Do *et al. Genome Biology* (2020) 21:153
https://doi.org/10.1186/s13059-020-02059-3

Genome Biology

RESEARCH Open Access

Allele-specific DNA methylation is increased in cancers and its dense mapping in normal plus neoplastic cells increases the yield of disease-associated regulatory SNPs

Catherine Do[1,2*], Emmanuel L. P. Dumont[1,2], Martha Salas[1,2], Angelica Castano[1,2], Huthayfa Mujahed[3], Leonel Maldonado[4], Arunjot Singh[5], Sonia C. DaSilva-Arnold[6], Govind Bhagat[7,8], Soren Lehman[3], Angela M. Christiano[9], Subha Madhavan[10], Peter L. Nagy[11], Peter H. R. Green[8], Rena Feinman[1,2,10], Cornelia Trimble[4], Nicholas P. Illsley[6], Karen Marder[12,13], Lawrence Honig[12,13], Catherine Monk[14], Andre Goy[1,2,10], Kar Chow[1,2,10], Samuel Goldlust[1,2], George Kaptain[1,2], David Siegel[1,2,10] and Benjamin Tycko[1,2,10*]

* Correspondence: catherine.do@hmh-cdi.org; benjamin.tycko@hmh-cdi.org
[1] Hackensack-Meridian Health Center for Discovery and Innovation, Nutley, NJ 07110, USA
Full list of author information is available at the end of the article

Abstract

Background: Mapping of allele-specific DNA methylation (ASM) can be a post-GWAS strategy for localizing regulatory sequence polymorphisms (rSNPs). The advantages of this approach, and the mechanisms underlying ASM in normal and neoplastic cells, remain to be clarified.

Results: We perform whole genome methyl-seq on diverse normal cells and tissues and three cancer types. After excluding imprinting, the data pinpoint 15,112 high-confidence ASM differentially methylated regions, of which 1838 contain SNPs in strong linkage disequilibrium or coinciding with GWAS peaks. ASM frequencies are increased in cancers versus matched normal tissues, due to widespread allele-specific hypomethylation and focal allele-specific hypermethylation in poised chromatin. Cancer cells show increased allele switching at ASM loci, but disruptive SNPs in specific classes of CTCF and transcription factor binding motifs are similarly correlated with ASM in cancer and non-cancer. Rare somatic mutations affecting these same motif classes track with de novo ASM. Allele-specific transcription factor binding from ChIP-seq is enriched among ASM loci, but most ASM differentially methylated regions lack such annotations, and some are found in otherwise uninformative "chromatin deserts."

Conclusions: ASM is increased in cancers but occurs by a shared mechanism involving disruptive SNPs in CTCF and transcription factor binding sites in both normal and neoplastic cells. Dense ASM mapping in normal plus cancer samples reveals candidate rSNPs that are difficult to find by other approaches. Together with GWAS data, these rSNPs can nominate specific transcriptional pathways in susceptibility to autoimmune, cardiometabolic, neuropsychiatric, and neoplastic diseases.

Background

Genome-wide association studies (GWAS) have implicated numerous DNA sequence variants, mostly single nucleotide polymorphisms (SNPs) in non-coding regions, as candidates for mediating inter-individual differences in disease susceptibility. However, to promote GWAS statistical signals to biological true-positives, and to identify the functional sequence variants that underlie these signals, several obstacles need to be overcome. Multiple statistical comparisons demand stringent thresholds for significance, $p < 5 \times 10^{-8}$ for a GWAS, and this level can lead to the rejection of biological true-positives with subthreshold p values [1]. A more fundamental challenge is identifying the causal regulatory SNPs (rSNPs) among the typically large number of variants that are in linkage disequilibrium (LD) with a GWAS peak SNP. Combined genetic-epigenetic mapping can address these challenges. In particular, identification of non-imprinted allele-specific CpG methylation dictated by cis-acting effects of local genotypes or haplotypes (sometimes abbreviated as hap-ASM but hereafter referred to simply as ASM) led us and others to suggest that mapping this type of allelic asymmetry could prove useful as a “post-GWAS” method for localizing rSNPs [2–12]. The premise is that the presence of an ASM DMR can indicate a bona fide regulatory sequence variant (or regulatory haplotype) in that genomic region, which declares itself by conferring the physical asymmetry (i.e., ASM) between the two alleles in heterozygotes. ASM mapping and related post-GWAS approaches such as allele-specific chromatin immunoprecipitation-sequencing (ChIP-seq) [13, 14] can facilitate genome-wide screening for disease-linked rSNPs, which can then be prioritized for functional studies. However, the unique advantages of ASM mapping, and its potential non-redundancy with other post-GWAS mapping methods, remain to be clarified.

