ARTICLE

Mechanisms and Disease Associations of Haplotype-Dependent Allele-Specific DNA Methylation

Catherine Do,[1,*] Charles F. Lang,[1] John Lin,[1] Huferesh Darbary,[1] Izabela Krupska,[1] Aulona Gaba,[1,2] Lynn Petukhova,[3] Jean-Paul Vonsattel,[4] Mary P. Gallagher,[5] Robin S. Goland,[6] Raphael A. Clynes,[7] Andrew Dwork,[4] John G. Kral,[8] Catherine Monk,[9] Angela M. Christiano,[10] and Benjamin Tycko[1,4,*]

Haplotype-dependent allele-specific methylation (hap-ASM) can impact disease susceptibility, but maps of this phenomenon using stringent criteria in disease-relevant tissues remain sparse. Here we apply array-based and Methyl-Seq approaches to multiple human tissues and cell types, including brain, purified neurons and glia, T lymphocytes, and placenta, and identify 795 hap-ASM differentially methylated regions (DMRs) and 3,082 strong methylation quantitative trait loci (mQTLs), most not previously reported. More than half of these DMRs have cell type-restricted ASM, and among them are 188 hap-ASM DMRs and 933 mQTLs located near GWAS signals for immune and neurological disorders. Targeted bis-seq confirmed hap-ASM in 12/13 loci tested, including *CCDC155, CD69*, *FRMD1*, *IRF1, KBTBD11*, and *S100A*-ILF2*, associated with immune phenotypes, *MYT1L, PTPRN2, CMTM8* and *CELF2*, associated with neurological disorders, *NGFR* and *HLA-DRB6*, associated with both immunological and brain disorders, and *ZFP57*, a *trans*-acting regulator of genomic imprinting. Polymorphic CTCF and transcription factor (TF) binding sites were over-represented among hap-ASM DMRs and mQTLs, and analysis of the human data, supplemented by cross-species comparisons to macaques, indicated that CTCF and TF binding likelihood predicts the strength and direction of the allelic methylation asymmetry. These results show that hap-ASM is highly tissue specific; an important *trans*-acting regulator of genomic imprinting is regulated by this phenomenon; and variation in CTCF and TF binding sites is an underlying mechanism, and maps of hap-ASM and mQTLs reveal regulatory sequences underlying supra- and sub-threshold GWAS peaks in immunological and neurological disorders.

Introduction

Statistical evidence from genome-wide association studies (GWASs) has implicated numerous DNA sequence variants, mostly SNPs, as candidates for inter-individual phenotypic differences and disease susceptibility. However, most of these variants reside in non-coding regions, and how they result in differences in phenotypes is not well understood. Moreover, multiple statistical comparisons demand stringent thresholds for significance, $p < 5 \times 10^{-8}$ for a GWAS,[1] and this level probably leads to the rejection of many biological true positives with sub-threshold p values. In fact, diseases interrogated with well-powered GWASs demonstrate that the majority of risk alleles have small effect sizes and would not achieve genome-wide significance in more moderately sized cohorts.[2] Therefore, methods to provide biological validity to both supra- and sub-threshold variants are a high priority. A combined genetic-epigenetic approach can help to address this challenge. In particular, identification of haplotype-dependent allele-specific methylation (hap-ASM) in the human genome by our group (Kerkel et al.[3]) and by Schalkwyk et al.[4] and Hellman and Chess,[5] followed by others,[6–15] led to suggestions that it might contribute to inter-individual phenotypic variation and that co-mapping this type of allelic asymmetry with GWAS data could prove useful for promoting GWAS statistical signals to biological true positives.[16,17]

ASM is a hallmark of two different phenomena: genomic imprinting, where the methylation of an allele is determined by its parent-of-origin, and (non-imprinted) hap-ASM, in which the local sequence context acts in *cis* to dictate the methylation status of local CpGs.[17] Hap-ASM can be assessed either directly by bisulfite sequencing (bis-seq) in heterozygotes or by methylation quantitative trait loci (mQTL) analysis, which correlates net methylation of single CpGs with genotypes at nearby SNPs. Mapping hap-ASM and mQTLs and superimposing these maps on GWAS data can support the biological relevance of GWAS peaks, the hypothesis being that detection of hap-ASM or an mQTL near a GWAS peak suggests the presence of a bona fide regulatory SNP or haplotype, which reveals its presence by conferring a physical asymmetry between the two alleles in heterozygotes. Additional evidence, including experiments in animal models, is needed for a complete understanding of a given locus, but the

[1]Institute for Cancer Genetics, Columbia University, New York, NY 10032, USA; [2]Department of Obstetrics and Gynecology, Vienna Medical University, Vienna 1090, Austria; [3]Departments of Epidemiology and Dermatology, Columbia University, New York, NY 10032, USA; [4]Departments of Pathology and Cell Biology, Columbia University, New York, NY 10032, USA; [5]Division of Pediatric Endocrinology, Diabetes and Metabolism, Columbia University, New York, NY 10032, USA; [6]Naomi Berrie Diabetes Center, Columbia University, New York, NY 10032, USA; [7]Departments of Pathology, Medicine, and Dermatology, Columbia University, New York, NY 10032, USA; [8]Departments of Surgery, Medicine, and Cell Biology, SUNY Downstate Medical Center, Brooklyn, NY 11203, USA; [9]Departments of Psychiatry and Behavioral Medicine and Obstetrics and Gynecology, Columbia University, New York, NY 10032, USA; [10]Departments of Dermatology, Genetics, and Development, Columbia University, New York, NY 10032, USA
*Correspondence: cd2695@columbia.edu (C.D.), bt12@columbia.edu (B.T.)
http://dx.doi.org/10.1016/j.ajhg.2016.03.027.

combined hap-ASM/mQTL/GWAS method, and related methods such as eQTL/GWAS analysis,[18] allows genome-wide screening for regulatory loci, which can then be prioritized for such studies.

Understanding the mechanisms of hap-ASM could provide additional insights. Previously, we documented examples of genes with hap-ASM in which the differentially methylated regions (DMRs) are discrete in size (1 to 2 kb) and precisely overlap with binding sites for the insulator protein CTCF,[14] and we proposed a model for hap-ASM in which polymorphisms in CTCF binding sites abrogate CTCF binding in a haplotype-dependent manner and lead to preferential CpG methylation of the unoccupied allele.[14,17] Here, we test this mechanism by genome-wide and fine-mapping of CpG methylation patterns in human tissues, supplemented by cross-species comparisons of methylation patterns in *Homo sapiens* and macaques. In parallel, we identify examples of strong hap-ASM DMRs and mQTLs in T cells and brain, many of which are tissue specific, not previously reported, and located near supra- and sub-threshold GWAS peaks for immunological and neurological diseases. Lastly, we find that an important *trans*-acting regulator of genomic imprinting is in turn regulated by hap-ASM.

Material and Methods

Human and Macaque Tissues

The human tissues and cell types for this study are in Table S1. CD3-positive T lymphocytes were isolated by negative selection (RosetteSep, Sigma) from peripheral blood samples obtained with informed consent. The GM12878 cell line DNA was purchased from Coriell. All other human tissues were obtained from autopsies. The mQTL discovery set included samples from 54 total CD3-positive T cells, 44 temporal cortex (TC) gray matter samples, 18 NeuN-positive TC neurons isolated by FANS (see below), 22 non-neuronal brain cells (NeuN-negative from FANS; referred to hereafter as glia), and 37 placenta samples (tissue taken from immediately below the fetal surface and therefore not contaminated by maternal decidua). The hap-ASM discovery set for Methyl-Seq, designed to represent a diverse human tissues, included three T cell, three brain, two heart, two liver, one lung, and two placenta samples. Informed consent was given for the human samples and institutional review board (IRB) approval obtained. The rhesus and bonnet macaque (*Macaca mulatta* and *radiata*) tissues, including peripheral blood and liver, were obtained from necropsies. The maintenance of the macaque colony was approved by the Institutional Animal Care and Use Committee of SUNY Downstate Medical Center.

Isolation of Neuronal and Glial Nuclei by Fluorescence-Activated Nuclear Sorting

Frozen tissue (0.25 to 0.5 g) was homogenized for 1 min on ice, then layered over a 60% sucrose cushion and centrifuged at 28,500 rpm for 2.5 hr at 4 degrees, as described by Matevossian and Akbarian.[19] The pelleted nuclei were re-suspended and incubated with anti-NeuN-Alexa 488 conjugated antibody (Millipore cat# MAB377X; RRID: AB_2149209) for 1 hr at 4°C prior to FANS. Both neuronal-enriched NeuN-positive and glial-enriched NeuN-negative samples were recovered. Enrichment of NeuN-positive nuclei was confirmed by using an aliquot of the FANS samples to prepare cytospin slides and visualized by fluorescence microscopy. To assist nuclei visualization, the DNA was stained with To-Pro3. Only samples that showed >95% Neu-N-positive cells on analysis by NeuN immunofluorescence on cytospin slides were utilized for DNA extraction. DNA was prepared from the flow-sorted nuclei by standard SDS/proteinase-K lysis followed by precipitation in 80% isopropanol with glycogen carrier.

Illumina 450K Methylation Profiling and SNP Genotyping for Identifying mQTLs

Genomic DNA (500 ng) was used as per the manufacturer's instructions for HumanMethylation450 Beadchips (Illumina), with all assays performed at the Roswell Park Cancer Institute (RPCI) Genomics Shared Resource. Data were processed with Genome Studio, which calculates the fractional methylation (AVG_Beta) at each queried CpG, after background correction, normalization to internal control probes, and quantile normalization. All probes mapping to the X or Y chromosome were removed, along with probes that queried CpGs overlapping the positions of known common DNA variants as reported in dbSNP build 138 (allele frequency $\geq$ 1%), because these CpGs are destroyed by the SNP itself. As recommended by Illumina, AVG_Beta values with a detection p value > 0.05 were excluded from the analysis and replaced by missing values. A small number of probes (<0.06%) with more than 20% missing values were excluded. In parallel, the DNA samples were genotyped on Illumina HumanOmni2.5 Beadchips, followed by initial data processing with Genome Studio. SNPs were annotated with dbSNP138.

We mapped mQTLs in each tissue and cell type using combined SNP genotyping and 450K methylation data. Only SNPs with at least three samples per genotype (AA, AB, and BB, where A is the reference and B the alternate allele), with frequencies in Hardy-Weinberg equilibrium, were analyzed. Exact tests of Hardy-Weinberg equilibrium for two allele markers were performed with R "genetics" package and SNP genotypes with a p value < 0.05 were considered in disequilibrium and excluded from further analyses. Previous studies have found that methylation at mQTLs correlates best with nearby SNPs, with the correlations decaying rapidly over 1–2 kb and the majority of mQTLs located within 100 kb.[8,12,15] Because our data confirmed this finding, we focused on CpGs located within 75 kb on either side of heterozygous SNPs (150 kb windows) to assess SNP-CpG methylation correlations. The SNPs and CpGs meeting these criteria are referred to hereafter as index SNPs and informative CpGs. For mQTLs, we required that the fractional methylation of a given CpG should be a linear function of the genotype of the index SNP, where a numeric value has been assigned to each genotype as follows: AA = 0, AB = 1, and BB = 2. The coefficient β of the function reflects the additive effect of one alternative allele, such that 2β is the methylation difference between allele A and B. Thus, we searched for mQTLs using linear regression to model the relationship between the fractional methylation (AVG_Beta) and index SNP genotype. Each index SNP-CpG pair within the 150 kb window was tested. We defined mQTLs as CpGs with an R-squared ≥ 0.5 (which reflects the goodness of the fit to the linear function), a coefficient $\beta \geq 0.1$, and p value corrected for multiple testing using Benjamini-Hochberg method < 0.05 (FDR at 5%). The importance of using each of these criteria is explained further in Figure S1. We ranked the mQTLs by strength of allelic asymmetry using the geometric mean of the difference in fractional methylation between the A and B alleles, and the R-squared value. Potential batch effects in the 450K data were assessed by multivariate linear regression including the batch and genotype as explanatory covariates. Minimal batch effects were observed, such that 94% of the mQTLs remained significant after batch correction. For testing mechanistic hypotheses by enrichment analyses, to avoid bias due to multiple mQTL CpGs within a small window, and to parallel our deep bis-seq data where the average DNA fragment length was 250 bp, consecutive mQTL CpGs within 250 bp were considered as a single DMR.

In the data from the brain samples, half of which were neuropathologically normal and the other half affected by late-onset Alzheimer disease (AD [MIM: 104300]), we ruled out the effect of AD status on the identified mQTLs by performing multivariate linear regression, including the genotype, the disease status, and the interaction term between disease status and genotype as explanatory covariates. The coefficient β of the disease reflects the overall effect of AD on net methylation, and the coefficient β of the interaction term reflects the effect of AD on the correlation between methylation and genotype. Thus, a significant coefficient β of the disease suggests a methylation difference between AD versus control samples, and a significant coefficient β of the interaction term suggests that mQTL presence or strength is different between AD and controls. All analyses were performed with R.

Agilent SureSelect Methyl-Seq for Mapping ASM

We used the Agilent SureSelect Methyl-Seq DNA hybrid capture kit, followed by Nextgen bis-seq. In this protocol, targeted regions (total of 3.7M CpGs) include RefGenes, promoter regions, CpG islands, CpG island shores, shelves, and DNase I hypersensitive sites. DNA was sheared to an average size of 250 bp and bisulfite converted with the EZ DNA methylation kit (Zymo). Paired-end reads (150 bp) were generated with an Illumina HiSeq2000 sequencer. One of the brain samples was relatively under-represented in the library, so additional sequences were generated on a MiSeq sequencer to improve the coverage for this sample. After trimming for low-quality bases (Phred score < 30) and reads with a length < 40 bp with TrimGalore, the reads were aligned to the human genome (GRCh37) using Bismark[20] and duplicate reads were removed with Samtools. Coverage metrics were calculated with Picard tools and SNP calling was performed with BisSNP.[21] Genotyping was carried out with human genome GRCh37 and dbSNP137 as references. We filtered out heterozygous SNPs with fewer than ten reads per allele. Bisulfite treatment converts unmethylated C residues to T, whereas methylated C residues are not converted. Therefore, for C/T and G/A SNPs (depending on the strand), the distinction between the alternate allele and bisulfite conversion is not possible and these SNPs are not informative for hap-ASM analysis. However, because Agilent SureSelect captures negative stranded DNA fragments, only G/A SNPs needed to be filtered out. ASM calling was performed with Bismark, after separating the valid SNP-containing reads by allele. Informative SNPs were defined as non-G/A heterozygous SNPs that passed BisSNP criteria and were covered by more than ten reads per allele. Informative regions were defined as regions with overlapping reads covering at least one informative SNP. To further increase the stringency and accuracy of ASM calling, only regions with at least three CpGs covered by more than ten reads per allele were considered. ASM CpGs were then defined as CpGs with Fisher's exact test p value < 0.05. Hap-ASM regions were defined as regions with $\geq 20\%$ methylation difference after averaging all CpGs covered, and a Wilcoxon p value corrected for multiple testing by Benjamin-Hochberg method < 0.05 (FDR at 5%). We ranked the ASM regions by allelic asymmetry using the geometric mean of the methylation difference, number of ASM CpGs, and percentage of ASM CpGs among all covered CpGs. Data post-processing was performed by R. Although this study focuses on hap-ASM, genomic imprinting also produces ASM, affecting approximately 100 DMRs. Therefore, we used the GeneImprint database to flag all known imprinted chromosomal domains, which additionally served as positive internal controls for ASM detection in our experiment. ASM regions within ±75 kb of the transcription starting site of known imprinted genes were not considered as hap-ASM and are listed separately in Table S5.

Validations and Fine-Mapping of Hap-ASM DMRs via Targeted Bis-Seq

Targeted bis-seq was utilized for validation and fine-mapping of hap-ASM and mQTL regions. Primers (Table S2) were designed in MethPrimer, and bisulfite-converted DNA was amplified by PCR, followed by either Sanger or Nextgen (Illumina MiSeq) sequencing. Sample preparation for MiSeq was performed on a Fluidigm AccessArray high-throughput PCR machine with sample bar-codes incorporated in a second round of PCR, as described.[14] PCRs for Sanger and MiSeq were performed in triplicate and pooled to ensure sequence complexity. For Sanger sequencing, PCR products were cloned using the TopoTA Cloning System (Invitrogen) as described.[22] ASM was assessed when the coverage was at least 10 clones per allele for Sanger sequencing and 100 DNA fragments for MiSeq sequencing. Although the absolute differences between methylation of the two alleles are not exaggerated by very deep sequencing, the p values tend to zero as the number of reads increases. Therefore, to allow comparisons with the Sanger sequencing data, for the MiSeq data we carried out bootstrapping (1,000 random samplings, 20 reads per allele). Significance of methylation differences between alleles was assessed via the Wilcoxon test. Samplings and bootstrapping were performed with R. For graphical representations of the MiSeq data, one representative sample is shown.

