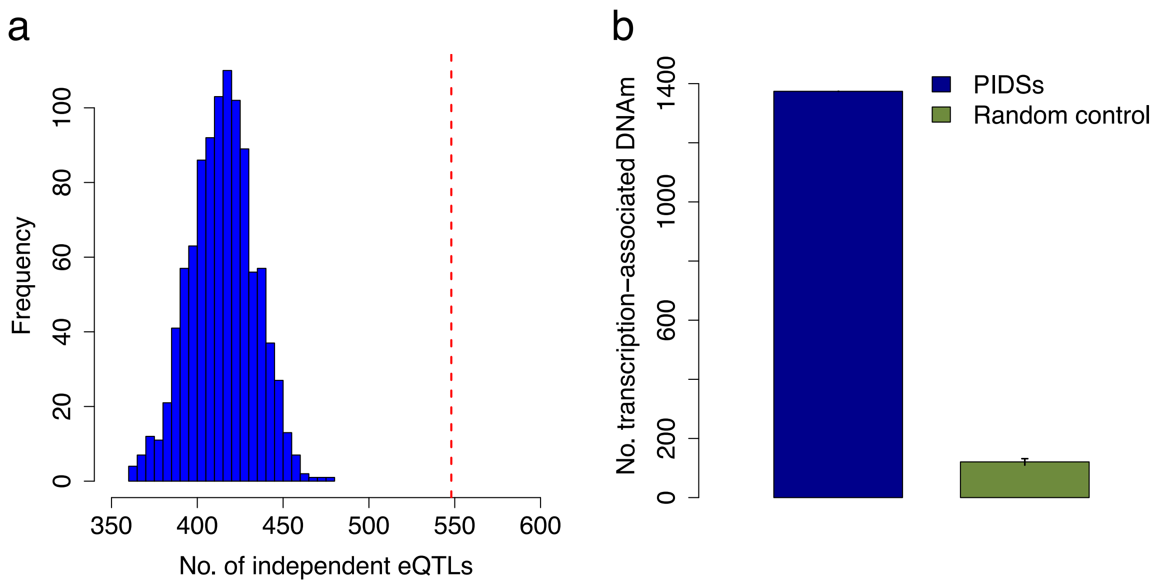

Figure 5 Enrichment of eQTLs or transcription-associated DNAm sites in PIDS regions of the Pm-PAI genes. Panel a): the number of independent cis-eQTLs ($P_{\mathrm{eQTL}} < 5\times10^{-8}$) located in PIDS regions of the Pm-PAI genes. The red dash line represents the observed number and the blue histogram represents the distribution of 1000 control sets. Panel b): the number of transcription-associated DNAm sites located in PIDS regions of the Pm-PAI genes. The blue bar represents the observed number and the green bar represents the mean of 1000 control sets. The error bar represents the standard deviation estimated from the control sets.