Genome-wide analysis of ASM by methylation sequencing (methyl-seq) is also yielding insights to the fundamental mechanisms that shape DNA methylation patterns. Our previous data using bisulfite sequence capture (Agilent SureSelect) revealed ASM DMRs and methylation quantitative trait loci (mQTLs) in human brain cells and tissues, and in T lymphocytes, and uncovered a role for polymorphic CTCF and transcription factor (TF) binding sites in producing ASM [8]. Others have pursued similar approaches with progressively greater genomic coverage [10, 11], with substantial though partial overlap in the resulting lists of ASM DMRs [9], and with consistent conclusions regarding the importance of destructive SNPs in CTCF and TF binding sites as a mechanism underlying ASM. However, since ASM is often tissue-specific and its mapping requires heterozygotes at one or more “index SNPs” in the DMR, constraints from the numbers of individuals and numbers of cell types have limited the harvest of high-confidence ASM DMRs. These factors have in turn limited the assessment of specific classes of TF and CTCF binding sites for their involvement in ASM and limited the yield of candidate rSNPs in disease-associated chromosomal regions. Further, while some studies of cancer samples have been done using targeted methyl-seq [15–18], the genome-wide features and mechanisms of ASM in human neoplasia have yet to be clarified.

To address these issues, we have expanded our previous methyl-seq dataset and carried out whole genome bisulfite sequencing (WGBS) on a new large series of human samples spanning a range of normal tissues and cell types from multiple individuals, plus three types of human cancers—multiple myeloma, B cell lymphoma, and glioblastoma. We identify high-confidence ASM DMRs using stringent criteria, perform extensive validations, apply a multi-step analytical pipeline to compare mechanisms of ASM in normal and cancer cells, and assess the unique strengths of dense ASM mapping for finding mechanistically informative disease-associated rSNPs.

Results

Mapping of high-confidence ASM regions in normal and neoplastic human samples

The biological samples in this study are listed in Additional file 1: Table S1, and our approaches for identifying ASM DMRs, testing mechanisms, and nominating disease-associated rSNPs are diagrammed in Additional file 2: Figure S1A, B. The sample set included diverse tissues and purified cell types from multiple individuals, with an emphasis on immune system cells, brain, carcinoma precursor lineages, and several other normal tissues and cell types, plus a set of primary cancers including multiple myeloma, B cell lymphoma, and glioblastoma multiforme (GBM) (Additional file 1: Table S1 and Additional file 2: Figure S2). Since placental trophoblast has epigenetic and biological similarities to cancers [19–23], we also analyzed unfractionated placental tissue (chorionic plate) and purified placental trophoblast. Agilent SureSelect methyl-seq is a sequence capture-based method for genome-wide bisulfite sequencing that covers 3.7 million CpGs, located in all RefSeq genes and concentrated in promoter regions, CpG islands, CpG island shores, shelves, and DNAse I hypersensitive sites. We previously applied this method to 13 human samples [8], and for the current study, we added samples so that the final SureSelect series includes 24 samples of normal tissues and purified cell types, plus one lymphoblastoid cell line (LCL; GM12878). All samples were from different individuals, except for a trio among the brain samples consisting of one frontal cortex (Brodmann area BA9) and two temporal cortex samples (BA37 and BA38) from the same autopsy brain (Additional file 1: Table S1).

To further increase the number of samples and cell types, to obtain complete genomic coverage, and to include cancer samples, we performed WGBS on 81 human samples. As listed in Additional file 1: Table S1 and Additional file 2: Figure S2, the non-cancer tissues and cell types included a set of immune system cells (T cells, B cells, monocyte/macrophages, whole blood and whole reactive lymph node) from multiple individuals, whole and fractionated samples including purified villous cytotrophoblast and extravillous trophoblast from a term placenta, several normal liver samples, primary bladder and mammary epithelial cells from multiple individuals, and whole and fractionated (NeuN-positive neurons and NeuN-negative glia) samples from cerebral cortex of multiple autopsy cases, plus the GM12878 LCL. The WGBS series included 16 primary human cancers, comprising 3 B cell lymphomas, 7 multiple myeloma cases (CD138+ cells from bone marrow aspirates), and 6 cases of GBM. While the two series were mostly distinct, 5 of the non-cancer samples were analyzed by both SureSelect and WGBS (Additional file 1: Table S1).