Bis-Seq Analysis of 5mC and 5hmC

To assess the relative contributions of 5mC and 5hmC to ASM, we used the TrueMethyl 6 kit (CEGX), according to the instructions of the manufacturer. This chemical conversion-based approach uses bis-seq of multiple clones to separately score 5mC-only, so that the percent contribution of 5hmC to net methylation at each CpG can be inferred from the difference of net methylation observed using oxy-bis-seq and traditional bis-seq. For this purpose, we required at least 12 clones per allele for both oxy-bis-seq and traditional bis-seq.

Annotation of Hap-ASM and mQTL Regions and Bioinformatic Enrichment Analysis

To annotate hap-ASM loci and mQTLs, we defined small (200 bp, 500 bp) and large (150 kb) windows centered on each informative region for the hap-ASM DMRs and centered on each informative CpG for the mQTLs. The small windows were used to assess mechanistic hypotheses involving local sequence elements and chromatin states, and the large windows were used as an approximation of haplotype blocks for assessing proximity of hap-ASM DMRs and mQTLs to GWAS peaks. From the UCSC Genome

Browser (GRCh37 assembly), we downloaded RefSeq annotations, CpG islands, conserved elements, repetitive elements, and CTCF and TF peaks and motif occurrences in all cell lines from ENCODE and related projects (GEO: GSE29611, GSE27584, GSE30263, GSE31477).[23,24] Chromatin states in lung, heart, liver, placenta, T cell, brain, astrocyte, and H9-neurons were downloaded from the Roadmap Epigenomics project (GEO: GSE18927). We aggregated the 70 cell lines queried for CTCF binding peaks into 17 cell types (Table S3) and classified CTCF peaks by strength according to the 25th, 50th, and 75th percentile of all peak values. Cell restricted, multi-cell, and pan-cell peaks were defined with the 25th, 50th, and 75th percentile of the number of cell types. For analyzing CTCF binding site motifs, we scored occurrences of the canonical sequence identified by ENCODE, as well as de novo predicted motifs identified by Kheradpour and Kellis.[23] GWAS traits and associated SNPs were downloaded from NHGRI. We used BedTools to intersect the genomic coordinates of our informative and hap-ASM/mQTL regions to the coordinates of the annotation sets, using 150 kb windows for GWAS traits/SNPs and gene name annotations and 200 bp and 500 bp windows for annotations of local sequences and chromatin features.

To test whether hap-ASM occurs at specific types of sequences more often than random expectation, we used univariate logistic regression with hap-ASM or mQTL as the dependent variable and the tested sequence feature as the explanatory covariate. Because hap-ASM regions have different sizes, to avoid bias due to a higher probability of longer regions to overlap with regulatory elements, enrichment analysis was carried out with fixed windows. To test for robustness, we ran each analysis with two different window sizes, 200 bp and 500 bp, centered on each DMR. Because the denominator for the enrichment analyses depends on the platform, we performed all enrichment analyses separately for hap-ASM DMRs and mQTLs. To test the effect of ASM CpG density, hap-ASM regions were categorized as three levels, with 3, 4–5, and >6 ASM CpGs. The tested regulatory element was defined as the dependent variable and hap-ASM, categorized according to the number of hap-ASM CpGs, was the explanatory covariate. Multivariate logistic regression was used to adjust for the number of CpGs, so that CpG-rich hap-ASM regions would not be compared to CpG poor regions. For enrichment analysis of polymorphic CTCF binding sites and TF binding sites (TFBS) among hap-ASM regions, polymorphic sites were defined as motif occurrences in a 200 bp window containing at least one informative SNP. For the 500 bp window and mQTLs, polymorphic binding sites was defined as motif occurrences containing at least one SNP with minor allele frequency ≥ 0.2, because based on our sample size, rarer SNPs were unlikely to be sufficiently informative. The effect of polymorphisms on CTCF and TF binding likelihood was estimated for each allele by their position weight matrix (PWM) score:

$$\sum_{i,j} p_{i,j} \log\left(p_{i,j} \big/ p_b\right).$$

Here $p_{i,j}$ *is* the probability of the nucleotide for each position in the PWM from the ENCODE data, and p_b the nucleotide background frequency assuming equal probabilities of each nucleotide ($p_b = 0.25$). For motif occurrences with a PWM score > 3, correlations between allelic difference of methylation and difference of PWM score were assessed via linear regression.

To assess DMR boundaries, we used our T cell mQTL dataset. Because estimation of the boundaries is limited by 450K CpG coverage, we looked for mQTL CpGs in CpG-rich regions, with at least one CpG in the proximate 500 bp, one CpG between 500 bp and 1,000 bp, one CpG between 1,000 bp and 2,000 bp, and one CpG after 2,000 bp, upstream and downstream of the index CpG. The boundaries of mQTLs were defined as at least two consecutive CpGs whose methylation lacked significant correlation with the index SNP. Fine mapping of hap-ASM DMRs directly via the Agilent Methyl-Seq data was performed on seven hap-ASM regions for which the 2 kb upstream and downstream flanking regions contained at least one heterozygous SNP in samples with hap-ASM. For eQTL enrichment analysis, genes in 150 kb windows spanning ASM DMRs and mQTLs were annotated with the eQTL browser. The distance to eQTLs was defined as the distance to the transcriptional start site of the genes showing eQTLs. Analyses were performed with R and STATA statistical software.

Results

Methyl-Seq in Multiple Primary Human Tissues Produces Maps of Hap-ASM

The terms mQTL and hap-ASM are related, but not synonymous. Although they both describe the same class of allelic asymmetry, in which the DNA methylation on each allele is sequence dependent, they are mapped by different strategies (Material and Methods and Figure S1). To test pan-tissue mechanisms and identify ASM DMRs near statistical peaks from diverse GWA studies, our Methyl-Seq sample set included diverse human tissues, including brain, T cells, placenta, liver, heart, and lung from different individuals (Table S1). In contrast, our array-based approach for mapping mQTLs utilized larger numbers of samples, concentrating on T cells, brain, and, in a smaller set, placentas. In total, the two approaches provided information on 3.7 million CpGs in the Methyl-Seq data to directly identify ASM, and 485,000 CpGs in the array-based methylation data, which we used with Illumina 2.5M SNP array data to identify mQTLs.

By Methyl-Seq, we obtained a median fragment length of 200 bp and mean depth of coverage in the targeted regions ranging from 50 to 94 across the samples, with 80% of these regions covered by more than 20 reads (Figure S2). We were able to query 278,897 CpGs located in the same fragments as 44,851 index SNPs. Among them, 42,904 CpGs (15%) showed significant asymmetry of fractional methylation between the two alleles by Fisher's exact test at $p < 0.05$. Asymmetric methylation resulting from the loss of CpG sites due to the presence of SNPs accounted for 30% of these CpGs (12,856 CpGs). Such sites can be interesting for gene regulation, but for our subsequent analyses we did not consider these genetically polymorphic CpGs as having ASM. Thus, from this experiment we brought forward 30,048 CpGs with bona fide ASM for bioinformatic enrichment analyses.

Because analyzing epigenomic data both at the level of the individual CpG and the differentially methylated region (DMR) improves specificity and is crucial for testing mechanisms,[25] we used both approaches. To identify strong ASM DMRs, ASM regions were defined by at least three CpGs with significant allelic asymmetry in fractional methylation (Fisher's exact test $p < 0.05$). We further required at least two contiguous CpGs with ASM, and an absolute difference in fractional methylation of $\geq 20\%$ between alleles after averaging over all covered CpGs in the DMR. The significance threshold for the difference in fractional methylation between alleles was set at a Wilcoxon test p value with multiple testing correction < 0.05 (FDR at 5%). With these cut-offs, we found 795 strong hap-ASM regions, representing 2% of all informative SNP-containing regions (Figure S2 and Table S4). Some of the hap-ASM DMRs overlapped with promoter sequences near transcriptional start sites, but most were intergenic or intronic, with a modest enrichment in intergenic regions (OR = 1.3, p = 0.001). CpG islands were represented among the hap-ASM DMRs but were not found to be enriched (Figure S3). Interestingly from an evolutionary perspective, conserved sequence elements were significantly under-represented among hap-ASM DMRs compared to polymorphic regions without hap-ASM (OR = 0.74, $p = 9 \times 10^{-5}$).

In the above analyses we deliberately excluded all CpG-containing reads that mapped to known imprinted chromosomal domains. However, we identified other clusters of reads with ASM mapping within 150 kb windows centered on 47 known imprinted genes (Table S5). This finding of multiple imprinted loci, representing a 5-fold increase over random expectation ($p = 6 \times 10^{-29}$), serves as a positive internal control for our overall experiment. As supported by our targeted bis-seq in series of heterozygous individuals (below), the large majority of ASM loci revealed by the Methyl-Seq data in fact represent hap-ASM, not imprinting.

Identification and Mapping of Hap-ASM in *ZFP57*, a *trans*-Regulator of Genomic Imprinting

Hap-ASM in a gene with interesting biological functions can provide a concrete example of our Methyl-Seq and mQTL findings. We used targeted bis-seq to fine-map a hap-ASM DMR and mQTL that our data had revealed 3.5 kb upstream the transcription start site of *ZFP57* (MIM: 612192) on chromosome 6 (Figure 1 and Table 1). This gene, in the KRAB domain-containing ZNF superfamily, is not imprinted (confirmed by binomial test in our series of heterozygotes), but it codes for a TF that is an important *trans*-acting regulator of DNA methylation imprints.[26,27] The bis-seq data from a contig of 12 amplicons, each covering a high frequency (i.e., informative) SNP and at least three CpGs, showed that hap-ASM in this region spans a discrete DMR of 2 kb, with allelic asymmetry in 6/6 T cell samples tested (Figure 1). This DMR is located between a GWAS SNP associated at sub-threshold significance with migraine (MGR1 [MIM: 157300]; $p = 9 \times 10^{-6}$) and another with multiple sclerosis (MS [MIM:126200]; at supra-threshold significance, $p = 10^{-17}$). Additional validations of hap-ASM in T cells and brain samples are in Figure 2, and our mapping of the hap-ASM region in brain revealed the same DMR as in T cells (Figure S4). Of mechanistic importance (see below), these data show that the *ZFP57* hap-ASM DMR precisely overlaps an ENCODE CTCF binding peak (Figures 1 and S4). To further assess tissue specificity, we carried out targeted Nextgen bis-seq in a wider range of tissues from multiple individuals and found that hap-ASM is strong in T cells and brain, weaker in colon, heart, liver, lung, and monocytes, and absent in placenta, where the DMR is biallelically hypomethylated (Figure S4).

Pan-tissue and Tissue-Specific mQTLs in T Lymphocytes, Brain, and Placenta

We next assessed *cis*-regulated mQTLs in T cells, brain temporal cortex (TC) gray matter, FANS-isolated NeuN-positive TC neurons, FANS-isolated non-neuronal TC cells (NeuN-negative; hereafter referred to as glia, but understood to include other cell types such as microglia), and placenta. Overall, 138 samples were included in this discovery phase (Table S1). Because previous reports have shown that CpG methylation at mQTLs correlates most strongly with nearby SNPs,[8] we restricted our analyses to CpGs in 150 kb windows centered on each index SNP. This window size is in fact sufficient to capture most *cis*-regulated mQTLs, as indicated by the decay curve of the correlation coefficients in our data (Figure S5). Overall, 451,419 CpGs were informative in T cells, 450,805 in brain TC, 444,047 in neurons, 447,867 in glia, and 450,631 in placenta. Using linear regression corrected for multiple testing, we assessed correlations between genotypes and methylation for each index SNP-CpG pair, defining an mQTL by p value and effect size, namely a significant linear correlation (p value corrected for multiple testing < 0.05), R-squared ≥ 0.5, reflecting the goodness of fit to the linear model, and a regression β coefficient associated with genotype ≥ 0.1, corresponding to a difference in net methylation between the two alleles ≥ 0.2. The importance of combining statistical significance and effect size criteria is highlighted by the examples in Figure S1. Using these criteria, we identified 1,440 mQTL CpGs in T cells, 737 in unfractionated brain TC, 364 in neurons, 867 in glia, and 866 in placenta (Figure S2 and Tables S6–S10). Like the hap-ASM loci, the mQTLs were enriched in intergenic CpGs (OR = 1.9, $p = 2.1 \times 10^{-63}$), with promoter CpGs being present but relatively under-represented (OR = 0.68, $p = 5.5 \times 10^{-21}$; Figure S3).

We found marked tissue specificity among the sets of mQTLs (Figures 1, 2, and S6). As expected, most of the mQTLs in unfractionated brain TC were also found in purified neurons or glia (61%), but only 28% of mQTLs in brain tissue and purified brain cells overlapped with mQTLs in T cells, and 12% with mQTLs in the placentas. To further assess cell type specificity, we superimposed informative CpGs in neurons and glia and identified 188 neuron-restricted (or neuron-stronger) and 602 glia-restricted (glia-stronger) mQTL CpGs (Figure 2). Because most of our neuron and glia preparations were paired samples from the same brains, we were able to perform a sub-analysis including only the paired samples, which confirmed the

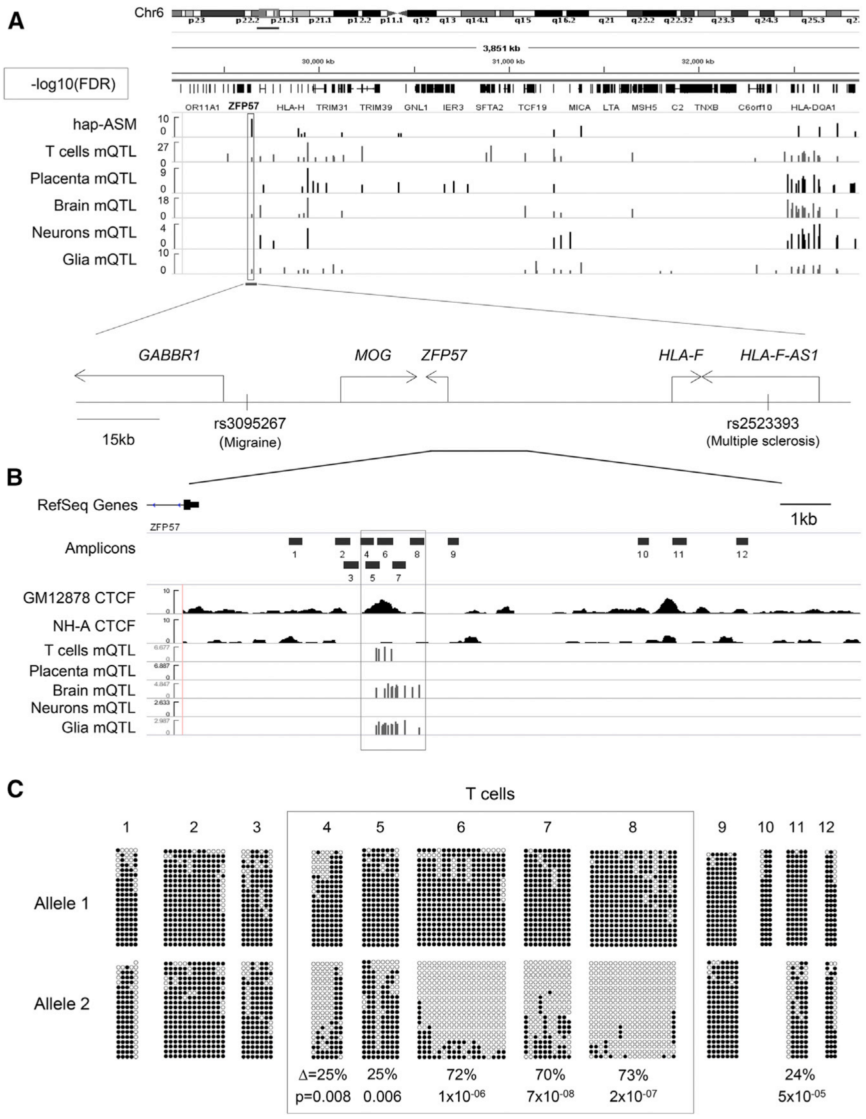

Figure 1. Hap-ASM DMRs and mQTLs in the HLA Region on Chromosome 6, and Targeted Bis-Seq Defining a Hap-ASM DMR Upstream of *ZFP57*

(A) Map of hap-ASM and mQTLs on chromosome band 6p22.1. The tracks show $-\log_{10}$ q values for hap-ASM DMRs in the overall Methyl-Seq dataset and for mQTLs in each of the indicated tissues. Although there are some pan-tissue mQTLs, most are tissue specific. As explained in the text, the net yield of loci from the hap-ASM and mQTL approaches is additive.

