Published in final edited form as:

Science. 2018 September 28; 361(6409): . doi:10.1126/science.aar3146.

Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci

Vitor Onuchic[#1,2,3,4,†], **Eugene Lurie**[#1,3,4], **Ivenise Carrero**[1,3], **Piotr Pawliczek**[1,3], **Ronak Y. Patel**[1,3], **Joel Rozowsky**[5], **Timur Galeev**[5], **Zhuoyi Huang**[1,6], **Robert C. Altshuler**[4,7,8], **Zhizhuo Zhang**[7,8], **R. Alan Harris**[1,3,4], **Cristian Coarfa**[1,3,4], **Lillian Ashmore**[1,2,3], **Jessica W. Bertol**[9], **Walid D. Fakhouri**[9], **Fuli Yu**[1,2,6], **Manolis Kellis**[4,7,8], **Mark Gerstein**[5], and **Aleksandar Milosavljevic**[1,2,3,4,‡]

[1]Molecular and Human Genetics Department, Baylor College of Medicine, Houston, TX, USA.

[2]Program in Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX, USA.

[3]Epigenome Center, Baylor College of Medicine, Houston, TX, USA.

[4]NIH Roadmap Epigenomics Project.

[5]Program in Computational Biology and Bioinformatics, Department of Molecular Biophysics and Biochemistry, and Department of Computer Science, Yale University, New Haven, CT, USA.

[6]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA.

[7]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA.

[8]Broad Institute of Harvard University and Massachusetts Institute of Technology, Cambridge, MA, USA.

[9]Center for Craniofacial Research, Department of Diagnostic and Biomedical Sciences, School of Dentistry, University of Texas Health Science Center at Houston, Houston, TX, USA.

[#] These authors contributed equally to this work.

Abstract

[‡]Corresponding author. amilosav@bcm.edu.

Author contributions: V.O., E.L., and A.M. conceptualized research goals and designed studies; R.A.H and C.C. performed collection and organization of sequencing datasets; C.C., P.P., and R.Y.P. performed processing of sequencing datasets; V.O., E.L., P.P, R.Y.P., J.R., T.G., Z.H., R.C.A., Z.Z., and L.A. performed data curation; Z.H. created and performed variant calling pipelines with supervision from F.U.; J.R., T.G., R.C.A., and Z.Z. provided software and participated in allelic histone calling with supervision from M.K. and M.G.; V.O., E.L., and I.C. performed bioinformatics, statistical analyses, and presentation of data with supervision from A.M.; J.W.B. performed luciferase assay validation experiments with supervision from W.D.F.; E.L. performed validation experiments on external datasets; L.A. developed the online resource; V.O., E.L., and A.M. wrote the original draft; E.L., M.K., and A.M. revised and edited the manuscript.

[†]Present address: Illumina, San Diego, CA, USA.

Competing Interests: The authors declare no competing interests.

Data and materials availability: All allelic datasets are available on: https://genboree.org/genboreeKB/projects/allelic-epigenome. Code utilized for bioinformatic analyses can be found at: https://github.com/BRL-BCM/allelic_epigenome. Utilized whole genome sequencing, whole genome bisulfite sequencing, and ChIP-sequencing reads are stored on the Roadmap Epigenomics BioProject page (accessions: PRJNA34535 and PRJNA259585) and dbGaP (accessions: phs000791.v1.p1,phs000610.v1.p1, and phs0007000.v1.p1).

To assess the impact of genetic variation in regulatory loci on human health, we construct a high-resolution map of allelic imbalances in DNA methylation, histone marks, and gene transcription in 71 epigenomes from 36 distinct cell and tissue types from 13 donors. Deep whole-genome bisulfite sequencing of 49 methylomes reveals sequence-dependent CpG methylation imbalances at thousands of heterozygous regulatory loci. Such loci are enriched for stochastic switching, defined as random transitions between fully methylated and unmethylated states of DNA. The methylation imbalances at thousands of loci are explainable by different relative frequencies of the methylated and unmethylated states for the two alleles. Further analyses provide a unifying model that links sequence-dependent allelic imbalances of the epigenome, stochastic switching at gene regulatory loci, and disease-associated

A majority of imbalances in DNA methylation between homologous chromosomes in humans are associated with genetic variation *in cis*, where the genetic variants affect the methylation state of neighboring cytosines on the same chromosome (1,2). Such sequence-dependent allele-specific methylation (SD-ASM) affects at least 8-10% of the autosomal genome (3–5). SD-ASM is an ideally “controlled” natural experiment providing information about consequences of genetic variation *in cis* because both “case” and “control” loci can be found within an individual on homologous chromosomes within the same cellular and nuclear environment.

In contrast to association-based methods, such as expression quantitative trait loci (eQTLs) and methylation quantitative trait loci (mQTLs), employed to establish the functional effects of common (>5% minor allele frequency) non-coding variants (6), allelic imbalances (AIs) can be established by profiling a single sample, revealing the functional effects in *cis* of both relatively low-risk common variants and highly penetrant disease-causing rare and de novo variants (1, 2, 5, 7–10).

While the analyses of SD-ASM traditionally involved measurement of average methylation levels across many cells, the epigenome is known to exhibit stochastic cell-to-cell variation and even variation between the two chromosomes within the same nucleus (11). Whole-genome bisulfite sequencing (WGBS) has the potential to provide insights into stochastic variation at the ultimate single-chromosome level of resolution, as it provides information about the genetic variant and methylation of neighboring cytosines within the same sequencing read that comes from a single chromosome. Sequencing-based studies of DNA methylation have revealed pervasive stochastic epigenetic polymorphisms within autosomal loci (12, 13).

Methylation patterns evolve along predictable trajectories during normal development (14), being highly stochastic in early metastable stages and stabilizing within differentiated tissues, resulting in mosaicism (15). The methylation patterns are maintained in stem cells through a dynamic stochastic epigenetic switching equilibrium (an ergodic process), providing both epigenetic buffering of environmental noise and responsiveness to specific transcription factors (12). In contrast, more differentiated cells (12), and tumor cells (13, 16), appear to employ the more error-prone mechanism of direct replication of CpG methylation patterns which is associated with clonality and mosaicism. Despite the stochastic nature of DNA methylation changes during development, in response to environmental input and in human diseases, the effects of genetic variation on stochasticity and mosaicism of the epigenome remain unexplored.

Results

Patterns of allelic imbalances across epigenomic marks

To explore the effects of genetic variation on the epigenome, the NIH Roadmap Epigenomics Project (17) has now completed whole-genome sequencing (WGS) on genomes of 13 donors and published NIH Roadmap reference epigenomes from 71 combined samples that collectively represent 27 distinct tissue types and 9 cell types (fig. S1). For accurate identification of heterozygous genomic loci, we sequenced the donor genomes (18). Eight assays were included in most of the samples and used for AI detection: WGBS, RNA-seq, and ChIP-seq for 6 different histone marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3, and H3K27ac) (fig. S1). Allele-specific methylation (ASM) analysis was performed at heterozygous single-nucleotide polymorphism (SNP) loci within the 49 WGBS methylomes using a threshold of methylation difference of >30% between alleles and by estimating significance by Fisher’s exact test on the counts of methylated and unmethylated cytosines observed on the same sequencing read with each of the two SNP alleles (fig. S2A) (18). The identification of AIs for histone marks and transcription was performed using the AlleleSeq pipeline (fig. S2B) (18, 19).

Considering the AIs in all the marks, the imbalances in DNA methylation were by far the most abundant (table S1 and fig. S3), largely due to the genome-wide distribution of DNA methylation, in contrast to the uneven genomic distribution of other marks. Among the histone marks, H3K27ac had more imbalance calls than others (table S1), in part due to deeper ChIP-seq coverage for H3K27ac (table S1 and fig. S3). At promoters, H3K27ac and H3K4me3 marks were more abundant on the allele with less DNA methylation (Fig. 1A). Conversely, H3K9me3 signal was more abundant on the allele with more methylation in promoters (Fig. 1A). At enhancers H3K27ac tended to occur more often on the allele with less DNA methylation (Fig. 1A). We also detected, at high specificity, enrichment of AIs in methylation and coordinated changes in transcription and histone marks within a majority of those imprinted loci that included a heterozygous SNP (figs. S4 and S5) (18).