Numbers of mapped reads and depth of sequencing are in Additional file 1: Table S1, and numbers of informative (heterozygous) SNPs are in Additional file 2: Figure S2. As a quality control, we performed Principle Component Analysis (PCA) using net methylation values for CpGs informative in both SureSelect and WGBS. This procedure revealed the expected segregation of samples according to cell and tissue type and cancer vs non-cancer status. It also revealed some expected findings for cell lineages, for example highlighting both similarities and differences in methylation patterns in the brain cells (whole cerebral cortex, glia, neurons) and the GBMs (Additional file 2: Figure S3A). As another aspect of the quality control, when the same biological samples were analyzed on SureSelect and WGBS, the corresponding data points clustered closely together by PCA (e.g., the GM12878 LCL; Additional file 2: Figure S3A).

Our analytical pipeline (Additional file 2: Figure S1A, B) includes steps to identify and rank ASM DMRs for strength and confidence and utilize the resulting maps, together with public ENCODE and related data, for investigating mechanisms. For ASM calling, we separated the SureSelect and WGBS reads by alleles using SNPs that were not destroyed by the bisulfite conversion, and defined ASM DMRs by at least 3 CpGs with significant allelic asymmetry in fractional methylation (Fisher's exact test $p < 0.05$). We further required at least 2 contiguous CpGs with ASM, an absolute difference in fractional methylation of $\geq$ 20% between alleles after averaging over all covered CpGs in the DMR, and an overall difference in fractional methylation between alleles passing a Benjamini-Hochberg (B-H) corrected Wilcoxon p value (false discovery rate, FDR) < .05 (Additional file 2: Figure S1A, B).

Using these cut-offs, we found a good yield of recurrent ASM regions (Additional file 2: Figure S3B), but also many more loci with ASM seen in only one sample (Additional file 2: Figure S4A). We utilized such rare or "private" ASM loci for analyzing per-sample ASM frequencies, but for our downstream analyses focused on mechanisms and disease associations, we required ASM in at least two samples. Using these stringent criteria, in the combined SureSelect and WGBS dataset, after removing known imprinted loci (see below), we found 15,112 recurrent ASM DMRs, tagged by 17,931 index SNPs, representing 0.7% of all informative SNP-containing regions with adequate sequence coverage. These data are tabulated using the ASM index SNPs as unique identifiers, and annotated for strength of allelic methylation differences, presence or absence of ASM for each of the various types of samples, chromatin states, TF binding motifs, LD of the ASM index SNPs with GWAS peak SNPs, and other relevant parameters, in Additional file 3: Table S2, with parameter definitions in Additional file 4: Table S3.

ASM in imprinted chromosomal regions

While this study focuses mainly on non-imprinted (genotype- or haplotype-dependent) ASM, genomic imprinting also produces ASM, due to parent-of-origin dependent DNA methylation affecting a small number of imprinted chromosomal domains (~ 150 genes). Therefore, we used the GeneImprint database ([24]; see "Availability of data and materials" section) and manual annotations from the literature to flag imprinted gene regions, many of which showed ASM in the SureSelect and WGBS data, thus serving as positive internal controls for ASM detection (Additional file 5: Table S4). Since a hallmark of parent-of-origin dependent ASM (i.e., imprinting) is 50/50 allele switching between individuals, to test for possible novel imprinted loci, we assessed allele switching frequencies for all loci that showed ASM in non-cancer samples from 10 or more different individuals, after excluding known imprinted regions ("Materials and methods" section). The number of ASM DMRs decreases steeply when they are required to be found in many individuals (Additional file 2: Figure S4) because identifying such loci requires both a high number of informative individuals and highly recurrent

ASM. Accordingly, among the non-cancer samples, 324 ASM DMRs (corresponding to 367 index SNPs) outside of known imprinted regions were identified as showing significant ASM in more than 10 individuals. Only 11/324 (3%) of this group of DMRs showed allele switching at a frequency of greater than or equal to 20% of individuals. In comparison, among ASM DMRs identified in our dataset and located in or near known imprinted genes, a large majority (26/29; 90%) showed high-frequency allele switching, with an approximately 50:50 ratio, as expected for parental imprinting. These results show that WGBS is a robust method for detecting imprinted chromosomal regions, and at the same time indicate that most of the ASM loci identified by this genome-wide approach reflect non-imprinted ASM. Interestingly, even among the small number of highly recurrent ASM loci with frequent allele switching in normal cells and tissues and located outside of validated imprinted domains, some have been reported as imprinted in humans with inconsistent findings or variability (e.g., *IGF2R*, *IGF1R*). This small group of loci (Additional file 6: Table S5) are not pursued further here but will be of interest for future testing of parent-of-origin dependent behavior.