(B) Zoomed-in view of *ZFP57*, showing relevant ENCODE tracks and the amplicons utilized for targeted bis-seq.

(C) Results of targeted bis-seq showing that the *ZFP57* hap-ASM DMR spans approximately 2 kb in T cells. The major DMR (amplicons in the gray rectangle) precisely overlaps a strong CTCF ChIP-seq peak and is located between two GWAS SNPs, one associated with migraine ($p = 9 \times 10^{-6}$) and the other with multiple sclerosis ($p = 10^{-17}$). A weaker DMR is found overlapping a second CTCF peak (amplicon 11). Circles represent consecutive CpGs, with each line being a unique read. White circles are unmethylated CpGs and black circles are methylated CpGs. Alleles 1 and 2 represent the methylated and unmethylated alleles in these heterozygous samples. Wilcoxon p values and methylation differences were calculated by bootstrapping (1,000 sampling of 20 reads per allele) and are indicated only for significant hap-ASM (Δ fractional methylation > 0.2, >3 ASM CpGs, and $p < 0.05$). One representative random sample of each allele (20 reads per allele) is shown. Validations and fine-mapping of this DMR in other tissues and cell types are in Figures 2 and S4. Δ, absolute difference in percentage of methylation between alleles in heterozygous samples; q value, p value corrected for multiple testing using the Benjamini-Hochberg method; values are from bootstrapping.

cell type specificity (Figure S6). This situation is illustrated by our targeted bis-seq data for a GWAS peak-associated hap-ASM region upstream of *NGFR* (MIM: 162010), coding for the NGF receptor that binds neurotrophins.[28] The results confirmed ASM that is strong in neurons, moderately strong in T cells, and weak or absent in glial cells and placenta (Figure S7). As we show below, allele-specific TF binding site occupancy is one mechanism of hap-ASM and mQTLs, and it can partly explain this tissue specificity. However, differences in global methylation levels between tissues probably also play a role. Such differences are revealed in our data by comparing the mean CpG methylation levels genome-wide in T cells, brain, and placenta, expressed as Kernel density plots in Figure S2, which show that the most divergent tissue for mQTLs and hap-ASM, the placenta, has global CpG hypomethylation compared to the other two tissues.

As an important technical and biological point, among the 44 adult TC gray matter samples in this experiment, 22 were without significant neuropathological changes and 22 showed neuropathological findings of moderate to severe late-onset AD. To test for possible AD-specific effects on DNA methylation in the brain mQTLs, we carried out multivariate linear regression including the genotype, the disease status, and the interaction term between disease status and genotype. This procedure revealed no significant AD-specific effects on DNA methylation at the mQTL CpGs, and no differential genotype effects on methylation levels in AD compared to controls (Tables S7–S9). This finding of little influence of AD neuropathology on DNA methylation patterns is consistent with the low yield of differentially methylated loci in prior case-control studies,[29–32] particularly those passing a p value corrected for multiple testing $< .05$ and fractional

Table 1. Examples of Disease-Relevant Loci with Hap-ASM or mQTLs

Primary Dataset and Ranking for Strength of Hap-ASM or mQTL	SNP in Phase with Hap-ASM or mQTL[a]	Closest Genes in 150 kb Window	Targeted Bis-Seq Validations[b]	GWAS Signals in 150 kb Window Centered on Hap-ASM DMR or mQTL		
Immune System						
hap-ASM (15); mQTL (33)	chr1: rs9330298; cg08477332	*S100A**, *SNAPIN*, *ILF2*	brain: 7/8; T cell: 7/7	rs7536700; myeloma; $p = 4.00 \times 10^{-6}$	–	–
hap-ASM (107); mQTL (302)	chr5: rs2549004; cg21138405	*IRF1*, *IL5*, *RAD50*	brain: 1/7; T cell: 9/12	rs11745587; asthma; $p = 2.00 \times 10^{-6}$	rs12521868; CD; $p = 1.00 \times 10^{-20}$	rs2188962; IBD; $p = 1.00 \times 10^{-52}$

hap-ASM (179)	chr6: rs1565443; haplotype:
	rs1565441;
	rs1565442;
	rs1565443;
	rs6937877;
	rs9364402
		FRMD1	brain: 0/10;
			T cell: 7/15;
			T cell: 7/7
				rs1473500 ; Immune resp.; $p = 3.00 \times 10^{-7}$	–	–

hap-ASM (183)	chr12: rs12304510	*CLECL1*, *CD69*, *CLEC2D*	brain: 0/7; T cell: 5/9	rs10466829; MS; $p = 1.00 \times 10^{-8}$	rs4763879; T1D; $p = 2.00 \times 10^{-11}$	rs11052552; T1D; $p = 7.00 \times 10^{-7}$
hap-ASM (45)	chr19: rs10411630	*CCDC155*, *DKKL1*, *TEAD2*	brain: 5/8; T cell: 2/3	rs2303759; MS; $p = 5.00 \times 10^{-9}$	–	–
hap-ASM (137)	chr19: rs12975442	*EVI5L*, *MAP2K7*, *TGFBR3L*	GM12878; LBCL	rs558718; HIV-1 control; $p = 4.00 \times 10^{-6}$	–	–
Immune System and Brain						
hap-ASM (61); mQTL[c] (23)	chr6: rs78274956; cg05844871	*HLA-DRB6*	brain: 9/9; T cell: 1/6	rs4530903; SCZD; $p = 5.00 \times 10^{-6}$	rs9271192; AD; $p = 3.00 \times 10^{-12}$	rs3828840; MS; $p = 5.00 \times 10^{-15}$
hap-ASM (16); mQTL (279)	chr6: rs2747429; cg16885113	*ZFP57*, *MOG*, *HLA-F*	brain: 4/4; T cell: 5/5	rs3095267; migraine; $p = 9.00 \times 10^{-6}$	rs2523393; MS; $p = 1.00 \times 10^{-17}$	rs9258260; CD; $p = 2.00 \times 10^{-10}$

mQTL (58)	chr17: rs2412102; cg10163794	*PHB*, *NGFR*	neur: 4/4;
			glia: 0/2;
			T cell: 4/4
				rs16948200 ; immune resp.; $p = 2.00 \times 10^{-8}$	rs1035050; bipolar; $p = 9.00 \times 10^{-6}$	–

Brain						
hap-ASM (440)	chr2: rs11684605	*PXDN*, *MYT1L*	brain: 4/8; T cell: 0/8	rs6735179; antipsychotic Rx; $p = 1.00 \times 10^{-7}$	–	–
hap-ASM (65)	chr3: rs9838223	*CMTM8*, *GPD1L*	brain: 7/11; T cell: 0/5	rs9825310; alcohol dep.; $p = 8.00 \times 10^{-6}$	rs4380451; bipolar; $p = 4.00 \times 10^{-6}$	–
hap-ASM (527)	chr7: rs35487364	*PTPRN2*	brain: 2/12; T cell: 0/6	rs6459804; bipolar, SCZD; $p = 8.00 \times 10^{-6}$	–	–
mQTL (116)	chr22: rs7893689; cg04065885	*CELF2*	glia: 6/6	rs201119; AD; $p = 1.00 \times 10^{-8}$[d]	rs2242451; AD; $p = 0.003$[d]	–

To test the validity of our criteria for calling hap-ASM and mQTLs, we deliberately chose loci in a wide range of strength of allelic asymmetry (rankings) and with at least one nearby GWAS signal associated with a disease-relevant tissue or cell type. The results of bis-seq (numbers of heterozygotes with allelic asymmetry in methylation of total numbers of heterozygotes) highlight tissue specificity of the hap-ASM. In addition, two of these regions showed inter-individual differences in the presence or absence of hap-ASM. For one of them, the *FRMD1* region, we showed that extended haplotypes rather than the index SNP alone dictate methylation asymmetry (Figure S11). Many of these loci are in genomic regions with sub-threshold GWAS signals, and in these situations, the epigenetic data provide independent evidence for biological significance. GWA data were obtained from the NHGRI-EBI GWAS Catalog. References for GWAS are listed in the Supplemental Data, as well as additional references for candidate genes. Abbreviations and OMIM numbers are as follows: AD, Alzheimer's disease; Alcohol dep., alcohol dependence; antipsychotic Rx, response to antipsychotic therapy (MIM: NA); bipolar, bipolar affective; CD, Crohn disease; HIV-1 (HIV-1, susceptibility to [MIM: 609423]); Immune response, immune response to smallpox (MIM: NA); IBD, inflammatory bowel disease (IBD1 [MIM: 266600)]); MS, multiple sclerosis; Myeloma, multiple myeloma; SLE, systemic lupus erythematosus (SLE [MIM: 152700]); SCZD, schizophrenia; T1D, type I diabetes mellitus; *ALG12* (MIM: 607144), *ARC* (MIM: 604223), *ARL11* (MIM: 609351), *BRD1* (MIM: 604589), *DKKL1* (MIM: 605418), *EBPL* (MIM: NA), *EVI5L* (MIM: NA), *GPD1L* (MIM: 611778), *HLA-F* (MIM: 143110), *IL5* (MIM: 147850), *JRK* (MIM: 603210), *MAP2K7* (MIM: 603014), *MOG* (MIM: 159465), *PHB* (MIM: 176705), *PSCA* (MIM: 602470), *RAD50* (MIM: 604040), *RCBTB1* (MIM: 607867), *SNAPIN* (MIM: 607007), *TEAD2* (MIM: 601729), *TGFBR3L* (MIM: NA), *ZBED4* (MIM: 612552), NA, not available.

[a] Highest ranking index SNPs and CpGs are listed here; see Tables S4 and S6–S10 for complete lists and SNPs in phase with mQTLs.

[b] Heterozygotes assessed by Agilent Methyl-Seq and/or targeted bis-seq.

[c] The 150 kb window centered on this validated DMR contains 29 additional mQTLs in brain including cg20532376, cg00119778, cg13972202, cg10466124, cg18111114, cg00103771, cg07984380, cg13910785, cg19575208, cg08845336, and cg00598125 (Table S7).

[d] AD-associated SNPs were reported in Lee et al.[67] and Wijsman et al.[68]

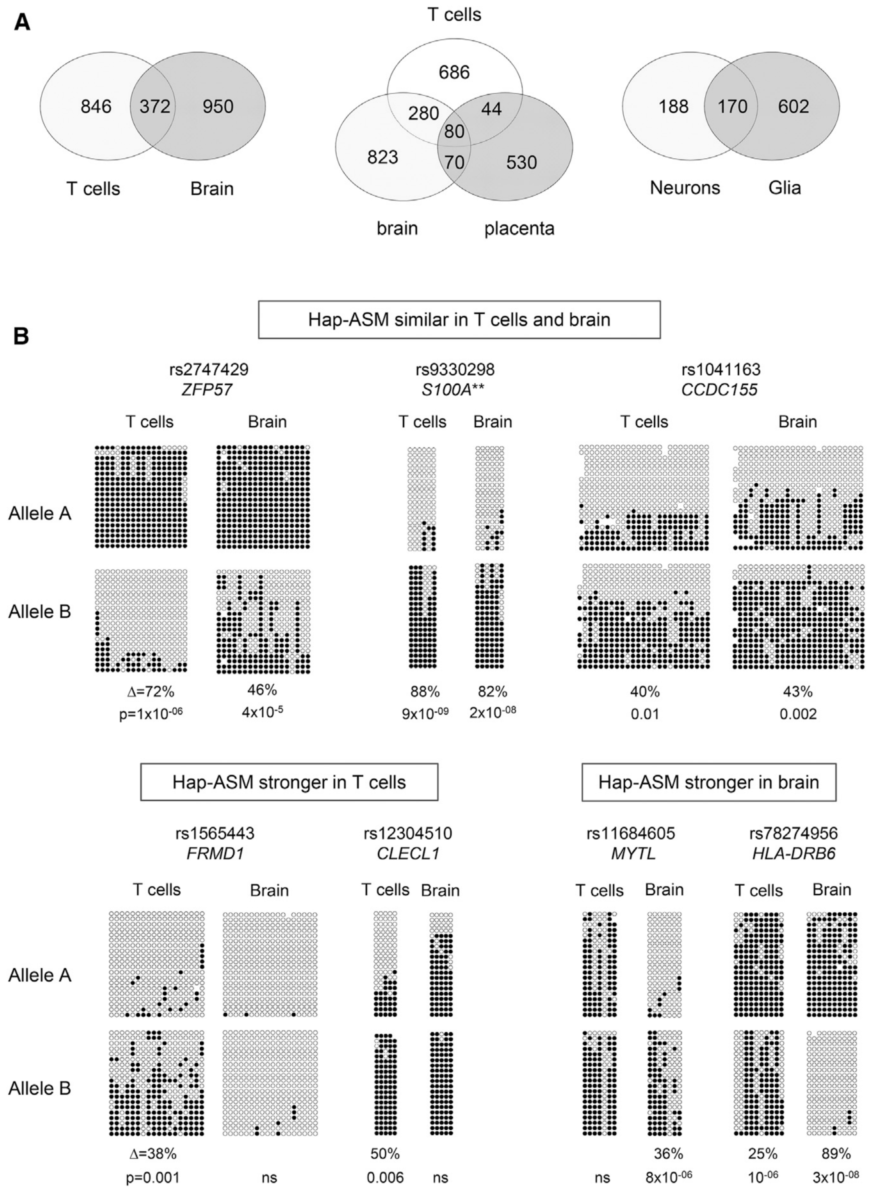

Figure 2. Genome-wide Data and Targeted Bis-seq Highlight Pan-tissue and Tissue-Restricted Hap-ASM

(A) Venn diagrams of mQTLs identified in placenta, brain, T cells, neuron, and glia, showing that although pan-tissue mQTLs can be found, the majority of mQTLs are tissue or cell type specific. Only mQTLs with CpGs-SNP pairs that were informative in every tissue are considered here. Thus, the numbers in these diagrams are somewhat smaller than the total mQTLs listed in the Tables S6–S10.

(B) Validation of hap-ASM regions identified in the *S100A*** cluster, *ZFP57*, *CCDC155*, *FRMD1*, *CLECL1*, *PXDN*, and *HLA-DRB***. Whereas hap-ASM in the *S100A*** cluster, *ZFP57*, and *CCDC155* is robust in both T cells and brain, the hap-ASM DMRs in *FRMD1* and *CLECL1* are strong in T cells but weak or absent in brain. Conversely, hap-ASM in *PXDN* and *HLA-DRB6* is significant in brain but not or only weakly in T cells. Alleles A and B represent, respectively, the reference and alternative allele of one heterozygous sample. The p values were calculated as in Figure 1. Table 1 lists the number of heterozygous samples subjected to bis-seq. Figures 3, 4, and S10–S17 show bis-seq validations in additional samples and loci.

methylation difference ≥ 0.1, which corresponds to the sensitivity of Illumina Beadchip arrays.[33,34] The AD-associated CpGs from the largest of those studies[30–32] showed no significant differences in methylation between the AD case and control subjects in our dataset (differences in mean fractional methylation ranging from 0.00003 to 0.06, with 95% less than 0.03, consistent with Lunnon et al.,[31] and none passing FDR at 0.5; Table S12). Nonetheless, as we discuss below, our mQTL and hap-ASM data from brain cells are in fact useful as an adjunct to GWAS data for mapping genes and regulatory sequences that underlie inter-individual differences in AD susceptibility.

ASM and mQTL Screens Are Complementary Approaches

When we compared the informative regions between our mQTL and Methyl-Seq datasets, we found little overlap, with only 4.5% of the informative 450K CpGs queried by Methyl-Seq and, conversely, 40% of the informative Methyl-Seq regions covering at least one informative 450K CpG. Only 2% of the array-informative CpGs and 16% of Methyl-Seq-informative regions were assessed by the same informative SNPs in both approaches Thus, as shown for one chromosomal region in Figure 1A and summarized globally in Figure S2, combining the two methods gives an additive yield, allowing a more comprehensive profiling of hap-ASM, and its surrogate mQTLs, through the genome.