We next evaluated the extent of reported SD-ASM. Consistent with genetic effects *in cis* (6, 20–22), co-occurrence of ASM at the same heterozygous locus across different samples was higher than expected by chance under a permutation-based null model (fig. S6A). The degree of co-occurrence of ASM tended to be higher for pairs of samples across tissues of the same individual than between pairs from the same tissue across different individuals, which was higher than for samples without matching tissue or individual (fig. S6B). Low concordance in ASM calls between individuals may be due to local haplotype context, epigenetic drift, or other non-genetic factors (3, 4, 6, 20, 22, 23). Gaussian mixture modeling (18) showed that allelic differences in methylation (above the 30% threshold) at heterozygous SNPs had tendency to occur in the same direction (the same allele showing higher methylation than the other) across pairs of samples (fig. S6, C to E).

In order to increase the power to detect SD-ASM at high sensitivity, we pooled the reads across all 49 methylomes and applied the same detection method as for individual samples (fig. S7A) (18). The deep coverage of the combined set (1,691-fold total coverage in bisulfite sequencing reads in the combined set of 49 methylomes) increased our power to detect those sequence-associated allelic imbalances that were detectable across different tissues and donors (fig. S7, B to D), while our power to detect tissue-dependent and donor-dependent SD-ASM was reduced (fig. S8, A and B). The number of accessible heterozygous loci (those having at least 6 counts per allele), for SD-ASM determination, after pooling rose to 4,913,361, increasing our SD-ASM mapping resolution—measured as an average distance between "index hets"—to 600bp. At the 30% methylation difference default threshold allelic imbalances were detected at 5% of index hets; lowering the threshold to 20%, a total of ~8% index hets showed AIs.

Sensitivity of the methylome to genetic variation varies across classes of genomic elements

We next explored whether SD-ASM had the tendency to occur within any particular type of genomic element. Using the reads pooled across the 49 methylomes, we observed depletion of SD-ASM within promoters containing CpG islands (Fig. 1B), as well as within CpG islands in general (Fig. 1C), consistent with observations that ASM is depleted in CpG islands (4, 21) and that mQTLs are depleted within promoters of genes within CpG islands (24, 25) and that expression Quantitative Trait Methylation is enriched within CpG island shores and not in CpG islands themselves (26). In contrast, and mirroring previous mQTL patterns (25), promoters of genes not in CpG islands showed high levels of SD-ASM (Fig. 1D). We also observed enrichment of SD-ASM downstream from the promoter and into the gene body (Fig. 1, B and 1D) and positive association between allele-specific expression (ASE) and ASM over exons (Fig. 1A), consistent with higher methylation of actively transcribed regions including those on the X chromosome (27) and with the enrichment of mQTLs in regions flanking the transcription start site (TSS) (23, 28). One factor contributing to the ASM, particularly near the transcription start sites (Fig. 1D), may be the presence of transcriptional regulatory signals (29).

SD-ASM was also highly enriched within enhancers (Fig. 1E), consistent with previous reports (24, 28). The abundance of transcription factor (TF) binding sites within enhancers, suggests that SD-ASM may result from disruption of transcription factor binding (23). Under that assumption, our data suggests that TF binding at CpG islands and CpG-rich promoters is buffered against genetic perturbations, while the TF binding to non-CpG promoters and enhancers is most sensitive. A somewhat puzzling mild depletion of SD-ASM in the flanking regions of enhancers was also observed (Fig. 1E), also suggesting buffering in those regions.

SD-ASM is attributable to differences between allele-specific epiallele frequency spectra

We next asked if the lack of buffering at SD-ASM loci may result in excess stochasticity and metastability, defined by the presence of more than one stable state, each stable state corresponding to an epiallele (single chromosome methylation pattern). To answer this question, we made use of the deep combined WGBS read coverage across 49 methylomes (table S2) and of the fact that each read relates a single variant to a single epiallele. We assessed epialleles by scoring the methylation status of four homozygous CpG sites ($4^2 = 16$ possible epialleles) that were the closest to each index het in individual WGBS reads (13, 14) (Fig. 2A). (We note that our use of the term "epiallele" follows the most recent usage (12, 13) and does not comply with the original definition (30) that implies inter-generational inheritance. Our use of the term "metastability" is consistent with its use in dynamical systems theory and does not imply inheritance of an epiallele during cell division.)

To quantify the amount of stochasticity at index het loci, we employed Shannon entropy (18). The entropy values ranged from 0 to 4: an even distribution of frequencies across the 16 possible epiallele patterns produces a maximum entropy score of 4 bits; whereas a complete absence of stochasticity due to maximal "buffering" implies just one epiallele with nonzero frequency and an entropy score of 0 bits. To assess quantitatively any differences in buffering (lack of sensitivity to genetic variation) between SD-ASM and control loci, we identified SD-ASM loci that had sufficient coverage and a close index het without ASM and compared entropies. A total of 6,619 (2.7%) of 241,360 loci with SD-ASM met the two criteria (18). A striking difference in entropy was observed, providing a quantitative assessment of the higher stochasticity at the SD-ASM vs. control loci (Fig. 2B).

We next examined enrichment for epigenetic polymorphisms at SD-ASM loci. We estimated the number of frequent epialleles for each locus by sorting the epialleles from the most to the least frequent and identified the minimal-size "top-list" of epialleles that accounted for at least 60% of all the reads with ascertained epialleles. In contrast to the control loci, which typically had only one high-frequency epiallele on the "top-list" and were therefore not epigenetically polymorphic, SD-ASM loci showed multiple frequent epialleles, in most cases just two (Fig. 2C). By examining the top pairs of epialleles, we found that 71.7% of the pairs consisted of one that was completely methylated and another completely unmethylated (Fig. 2C). This is concordant with previous reports of biphasic (fully methylated and fully unmethylated) distributions of methylation in amplicons with high interindividual methylation variance and in PCR clones with bimodal methylation patterns (3, 31). Strikingly, allelic imbalances at SD-ASM loci could be traced to shifts in epiallele frequency spectra between alleles, typically shifts in relative frequencies of the fully methylated and fully unmethylated epialleles (Fig. 2C). We validated the observed excess of stochasticity and the enrichment for the biphasic pattern at SD-ASM loci using an independent WGBS dataset from ENCODE (fig. S9, A-C) (18).

We next quantified the relationship between genetic variation and stochastic epialleles. At each locus we estimated the probabilities of epialleles for each allele (higher probabilities indicated by thicker arrows in Fig. 2, C and D). We then quantified the degree to which genetic alleles determine epiallele frequencies using a Coefficient of Constraint (18), an information-theoretic measure that is a generalization of the rsquared Coefficient of Determination commonly used in genetics and is more appropriate for quantifying genetic determination of stochastic phenotypes. A larger value for the Coefficient of Constraint value signifies that epigenetic variation is more constrained (determined by) genetic variation *in cis*. Intuitively, a larger Coefficient of Constraint indicates larger difference in the epiallelic frequency spectra corresponding to the two alleles, implying higher degree of determination of epiallele frequency spectra by the genetic alleles (Fig. 2D).

There are two general mechanistic models that could explain the effect of sequence variation *in cis* on epiallele frequency spectra. The ergodic/periodic model stipulates ongoing switching between metastable states, the transitions being stochastic with a possible component of periodicity such as circadian oscillations. If a sufficient number of stochastic transitions from one epiallele to another occur, that epiallele frequency spectrum depends largely on the sequence-dependent shape of the current energy landscape (i.e., state transition probabilities) and not on the epigenetic memory of past events (Fig. 2E). In contrast, the mosaic model stipulates that epialleles are stably transmitted over time and even during cell division, being “frozen” after a period of initial metastability into one of the stable states. Importantly, both models entail a period of metastability, whether past (mosaicism) or current (ergodic/periodic model).

CTCF binding loci show sequence-dependent stochastic switching and looping

Because of its association with DNA methylation at a large number of binding sites, we next examined the role of CCCTC-binding factor (CTCF) in creating the metastable states that correspond to epialleles. Metastability is known to be created by positive (including double negative) feedback loops (32) that in our case also include interactions in *cis*, such as the protection against DNA methylation by CTCF binding and reciprocal preference of CTCF for unmethylated DNA (33). The first indication of the role of CTCF binding in metastability came from the observation that the heterozygote with the larger Coefficient of Constraint (G/C het) also showed larger differences in predicted CTCF binding affinity between the two alleles than the other (T/C het) (Fig. 2D). Considering that the Coefficient of Constraint is proportional to the differences in epiallele frequency spectra for the two alleles (identical epiallele frequency spectra resulting in Coefficient Constraint value of 0), this observation suggested a positive correlation between the Coefficient of Constraint and the differences in CTCF binding affinity for the two genetic alleles, which was indeed observed (Fig. 3A). In terms of the epigenetic landscape distortion due to genetic variation, we see that sequence variants that show larger differences in CTCF binding affinity also show a greater differences in their epigenetic (energy) landscapes, as reflected in the more prominent shifts between alleles in their occupancy of metastable states (as measured by higher values of Coefficient of Constraint) (Fig. 3A, top). As CTCF binding and demethylation of its binding site are mutually reinforcing (forming a positive feedback loop and a metastable state), the model also predicts that the variants associated with higher CTCF binding affinities will show lower methylation, which is indeed the case (Fig. 3A, bottom) as previously observed (23, 24). Taken together, these results suggest sequence-dependent stochastic epigenetic switching between metastable states that is mediated by CTCF binding.