Validations by cross-platform comparisons and targeted methyl-seq

Consistency in the methylation profiles of genomic regions covered by both SureSelect and WGBS is shown in Additional file 2: Figure S3 series-wide and in Additional file 2: Figure S5 for single DNA samples analyzed by both methods. In addition, tracks of net methylation comparing both methods in the same biological sample revealed similar patterns in regions that were covered by both methods (Additional file 2: Figure S6). In the overall series, within the fraction of the genome that was adequately covered by both methods and contained informative SNPs, we found 2005 (49.1%) shared ASM "hits." This substantial but partial overlap is expected, given that most ASM loci show a significant allelic methylation bias in some but not all individuals (Additional file 3: Table S2). In addition, some ASM DMRs passed our stringent criteria in SureSelect but not in WGBS due to the greater sequencing depth of SureSelect in some regions. The pairwise correlation of allelic methylation difference between the 5 samples assessed both by SureSelect and WGBS showed that a majority (80%) of the 1203 "discordant" but adequately covered ASM SNPs were suggestive but sub-threshold, showing either less than 3 CpGs passing ASM criteria or a sub-threshold *p* value due to lack of depth and spatial coverage, or an allelic methylation difference in the same direction with magnitude > 10% but less than 20%.

To assess the true-positive rate for ASM calling, we selected 27 ASM DMRs, distributed through the range of ASM strength and confidence scores, for targeted bisulfite sequencing (bis-seq). As summarized in Additional file 7: Table S6, this validation procedure confirmed the presence of ASM in two or more independent biological samples, outside of those utilized for the genome-wide series, with no discordance in the observed direction of the allelic methylation bias between the genome-wide methylation sequencing data and the targeted bis-seq, in 22 of the 27 DMRs assayed (examples in Additional file 2: Figure S7-S10). For 4 of the 5 remaining loci, the presence of ASM was confirmed by targeted bis-seq using DNA from a sample (index case) that had shown ASM in the primary SureSelect or WGBS series. The single non-validated ASM DMR had a weak overall rank, but other examples in the lower tertile of ranks were validated (Additional file 7: Table S6). Since our calculation of overall rank incorporates both the ASM strength and the percentage of heterozygotes showing ASM, rare individuals can show strong ASM even in loci with weak overall ranks (Additional file 2: Figure S10). The high overall validation rate, 81% using biological samples outside the primary genome-wide series, and 96% including independent technical validations on index samples from the primary series, indicates a high true-positive rate of the genome-wide data. Nonetheless, although the current dataset provides dense and reliable maps of ASM, it is still non-saturating; with inclusion of more individuals and greater sequencing depth, more ASM DMRs will be identified—particularly those tagged by rare SNPs or having a narrow cell type-specificity or low ASM magnitude.

ASM is increased in cancers due to widespread allele-specific CpG hypomethylation and focal allele-specific CpG hypermethylation in regions of poised chromatin

As shown in Fig. 1a, when the numbers of ASM index SNPs per sample based on WGBS were normalized to the numbers of informative SNPs and the samples classified by normal vs cancer status, the number of ASM DMRs in the cancers overall was on average 5-fold greater than in the overall group of non-neoplastic samples (Wilcoxon $p = 1.7 \times 10^{-7}$). The differences in frequency of ASM between cancer and non-cancer were greater when compared using cell lineage-matched samples, non-neoplastic B cells for comparing to the B cell lymphomas and multiple myelomas, and non-neoplastic glial cells for comparing to the GBMs. Compared to these lineage-matched normal cell types, the average fold increase in ASM was 5-fold for multiple myeloma, 8.5-fold for the B cell lymphomas, and 9-fold for the GBMs.

Since placental trophoblast is unique in having a cancer-like epigenomic profile [21–23], we also compared the per-sample ASM frequencies in cancers vs all normal somatic cells and tissues, excluding placenta and trophoblast, which revealed a 7-fold greater frequency of ASM in the cancers (Wilcoxon $p = 1.4 \times 10^{-9}$), The EBV-transformed lymphoblastoid line (GM12878), which we had included to allow a direct reference to ENCODE data, showed a frequency of ASM in the mid-neoplasia range (Fig. 1a), which is important since much existing allele-specific mapping data, including expression and methylation quantitative trait loci (eQTLs, meQTLs) and allele-specific TF and CTCF binding by ChIP-seq (ASB) are from LCLs.