Hap-ASM and mQTLs Highlight Regulatory Regions near Supra- and Sub-threshold GWAS Peaks Associated with Immunological and Neurological Disorders

Overlapping our data with the GWAS catalog, we found 397 strong hap-ASM regions from our multi-tissue experiment and 778 T cell, 395 brain (unfractionated TC), 204 neuron, 483 glia, and 474 placenta mQTLs with at least one GWAS SNP in a 150 kb window centered on the mQTL or hap-ASM DMR. These loci are listed and ranked by strength of allelic methylation asymmetry in Tables S4 and S6–S10. To focus on the loci of greatest interest, we next narrowed our search to GWA traits relevant to the tissue or cell type in which we had identified the hap-ASM or mQTL. This procedure revealed 61 hap-ASM DMRs and 257 T cell, 148 brain, 81 neuron, and 179 glia mQTLs located within

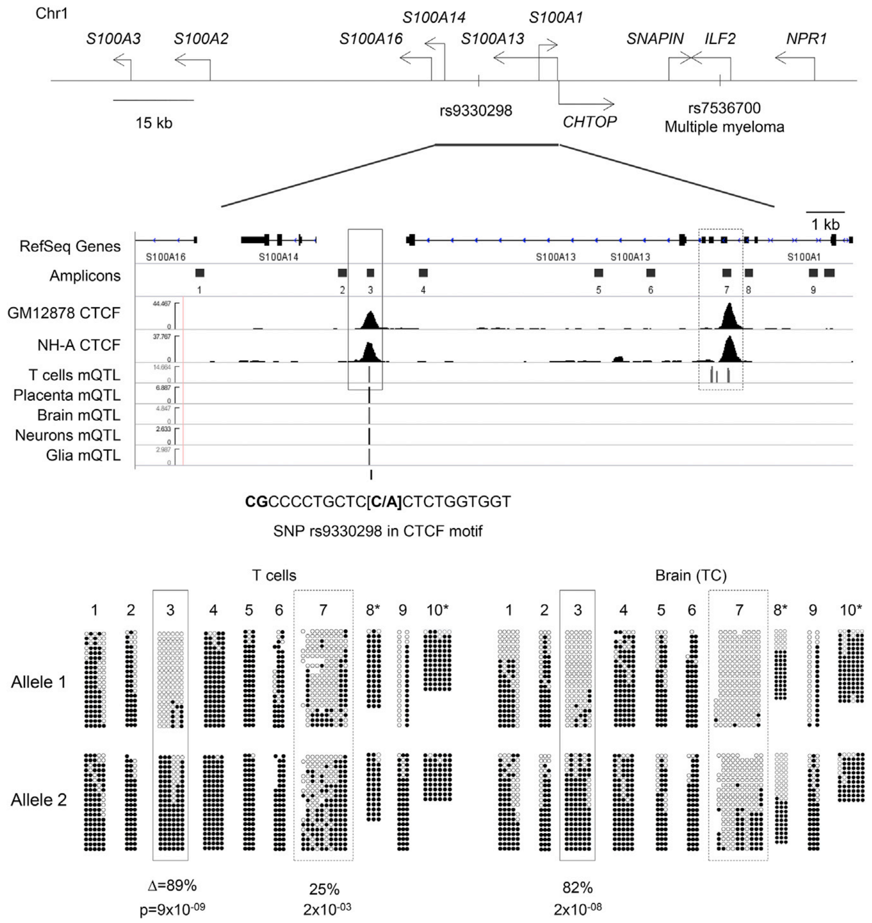

Figure 3. Fine-Mapping of Hap-ASM in the S100A*-ILF2 Region, Containing a Sub-threshold GWAS Signal for Multiple Myeloma, Reveals DMRs Overlapping with CTCF Binding Sites

The map and genome browser tracks show a strong hap-ASM DMR identified by both Methyl-Seq and mQTL analysis. The DMR is validated by targeted bis-seq in an intergenic region of the *S100A** gene cluster, located 70 kb from a subthreshold GWAS peak ($p = 4 \times 10^{-6}$) for multiple myeloma. The DMR (solid rectangle) overlaps exactly with a CTCF ChIP-seq peak, with the index SNP located in the CTCF binding motif. A second weaker hap-ASM region identified by the fine-mapping in T cells and by mQTL analysis, overlaps another CTCF peak (dashed rectangle). Amplicon 9 contains a SNP CpG eliminating the CpG on the hypomethylated allele. Nextgen bis-seq (MiSEQ) was performed for all amplicons, except amplicons 8 and 10 (*, Sanger sequencing). For the MiSEQ data, the p values were calculated as in Figure 1. For Sanger sequencing, we required at least ten clones per allele.

75 kb of tissue-relevant GWAS peaks (Tables S4 and S6–S10). Hap-ASM loci were often located near GWAS peaks, but were not specifically enriched for such locations. However, the combined larger sets of brain and T cell mQTLs were moderately but significantly enriched within 150 kb windows centered on GWAS peaks ($OR = 1.3$, $p = 1.8 \times 10^{-9}$).

Targeted Bis-Seq of Hap-ASM and mQTL Regions Validates the Genome-wide Data, Localizes DMR Boundaries, and Reveals Inter-individual Variability

We next selected additional genes for validations and fine-mapping of their DMRs using Nextgen and Sanger bis-seq. As shown in Table 1, these loci included six hap-ASM DMRs near immune-related GWAS peaks, four hap-ASM DMRs near GWAS peaks for neurological phenotypes, and three hap-ASM DMRs in chromosomal regions with multiple GWAS peaks for both immune- and brain-related phenotypes. The results confirmed hap-ASM in 12/13 of these loci (Figures 1, 2, 3, 4, S4, S7, and S10–S17), including 7 chromosomal regions with supra-threshold GWAS signals of $p < 5 \times 10^{-8}$ and 5 chromosomal regions with sub-threshold GWAS signals ($p = 9 \times 10^{-6}$ to $p = 1 \times 10^{-7}$). Among these examples are candidate susceptibility loci for AD, bipolar disorder (BPAD [MIM: 125480]), multiple sclerosis, type I diabetes mellitus (IDDM [MIM: 222100]), Crohn disease (IBD1 [MIM: 266600]), immune response to a vaccine, and other relevant medical phenotypes. Although several of these hap-ASM DMRs were detected in multiple tissues, in our validations comparing brain and T cell samples, we found that hap-ASM in the *CLECL1* (MIM: 607467) and *FRMD1* (MIM: NA) regions is mostly or entirely T cell specific, whereas hap-ASM in the *MYT1L* (MIM: 613084) and *CMTM8* (MIM: 607891) regions is mostly or entirely brain specific (Table 1). In addition, validation of hap-ASM in the *NGFR* region in paired neuron and glia from the same individuals showed strong hap-ASM in neurons and weaker or absent hap-ASM in glia (Figure S7). Thus, hap-ASM can provide localizing data and support for the biological relevance of GWAS signals in a tissue- and cell-type-specific manner.

Interestingly, for some of the DMRs, hap-ASM was observed only in a subset of the heterozygous samples from a given tissue or cell type (Table 1). Such variation might be explained either by environmental effects on the epigenetic patterns or by a genetic mechanism in which the combined effect of several SNPs in a haplotype block, not just the index SNP, could dictate the presence or absence of hap-ASM. To test the latter possibility, we focused on the *FRMD1* region, which showed hap-ASM in about half of the heterozygous T cell samples. We genotyped 10 heterozygous individuals using a 600 bp amplicon centered on the index SNP, with cloning and Sanger sequencing of the clones. Based on the resulting phased haplotypes, we found that 5/5 of the samples with hap-ASM shared the same extended haplotype including

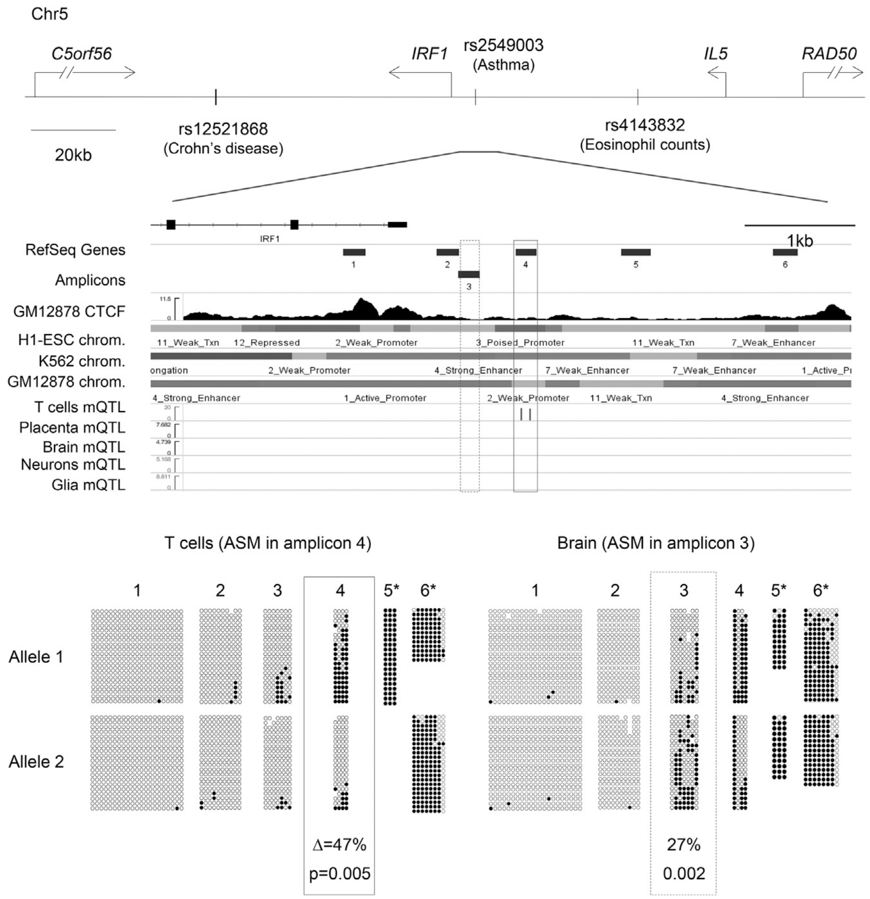

Figure 4. Hap-ASM in the Upstream Promoter Region of the Immune Phenotype-Associated *IRF1* Gene

The map and browser tracks show the two hap-ASM DMRs identified from Agilent Methyl-Seq (black rectangles) in the region immediately upstream of *IRF1*. The 150 kb window centered on the hap-ASM DMRs contains immune-related GWAS peaks tagged by SNPs rs12521868 and rs4143832 (GWAS p values 10^{-20} and 10^{-10}, respectively). Hap-ASM is consistently observed in T cells and is much weaker in brain, which confirms the presence of mQTL only in T cells. Unlike the hap-ASM in the *S100A** cluster, here the "shoulders" around the hap-ASM are biphasic, *low* methylated and *high* methylated, regions. The DMR is coincident with a dynamic chromatin region, poised in embryonic stem cells and active in differentiated cells. Representative samples of each queried tissue are shown. p values were calculated as in Figure 1. Abbreviations are as follows: *, Sanger sequencing; GM12878, lymphoblastoid cell line, H1-ESC, human embryonic stem cell, K562, leukemia cell line, NS, not significant; chrom., chromatin states from the Roadmap Epigenomics project.

rs1565441, rs1565442, rs1565443, rs6937877, and rs9364402 in the heterozygous configuration AAAA/BBBB (Figure S11). Conversely, none of the samples that were negative for hap-ASM had this haplotype configuration, instead having other configurations such as AAAA/AABBB (Figure S11). To confirm this finding, we carried out additional bis-seq and genotyping in two pairs of monozygotic twins heterozygous for the index SNP, in which one pair showed hap-ASM and the other did not. The pair with hap-ASM carried the AAAAA/BBBBB haplotype configuration, whereas the pair that was negative for hap-ASM did not (Figure S11). These results show that extended haplotypes, rather than single proximal SNPs, dictate the allelic asymmetry at some hap-ASM DMRs.

To exclude imprinting as an explanation for ASM at any of these loci, we tested the probability of ASM under the imprinting hypothesis, namely that methylation of an allele is determined by the parent of origin, not the local haplotype. Using genotypes of the index SNPs, the imprinting hypothesis, which predicts a random, genotype-independent pattern of ASM, was rejected (binomial test p value < 0.05) for 10 of the 12 loci with validated ASM, and not conclusive for the remaining two DMRs because of the small number of heterozygotes. This outcome gives strong confidence that the large majority of ASM DMRs identified by our genome-wide screen indeed represent hap-ASM, not genomic imprinting.

Lastly, because the "sixth base" 5-hydroxy-methylcytosine (5hmC) is abundant in brain and especially neurons, we carried out oxy-bis-seq for three brain-associated hap-ASM DMRs. This procedure showed that both 5mC and 5hmC contribute to hap-ASM, making contributions to the allelic asymmetry in the same direction (Figure S18).

DMR Mapping and Bioinformatic Enrichment Analyses Support Variation in CTCF and TF Binding Sites as a Mechanism Underlying Hap-ASM and mQTLs

Fine-Mapping of Hap-ASM Shows Discrete DMRs Overlapping with Regulatory Sequences

Determining the boundaries of DMRs is crucial for testing mechanistic hypothesis, but most prior studies have not defined such boundaries. The examples of hap-ASM that we mapped previously,[14] and the *ZFP57* DMR described above, are all small discrete regions, 1 to 2 kb in size. To see whether these findings would generalize, we asked whether our 450K data had sufficient CpG coverage to define DMR boundaries. Inferring DMR sizes from these data is limited by the CpG coverage of the Beadchips, so to overcome this limitation, we analyzed our largest 450K dataset, from the T cells, and focused on the 270 mQTL CpGs in the most CpG-rich, and thus most informative, regions (134 mQTLs; see Material and Methods). Putative boundaries of mQTLs were defined as the presence of at least two consecutive CpGs lacking significant correlations with the index SNPs. By this procedure, the median DMR length was 720 bp, with 90% of DMRs spanning less than 2,000 bp (Figure S8). We observed higher methylation levels in the flanking CpGs for 38% of the DMRs, with the mQTL region corresponding to a low methylated well. For 15% of the DMRs, the flanking CpGs showed lower methylation levels (inverted well), and for 47% we observed a biphasic pattern, with one methylated and one unmethylated "shoulder" (Figure S8).

Taking a different approach, we used our Methyl-Seq data and searched for hap-ASM regions for which the 2,000 bp upstream and downstream flanking regions contained at least one heterozygous SNP in the samples with hap-ASM. We identified 7 such informative regions, and for 5 of them the DMR size was less than 2,000 bp (Figure S9). In addition, we used targeted bis-seq to map the boundaries of three additional hap-ASM regions from our genome-wide discovery set. We designed contigs of bis-seq amplicons upstream and downstream of DMRs in the *S100A** cluster (chr1), *PXDN* (chr2), and *IRF1* (chr5) regions (*S100A13* [MIM: 601989], *S100A14* [MIM: 607986], *PXDN* [MIM: 605158], *IRF1* [MIM: 147575]). We required each amplicon to cover a high-frequency SNP and at least three CpGs. From the resulting data, the DMRs in all three regions were discrete, spanning from 1 to 2 kb of DNA and having well-defined boundaries (Figures 3, 4, and S10). Like the *ZFP57* DMR, the DMR in the *S100A** region was identified in both mQTL and Methyl-Seq discovery sets and precisely coincides with a CTCF ChIP-seq peak. Moreover, in this DMR the index SNP is in a CTCF binding site sequence motif (Figure 3). As indicated by ENCODE data, the *PXDN-MYTL* DMR coincides with polymorphic MYC TF and CTCF binding motifs. Both motifs are associated with cell-type-restricted ChIP-seq peaks, which can explain the cell-type-specific hap-ASM at this locus (Figure S10). Although the *IRF1-IL5* (MIM: 147850) hap-ASM DMRs and mQTL did not overlap a CTCF site or known TF binding site, they were located in a region of poised chromatin (Figure 4). These examples highlight genetic variations in DNA regulatory elements as a potential mechanism for some, though perhaps not all, examples of hap-ASM, a hypothesis that we test globally below.