Because the CTCF transcription factor establishes chromatin loops (34), we asked if the allelic state of methylation also coincided with allelic looping. Toward this goal, we utilized a study (35) reporting heterozygous SNP loci that associate both with allelic CTCF binding and allelic chromatin looping, as determined by Chromatin Interaction Analysis by Paired-

End Tag Sequencing (ChIA-PET). Indeed, a total of 44 of those SNP loci were also present in our dataset. Comparing our signals for the methylation state of CTCF binding sites with the predicted CTCF motif disruption scores suggested that SD-ASM is a more accurate indicator of allelic CTCF binding and looping than the motif disruption score (Fig. 3B) (18).

Transcription factor binding sites show sequence-dependent shifts in epiallele frequency spectra and allelic imbalances

Analyses of ASM at regulatory elements and eQTLs revealed associations between ASM and allele-specific histone marks with downstream allele-specific transcription (fig. S10, A to F) (18). These results complemented previous studies (22, 23) and suggested involvement of allele-specific TF binding and co-factors in ASM. To examine the role of allele-specific TF binding, we focused on the set of 377 TFs assessed for binding affinity using the high-throughput systematic evolution of ligands by exponential enrichment (SELEX) method (36). As for CTCF, we identified the subset of binding motif loci in a heterozygous state with two frequent epialleles and examined the correlation between Coefficient of Constraint and difference in predicted allelic binding affinities across these loci for each TF (table S3). Because of the relatively small number of such loci per TF, only 13 showed significant individual p-values (t-test, $P < 0.05$), with only CTCF surviving Bonferroni correction (for testing 377 TFs; table S3). However, a majority (11 of 13) of the TFs that showed individually significant correlation also showed positive correlation ($P = 0.01$, Binomial test), consistent with the pattern observed for CTCF where larger differences in TF binding affinities correspond to larger distortions in the configuration of metastable states within the landscape (table S3).

Likewise, we next examined for all 377 transcription factors whether disruptions of their predicted binding sites associated with methylation imbalances. A majority (241) showed SD-ASM enrichment within their binding motifs compared to flanking loci (500bp on each side) (Fig. 3C and table S4), suggesting that transcription factor binding associates with allelic DNA methylation. The SD-ASM outside of the examined motifs may be attributable to sequence variation within undiscovered binding loci, within motifs of non-coding RNAs, or within loci in physical proximity/contact with regions of perturbed TF activity.

We then examined the relation between allelic differences in motif strengths and methylation levels at SD-ASM loci (18). We observed that for more than half of the transcription factors tested (207), there was an association between motif strength and level of methylation (Fig. 3C). Most TFs (159) showed gain in methylation on the allele with the disrupted motif, consistent with the TF binding either protecting a region from passive methylation (37) or causing active demethylation (Fig. 3C) (38). In contrast, a smaller number (48) of TFs, including members of TF families that recruit methyl-transferases such as the ETS-domain transcription factor family members (39, 40), showed loss of methylation on the allele with the disrupted motif (Fig. 3, C and D). About a quarter of TFs that show enrichment for SD-ASM show no bias in directionality (table S4), the lack of bias being explainable by contextual behavior at different binding loci, such as for Nuclear factor of activated T-cells 1 (41), or due to competing TFs at overlapping motifs. Our results support that transcription factor motif sequences are predictive of proximal CpG methylation levels (23, 42, 43).

We sought to validate the downstream functional consequences of SD-ASM variants, with predicted allelic differences in transcription factor binding, using a luciferase assay. We prioritized cis-overlapping motifs (CisOMs) including those of cMYC proto-oncogene (cMYC) and tumor suppressor p53 (TP53) that show competitive binding at many loci (44) because CisOMs provide one of the mechanisms of metastability (Fig. 3E) and also those that may have consequences for human disease (table S5). All four SNP validations showed allelic effects on luciferase expression, including two SNPs within CisOMs for cMYC and P53 and some falling within disease-associated loci (fig. S11) (18), suggesting that SD-ASM helps identify those disease-associated variants that also have functional consequences.

SD-ASM is enriched near disease-associated loci

We observed that heterozygous variants with SD-ASM were enriched in the neighborhood of variants previously reported as significant in genome-wide association studies (GWAS) of common disease (22, 23, 45) (Fig. 4A). The enrichment was stronger around GWAS variants that have been replicated in multiple studies vs. those that have not. To explore more specifically the role of enhancers, we performed a similar enrichment analysis focusing only on GWAS and SD-ASM variants overlapping enhancer elements. Enhancers containing replicated GWAS variants were significantly ($P < 0.0001$, Chi-square test) more likely to also contain a variant with SD-ASM than enhancers that did not contain replicated GWAS variants (Fig. 4B). Taken together, these results indicate that allelic imbalances provide information about the role of specific loci in common diseases, pointing to the loci that are sensitive to the effects of genetic variation and have functional effects. The enrichment of both GWAS loci and allelic imbalances at enhancers, and sensitivity of TF binding to genetic variation discussed in previous sections, provide a mechanistic link between allelic imbalances and GWAS associations.

Variants showing SD-ASM are under purifying selection

Because the variants with large effects are under purifying selection, they tend to be rare with frequencies below the detection threshold of association studies such as GWAS, mQTL, and eQTL. In contrast, allelic imbalances may provide evidence for functional effects even for rare variants that may be detected in only one individual. Based on previous studies that have utilized signatures of purifying selection such as shifts toward smaller derived allele frequency (DAF) to identify functional variants (46, 47), we would expect that ASM variants would also tend to have lower DAF than those without ASM. Therefore, we obtained DAF estimates from the 1000 Genomes Project (48), ignoring variants that overlapped regions with low accessibility to variant calling. We observed that in nearly every sample in our dataset, heterozygous variants with ASM were significantly ($P < 0.05$, Chi-squared test) more likely to have DAF smaller than 1%, than those without ASM (methylation difference between alleles < 5%) (Fig. 4C). Overall, this analysis found ~130 (median) more rare (DAF < 1%) variants than expected among those with ASM per individual methylome, providing a lower bound on the number of those under purifying selection per individual. When we repeated the analysis for enhancer regions, strong signal was again observed (Fig. 4D), suggesting a median excess of at least 26 enhancer variants under purifying selection per individual.

The lower bounds from individual samples may underestimate the extent of purifying selection due to under-detection of SD-ASM. We therefore investigated whether an enrichment for rare variants could also be seen for those variants associated with SD-ASM from the combined dataset, using neighboring variants as controls (18). We observed that the chance of a locus having SD-ASM decreased as the derived allele frequency increased (Fig. 4E; notice that there were very few variants with DAF > 50%, causing high variance and large confidence intervals). We further tested whether there was a significant enrichment for variants with DAF < 1%, among those with SD-ASM, and found that such enrichment was indeed significant (odds ratio 1.18; P < 0.0001, Chi-squared test; Fig. 4F). That enrichment represents an excess of 2,184 rare variants among those with SD-ASM compared to controls. Considering that this observed excess represents a set of 11 genomes (9 individuals and 2 cells lines), we estimate at least ~200 variants with SD-ASM under purifying selection per individual donor.

Discussion

Taken together, our findings suggest a mechanistic link between sequence-dependent allelic imbalances of the epigenome, stochastic switching at gene regulatory loci, and disease-associated genetic variation. Our allelic epigenome map reveals CpG methylation imbalances exceeding 30% differences at 5% of the loci, which is more conservative than previous estimates in the 8-10% range (3, 4); similar value (8%) is observed in our dataset when we lowered our threshold for detecting allelic imbalance to 20% methylation difference between the two alleles. We observe an excess of rare variants among those showing ASM, suggesting that an average human genome harbors at least ~200 detrimental rare variants that also show ASM.

The methylome's sensitivity to genetic variation is unevenly distributed across the genome. The higher buffering of CpG islands, and associated promoters, may be due to their utilization by housekeeping genes. Conversely, other promoters and enhancers may show lower buffering, due to their presence in tissue-specific genes that tend not to be associated with CpG islands (49). Our findings are consistent with the evolutionary advantages that may stem from the buffering of housekeeping genes against the effects of random mutations, while still retaining the potential for evolutionary innovation through changes in the regulation of less essential genes with tissue-specific expression patterns, and refine reports suggesting that *all* promoters show epigenetic buffering (50). Highest sensitivity to genetic variation at enhancers, and potentially perturbations in general, is consistent with observations of high cell-to-cell variability of their methylation (28, 51). Validated GWAS loci, and the enhancers within those loci, show enrichment for allelic imbalances, suggesting that sensitivity of those loci to genetic variation, and potentially also to environmental influences, may at least in part explain their role in common diseases.