Given the well-known trend toward lower genome-wide ("global") DNA methylation in human neoplasia [25, 26], to valuate mechanisms that could account for the gain of ASM in the cancers, we asked whether there might be an inverse correlation between global methylation levels and frequencies of ASM. Global genomic hypomethylation was found in the GM12878 LCL and in the three types of primary cancers in our series (Fig. 1a and Additional file 2: Figure S10). As expected from prior studies by us and others [8, 21, 23], the placental tissue and purified trophoblast also showed global hypomethylation. Kernel density plots showed diffuse hypomethylation with nearly complete loss of the high methylation peak (fractional methylation > 0.8) in lymphoma and myeloma compared to B cells, and a less dramatic but still significant hypomethylation in the GBMs compared to normal glia (Additional file 2: Figure S11).

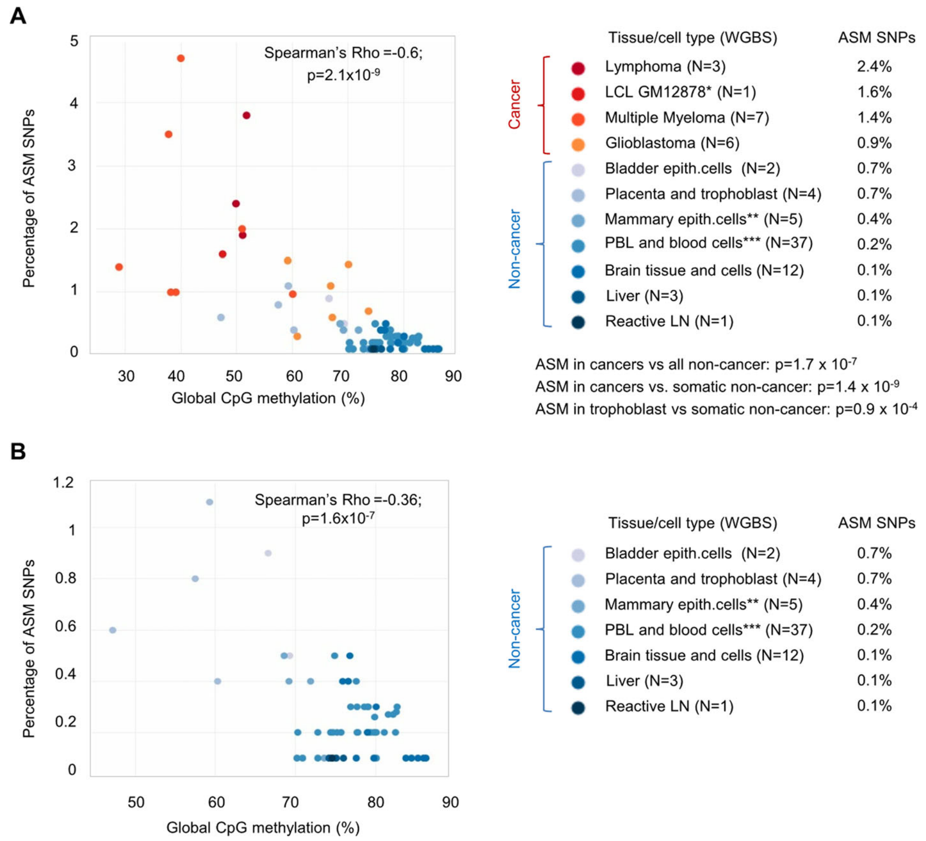

Fig. 1 ASM is increased in cancers and correlates with global DNA hypomethylation. **a** Relationship between global DNA methylation and the percentage of SNPs that reveal ASM in each sample, showing a strong inverse correlation between per sample ASM frequencies and global methylation levels. Cancer samples (color-coded in red scale) have higher ASM frequencies than nearly all the non-cancer samples (color-coded in blue scale), except for partial overlap of the GBM ASM frequencies with those of normal placental trophoblast and cultured bladder epithelial cells. When compared to lineage-matched normal cell types, there is a 5- to 9-fold higher frequency of ASM in the cancers. The EBV-immortalized but euploid GM12878 LCL shows global hypomethylation and a high frequency of ASM. **b** Zoomed-in graph showing an inverse correlation of ASM frequencies with global methylation, even when focusing on the non-cancer samples. *p* values are from Wilcoxon tests. *EBV-immortalized GM12878 LCL is grouped with the cancers. **Mammary epithelial cell lines (*N* = 3) and epithelium-rich normal breast tissue (*N* = 2). ***Includes purified T cells, B cells, and monocytes/macrophages (Additional file 1: Table S1)