Hap-ASM DMRs and mQTLs Are Enriched in Specific Chromatin States

The Epigenomics Roadmap consortium identified 15 chromatin states that show different average levels of DNA methylation, different degrees of evolutionary conservation, and differences in several other features.[35] For example, states associated with bivalent enhancers show a broad distribution of methylation levels, standard heterochromatin and heterochromatin associated with ZNF genes and repeats (here we use the standard ENCODE term for this class of sequences) are depleted for evolutionarily conserved elements, and so on. Each of the 15 chromatin states are enriched in specific histone marks: repressed Polycomb state in H3K27me3; bivalent enhancer chromatin in H3K27me3 and H3K4me1; heterochromatin in H3K9me3, and heterochromatin associated with ZNF genes and repeats in H3K36me3 and H3K9me3.[35] As shown in Figure S19, among both hap-ASM DMRs and mQTLs combined across tissues, we found significant enrichments in repressed Polycomb state, bivalent enhancers, enhancers, heterochromatin, and heterochromatin associated with ZNF genes and repeats. Conversely, significant under-representation was found for regions classified as active TSS state. Our analysis of the mQTL data showed an enrichment of PRC2-bound regions and enhancers, as well as an under-representation in active promoters in every queried cell and tissue, whereas enrichments of the other chromatin states were more tissue dependent (Figure S20). Given that polycomb-bound and bivalent enhancers are associated with developmentally regulated genes[36] and that polycomb-repressed states, enhancers, and heterochromatin are often tissue specific, whereas active promoters and transcribed states are more often constitutive,[35] these findings indicate that hap-ASM DMRs and mQTLs occur preferentially in developmentally regulated chromatin states and are relatively depleted in regions of constitutive chromatin.

Hap-ASM DMRs and mQTLs Are Enriched in Polymorphic and Tissue-Specific CTCF Binding Sites

Insulator sequences occupied by CTCF are a unique chromatin state that regulates promoter-enhancer interactions via chromatin looping. Because our previously published findings,[14] as well as recently published data from lymphoblastoid cell lines,[37] and the additional examples of hap-ASM loci described above implicate CTCF binding sites in hap-ASM, we next tested our hap-ASM and mQTL datasets overall for statistical enrichment in CTCF ChIP-seq peaks and polymorphic CTCF binding motifs. We defined strong peaks as the top 25th percentile in ENCODE ChIP-seq data. The results showed that hap-ASM DMRs are enriched in strong CTCF peaks that are present in one or multiple cell types ($OR = 2.2$, $p = 0.001$), but not in strong CTCF peaks identified in almost all (>75%) of the assessed cell types (Figure 5A). Similar results were obtained for our mQTL dataset, with a stronger enrichment in cell-type-restricted CTCF peaks than in pan-cell type CTCF peaks ($OR = 3.3$, $p = 1.8 \times 10^{-29}$ and $OR = 1.5$, $p = 3.4 \times 10^{-07}$, respectively) (Figure 5A). In contrast, weak CTCF peaks were not enriched in either of our datasets.

In our next level of testing, we focused on CTCF binding motifs. Because the number of canonical binding motif occurrences in the human genome is lower than the total number of CTCF ChIP-seq peaks, to increase the number of evaluable loci we combined ChIP-seq data from any tissue or cell type and included both the canonical motifs and the de novo predicted CTCF binding motifs identified by Kheradpour et al.[23] By this strategy we identified 4,164 informative regions, including 51 hap-ASM DMRs, overlapping with CTCF binding motif occurrences. We defined informative motifs as those overlapping an index SNP and divided these into three classes: (A) polymorphic motifs with or without a CpG, (B) CpG-containing invariant

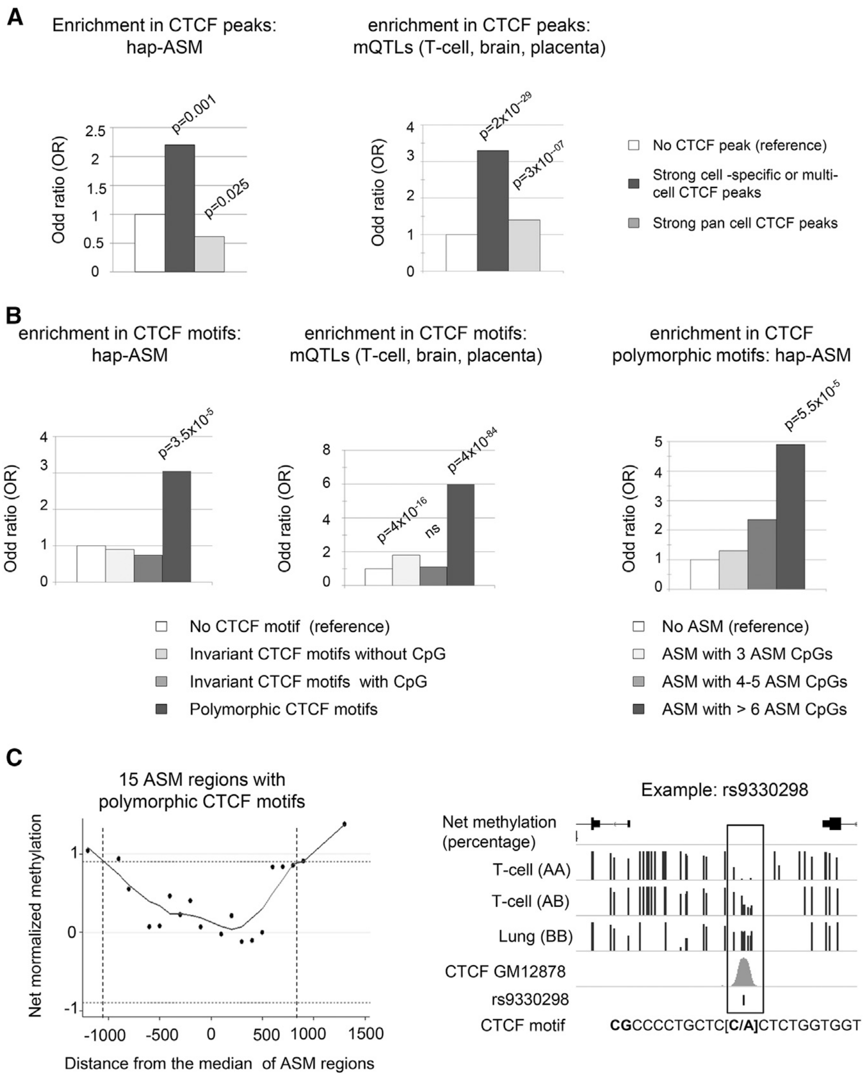

Figure 5. Hap-ASM DMRs and mQTLs Coincide with Low-Methylated Wells and Are Enriched in Polymorphic CTCF Binding Sites

(A) The graphs show a strong overall enrichment in polymorphic CTCF motifs among both hap-ASM regions and mQTLs. CTCF motifs were categorized into three classes: invariant CTCF motifs without CpG (considered as reference class), invariant CTCF motifs with CpG, and polymorphic CTCF motifs. Enrichment is expressed by the odds ratio (OR) on the y axis.

(B) Bar graph showing increasing enrichment of polymorphic CTCF motifs with the number of ASM CpGs in the DMR. To avoid potential bias due to different CpG density between the hap-ASM DMRs and the background, ORs are adjusted for the number of CpGs covered by at least 10 reads per allele. A positive correlation is observed between OR and the number (i.e., local density) of CpGs with ASM.

(C) The graph shows the averaged net methylation of all CpGs within ±1,000 bp of the center of all hap-ASM DMRs that overlap polymorphic CTCF motifs. This result indicates that hap-ASM DMRs associated with CTCF sites are discrete and occur in low methylated "wells." Values are binned by 100 bp (dots) and Lowess smoothing performed (curve). The map and net methylation data show, as an example, the hap-ASM region associated with SNP rs9330298 in the *S100A** gene cluster. The hap-ASM overlaps a polymorphic CTCF site; in the homozygous AA sample, this is a low-methylated region, reflecting CTCF binding, surrounded by hypermethylated "shoulders." A, reference allele; B, alternate allele.

motifs, and (C) invariant motifs without a CpG. We expected class C motifs to have the least association with ASM, and we therefore used them as a reference in logistic regression. As shown in Figure 5B, hap-ASM DMRs indeed proved to be enriched only in polymorphic CTCF motifs (OR = 3.1; p = 3.5 × 10^{-05}). We next grouped the hap-ASM regions into three categories with increasing ASM CpG-content (3–4, 4–5, and ≥6 ASM CpGs). Logistic regression was adjusted for the number of CpGs with good coverage so that CpG-rich hap-ASM regions are not compared to CpG poor background. As shown in Figure 5B, we found that enrichment in polymorphic CTCF motifs increases with the number of hap-ASM CpGs in the DMR. Similarly, mQTLs were strongly enriched in polymorphic CTCF motifs (OR = 7.8; p = 2.2 × 10^{-69}) (Figure 5B). The results were robust in each tissue type (Figure S21) and were stable when using larger 500 bp windows around each index SNP. These findings of strong enrichment in polymorphic but not invariant CTCF motifs, with a strong hap-ASM CpG density effect, supports allele-specific binding of CTCF as a sequence-dependent mechanism underlying a subset of hap-ASM loci. We then analyzed the averaged net methylation of all CpGs within ±1,000 bp of the center of hap-ASM DMRs that overlap polymorphic CTCF motifs (200 bp window). We found that these hap-ASM DMRs were coincident with low methylated "wells," which is consistent with selective CTCF binding to the unmethylated alleles (Figures 5C and S21).

Lastly, to ask whether hap-ASM regions with polymorphic CTCF binding sites in fact exhibit allele-specific binding of CTCF, we used available whole-genome sequencing (1000 Genomes project; SRA: SRA029810, SRA175417, SRA062045) and ChIP-seq data (ENCODE; GEO: GSM733752) in the GM12878 lymphoblastoid cell line. Among our hap-ASM DMRs with polymorphic CTCF motifs, two loci were informative (heterozygous SNP genotype and coverage ≥ 10×) in the GM12878 cells. We performed Sanger bis-seq on GM12878 cell DNA and indeed found hap-ASM at these loci (Figure S22). In other words, although WGS data contained DNA fragments mapping to both alleles for rs9330298 and rs12975442 in this cell line, the ChIP-seq data revealed allele-specific binding of CTCF, with most of the CTCF-bound DNA fragments aligning to the unmethylated alleles ($p = 1.2 \times 10^{-11}$ and $p = 1.9 \times 10^{-07}$, respectively, allelic bias tested by binomial test), and with a weaker ChIP-seq asymmetry underlying the weaker but significant hap-ASM for rs12975442 (Figure S22).

Hap-ASM DMRs and mQTLs Are Enriched in Polymorphic Binding Sites for a Group of TFs

Allele-specific TF binding occurs at up to 5% of genomic sites,[38] and TF occupancy might play a role in shaping DNA methylation patterns.[39] We therefore asked whether hap-ASM loci might also be enriched in polymorphic TF binding sites other than CTCF. We overlapped our data to all TF canonical motif occurrences in ENCODE data and found 9,409 informative occurrences. Similar to our strategy for CTCF, we grouped these TF motifs into three classes: (A) polymorphic, (B) invariant with CpG sites, and (C) invariant without CpG. From our Methyl-Seq data, we found 19 TFs, including MYC and MEF2, with an enrichment OR $\geq$ 1.5 among hap-ASM loci (Figure S23 and Table S11). Overall, hap-ASM was strongly and significantly enriched in polymorphic motifs of at least one of these TFs ($OR = 2.9$, $p = 4.9 \times 10^{-7}$). Supporting a role for allele-specific binding of TFs in hap-ASM, four of these factors—MYC, AP1, CEBPB, and GATA1—can show methylation-sensitive binding.[40–44] In our mQTL data, among informative TF motif occurrences, we identified 34 polymorphic TF motifs with an $OR > 1.5$ in T cells, 34 in placenta, and 30 in brain TC (Figure S23 and Table S11). Some of these TFs, such as AP1, CREB, YY1, ZNF143, and ZNF263, are found in each tissue, but others are more tissue specific. Overall, as shown in Figure S23, we found strong enrichment of the polymorphic motifs of at least one of these TFs among mQTLs in each tissue ($OR = 5.1$, $p = 8.3 \times 10^{-56}$ in T cells, $OR = 4.9$, $p = 7.1 \times 10^{-40}$ in placenta, and $OR = 3.8$, $p = 1.1 \times 10^{-18}$ in brain TC). These results were also found in neurons and glia (Figure S23) and were robust when using the 500 bp window size. Next, the analysis of net methylation of the flanking regions of hap-ASM DMRs that overlap polymorphic TFs motifs (200 bp window) showed, once again, that that these hap-ASM DMRs were coincident with low methylated "wells," although more heterogeneity is observed compared to CTCF (Figure S24).

To further test allele-specific CTCF and TF binding as a mechanism for hap-ASM, we estimated the binding likelihood to each allele at polymorphic CTCF and TF motif occurrences within 200 bp windows of hap-ASM DMRs, based on probability weight matrices from ENCODE. As shown in Figure 6, for CTCF, the difference of PWM score between alleles was significantly anti-correlated with the difference of methylation, suggesting that lack of CTCF occupancy is mostly associated with hypermethylation and that SNPs with higher disruptive effects are associated with stronger hap-ASM. For TF polymorphic motif occurrences, we identified two sets of TFs, one set for which relative hypermethylation was associated with low binding likelihood, suggesting protection against methylation, and a set where hypermethylation was associated with high binding likelihood, suggesting gains of methylation mediated by TF occupancy (Figure 6).

Cross-species Comparisons of Methylation Patterns Validate the CTCF Hypothesis and Point to Variation in CTCF Binding Sites as a Mechanism for Gains and Losses of Hap-ASM in Evolution

As noted above, hap-ASM DMRs are relatively enriched in non-conserved sequence elements. Capitalizing on this observation, as an additional test of the CTCF hypothesis, we used a cross-species approach in which we carried out targeted bis-seq in PBL and liver from macaque monkeys at five loci orthologous to human loci with CTCF motifs. We chose these loci such that the macaque sequences diverged from human sequences at critical base pairs in these motifs. As shown in Figures 7, S25, and S26, the results at each of the five DMRs confirmed the expected negative correlation between the methylation levels (as determined by bis-seq) and CTCF binding likelihood (as indicated by the PWM score) when comparing each human allele and the monkey alleles.