Overall, our results suggest an explanation for the conservation of "intermediate methylation" states at regulatory loci (52). These "intermediate methylation" states reflect the relative frequencies of fully methylated and fully unmethylated epialleles corresponding to biphasic On/Off switching patterns (31) at regulatory loci marked with SD-ASM.

Moreover, our analyses reveal that the SD-ASM is explainable by allele-specific switching patterns at thousands of heterozygous loci.

Waddington's Epigenetic Landscape has served as a guiding metaphor for the emergence of cellular identities during development. The Landscape has until now been an abstract construct (53), disconnected from the mechanistic function of gene regulation. Our energy landscapes (Figs. 2D and 3E) may be interpreted as a special type of Epigenetic Landscape model of gene regulatory interactions *in cis* where epialleles correspond to metastable states (attractors) within the landscape (30). As cells transition into a more differentiated state, the cellular epigenome enters a buffered "valley" at a bifurcation point in the landscape (Fig. 2E). Because our current dataset is static, it precludes us from being able to distinguish between the mosaic and stochastic/periodic models that characterize the more and less differentiated states respectively.

Consistent with proposed theoretical models that define the Landscape in terms of potential energy functions (53), and postulate local attractors created by positive feedback loops (32), our model involves interactions between one or more TFs, DNA methylation, and likely other epigenomic marks. Competitive binding of TFs at CisOMs may be one of the mechanisms of metastability. By putting the metastable states within the Landscape in correspondence with epialleles and TF binding, we bring Waddington's Landscapes into correspondence with assayable and quantifiable epiallele frequency spectra and with specific mechanisms of gene regulation.

Our findings are consistent with the role for bistable switching and stochasticity in bacterial gene regulation (54) and extend stochasticity to eukaryotic cells in vivo within their natural tissue context across a diversity of human tissues. One obvious question is the possible purpose of stochasticity at gene regulatory loci across both domains life. The sharp contrast between the "digital" nature of regulatory elements and the "analog" nature of concentrations of upstream TFs, and downstream gene products, suggests that we may be observing a naturally evolved system similar to von Neumann's "stochastic computer", an early hybrid analog-digital computer design that can implement regulatory circuits within control systems where analog quantities are not encoded using the usual binary or decimal system but are encoded as fractions of "On" states in a stochastic series of "On" and "Off" states (55). In contrast to a stand-alone stochastic computer, there is a multitude of cells within a tissue, raising the question about the role of intercellular communication within tissue microenvironment in establishing a dynamic equilibrium resulting in observed methylation averages.

To promote further data analyses and experimental work by the community, we provide an Allelic Epigenome Atlas, a collection of annotations, including allelic imbalance scores, for all of the ~4.9M heterozygous loci analyzed here (18). The breadth of the genome-wide allelic imbalance scores available in the Allelic Epigenome Atlas, which includes the multitude of different individual tissue and cell types profiled herein, may serve as additional layers of functional evidence for prioritization of disease-causal candidate variants, including subthreshold GWAS variants in implicated regions and non-coding variants detected by clinical whole-genome sequencing.

Materials and Methods

Heterozygous SNP loci were identified using a joint variant calling pipeline on WGS datasets from 13 donors from the NIH Roadmap Epigenomics Project. ChIP-seq and RNA-seq datasets from a total of 71 various tissues of these donors were utilized to detect allelic imbalances in histone marks and transcription as described (19). Allele-specific methylation was detected using an in-house script on 49 WGBS datasets to compare counts of methylated and unmethylated cytosines proximal to each allele. Different regulatory regions and GWAS and eQTL loci were tested for enrichment of allelic imbalances using nearby heterozygous SNPs without allelic imbalance as controls. Differences in methylation between alleles were compared to the alleles' predicted TF motif strengths. Epialleles were analyzed by quantifying the methylation patterns of the 4 closest CpGs to the alleles in single WGBS reads and comparing these patterns between alleles. Shannon entropy was calculated at different loci using epiallele patterns to quantify stochasticity at SD-ASM loci. Coefficient of constraints were calculated at SD-ASM loci to determine the constraint of epiallele polymorphisms by genetic variants.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank I. Golding for helpful comments and reviewing the paper and Y. Ruan for assisting with allelic looping datasets.

Funding: A.M. acknowledges support from the Common Fund of the National Institutes of Health (Roadmap Epigenomics Program, grant U01 DA025956). M.G. acknowledges support from the National Human Genome Research Institute (grant 5U24HG009446-02). W.D.F. acknowledges support from the National Institute of General Medical Sciences (grant GM122030-01). M.K. acknowledges support from the National Institute of Mental Health (grant R01 MH109978) and the National Human Genome Research Institute (grants U01 HG007610 and R01 HG008155).

REFERENCES AND NOTES

1. Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, Hod E, Li K, Murty VV, Schupf N, Vilain E, Morris M, Haghighi F, Tycko B, Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. Nat. Genet 40, 904–908 (2008). doi: 10.1038/ng.174 Medline [PubMed: 18568024]
2. Schalkwyk LC, Meaburn EL, Smith R, Dempster EL, Jeffries AR, Davies MN, Plomin R, Mill J, Allelic skewing of DNA methylation is widespread across the genome. Am. J. Hum. Genet 86, 196–212 (2010). doi:10.1016/j.ajhg.2010.01.014 Medline [PubMed: 20159110]
3. Zhang Y, Rohde C, Reinhardt R, Voelcker-Rehage C, Jeltsch A, Non-imprinted allele-specific DNA methylation on human autosomes. Genome Biol. 10, R138 (2009). doi:10.1186/gb-2009-10-12-r138 [PubMed: 19958531]
4. Gertz J, Varley KE, Reddy TE, Bowling KM, Pauli F, Parker SL, Kucera KS, Willard HF, Myers RM, Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. PLOS Genet. 7, e1002228 (2011). doi:10.1371/journal.pgen.1002228 [PubMed: 21852959]
5. Hellman A, Chess A, Extensive sequence-influenced DNA methylation polymorphism in the human genome. Epigenetics Chromatin 3, 11 (2010). doi:10.1186/1756-8935-3-11 [PubMed: 20497546]

6. Bell CG, Finer S, Lindgren CM, Wilson GA, Rakyan VK, Teschendorff AE, Akan P, Stupka E, Down TA, Prokopenko I, Morison IM, Mill J, Pidsley R, Deloukas P, Frayling TM, Hattersley AT, McCarthy MI, Beck S, Hitman GA; International Type 2 Diabetes 1q Consortium, Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. PLOS ONE 5, e14040 (2010). doi:10.1371/journal.pone.0014040 [PubMed: 21124985]
7. Tycko B, Allele-specific DNA methylation: Beyond imprinting. Hum. Mol. Genet 19 (R2), R210–R220 (2010). doi:10.1093/hmg/ddq376 [PubMed: 20855472]
8. Waszak SM, Delaneau O, Gschwind AR, Kilpinen H, Raghav SK, Witwicki RM, Orioli A, Wiederkehr M, Panousis NI, Yurovsky A, Romano-Palumbo L, Planchon A, Bielser D, Padioleau I, Udin G, Thurnheer S, Hacker D, Hernandez N, Reymond A, Deplancke B, Dermitzakis ET, Population variation and genetic control of modular chromatin architecture in humans. Cell 162, 1039–1050 (2015). doi:10.1016/j.cell.2015.08.001 [PubMed: 26300124]
9. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, Lewellen N, Myrthil M, Gilad Y, Pritchard JK, Identification of genetic variants that affect histone modifications in human cells. Science 342, 747–749 (2013). doi:10.1126/science.1242429 [PubMed: 24136359]
10. Sun W, Poschmann J, Cruz-Herrera Del Rosario R, Parikshak NN, Hajan HS, Kumar V, Ramasamy R, Belgard TG, Elanggovan B, Wong CCY, Mill J, Geschwind DH, Prabhakar S, Histone acetylome-wide association study of autism spectrum disorder. Cell 167, 1385–1397.e11 (2016). doi:10.1016/j.cell.2016.10.031 [PubMed: 27863250]
11. Lyon MF, Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature 190, 372–373 (1961). doi:10.1038/190372a0 [PubMed: 13764598]
12. Shipony Z, Mukamel Z, Cohen NM, Landan G, Chomsky E, Zeliger SR, Fried YC, Ainbinder E, Friedman N, Tanay A, Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. Nature 513, 115–119 (2014). doi:10.1038/nature13458 [PubMed: 25043040]
13. Landan G, Cohen NM, Mukamel Z, Bar A, Molchadsky A, Brosh R, Horn-Saban S, Zalcenstein DA, Goldfinger N, Zundelevich A, Gal-Yam EN, Rotter V, Tanay A, Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nat. Genet 44, 1207–1214 (2012). doi:10.1038/ng.2442 [PubMed: 23064413]
14. Florio E, Keller S, Coretti L, Affinito O, Scala G, Errico F, Fico A, Boscia F, Sisalli MJ, Reccia MG, Miele G, Monticelli A, Scorziello A, Lembo F, Colucci-D’Amato L, Minchiotti G, Avvedimento VE, Usiello A, Cocozza S, Chiariotti L, Tracking the evolution of epialleles during neural differentiation and brain development: D-Aspartate oxidase as a model gene. Epigenetics 12, 41–54 (2017). doi:10.1080/15592294.2016.1260211 [PubMed: 27858532]
15. Ginart P, Kalish JM, Jiang CL, Yu AC, Bartolomei MS, Raj A, Visualizing allele-specific expression in single cells reveals epigenetic mosaicism in an H19 loss-of-imprinting mutant. Genes Dev. 30, 567–578 (2016). doi:10.1101/gad.275958.115 [PubMed: 26944681]
16. Siegmund KD, Marjoram P, Woo YJ, Tavaré S, Shibata D, Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. Proc. Natl. Acad. Sci. U.S.A 106, 4828–4833 (2009). doi:10.1073/pnas.0810276106 [PubMed: 19261858]
17. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein C, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K-H, Feizi S, Karlic R, Kim A-R, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai L-H, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M; Roadmap Epigenomics Consortium, Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015). doi:10.1038/nature14248 [PubMed: 25693563]
18. Materials and methods are available as supplementary materials.

19. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, Bhardwaj N, Rubin M, Snyder M, Gerstein M, AlleleSeq: Analysis of allele-specific expression and binding in a network framework. Mol. Syst. Biol 7, 522 (2011). doi: 10.1038/msb.2011.54 [PubMed: 21811232]

20. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, Lin S, Lin Y, Jung I, Schmitt AD, Selvaraj S, Ren B, Sejnowski TJ, Wang W, Ecker JR, Human body epigenome maps reveal noncanonical DNA methylation variation. Nature 523, 212–216 (2015). doi:10.1038/nature14465 [PubMed: 26030523]

21. Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, Yen C-A, Lin S, Lin Y, Qiu Y, Xie W, Yue F, Hariharan M, Ray P, Kuan S, Edsall L, Yang H, Chi NC, Zhang MQ, Ecker JR, Ren B, Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature 518, 350–354 (2015). doi:10.1038/nature14217 [PubMed: 25693566]

22. Hutchinson JN, Raj T, Fagerness J, Stahl E, Viloria FT, Gimelbrant A, Seddon J, Daly M, Chess A, Plenge R, Allele-specific methylation occurs at genetic variants associated with complex disease. PLOS ONE 9, e98464 (2014). doi:10.1371/journal.pone.0098464 [PubMed: 24911414]

23. Do C, Lang CF, Lin J, Darbary H, Krupska I, Gaba A, Petukhova L, Vonsattel J-P, Gallagher MP, Goland RS, Clynes RA, Dwork A, Kral JG, Monk C, Christiano AM, Tycko B, Mechanisms and disease associations of haplotype-dependent allele-specific DNA methylation. Am. J. Hum. Genet 98, 934–955 (2016). doi:10.1016/j.ajhg.2016.03.027 [PubMed: 27153397]

24. Bell CG, Gao F, Yuan W, Roos L, Acton RJ, Xia Y, Bell J, Ward K, Mangino M, Hysi PG, Wang J, Spector TD, Obligatory and facilitative allelic variation in the DNA methylome within common disease-associated loci. Nat. Commun 9, 8 (2018). doi:10.1038/s41467-017-01586-1 [PubMed: 29295990]

25. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, Falconnet E, Bielser D, Gagnebin M, Padioleau I, Borel C, Letourneau A, Makrythanasis P, Guipponi M, Gehrig C, Antonarakis SE, Dermitzakis ET, Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife 2, e00523 (2013). Medline [PubMed: 23755361]

26. Gutierrez-Arcelus M, Ongen H, Lappalainen T, Montgomery SB, Buil A, Yurovsky A, Bryois J, Padioleau I, Romano L, Planchon A, Falconnet E, Bielser D, Gagnebin M, Giger T, Borel C, Letourneau A, Makrythanasis P, Guipponi M, Gehrig C, Antonarakis SE, Dermitzakis ET, Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. PLOS Genet. 11, e1004958 (2015). doi:10.1371/journal.pgen.1004958 [PubMed: 25634236]

27. Hellman A, Chess A, Gene body-specific methylation on the active X chromosome. Science 315, 1141–1143 (2007). doi:10.1126/science.1136352 [PubMed: 17322062]

28. Cheung WA, Shao X, Morin A, Siroux V, Kwan T, Ge B, Aïssi D, Chen L, Vasquez L, Allum F, Guénard F, Bouzigon E, Simon M-M, Boulier E, Redensek A, Watt S, Datta A, Clarke L, Flicek P, Mead D, Paul DS, Beck S, Bourque G, Lathrop M, Tchernof A, Vohl M-C, Demenais F, Pin I, Downes K, Stunnenberg HG, Soranzo N, Pastinen T, Grundberg E, Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome. Genome Biol. 18, 50 (2017). doi:10.1186/s13059-017-1173-7 [PubMed: 28283040]

29. Park SG, Hannenhalli S, Choi SS, Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. BMC Genomics 15, 526 (2014). doi:10.1186/1471-2164-15-526 [PubMed: 24964727]

30. Rakyan VK, Blewitt ME, Druker R, Preis JI, Whitelaw E, Metastable epialleles in mammals. Trends Genet. 18, 348–351 (2002). doi:10.1016/S0168-9525(02)02709-9 [PubMed: 12127774]

31. Martos SN, Li T, Ramos RB, Lou D, Dai H, Xu J-C, Gao G, Gao Y, Wang Q, An C, Zhang X, Jia Y, Dawson VL, Dawson TM, Ji H, Wang Z, Two approaches reveal a new paradigm of ‘switchable or genetics-influenced allele-specific DNA methylation’ with potential in human disease. Cell Discov. 3, 17038 (2017). doi:10.1038/celldisc.2017.38 [PubMed: 29387450]

32. Davila-Velderrain J, Martinez-Garcia JC, Alvarez-Buylla ER, Modeling the epigenetic attractors landscape: Toward a post-genomic mechanistic understanding of development. Front. Genet 6, 160 (2015). doi:10.3389/fgene.2015.00160 [PubMed: 25954305]

33. Maurano MT, Wang H, John S, Shafer A, Canfield T, Lee K, Stamatoyannopoulos JA, Role of DNA methylation in modulating transcription factor occupancy. Cell Reports 12, 1184–1195 (2015). doi:10.1016/j.celrep.2015.07.024 [PubMed: 26257180]

34. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y, Lu Y, Wu Y, Jia Z, Li W, Zhang MQ, Ren B, Krainer AR, Maniatis T, Wu Q, CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell 162, 900–910 (2015). doi:10.1016/j.cell.2015.07.038 [PubMed: 26276636]

35. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, Michalski P, Piecuch E, Wang P, Wang D, Tian SZ, Penrad-Mobayed M, Sachs LM, Ruan X, Wei C-L, Liu ET, Wilczynski GM, Plewczynski D, Li G, Ruan Y, CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell 163, 1611–1627 (2015). doi:10.1016/j.cell.2015.11.024 [PubMed: 26686651]

36. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T, Taipale J, DNA-binding specificities of human transcription factors. Cell 152, 327–339 (2013). doi:10.1016/j.cell.2012.12.009 [PubMed: 23332764]

37. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee B-K, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA, The accessible chromatin landscape of the human genome. Nature 489, 75–82 (2012). doi:10.1038/nature11232 [PubMed: 22955617]

38. Feldmann A, Ivanek R, Murr R, Gaidatzis D, Burger L, Schübeler D, Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. PLOS Genet. 9, e1003994 (2013). doi:10.1371/journal.pgen.1003994 [PubMed: 24367273]

39. Hervouet E, Vallette FM, Cartron PF, Dnmt1/Transcription factor interactions: An alternative mechanism of DNA methylation inheritance. Genes Cancer 1, 434–443 (2010). doi:10.1177/1947601910373794 [PubMed: 21779454]