Across the entire series of cancer and non-cancer samples, we found a strongly significant anti-correlation (i.e., inverse correlation) between per-sample ASM frequencies and global CpG methylation levels (Spearman's rho = – 0.6; $p = 2.1 \times 10^{-9}$). Importantly from a technical standpoint, this fundamental result was confirmed when we restricted our analysis to the WGBS data from a single sequencing facility using a single library preparation protocol (Additional file 2: Figure S12A). Arguing for global hypomethylation, not the malignant phenotype per se, as a main driving factor for increased ASM, the EBV-immortalized but euploid GM12878 LCL showed global hypomethylation and a high frequency of ASM, and among the non-neoplastic and non-immortalized samples, those that were relatively hypomethylated, namely the placental trophoblast and, to a lesser degree, the bladder epithelial cells that had been expanded in tissue culture, showed relatively higher per-sample frequencies of ASM (Fig. 1b).

To investigate how global hypomethylation could lead to increased ASM in cancers, we assessed the absolute and relative methylation levels of each of the two alleles across instances of ASM in the cancer samples, comparing myelomas and lymphomas to non-neoplastic B cells, and GBMs to normal glial cells. For each comparison, only the ASM-tagging index SNPs that were informative (heterozygous) in both cell types were considered, and we focused on loci showing ASM in the cancers but not in the cell lineage-matched non-neoplastic samples. We assessed the relative methylation levels of the low and high methylated alleles of these instances using a mixed linear model to estimate the average methylation level of each allele in each cell type taking into account the ASM magnitude in each cell type and the difference in ASM magnitude between cell types. As shown in Fig. 2 and Additional file 2: Figure S13, this approach revealed that the average configuration was a relative loss of methylation (LOM) on one allele in the cancers. In 72% of cancer-only ASM occurrences in myelomas, 76% in lymphomas, and 49% in GBMs, a strongly "hypermethylated/hypermethylated" configuration of the two alleles ("black/black") in non-cancer became a "hypomethylated/hypermethylated" ("white-gray/black") configuration in cancer (Fig. 2). The terminology here is a practical shorthand: "LOM" does not mean to imply that the normal cell types evolve into cancers; it simply indicates the direction of the change in comparing the allelic methylation levels in the cancer vs cell lineage-matched non-cancer samples. Similarly, "cancer-only ASM" does not mean to imply that ASM at a given locus will never be detected in any non-cancer sample in future studies; it simply refers to the loci that have ASM in one or more cancer samples and in none of the non-cancer samples in the current dataset.

While the inverse correlation between per-sample ASM frequencies and global methylation is unequivocal, a multivariate regression analysis suggested that additional mechanisms might also be at play. This analysis showed that the anti-correlation between global methylation and per-sample ASM frequencies is partly independent of neoplastic status ($p = 9.4 \times 10^{-5}$ after controlling for neoplastic status), and conversely, that the higher ASM frequencies in the cancers are only partly explained by global methylation levels ($p = 2.5 \times 10^{-4}$ after controlling for methylation levels). In fact, while most of the cancer-only ASM loci conformed to the allele-specific LOM model, we found smaller but still substantial sets of loci (16% to 25% in the three cancer types) in which ASM in the cancers reflected allele-specific gains of methylation (GOM), relative to a biallelic low methylation configuration of the same regions in the lineage-paired normal samples (Fig. 3 and Additional file 2: Figure S14).

To further characterize this interesting set of loci with allele-specific GOM in the cancers, we compared the genomic and regulatory features among these loci to the background features of all informative loci using logistic regressions. As a comparison, we performed the same analyses for ASM loci that showed allele-specific losses of methylation in the cancers. This procedure revealed very strong over-representation of the poised "bivalent" promoter state among the ASM DMRs with allele-specific GOM in the cancers, compared both to ASM loci overall (OR = 1.7; $p = 4.1 \times 10^{-26}$) and to ASM loci with allele-specific LOM in the cancers (OR = 40; $p < 10^{-999}$); Fig. 3. Poised promoters, as annotated by ENCODE chromatin state, are marked by the simultaneous presence of active histone marks, H3K4me3 and H3K4me2, and the repressive mark H3K27me3. Such regions are known to sometimes exist in a poised state in non-