Chromosomal Regions around Hap-ASM DMRs and mQTLs Are Enriched in eQTLs

Post-GWAS mapping strategies based on hap-ASM DMRs, mQTLs, and eQTLs each have technical advantages and disadvantages, making a combined approach desirable.[17] Thus, although the current study is focused on methylation, we were interested in overlaps of our data with eQTLs. Based on the eQTL browser, 9% of the hap-ASM regions and 50% of the mQTLs in the datasets described here lie in 150 kb windows that also contain at least one eQTL (the higher yield of mQTL-eQTL pairs probably reflecting the gene-centered design of the 450K methylation arrays, as compared to the more distributed genomic coverage of Agilent SureSelect Methyl-Seq). Among these parings are *ZFP57*, *CLECL1*, *HLA-DRB** (*HLA-DRB1* [MIM: 142857], *HLA-DRB6* [MIM: NA]), *S100A**, and *CCDC155* (MIM = NA). Moreover, using all informative genomic regions as background, we find that both hap-ASM DMRs and mQTLs are statistically enriched in eQTLs within 20 kb windows ($OR = 1.4$, $p = 0.008$; $OR = 1.6$, $p = 10^{-17}$, respectively). For mQTLs, the enrichment was higher for eQTLs within 1,000 bp ($OR = 2$, $p = 4.9 \times 10^{-21}$) and decreased for more distant eQTLs ($OR = 1.2$, $p = 3.4 \times 10^{-9}$ for eQTLs within 150 kb). Similar results were observed in each tissue. Likewise, the enrichment in eQTLs among hap-ASM regions was no longer significant overall after a distance of 40 kb. Thus, hap-ASM mapping can be a useful approach for identifying candidate genes affected by disease-associated genetic polymorphisms, especially when multiple distant SNPs, in linkage disequilibrium, are identified as GWAS signals. Importantly, enrichment analysis does not exclude specific instances in which the gene being regulated lies at a greater distance, such as when the key regulatory element is an insulator or enhancer, so to avoid

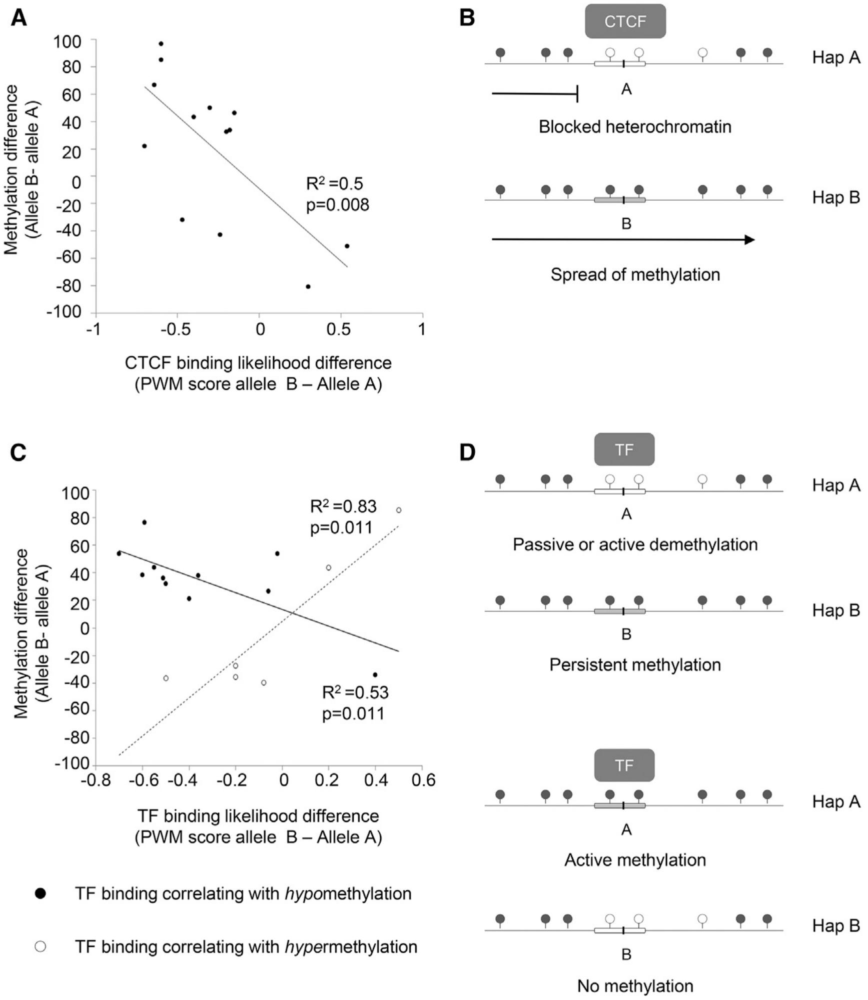

Figure 6. Methylation Differences between Alleles at Hap-ASM Loci Correlate with CTCF and TF Binding Likelihoods
(A) The graph shows methylation differences between alleles as a function of CTCF binding likelihood difference, estimated by PWM (position weight matrix) scores, and averaged across all hap-ASM regions within ±100 bp of a polymorphic CTCF motif occurrence. A significant negative correlation between methylation difference and CTCF binding likelihood is observed.
(B) CTCF acts as insulator and barrier element by blocking the spread of heterochromatin and CpG methylation. Genetic polymorphism in CTCF binding sites allow hypermethylation on the allele with lower or absent occupancy.
(C) The graph shows methylation differences between alleles as a function of TF binding likelihood difference, estimated by PWM scores, and averaged across all hap-ASM regions within ±100 bp of a polymorphic TF motif occurrences. Two subsets of TFs are observed: one with a significant negative correlation between methylation difference and TF binding likelihood difference (black circle) and one with a significant positive correlation (white circle).
(D) These results can be explained by postulating one class of TFs for which binding site occupancy is associated with demethylation of DNA and another class of TFs where binding is associated with gains of methylation. For these two classes of TF, disruption of their binding motifs by sequence polymorphisms is therefore associated with allele-specific DNA hypermethylation or hypomethylation, respectively.

missing such instances we used relatively large 150 kb windows here for tabulating and analyzing hap-ASM DMRs and mQTLs that map near GWAS peaks.

Discussion

These results have basic implications for understanding mechanisms of genetic-epigenetic interactions and practical value for validating GWAS peaks and homing in on causal variants. First considering the latter, identifying causal SNPs is essential for understanding disease pathogenesis but statistical associations by themselves cannot localize these sequences. Also, due to the small effect sizes of most risk alleles and the stringent significance thresholds necessitated by multiple testing in GWASs, many true-positive associations, including causal ones, can show sub-threshold ($>5 \times 10^{-8}$) p values. As we have proposed,[3,16,17] a strategy of overlapping hap-ASM/mQTL data with GWAS data can help to overcome these roadblocks. A number of recent studies have started to apply this promising approach,[3,4,6–15] but datasets analyzing disease-relevant primary human tissues, applying stringent criteria for hap-ASM and mQTLs, and using independent methods for validating the genome-wide profiling and defining the boundaries of DMRs are still sparse.

Here we have presented genome-wide maps of hap-ASM and mQTLs, focusing mainly on T lymphocytes and cerebral cortical brain tissue, including FANS-isolated neurons and glia, as well as a smaller set of term placentas. Key technical aspects of our study include (1) cell-type-specific profiling of both hap-ASM and mQTLs, leading to an increased net yield of disease-associated loci, (2) deep long-read Nextgen sequencing to maximize the yield of bona fide hap-ASM DMRs, (3) a definition of mQTLs based not only on p values but also on correlation coefficients and strength of the methylation asymmetry, and (4) independent validations of a diverse subset of the implicated regions by targeted bisulfite sequencing—with the high validation rate supporting the accuracy of our genome-wide datasets. The importance of defining mQTLs based not only on p values but also on correlation coefficients and strength of the methylation asymmetry is highlighted by our bis-seq data in Figure S27, which show that some

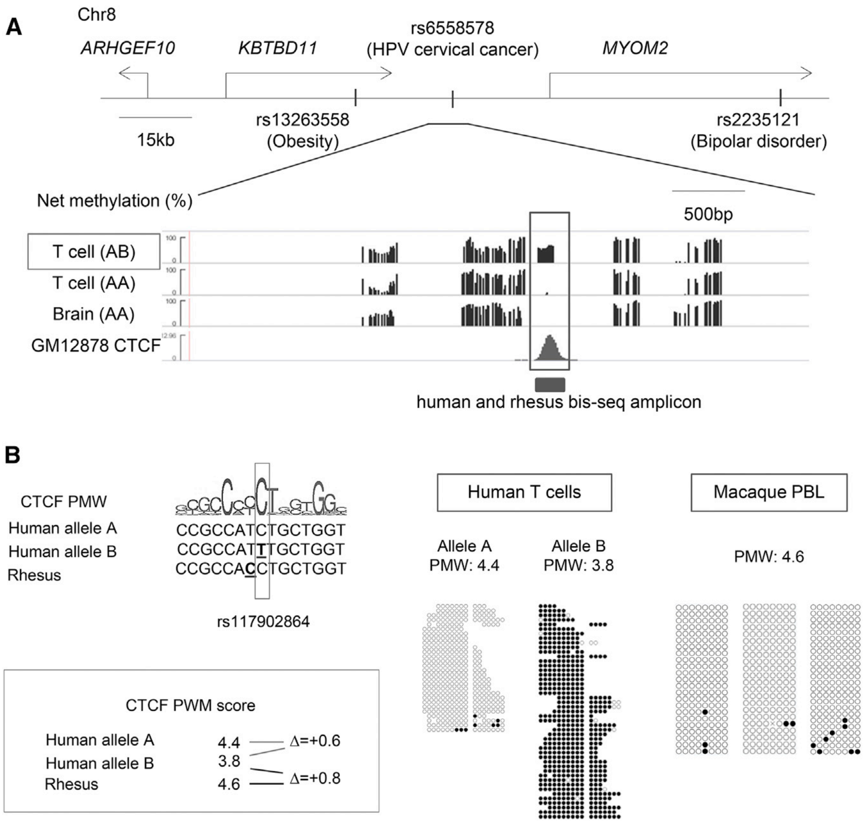

Figure 7. Comparison of Human and Macaque Methylation Patterns Supports the Correlation between Allele-Specific CTCF Binding Likelihood and Hap-ASM

(A) Map and graphs of net methylation in a 30 kb window containing a hap-ASM DMR encompassing SNP rs117902864. This SNP is relatively uncommon (minor allele frequency = 3%) and its detection in the only heterozygous T cell sample confirms the usefulness of the Methyl-Seq approach to identify rare hap-ASM DMRs. The hap-ASM overlaps a polymorphic CTCF binding site, in which the rare allele reduces binding likelihood (PWM score). Net methylation in homozygous samples showed that the DMR is located in a low-methylated region surrounded by high-methylated shoulders. The single heterozygous sample shows hypermethylation, consistent with a role for CTCF binding in maintaining the low-methylation level. GWAS significance levels are: rs6558578, HPV cervical cancer, $p = 7.00 \times 10^{-06}$; rs13263558, weight/body mass/obesity-related (obesity [MIM: 601665]) traits, $p = 3.00 \times 10^{-06}$; rs2235121, bipolar disorder and schizophrenia, $p = 8.00 \times 10^{-06}$.

(B) The sequence of the CTCF motif is shown for each human allele and for *Macaca mulatta*. The methylated allele (allele B) is associated with lower PWM score than allele A. No SNP is present in the CTCF motif instance in *Macaca mulatta* (rectangle). The macaque sequence differs from the human reference allele by one nucleotide (bold and underlined nucleotide), but this change does *not* alter CTCF binding likelihood. Sanger bis-seq of this region in PBL from two monkeys shows low methylation levels, similar to the methylation level observed in allele A in human T cells. Methylation data were obtained by MiSEQ for human and Sanger sequencing for macaque. Additional hap-ASM loci analyzed by this type of cross-species comparison are in Figures S25 and S26. Abbreviations are as follows: PWM, position weight matrix; A, reference allele; B, alternate allele.

candidate mQTLs from a prior study[15] that relied solely on statistical significance and therefore harvested a very large number of candidates (>10,000) are actually false positives. Nonetheless, our data validate a substantial group of loci from that study. Most importantly, our review of the prior studies[8,10,12,13,15] shows that more than half of the hap-ASM loci and mQTLs that we report here are newly discovered regions. Hap-ASM mapping requires both informative SNPs and CpGs, and we have shown that using our two complementary approaches increases the yield of identified regions. This seems to be true for other methods as well: using methyl-sensitive SNP array analysis (MSNP), we previously identified and fine mapped the DMRs for two novel examples of imprinted ASM and four hap-ASM loci.[3,14] Of these, the imprinted region in *VTRNA2-1* was informative in our Agilent Methyl-Seq data and the hap-ASM region in *CYP2A7* was queried here by the 450K Illumina methylation Beadchips, with these loci confirmed as ASM and mQTL, respectively. Likewise, Hutchinson et al. used MSNP to identify four hap-ASM loci.[7] None of them were informatively queried in our Methyl-Seq data and only one weak hap-ASM region (rs713875, difference between allele = 0.03) was queried by our methylation array data, which showed a subthreshold mQTL (difference between allele = 0.1, $R^2 = 0.4$, q value = 6×10^{-05}). In the future, deep whole-genome bisulfite sequencing with long reads is expected to add further information.

Supra- and Sub-threshold GWAS Peaks for which mQTL and Hap-ASM Mapping Supports Biological Relevance

Regarding the loci in which our findings help to promote GWAS signals to biological true positives, the gene lists are in Tables S4 and S6–S10 and examples are in Table 1. Methyl-Seq, but not the mQTL approach, can score hap-ASM in single heterozygous samples and thereby identify hap-ASM regions associated with rare SNPs. An interesting example is provided by the hap-ASM DMR that we detected between *KBTBD11* (MIM: NA) and *MYOM2* (MIM: 603509) on chromosome 8, for which we show maps and Methyl-Seq data in Figure 7. This DMR has strong hap-ASM in T cells and is situated close to a sub-threshold GWAS signal ($p = 7.0 \times 10^{-6}$) for susceptibility or resistance to HPV-associated cervical cancers (cervical cancer [MIM: 603956]).[45] The index SNP is rare (~1% MAF), which explains why this potentially important locus was detected in our hap-ASM data, but not in our screen for mQTLs.

Considering the other loci linked to immunological traits, multiple SNPs near the DMR tagged by SNP rs12304510 on chromosome 12 are associated with multiple sclerosis and type I diabetes.[46–48] Supporting the biological relevance of this DMR, the genes in this window, *CLECL1*, *CD69* (MIM: 107273), and *CLEC2D* (MIM: 605659), are expressed in hematopoietic cells and involved in T cell function.[49–51] Likewise, *IRF1* encodes a transcription factor that induces type I interferon and IL-27 production and is involved in Th17 cell differentiation, which play a critical role in autoimmune inflammation and viral infection.[52,53] The *IRF1*-linked hap-ASM DMR, which is stronger and more consistent in T cells than in brain (Table 1 and Figure 4), is located within 75 kb of GWAS SNPs associated with Crohn disease and asthma (asthma, bronchial [MIM: 600807]).[54–58] Also relevant to an important immunological phenotype, a GWAS SNP (rs1473500) located ~15 kb from the T-cell-specific hap-ASM DMR that we identified next to *FRMD1* on chromosome 6 was associated with secreted IL-2 in response to smallpox vaccine.[59] Here our supportive epigenetic data are particularly useful, because the GWAS signal was sub-threshold (3×10^{-7}). Another example is the *S100A** gene cluster and its associated hap-ASM DMR, which is also close to *ILF2* (MIM: 603181), encoding a subunit of nuclear factor of activated T cells (NFAT) that regulates IL-2 expression.[60] There is a sub-threshold GWAS signal ($p = 4.0 \times 10^{-06}$) for susceptibility to multiple myeloma (MIM: 254500) in this region, and our identification of a strong hap-ASM DMR within 50 kb of the GWAS peak helps to nominate this signal to a biological true positive, as well as raising the possibility of a role for *S100A** or *ILF2* in immune surveillance against multiple myeloma.

Another group of hap-ASM DMRs is relevant to both immunological and neurological phenotypes. The *HLA-DR** genes encode class II histocompatibility molecules expressed by antigen-presenting cells, and genetic variants in these genes are associated with autoimmune disorders.[48,61,62] However, there have also been reports of genetic associations of *HLA-DR* genes with neurological disorders including schizophrenia[63] and AD[64] and there is evidence suggesting a role for histocompatibility antigens in neuronal connectivity and synaptic plasticity.[65,66] A challenge in interpreting statistical associations in the HLA region is the LD structure, which makes it difficult to conclude that a disease-associated SNP is causal and not simply associated with a causal SNP several kb away. As shown in Figures 2 and S15, we have identified a hap-ASM DMR in *HLA-DRB6*, which remarkably is quite strong in brain samples while showing much weaker and less frequent allelic asymmetry in T cells. Thus, our hap-ASM data point to the CTCF-bound insulator element in the *HLA-DRB6* as a candidate for the causal regulatory element underlying the AD associations[64] in this genomic region. A recent study of 447 AD-affected brains and 293 control brains using 450K methylation arrays found that net DNA methylation in *HLA-DR** genes is associated with late-onset AD, with a statistical significance for the global association at the locus level, but with associations not passing multiple testing correction at the individual CpG level.[32] Based on our data, we propose that this result does not reflect altered DNA methylation due to pathological changes in AD brain tissue, but rather is an indication of *HLA-DR** as a susceptibility locus for this disease. Because of the hap-ASM in this region, this genetic susceptibility will manifest as small methylation differences between cases and controls in large 450K array-based methylation studies. In fact, one of the AD EWAS-associated CpGs in this region (Illumina ID cg13972202) was identified as an mQTL in our dataset (Table S12), and the other one is a SNP. Another gene associated with AD susceptibility where we observed hap-ASM in brain is *CELF2* (MIM: 602538) (Figure S17). GWAS associations between SNPs in this gene and late-onset AD were reported in *APOE* ε4 (MIM: 107741) homozygotes, and in vivo data from mice suggest that CELF2/CUGBP2, an RNA-binding protein, is involved in apoptosis in hippocampal neurons.[67–69]

Another example in this group of loci is the hap-ASM DMR upstream of *NGFR* on chromosome 17. This gene codes for the nerve growth factor receptor, which binds NGF and promotes neuronal survival, but this receptor is also expressed in T lymphocytes, for which NGF is mitogenic.[70] As shown in Figure S7, SNPs very close to the *NGFR*-linked hap-ASM DMR show near-threshold statistical associations not only with bipolar disorder, a brain-related phenotype, but also with a T-cell-related phenotype, the immune response to smallpox vaccine.[59,71] With regard to the other hap-ASM DMRs near GWAS peaks for neurological disorders, *CMTM8* is one of several chemokine-like factor genes located in a cluster on chromosome 3. This gene is expressed in brain, but its biological role is not yet known. The *CMTM8* hap-ASM DMR coincides with an intra-genic regulatory region marked by PRC2 occupancy in brain cells but not in T cells, and the strength of the allelic asymmetry is stronger in brain (Figure S13). GWAS SNPs associated with bipolar disorder and alcohol dependence (MIM: 103780), both sub-threshold, are within 15 kb of this DMR (Figure S13), so the epigenetic data help to promote this locus to a biological true positive. Likewise, *MYT1L*, Myelin Transcription Factor 1-Like, codes for a neural-specific TF expressed in fetal and adult brain. Deletions in chromosome band 2p25.3 suggest a role of this gene in intellectual disability[72] and a GWAS SNP in *MYT1L* has been associated, at near-threshold significance, with response to antipsychotic treatment.[73] The hap-ASM DMR in this gene overlaps with polymorphic MYC and CTCF motifs and is located 65 kb from this GWAS signal (Figure S10). A fourth interesting example, albeit with weaker and less frequent hap-ASM, is *PTPRN2* (MIM: 601698), which has been provisionally associated with bipolar disorder, schizophrenia (SCZD [MIM: 181500]), cocaine dependence, and depression (MDD [MIM: 608516])[74,75] (Figure S14).