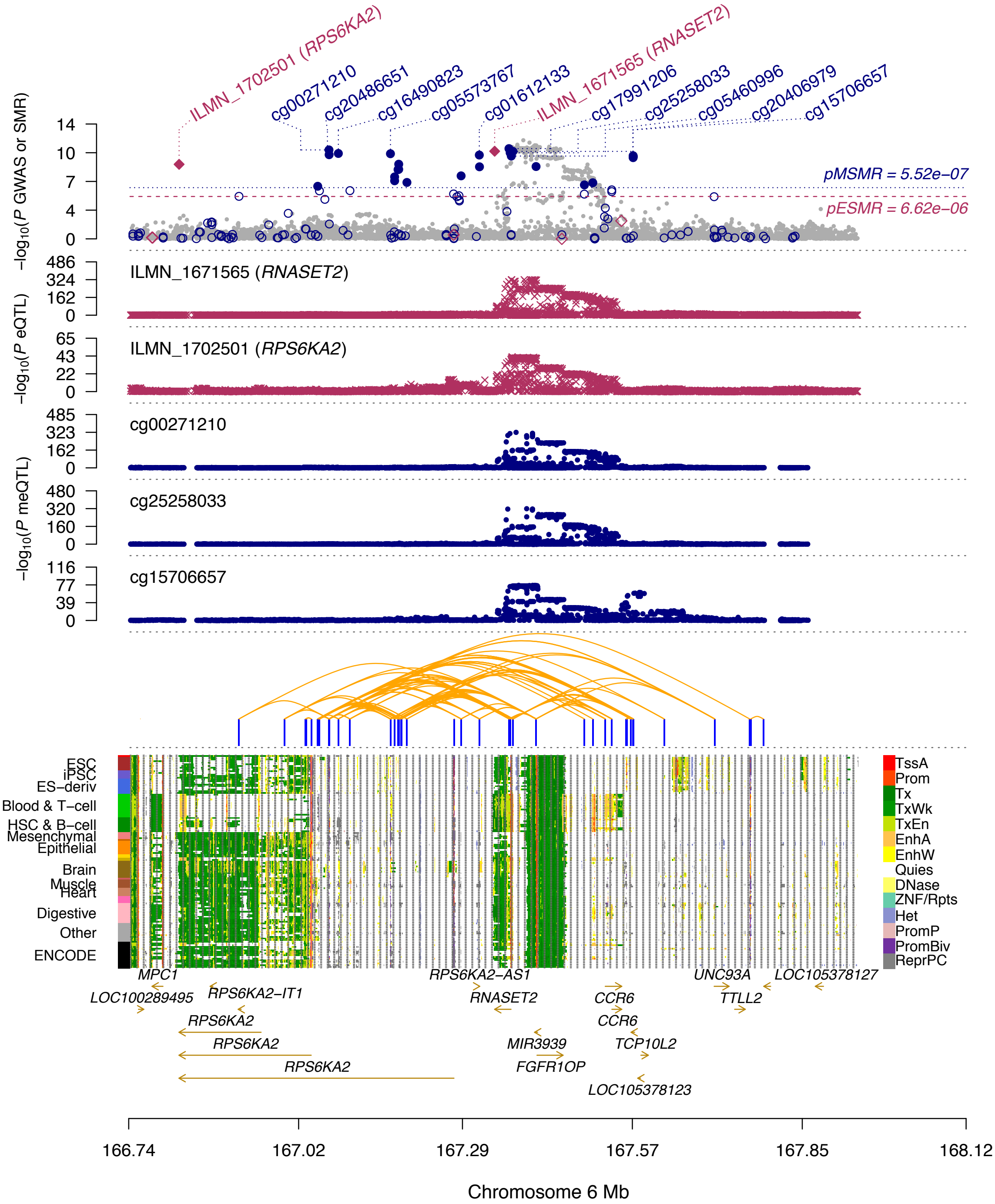

Figure 6 Prioritizing genes and functional regions at the *RPS6KA2* locus for Crohn's disease (CD). The top plot shows -log$_{10}$(P values) of SNPs from the GWAS meta-analysis (grey dots) for CD[48]. Red diamonds and blue circles represent -log$_{10}$(P values) from SMR tests for associations of gene expression and DNAm probes with CD, respectively. Solid diamonds and circles are the probes not rejected by the HEIDI test (P_{HEIDI}>0.01). The second and third plots show -log$_{10}$(P values) of SNP associations for the expression levels of probe ILMN_1671565 (tagging *RNASET2*) and ILMN_1702501 (tagging *RPS6KA2*), respectively, from the CAGE data. The fourth, fifth and sixth plots shows -log$_{10}$(P values) of SNP associations for the DNAm levels of probes cg00271210,

cg25258033, and cg15706657, respectively, from the mQTL meta-analysis. The panel on the bottom shows 14 chromatin state annotations (indicated by colours) inferred from 127 REMC samples (one sample per row) with the predicted PAIs annotated by orange curved lines on the top (see **Fig. S3a** for the overlap of the predicted PAIs with Hi-C data).