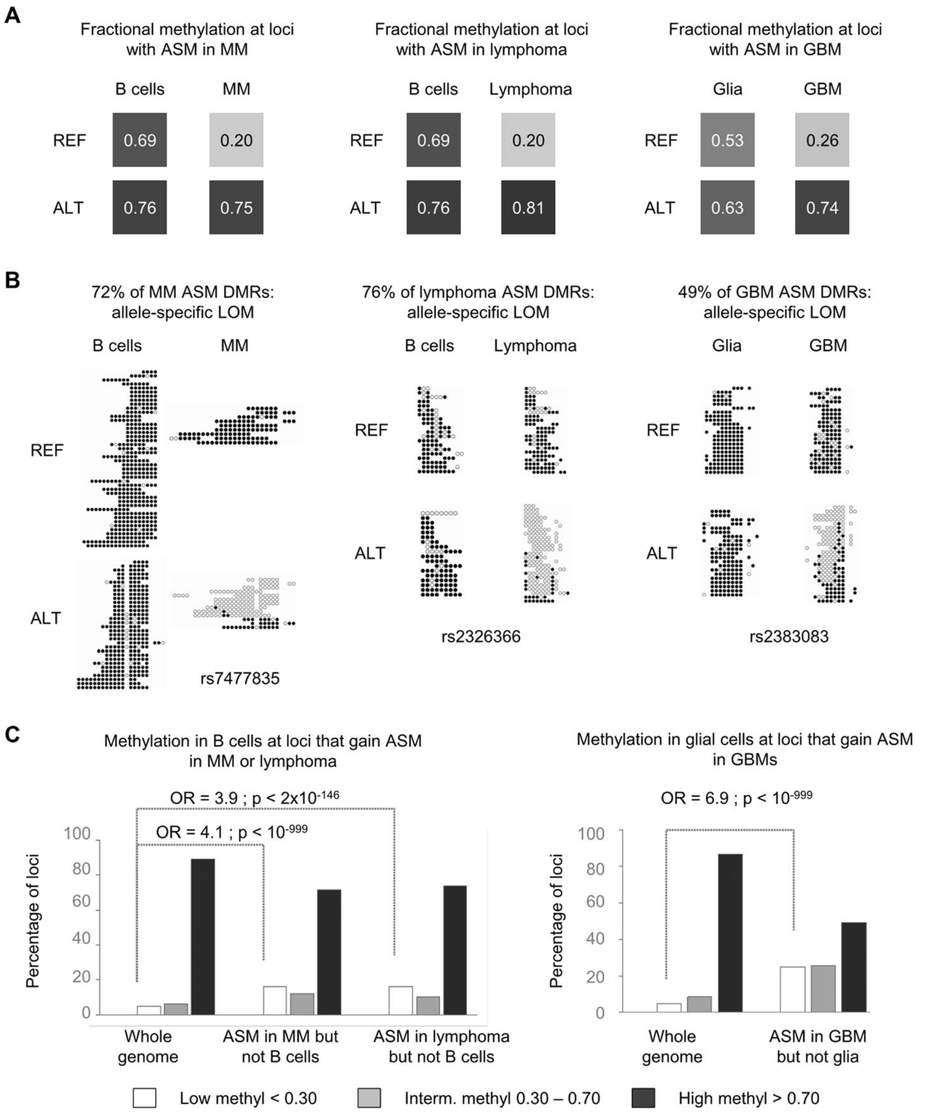

Fig. 2 Gains of ASM in cancers due to widespread allele-specific LOM. **a** Schematic showing the average configurations for allelic methylation levels in non-cancer and cancer samples at loci where ASM was observed only in cancer. Cancer samples are compared to lineage-matched non-cancer cell types. Average fractional methylation was estimated using a linear mixed model with random intercept and random slope (the "Materials and methods" section). For each sample type, the squares represent the model estimate of the average fractional methylation in the low and high methylated alleles. **b** Examples showing primary WGBS data. For the three types of cancers, the most frequent situation is an allele-specific LOM occurring in the cancers at loci that are highly methylated in the lineage-matched normal cell types. Methyl-seq reads separated by REF and ALT allele. Methylated CpGs are black and unmethylated CpGs are white. **c** Distribution of net methylation levels in normal B cells (left) and glia (right) grouped into 3 classes (low, intermediate and high methylation) at all informative CpGs (random expectation) and at CpGs where ASM was observed in the cancers but not in the matched non-cancer samples. While allele-specific LOM in cancer accounts for most instances of cancer-only ASM (black bars; high methylation in the normal samples), relative to the background of global hypomethylation in the cancers, it occurs less often than random expectation. In contrast, the smaller group of loci that have GOM leading to cancer-only ASM (white bars; low methylation in the normal samples) show a significant enrichment over random expectation, given the globally hypomethylated genomic background of the cancers