40. Hervouet E, Vallette FM, Cartron PF, Dnmt3/transcription factor interactions as crucial players in targeted DNA methylation. Epigenetics 4, 487–499 (2009). doi:10.4161/epi.4.7.9883 [PubMed: 19786833]

41. Nayak A, Glöckner-Pagel J, Vaeth M, Schumann JE, Buttmann M, Bopp T, Schmitt E, Serfling E, Berberich-Siebelt F, Sumoylation of the transcription factor NFATc1 leads to its subnuclear relocalization and interleukin-2 repression by histone deacetylase. J. Biol. Chem 284, 10935–10946 (2009). doi:10.1074/jbc.M900465200 [PubMed: 19218564]

42. Zeng H, Gifford DK, Predicting the impact of non-coding variants on DNA methylation. Nucleic Acids Res. 45, e99 (2017). doi:10.1093/nar/gkx177 [PubMed: 28334830]

43. Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, Blischak JD, Roux J, Pritchard JK, Gilad Y, Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLOS Genet. 10, e1004663 (2014). doi:10.1371/journal.pgen.1004663 [PubMed: 25233095]

44. Kin K, Chen X, Gonzalez-Garay M, Fakhouri WD, The effect of non-coding DNA variations on P53 and cMYC competitive inhibition at cis-overlapping motifs. Hum. Mol. Genet 25, 1517–1527 (2016). doi:10.1093/hmg/ddw030 [PubMed: 26908612]

45. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H, The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42 (D1), D1001–D1006 (2014). doi:10.1093/nar/gkt1229 [PubMed: 24316577]

46. De Silva DR, Nichols R, Elgar G, Purifying selection in deeply conserved human enhancers is more consistent than in coding sequences. PLOS ONE 9, e103357 (2014). doi:10.1371/journal.pone.0103357 [PubMed: 25062004]

47. Ward LD, Kellis M, Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science 337, 1675–1678 (2012). doi:10.1126/science.1225057 [PubMed: 22956687]

48. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR; 1000 Genomes Project Consortium, A global reference for human genetic variation. Nature 526, 68–74 (2015). doi:10.1038/nature15393 [PubMed: 26432245]

49. Zhu J, He F, Hu S, Yu J, On the nature of human housekeeping genes. Trends Genet. 24, 481–484 (2008). doi:10.1016/j.tig.2008.08.004 [PubMed: 18786740]

50. Maurano MT, Haugen E, Sandstrom R, Vierstra J, Shafer A, Kaul R, Stamatoyannopoulos JA, Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. Nat. Genet 47, 1393–1401 (2015). doi:10.1038/ng.3432 [PubMed: 26502339]

51. Gravina S, Dong X, Yu B, Vijg J, Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome. Genome Biol. 17, 150 (2016). doi:10.1186/s13059-016-1011-3 [PubMed: 27380908]

52. Elliott G, Hong C, Xing X, Zhou X, Li D, Coarfa C, Bell RJA, Maire CL, Ligon KL, Sigaroudinia M, Gascard P, Tlsty TD, Harris RA, Schalkwyk LC, Bilenky M, Mill J, Farnham PJ, Kellis M, Marra MA, Milosavljevic A, Hirst M, Stormo GD, Wang T, Costello JF, Intermediate DNA methylation is a conserved signature of genome regulation. Nat. Commun 6, 6363 (2015). doi:10.1038/ncomms7363 [PubMed: 25691127]

53. Jenkinson G, Pujadas E, Goutsias J, Feinberg AP, Potential energy landscapes identify the information-theoretic nature of the epigenome. Nat. Genet 49, 719–729 (2017). doi:10.1038/ng.3811 [PubMed: 28346445]

54. McAdams HH, Arkin A, It's a noisy business! Genetic regulation at the nanomolar scale. Trends Genet. 15, 65–69 (1999). doi:10.1016/S0168-9525(98)01659-X [PubMed: 10098409]

55. Alaghi A, Hayes JP, Survey of stochastic computing. ACM Trans. Embed. Comput. Syst 12, 1 (2013). doi:10.1145/2465787.2465794

56. Li H, Durbin R, Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009). doi:10.1093/bioinformatics/btp324 [PubMed: 19451168]

57. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA, From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43, 1–33 (2013). Medline [PubMed: 26270170]

58. Huang Z, Rustagi N, Veeraraghavan N, Carroll A, Gibbs R, Boerwinkle E, Venkata MG, Yu F, A hybrid computational strategy to address WGS variant analysis in >5000 samples. BMC Bioinformatics 17, 361 (2016). doi:10.1186/s12859-016-1211-6 [PubMed: 27612449]

59. Challis D, Yu J, Evani US, Jackson AR, Paithankar S, Coarfa C, Milosavljevic A, Gibbs RA, Yu F, An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics 13, 8 (2012). doi:10.1186/1471-2105-13-8 [PubMed: 22239737]

60. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ, A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet 43, 491–498 (2011). doi:10.1038/ng.806 [PubMed: 21478889]

61. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, McVean G, Lunter G; WGS500 Consortium, Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet 46, 912–918 (2014). doi:10.1038/ng.3036 [PubMed: 25017105]

62. Wang Y, Lu J, Yu J, Gibbs RA, Yu F, An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. Genome Res. 23, 833–842 (2013). doi:10.1101/gr.146084.112 [PubMed: 23296920]

63. Abyzov A, Urban AE, Snyder M, Gerstein M, CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 21, 974–984 (2011). doi:10.1101/gr.114876.110 [PubMed: 21324876]

64. Coarfa C, Yu F, Miller CA, Chen Z, Harris RA, Milosavljevic A, Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing. BMC Bioinformatics 11, 572 (2010). doi: 10.1186/1471-2105-11-572 [PubMed: 21092284]

65. Liu Y, Siegmund KD, Laird PW, Berman BP, Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. Genome Biol. 13, R61 (2012). doi:10.1186/gb-2012-13-7-r61 [PubMed: 22784381]

66. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). doi:10.1038/nature11247 [PubMed: 22955616]

67. Fang F, Hodges E, Molaro A, Dean M, Hannon GJ, Smith AD, Genomic landscape of human allele-specific DNA methylation. Proc. Natl. Acad. Sci. U.S.A 109, 7332–7337 (2012). doi: 10.1073/pnas.1201310109 [PubMed: 22523239]

68. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup, The sequence alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009). doi:10.1093/bioinformatics/btp352 [PubMed: 19505943]

69. Kuhn RM, Haussler D, Kent WJ, The UCSC genome browser and associated tools. Brief. Bioinform 14, 144–161 (2013). doi:10.1093/bib/bbs038 [PubMed: 22908213]

70. Guo W, Zhu P, Pellegrini M, Zhang MQ, Wang X, Ni Z, CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. Bioinformatics 34, 381–387 (2018). doi:10.1093/bioinformatics/btx595 [PubMed: 28968643]

71. Coetzee SG, Coetzee GA, Hazelett DJ, motifbreakR: An R/Bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics 31, 3847–3849 (2015). Medline [PubMed: 26272984]

72. Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, Iriarte B, Meng Y, Palmer CD, Esko T, Winckler W, Hirschhorn JN, Kellis M, MacArthur DG, Getz G, Shabalin AA, Li G, Zhou Y-H, Nobel AB, Rusyn I, Wright FA, Lappalainen T, Ferreira PG, Ongen H, Rivas MA, Battle A, Mostafavi S, Monlong J, Sammeth M, Mele M, Reverter F, Goldmann JM, Koller D, Guigo R, McCarthy MI, Dermitzakis ET, Gamazon ER, Im HK, Konkashbaev A, Nicolae DL, Cox NJ, Flutre T, Wen X, Stephens M, Pritchard JK, Tu Z, Zhang B, Huang T, Long Q, Lin L, Yang J, Zhu J, Liu J, Brown A, Mestichelli B, Tidwell D, Lo E, Salvatore M, Shad S, Thomas JA, Lonsdale JT, Moser MT, Gillard BM, Karasik E, Ramsey K, Choi C, Foster BA, Syron J, Fleming J, Magazine H, Hasz R, Walters GD, Bridge JP, Miklos M, Sullivan S, Barker LK, Traino HM, Mosavel M, Siminoff LA, Valley DR, Rohrer DC, Jewell SD, Branton PA, Sobin LH, Barcus M, Qi L, McLean J, Hariharan P, Um KS, Wu S, Tabor D, Shive C, Smith AM, Buia SA, Undale AH, Robinson KL, Roche N, Valentino KM, Britton A, Burges R, Bradbury D, Hambright KW, Seleski J, Korzeniewski GE, Erickson K, Marcus Y, Tejada J, Taherian M, Lu C, Basile M, Mash DC, Volpi S, Struewing JP, Temple GF, Boyer J, Colantuoni D, Little R, Koester S, Carithers LJ, Moore HM, Guan P, Compton C, Sawyer SJ, Demchok JP, Vaught JB, Rabiner CA, Lockhart NC, Ardlie KG, Getz G, Wright FA, Kellis M, Volpi S, Dermitzakis ET; GTEx Consortium, Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348, 648–660 (2015). doi:10.1126/science.1262110 [PubMed: 25954001]

73. Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R, Zhang AW, Parcy F, Lenhard B, Sandelin A, Wasserman WW, JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 44 (D1), D110–D115 (2016). doi:10.1093/nar/gkv1176 [PubMed: 26531826]

74. Fakhouri WD, Rahimov F, Attanasio C, Kouwenhoven EN, Ferreira De Lima RL, Felix TM, Nitschke L, Huver D, Barrons J, Kousa YA, Leslie E, Pennacchio LA, Van Bokhoven H, Visel A, Zhou H, Murray JC, Schutte BC, An etiologic regulatory mutation in IRF6 with loss- and gain-of-function effects. Hum. Mol. Genet 23, 2711–2720 (2014). doi:10.1093/hmg/ddt664 [PubMed: 24442519]

75. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H, Modeling sample variables with an experimental factor ontology. Bioinformatics 26, 1112–1118 (2010). doi:10.1093/bioinformatics/btq099 [PubMed: 20200009]

76. Eilbeck K, Lewis SE, Sequence ontology annotation guide. Comp. Funct. Genomics 5, 642–647 (2004). doi:10.1002/cfg.446 [PubMed: 18629179]

77. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA, Uberon, an integrative multi-species anatomy ontology. Genome Biol. 13, R5 (2012). doi:10.1186/gb-2012-13-1-r5 [PubMed: 22293552]

78. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF, Sabeti PC, Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell 165, 1519–1529 (2016). doi:10.1016/j.cell.2016.04.027 [PubMed: 27259153]

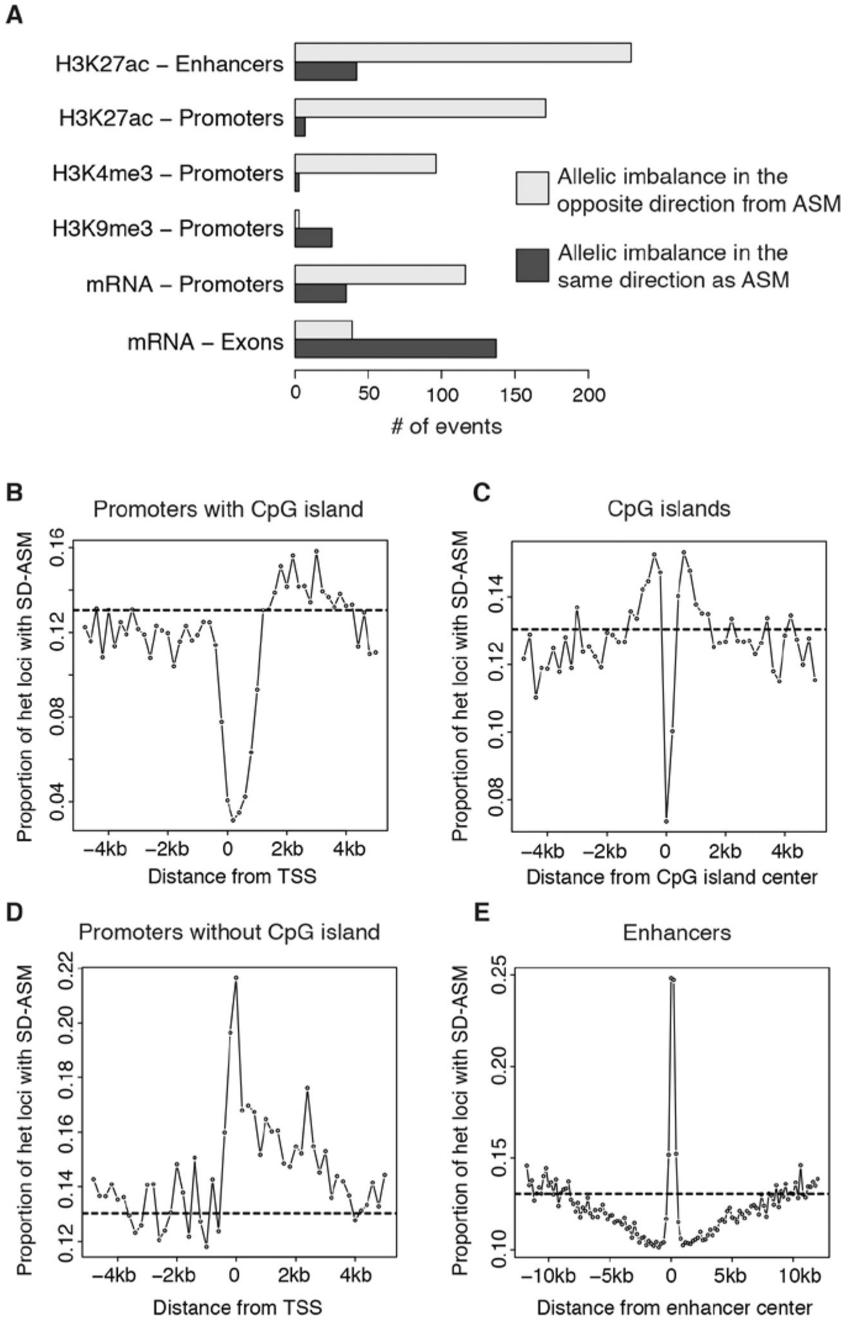

Fig. 1. Allelic imbalances vary depending on genomic region.

(**A**) Number of allelic imbalances in histone marks and transcription, overlapping ASM loci, over classes of genomic elements. (**B** to **E**) Proportions of SD-ASM loci over total heterozygous loci in 200bp bins near promoters, CpG islands, and enhancers.

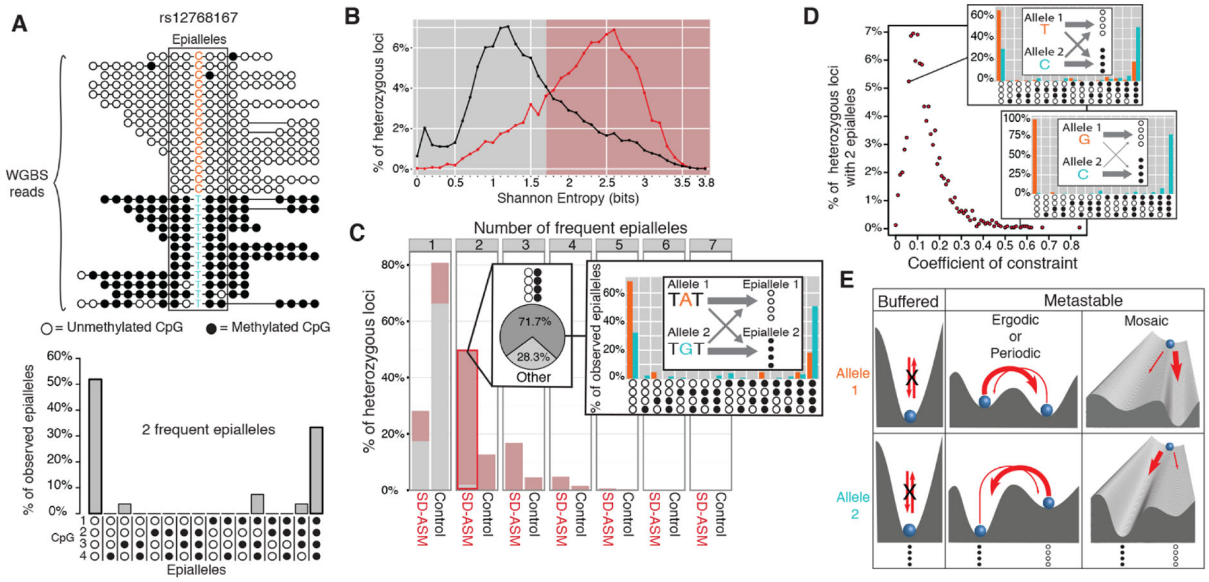

Fig. 2. Differences in epiallele frequency spectra causing SD-ASM.