Lastly, published GWASs for placenta-related traits such as intrauterine growth restriction, preeclampsia (PEE1 [MIM: 189800]), and premature birth are sparse to date and, probably for this reason, we did not find GWAS/mQTL overlaps for these conditions. Placentas are relatively under-represented in our current series, so anticipating such future GWASs, we are now collaborating with other investigators to expand the placenta data.

Special Implications of Hap-ASM in ZFP57

Relevant both to disease associations and more generally to epigenetic gene regulation, the strong hap-ASM DMR that we identified in *ZFP57* is of particular interest. *ZFP57* is located on chromosome 6 in the HLA region and encodes a TF that acts in *trans* to regulate genomic imprinting.[26,27] ZFP57 recruits DNMT1 and DNMT3a through its interaction with KAP1/TRIM28 and is thus required for acquisition and/or maintenance of a subset of DNA methylation imprints.[76] Loss of imprinting in several syndromes, particularly transient neonatal diabetes mellitus type 1, can be associated with mutations in *ZFP57*.[77,78] In addition, since a knockout of this gene in mice led to defects in cardiac development,[79] our finding of hap-ASM upstream of this gene suggests that it might be fruitful to test for associations of *ZFP57* haplotypes with human congenital heart defects (HDCA [MIM: 600001]). More broadly, our findings raise the possibility that some associations of non-coding SNPs in the HLA region with complex phenotypes might reflect not only altered expression of histocompatibility antigens, but also haplotype-dependent expression of *ZFP57*, mediated by or stabilized by hap-ASM, and possibly resulting in altered expression of imprinted genes.

Mechanisms of Hap-ASM: CTCF Sites, TF Sites, and Chromatin States

For several loci we carried out fine-mapping to define the boundaries of hap-ASM DMRs, a step that is crucial for testing mechanistic hypotheses. Overall, the results highlight a role for genetic variation in CTCF and TF binding sites as a mechanism underlying hap-ASM and mQTLs. Insulators act to establish chromatin loops, which can separate enhancers from promoters and block the spread of heterochromatin.[80] CTCF is a zinc-finger protein that is a key component of the insulator complex that anchors these loops.[81] At some imprinted loci, CTCF binding to DMRs is DNA methylation sensitive, thereby enforcing allele-specific expression of the imprinted genes,[82] and CTCF binding is also implicated in the formation or maintenance of low methylated regions.[83] We previously documented several examples of genes with hap-ASM in which the differentially methylated regions (DMRs) are discrete in size (~2 kb) and precisely overlap with binding sites for CTCF,[14] and we proposed a mechanism in which sequence polymorphisms in CTCF binding sites reduce or abrogate CTCF binding in a haplotype-dependent manner and lead to preferential CpG methylation on the unoccupied or less occupied allele.[14,17] Our bioinformatic analyses presented here, showing enrichment for polymorphic CTCF binding sites in hap-ASM/mQTL regions, and our cross-species comparisons strongly support this model. In addition, we found an overall enrichment for CTCF ChIP-seq peaks overlapping with hap-ASM DMRs, some of them without known or polymorphic motif instances. For one of these hap-ASM regions, *FRMD1*, we have shown that a combination of SNPs, all encompassed in a 450 bp segment of DNA, together determine the presence or absence of ASM. So, the simplest mechanistic possibility, which can be tested in future work, is that binding of CTCF-associated accessory factors might be influenced by these sequence variants, thus leading to allele-specific CTCF site occupancy and thereby allele-specific protection from CpG methylation, even though the CTCF site itself is not polymorphic. In addition to our model in which altered CTCF binding leads to hap-ASM, in principle at some loci hap-ASM could "come first," followed by allele-specific CTCF binding. We did not find *enrichment* in invariant CpG-containing CTCF binding sites among hap-ASM loci, but we did find examples of such loci. Therefore, it might be interesting to ask, in future studies, whether hap-ASM at invariant CpG-containing CTCF binding sites can be a primary cause of allele-specific binding of CTCF.

The polymorphic binding site model can also be made more general: allele-specific transcription factor (TF) binding can occur in up to 5% of the human genome,[38,84] changes in TF binding site occupancy can be responsible for active or passive demethylation of DNA,[39,83] and, for special TFs such as Pu.1, can mediate gains of methylation through interaction with DNMTs.[85] That polymorphisms in TF binding sequences can lead to hap-ASM has been suggested by a recent study using lymphoblastoid cells lines.[37] Taken together, our genome-wide, locus-specific, and cross-species data support both the CTCF-based and TF-based mechanisms, both for hap-ASM DMRs and for mQTLs, and in multiple disease-relevant human tissues and cell types. These results in turn provide a mechanism-based rationale for the strategy of using maps of hap-ASM and mQTLs for identifying pathogenic variants in disease-linked regulatory sequences.

Although polymorphic CTCF binding sites are enriched among hap-ASM DMRs and mQTLs, Tables S4 and S6–S10 show that they still account for only a minority of such loci. In this regard, our findings regarding chromatin states might be relevant. Those analyses point to allele-specific H3K9 methyltransferase recruitment and PRC2 binding as possible additional mechanisms. Regarding the H3K9 mark, zinc finger protein genes of the repetitive class constitute a super family of genes that form large heterochromatin regions targeted by H3K9 methyltransferase, SUV39H1.[86] SUV39H1 binds indirectly to DNA through the interaction between KRAB domain-containing proteins and KAP1 and recruits DNA methyltransferase to specific genomic sequences, inducing long-range repression through heterochromatin spread. In addition, one third of the ZNF proteins contain a KRAB domain and some have been shown to interact with KAP1, suggesting auto-regulatory loops.[87] Our TF motif analyses found enrichment in sites for several ZNF TFs, suggesting the disruption of KRAB-ZNF binding sites as a potential mechanism for hap-ASM. Likewise, poised chromatin is associated with genes regulated by the PRC2 complex, which is implicated in allele-specific chromatin repression at imprinted loci, and cross-talk between DNA methylation and the H3K27me3 histone mark in chromosomal regions bound by PRC2 has been described.[88–90] Although the sequences that recruit PRC2 in humans have not been identified, accumulating evidence supports a role for *cis*-regulatory sequences through interactions with recruitment factors,[91] and allele-specific binding of such accessory factors could explain our findings.

Gains and Losses of Hap-ASM in Species Divergence

Although we examined a small set of macaque loci, our data from comparing CTCF binding site sequences and CpG methylation patterns in human and macaques suggest an interesting line of future work on the possible role of hap-ASM in species divergence. In particular, the fact that CTCF binding likelihood tracks with methylation levels at non-conserved binding sites in macaques and humans suggests the possibility that evolution at such loci could account for differences in epigenetic patterning, chromatin organization, and phenotypic traits among primates. In fact, in our genome-wide analysis, only 20% of the polymorphic CTCF binding sites associated with hap-ASM were conserved at the sequence level between human and macaque, a result that is consistent with the general under-representation of evolutionarily conserved elements that we found among hap-ASM loci in humans, similar to what has also been described for eQTLs.[92]

Implications of Variable Hap-ASM among Individuals

We identified not only numerous tissue-restricted mQTLs and hap-ASM DMRs, but also hap-ASM DMRs that are present in some individuals but not others with the same genotype at the index SNP. Both in our Methyl-Seq data (Table S4) and among the loci that we chose for targeted bis-seq (Table 1), this finding of individual variation of hap-ASM is the rule rather than the exception. As we have demonstrated for hap-ASM in the *FRMD1* region, the combined effects of several SNPs in an extended haplotype, only partially captured by genotyping the index SNP, can explain some examples this variability (Figure S11). However, another possibility is that some haplotypes might be permissive for hap-ASM in heterozygotes only under certain environmental conditions. Altered CpG methylation patterns have been associated with diverse environmental factors,[93] and future large-scale genetic epidemiological and twin studies might reveal whether individual-restricted hap-ASM could be capturing some of these gene-environment interactions.

Accession Numbers

The accession numbers for the Methyl-Seq, Illumina HumanOmni2.5 Beadchips, and HumanMethylation450 Beadchips in this paper are GEO: GSE79144, GSE79148, GSE79254, and GSE79262.

Supplemental Data

Supplemental Data include 27 figures and 12 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.03.027.

Acknowledgments

This work was supported by NIH grants 1R01 MH092580-01, 1R01AG036040-01, 1R01AG035020-01, and DP3 DK094400-01. SNP genotyping and methylation Beadchip assays were performed by the Genomics Shared Resource of Roswell Park Cancer Institute, supported by National Cancer Institute (NCI) grant P30CA016056.

Received: October 16, 2015
Accepted: March 25, 2016
Published: May 5, 2016

Web Resources

1000 Genomes, http://www.1000genomes.org
eQTL Browser, http://www.ncbi.nlm.nih.gov/projects/gap/eqtl
Geneimprint, http://www.geneimprint.com/
GEO, http://www.ncbi.nlm.nih.gov/geo/
GWAS Catalog, http://www.ebi.ac.uk/gwas/
IGV, http://www.broadinstitute.org/igv/
MethPrimer, http://www.urogene.org/methprimer/
OMIM, http://www.omim.org/
Roadmap Epigenomics, http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html
UCSC Genome Browser, http://genome.ucsc.edu

References

1. Barsh, G.S., Copenhaver, G.P., Gibson, G., and Williams, S.M. (2012). Guidelines for genome-wide association studies. PLoS Genet. *8*, e1002812.
2. Park, J.H., Gail, M.H., Weinberg, C.R., Carroll, R.J., Chung, C.C., Wang, Z., Chanock, S.J., Fraumeni, J.F., Jr., and Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc. Natl. Acad. Sci. USA *108*, 18026–18031.
3. Kerkel, K., Spadola, A., Yuan, E., Kosek, J., Jiang, L., Hod, E., Li, K., Murty, V.V., Schupf, N., Vilain, E., et al. (2008). Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. Nat. Genet. *40*, 904–908.
4. Schalkwyk, L.C., Meaburn, E.L., Smith, R., Dempster, E.L., Jeffries, A.R., Davies, M.N., Plomin, R., and Mill, J. (2010). Allelic skewing of DNA methylation is widespread across the genome. Am. J. Hum. Genet. *86*, 196–212.
5. Hellman, A., and Chess, A. (2010). Extensive sequence-influenced DNA methylation polymorphism in the human genome. Epigenetics Chromatin *3*, 11.

6. Shoemaker, R., Deng, J., Wang, W., and Zhang, K. (2010). Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res. *20*, 883–889.
7. Hutchinson, J.N., Raj, T., Fagerness, J., Stahl, E., Viloria, F.T., Gimelbrant, A., Seddon, J., Daly, M., Chess, A., and Plenge, R. (2014). Allele-specific methylation occurs at genetic variants associated with complex disease. PLoS ONE *9*, e98464.
8. Plongthongkum, N., van Eijk, K.R., de Jong, S., Wang, T., Sul, J.H., Boks, M.P., Kahn, R.S., Fung, H.L., Ophoff, R.A., and Zhang, K. (2014). Characterization of genome-methylome interactions in 22 nuclear pedigrees. PLoS ONE *9*, e99313.
9. Smith, A.K., Kilaru, V., Kocak, M., Almli, L.M., Mercer, K.B., Ressler, K.J., Tylavsky, F.A., and Conneely, K.N. (2014). Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. BMC Genomics *15*, 145.
10. Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J., et al. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. *6*, e1000952.
11. Gamazon, E.R., Badner, J.A., Cheng, L., Zhang, C., Zhang, D., Cox, N.J., Gershon, E.S., Kelsoe, J.R., Greenwood, T.A., Nievergelt, C.M., et al. (2013). Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. Mol. Psychiatry *18*, 340–346.
12. Shi, J., Marconett, C.N., Duan, J., Hyland, P.L., Li, P., Wang, Z., Wheeler, W., Zhou, B., Campan, M., Lee, D.S., et al. (2014). Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nat. Commun. *5*, 3365.
13. Zhang, D., Cheng, L., Badner, J.A., Chen, C., Chen, Q., Luo, W., Craig, D.W., Redman, M., Gershon, E.S., and Liu, C. (2010). Genetic control of individual differences in gene-specific methylation in human brain. Am. J. Hum. Genet. *86*, 411–419.
14. Paliwal, A., Temkin, A.M., Kerkel, K., Yale, A., Yotova, I., Drost, N., Lax, S., Nhan-Chang, C.L., Powell, C., Borczuk, A., et al. (2013). Comparative anatomy of chromosomal domains with imprinted and non-imprinted allele-specific DNA methylation. PLoS Genet. *9*, e1003622.
15. Zhang, X., Moen, E.L., Liu, C., Mu, W., Gamazon, E.R., Delaney, S.M., Wing, C., Godley, L.A., Dolan, M.E., and Zhang, W. (2014). Linking the genetic architecture of cytosine modifications with human complex traits. Hum. Mol. Genet. *23*, 5893–5905.
16. Tycko, B. (2010). Mapping allele-specific DNA methylation: a new tool for maximizing information from GWAS. Am. J. Hum. Genet. *86*, 109–112.
17. Tycko, B. (2010). Allele-specific DNA methylation: beyond imprinting. Hum. Mol. Genet. *19* (R2), R210–R220.
18. Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M., et al. (2012). Patterns of cis regulatory variation in diverse human populations. PLoS Genet. *8*, e1002639.
19. Matevossian, A., and Akbarian, S. (2008). Neuronal nuclei isolation from human postmortem brain tissue. J. Vis. Exp. (20), 914.
20. Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.
21. Liu, Y., Siegmund, K.D., Laird, P.W., and Berman, B.P. (2012). Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. Genome Biol. *13*, R61.
22. Mendioroz, M., Do, C., Jiang, X., Liu, C., Darbary, H.K., Lang, C.F., Lin, J., Thomas, A., Abu-Amero, S., Stanier, P., et al. (2015). Trans effects of chromosome aneuploidies on DNA methylation patterns in human Down syndrome and mouse models. Genome Biol. *16*, 263.
23. Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. *42*, 2976–2987.
24. Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J., et al. (2014). Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. USA *111*, 6131–6138.
25. Michels, K.B., Binder, A.M., Dedeurwaerder, S., Epstein, C.B., Greally, J.M., Gut, I., Houseman, E.A., Izzi, B., Kelsey, K.T., Meissner, A., et al. (2013). Recommendations for the design and analysis of epigenome-wide association studies. Nat. Methods *10*, 949–955.
26. Li, X., Ito, M., Zhou, F., Youngson, N., Zuo, X., Leder, P., and Ferguson-Smith, A.C. (2008). A maternal-zygotic effect gene, Zfp57, maintains both maternal and paternal imprints. Dev. Cell *15*, 547–557.
27. Takikawa, S., Wang, X., Ray, C., Vakulenko, M., Bell, F.T., and Li, X. (2013). Human and mouse ZFP57 proteins are functionally interchangeable in maintaining genomic imprinting at multiple imprinted regions in mouse ES cells. Epigenetics *8*, 1268–1279.
28. Nykjaer, A., Willnow, T.E., and Petersen, C.M. (2005). p75NTR–live or let die. Curr. Opin. Neurobiol. *15*, 49–57.
29. Bakulski, K.M., Dolinoy, D.C., Sartor, M.A., Paulson, H.L., Konen, J.R., Lieberman, A.P., Albin, R.L., Hu, H., and Rozek, L.S. (2012). Genome-wide DNA methylation differences between late-onset Alzheimer's disease and cognitively normal controls in human frontal cortex. J. Alzheimers Dis. *29*, 571–588.
30. De Jager, P.L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L.C., Yu, L., Eaton, M.L., Keenan, B.T., Ernst, J., McCabe, C., et al. (2014). Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat. Neurosci. *17*, 1156–1163.
31. Lunnon, K., Smith, R., Hannon, E., De Jager, P.L., Srivastava, G., Volta, M., Troakes, C., Al-Sarraj, S., Burrage, J., Macdonald, R., et al. (2014). Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. Nat. Neurosci. *17*, 1164–1170.
32. Yu, L., Chibnik, L.B., Srivastava, G.P., Pochet, N., Yang, J., Xu, J., Kozubek, J., Obholzer, N., Leurgans, S.E., Schneider, J.A., et al. (2015). Association of brain DNA methylation in SORL1, ABCA7, HLA-DRB5, SLC24A4, and BIN1 with pathological diagnosis of Alzheimer disease. JAMA Neurol. *72*, 15–24.
33. Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., and Gunderson, K.L. (2009). Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics *1*, 177–200.
34. Marabita, F., Almgren, M., Lindholm, M.E., Ruhrmann, S., Fagerström-Billai, F., Jagodic, M., Sundberg, C.J., Ekström, T.J., Teschendorff, A.E., Tegnér, J., and Gomez-Cabrero, D. (2013). An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics *8*, 333–346.

35. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

36. Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

37. Banovich, N.E., Lan, X., McVicker, G., van de Geijn, B., Degner, J.F., Blischak, J.D., Roux, J., Pritchard, J.K., and Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS Genet. *10*, e1004663.

38. Reddy, T.E., Gertz, J., Pauli, F., Kucera, K.S., Varley, K.E., Newberry, K.M., Marinov, G.K., Mortazavi, A., Williams, B.A., Song, L., et al. (2012). Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome Res. *22*, 860–869.

39. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

40. Prendergast, G.C., and Ziff, E.B. (1991). Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. Science *251*, 186–189.

41. Ogawa, C., Tone, Y., Tsuda, M., Peter, C., Waldmann, H., and Tone, M. (2014). TGF-β-mediated Foxp3 gene expression is cooperatively regulated by Stat5, Creb, and AP-1 through CNS2. J. Immunol. *192*, 475–483.

42. Uhm, T.G., Lee, S.K., Kim, B.S., Kang, J.H., Park, C.S., Rhim, T.Y., Chang, H.S., Kim, J., and Chung, I.Y. (2012). CpG methylation at GATA elements in the regulatory region of CCR3 positively correlates with CCR3 transcription. Exp. Mol. Med. *44*, 268–280.

43. Mann, I.K., Chatterjee, R., Zhao, J., He, X., Weirauch, M.T., Hughes, T.R., and Vinson, C. (2013). CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. Genome Res. *23*, 988–997.

44. Perini, G., Diolaiti, D., Porro, A., and Della Valle, G. (2005). In vivo transcriptional regulation of N-Myc target genes is controlled by E-box methylation. Proc. Natl. Acad. Sci. USA *102*, 12117–12122.

45. Miura, K., Mishima, H., Kinoshita, A., Hayashida, C., Abe, S., Tokunaga, K., Masuzaki, H., and Yoshiura, K. (2014). Genome-wide association study of HPV-associated cervical cancer in Japanese women. J. Med. Virol. *86*, 1153–1158.

46. Barrett, J.C., Clayton, D.G., Concannon, P., Akolkar, B., Cooper, J.D., Erlich, H.A., Julier, C., Morahan, G., Nerup, J., Nierras, C., et al.; Type 1 Diabetes Genetics Consortium (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet. *41*, 703–707.

47. Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C., Patsopoulos, N.A., Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S.E., et al.; International Multiple Sclerosis Genetics Consortium; Wellcome Trust Case Control Consortium 2 (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature *476*, 214–219.

48. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.

49. Ryan, E.J., Marshall, A.J., Magaletti, D., Floyd, H., Draves, K.E., Olson, N.E., and Clark, E.A. (2002). Dendritic cell-associated lectin-1: a novel dendritic cell-associated, C-type lectin-like molecule enhances T cell secretion of IL-4. J. Immunol. *169*, 5638–5648.

50. Germain, C., Bihl, F., Zahn, S., Poupon, G., Dumaurier, M.J., Rampanarivo, H.H., Padkjær, S.B., Spee, P., and Braud, V.M. (2010). Characterization of alternatively spliced transcript variants of CLEC2D gene. J. Biol. Chem. *285*, 36207–36215.

51. González-Amaro, R., Cortés, J.R., Sánchez-Madrid, F., and Martín, P. (2013). Is CD69 an effective brake to control inflammatory diseases? Trends Mol. Med. *19*, 625–632.

52. Taniguchi, T., Ogasawara, K., Takaoka, A., and Tanaka, N. (2001). IRF family of transcription factors as regulators of host defense. Annu. Rev. Immunol. *19*, 623–655.

53. Harden, J.L., Krueger, J.G., and Bowcock, A.M. (2015). The immunogenetics of psoriasis: A comprehensive review. J. Autoimmun. *64*, 66–73.

54. Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. *42*, 1118–1125.

55. Li, X., Howard, T.D., Zheng, S.L., Haselkorn, T., Peters, S.P., Meyers, D.A., and Bleecker, E.R. (2010). Genome-wide association study of asthma identifies RAD50-IL13 and HLA-DR/DQ regions. J. Allergy Clin. Immunol. *125*, 328–335.e11.

56. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., et al.; NIDDK IBD Genetics Consortium; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. *40*, 955–962.

57. Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A., et al.; International IBD Genetics Consortium (IIBDGC) (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature *491*, 119–124.

58. McGovern, D.P., Jones, M.R., Taylor, K.D., Marciante, K., Yan, X., Dubinsky, M., Ippoliti, A., Vasiliauskas, E., Berel, D., Derkowski, C., et al.; International IBD Genetics Consortium (2010). Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Hum. Mol. Genet. *19*, 3468–3476.

59. Kennedy, R.B., Ovsyannikova, I.G., Pankratz, V.S., Haralambieva, I.H., Vierkant, R.A., and Poland, G.A. (2012). Genome-wide analysis of polymorphisms associated with cytokine responses in smallpox vaccine recipients. Hum. Genet. *131*, 1403–1421.

60. Müller, M.R., and Rao, A. (2010). NFAT, immunity and cancer: a transcription factor comes of age. Nat. Rev. Immunol. *10*, 645–656.

61. Mero, I.L., Gustavsen, M.W., Sæther, H.S., Flåm, S.T., Berg-Hansen, P., Søndergaard, H.B., Jensen, P.E., Berge, T., Bjølgerud, A., Muggerud, A., et al.; International Multiple Sclerosis Genetics Consortium (2013). Oligoclonal band status in Scandinavian multiple sclerosis patients is associated with specific genetic risk alleles. PLoS ONE *8*, e58352.

62. Han, J.W., Zheng, H.F., Cui, Y., Sun, L.D., Ye, D.Q., Hu, Z., Xu, J.H., Cai, Z.M., Huang, W., Zhao, G.P., et al. (2009). Genome-wide association study in a Chinese Han population identifies

nine new susceptibility loci for systemic lupus erythematosus. Nat. Genet. *41*, 1234–1237.

63. Fanous, A.H., Zhou, B., Aggen, S.H., Bergen, S.E., Amdur, R.L., Duan, J., Sanders, A.R., Shi, J., Mowry, B.J., Olincy, A., et al.; Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2012). Genome-wide association study of clinical dimensions of schizophrenia: polygenic effect on disorganized symptoms. Am. J. Psychiatry *169*, 1309–1317.
64. Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al.; European Alzheimer's Disease Initiative (EADI); Genetic and Environmental Risk in Alzheimer's Disease; Alzheimer's Disease Genetic Consortium; Cohorts for Heart and Aging Research in Genomic Epidemiology (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. *45*, 1452–1458.
65. Elmer, B.M., and McAllister, A.K. (2012). Major histocompatibility complex class I proteins in brain development and plasticity. Trends Neurosci. *35*, 660–670.
66. Lee, H., Brott, B.K., Kirkby, L.A., Adelson, J.D., Cheng, S., Feller, M.B., Datwani, A., and Shatz, C.J. (2014). Synapse elimination and learning rules co-regulated by MHC class I H2-Db. Nature *509*, 195–200.
67. Lee, J.H., Cheng, R., Barral, S., Reitz, C., Medrano, M., Lantigua, R., Jiménez-Velazquez, I.Z., Rogaeva, E., St George-Hyslop, P.H., and Mayeux, R. (2011). Identification of novel loci for Alzheimer disease and replication of CLU, PICALM, and BIN1 in Caribbean Hispanic individuals. Arch. Neurol. *68*, 320–328.
68. Wijsman, E.M., Pankratz, N.D., Choi, Y., Rothstein, J.H., Faber, K.M., Cheng, R., Lee, J.H., Bird, T.D., Bennett, D.A., Diaz-Arrastia, R., et al.; NIA-LOAD/NCRAD Family Study Group (2011). Genome-wide association of familial late-onset Alzheimer's disease replicates BIN1 and CLU and nominates CUGBP2 in interaction with APOE. PLoS Genet. *7*, e1001308.
69. Pacini, A., Toscano, A., Cesati, V., Cozzi, A., Meli, E., Di Cesare Mannelli, L., Paternostro, F., Pacini, P., and Pellegrini-Giampietro, D.E. (2005). NAPOR-3 RNA binding protein is required for apoptosis in hippocampus. Brain Res. Mol. Brain Res. *140*, 34–44.
70. Raychaudhuri, S.P., Raychaudhuri, S.K., Atkuri, K.R., Herzenberg, L.A., and Herzenberg, L.A. (2011). Nerve growth factor: A key local regulator in the pathogenesis of inflammatory arthritis. Arthritis Rheum. *63*, 3243–3252.
71. Scott, L.J., Muglia, P., Kong, X.Q., Guan, W., Flickinger, M., Upmanyu, R., Tozzi, F., Li, J.Z., Burmeister, M., Absher, D., et al. (2009). Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. Proc. Natl. Acad. Sci. USA *106*, 7501–7506.
72. De Rocker, N., Vergult, S., Koolen, D., Jacobs, E., Hoischen, A., Zeesman, S., Bang, B., Béna, F., Bockaert, N., Bongers, E.M., et al. (2015). Refinement of the critical 2p25.3 deletion region: the role of MYT1L in intellectual disability and obesity. Genet. Med. *17*, 460–466.
73. Adkins, D.E., Aberg, K., McClay, J.L., Bukszár, J., Zhao, Z., Jia, P., Stroup, T.S., Perkins, D., McEvoy, J.P., Lieberman, J.A., et al. (2011). Genomewide pharmacogenomic study of metabolic side effects to antipsychotic drugs. Mol. Psychiatry *16*, 321–332.
74. Curtis, D., Vine, A.E., McQuillin, A., Bass, N.J., Pereira, A., Kandaswamy, R., Lawrence, J., Anjorin, A., Choudhury, K., Datta, S.R., et al. (2011). Case-case genome-wide association analysis shows markers differentially associated with schizophrenia and bipolar disorder and implicates calcium channel genes. Psychiatr. Genet. *21*, 1–4.
75. Yang, B.Z., Han, S., Kranzler, H.R., Farrer, L.A., and Gelernter, J. (2011). A genomewide linkage scan of cocaine dependence and major depressive episode in two populations. Neuropsychopharmacology *36*, 2422–2430.
76. Zuo, X., Sheng, J., Lau, H.T., McDonald, C.M., Andrade, M., Cullen, D.E., Bell, F.T., Iacovino, M., Kyba, M., Xu, G., and Li, X. (2012). Zinc finger protein ZFP57 requires its co-factor to recruit DNA methyltransferases and maintains DNA methylation imprint in embryonic stem cells via its transcriptional repression domain. J. Biol. Chem. *287*, 2107–2118.
77. Boonen, S.E., Mackay, D.J., Hahnemann, J.M., Docherty, L., Grønskov, K., Lehmann, A., Larsen, L.G., Haemers, A.P., Kockaerts, Y., Dooms, L., et al. (2013). Transient neonatal diabetes, ZFP57, and hypomethylation of multiple imprinted loci: a detailed follow-up. Diabetes Care *36*, 505–512.
78. Baglivo, I., Esposito, S., De Cesare, L., Sparago, A., Anvar, Z., Riso, V., Cammisa, M., Fattorusso, R., Grimaldi, G., Riccio, A., and Pedone, P.V. (2013). Genetic and epigenetic mutations affect the DNA binding capability of human ZFP57 in transient neonatal diabetes type 1. FEBS Lett. *587*, 1474–1481.
79. Shamis, Y., Cullen, D.E., Liu, L., Yang, G., Ng, S.F., Xiao, L., Bell, F.T., Ray, C., Takikawa, S., Moskowitz, I.P., et al. (2015). Maternal and zygotic Zfp57 modulate NOTCH signaling in cardiac development. Proc. Natl. Acad. Sci. USA *112*, E2020–E2029.
80. Burgess-Beusse, B., Farrell, C., Gaszner, M., Litt, M., Mutskov, V., Recillas-Targa, F., Simpson, M., West, A., and Felsenfeld, G. (2002). The insulation of genes from external enhancers and silencing chromatin. Proc. Natl. Acad. Sci. USA *99* (*Suppl 4*), 16433–16437.
81. Ong, C.T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246.
82. Bell, A.C., and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature *405*, 482–485.
83. Feldmann, A., Ivanek, R., Murr, R., Gaidatzis, D., Burger, L., and Schübeler, D. (2013). Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. PLoS Genet. *9*, e1003994.
84. Butter, F., Davison, L., Viturawong, T., Scheibe, M., Vermeulen, M., Todd, J.A., and Mann, M. (2012). Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. PLoS Genet. *8*, e1002982.
85. de la Rica, L., Rodríguez-Ubreva, J., García, M., Islam, A.B., Urquiza, J.M., Hernando, H., Christensen, J., Helin, K., Gómez-Vaquero, C., and Ballestar, E. (2013). PU.1 target genes undergo Tet2-coupled demethylation and DNMT3b-mediated methylation in monocyte-to-osteoclast differentiation. Genome Biol. *14*, R99.
86. Vogel, M.J., Guelen, L., de Wit, E., Peric-Hupkes, D., Lodén, M., Talhout, W., Feenstra, M., Abbas, B., Classen, A.K., and van Steensel, B. (2006). Human heterochromatin proteins form large domains containing KRAB-ZNF genes. Genome Res. *16*, 1493–1504.
87. Groner, A.C., Meylan, S., Ciuffi, A., Zangger, N., Ambrosini, G., Dénervaud, N., Bucher, P., and Trono, D. (2010). KRAB-zinc finger proteins and KAP1 can mediate long-range

transcriptional repression through heterochromatin spreading. PLoS Genet. *6*, e1000869.

88. Hagarman, J.A., Motley, M.P., Kristjansdottir, K., and Soloway, P.D. (2013). Coordinate regulation of DNA methylation and H3K27me3 in mouse embryonic stem cells. PLoS ONE *8*, e53880.

89. Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.M., et al. (2006). The Polycomb group protein EZH2 directly controls DNA methylation. Nature *439*, 871–874.

90. Umlauf, D., Goto, Y., Cao, R., Cerqueira, F., Wagschal, A., Zhang, Y., and Feil, R. (2004). Imprinting along the Kcnq1 domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes. Nat. Genet. *36*, 1296–1300.

91. Margueron, R., and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. Nature *469*, 343–349.

92. Popadin, K.Y., Gutierrez-Arcelus, M., Lappalainen, T., Buil, A., Steinberg, J., Nikolaev, S.I., Lukowski, S.W., Bazykin, G.A., Seplyarskiy, V.B., Ioannidis, P., et al. (2014). Gene age predicts the strength of purifying selection acting on gene expression variation in humans. Am. J. Hum. Genet. *95*, 660–674.

93. Hatchwell, E., and Greally, J.M. (2007). The potential role of epigenomic dysregulation in complex human disease. Trends Genet. *23*, 588–595.