(**A**) Example of an epiallele frequency spectrum (below) derived from observed epialleles in WGBS reads (top). (**B**) Histograms of Shannon entropy, in bits, for the epiallele frequency spectra for the hets showing SD-ASM (red) and the nearest (control) hets without SD-ASM (black). (**C**) Most heterozygous loci with two frequent epialleles show SD-ASM, have entropy larger than 1.7 bits (red portion of the bar), the two epialleles being biphasic (fully methylated or fully unmethylated) 71.7% of the time. The callout on the right provides an example of a het where the difference between epiallele frequency spectra of allele 1 (A, orange) and allele 2 (G, blue) explain SD-ASM. (**D**). Histogram of Coefficients of Constraint for SD-ASM loci with two frequent epialleles. The callouts illustrate an example het (T/C, top right callout) with a low, and another (G/C, bottom right callout) with a high Coefficient of Constraint. (**E**) Illustration of buffering in contrast to ergodic/periodic and mosaic metastability.

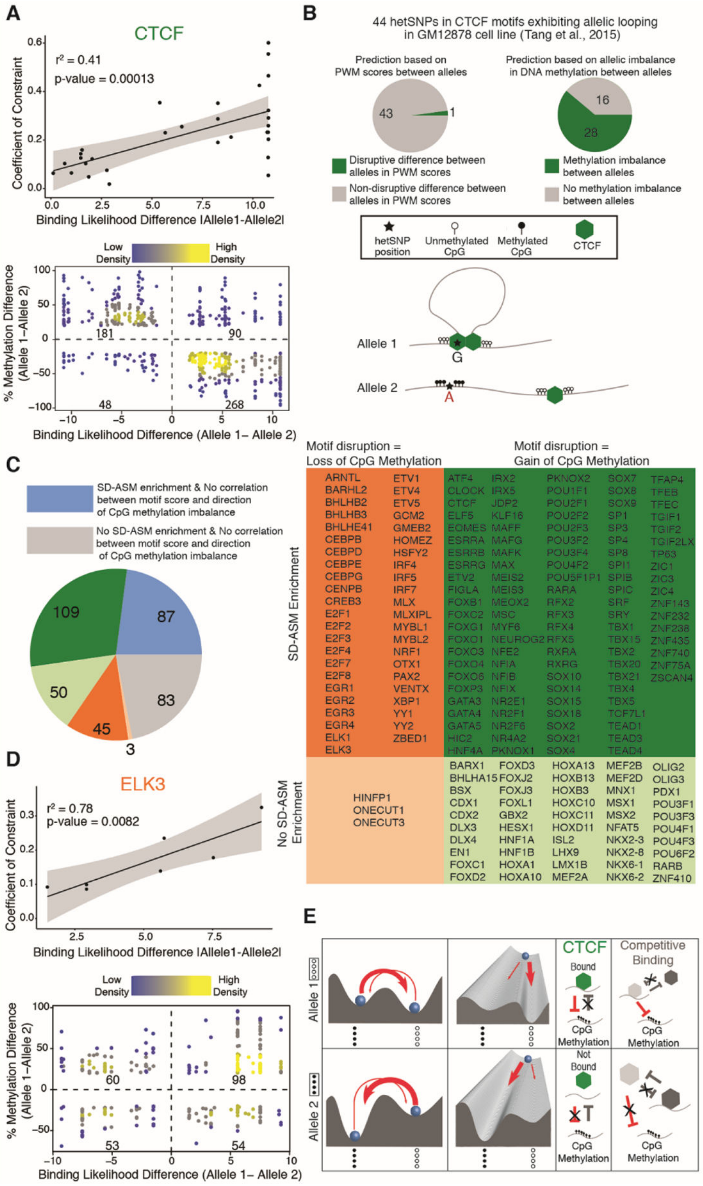

Fig. 3. Correlations between allelic differences in TF binding affinity, Coefficient of Constraint, and DNA methylation.

(**A**) (Top) Correlation between absolute CTCF binding affinity differences, based on position weight matrix scores (PWM), and the Coefficient of Constraint for predicted CTCF binding sites with SD-ASM, two frequent epialleles, and a biphasic methylation pattern. (Bottom) Correlation between CTCF binding affinity and DNA methylation at predicted CTCF binding sites. (**B**) SD-ASM is more predictive of allelic looping (28 true positive of 44 predictions) than motif disruption scores (1 true positive of 44 predictions). To control for specificity, thresholds were selected so that both methods predicted the same number (44) of hets to show allelic looping. (**C**) SD-ASM at binding sites of 377 TFs defined with the SELEX method. The pie chart on the left and the table on the right indicate both enrichments and directionality trends using a shared color code. (**D**) Top: Correlation between absolute ELK3 binding affinity differences and the Coefficient of Constraint for predicted binding sites with SD-ASM, two frequent epialleles, and a biphasic methylation pattern; Bottom: Correlation between ELK3 binding affinity and DNA methylation at predicted ELK3 binding sites (**E**) A mechanistic model of a sequence-dependent energy landscape with two metastable states, Allele 1 (top row) corresponding to a landscape where the most frequently occupied metastable state corresponds to a completely unmethylated epiallele and Allele 2 (bottom row) corresponding to a landscape where the most frequently occupied metastable state corresponds to a completely methylated epiallele. Putative positive feedback loops involving interactions between TF binding and binding site methylation are indicated for CTCF. An alternative model involving competitive binding of two transcription factors is indicated on the right. Significance of correlations tested using *t* test.

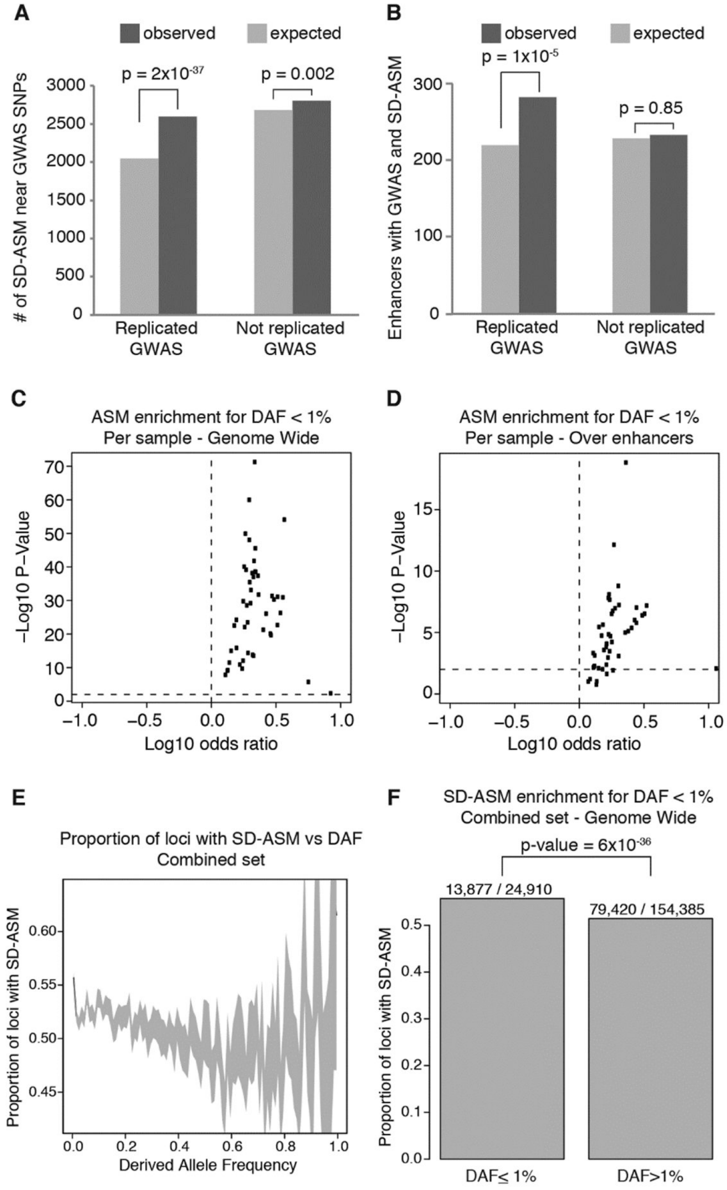

Fig. 4. Association of ASM with disease loci and purifying selection.

(**A** and **B**) Enrichment of ASM in the proximity of GWAS loci. ASM hets within 1 kilobase (Kb) of GWAS loci are compared to co-localized hets without ASM. (**C** to **F**) Evidence of purifying selection acting on rare variants with ASM. **[**(C) and (D)] Proportion of variants associated with ASM compared to those without ASM among the rare (DAF < 1%) variants across individual methylomes. [(E) and (F)] Proportion of loci with ASM over total heterozygous loci over windows of increasing DAF in the combined set of methylomes. (F) This bar-chart summary of the data in (E) shows the excess of SD-ASM variants among those with DAF < 1%. Chi-square tests used for significance of enrichments.