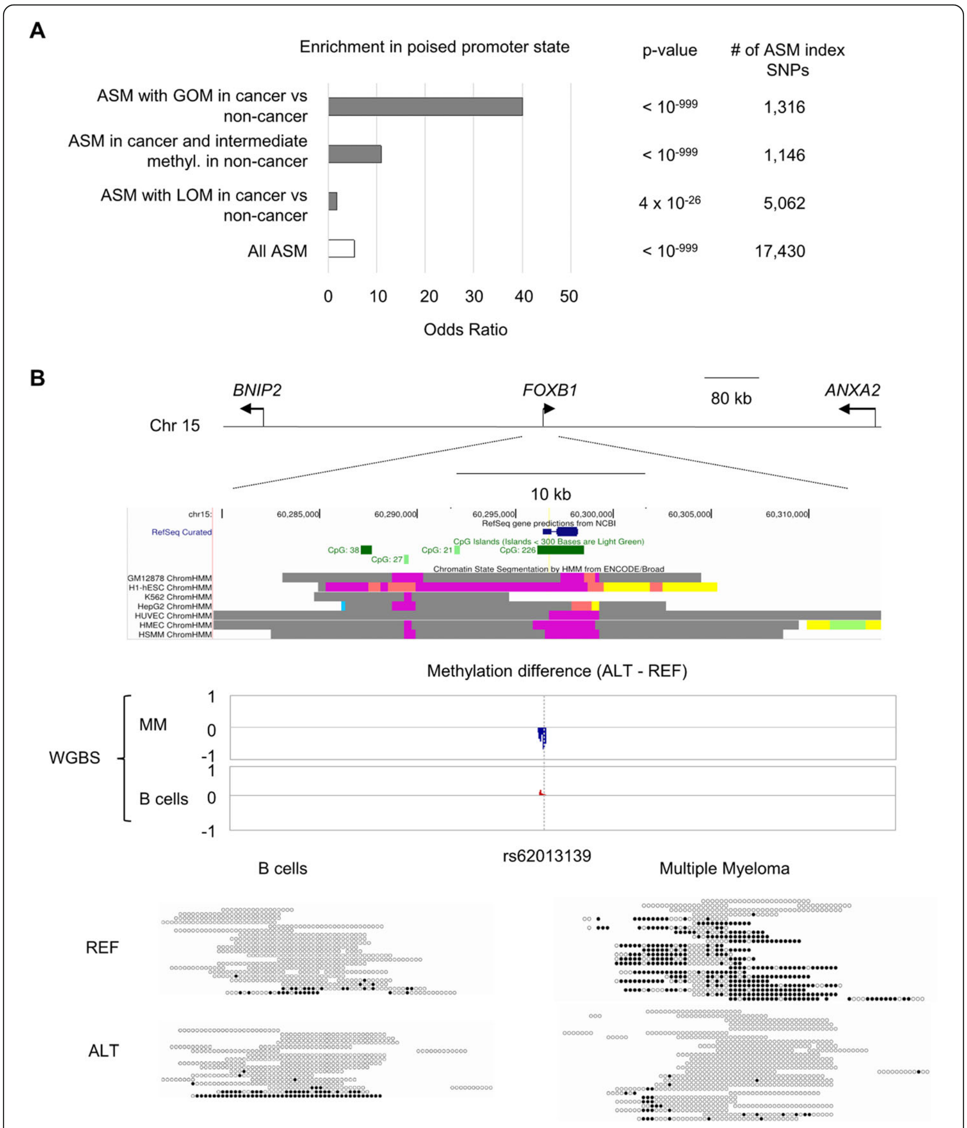

Fig. 3 Gains of ASM in cancers due to allele-specific GOM at loci in poised chromatin. **a** Graph showing enrichment in the poised promoter state as defined using ENCODE chromatin state segmentation by HMM. Although enrichment in poised promoter state is observed among ASM regions in general, this enrichment is dramatically increased among the subset of loci that show allele-specific GOM in cancers compared to cell lineage-matched normal samples. **b** Map of the *FOXB1* locus showing an example of allele-specific GOM in multiple myeloma overlapping a CpG-island region with a poised promoter chromatin state (color coded purple). Methylation differences between alleles (index SNP rs62013139) are shown as a genome browser track and as WGBS reads for CD138+ multiple myeloma cells from a bone marrow aspirate, which show strong ASM with hypermethylation of the REF allele, and a paired peripheral blood non-neoplastic B cell sample from the same patient, which shows very weak ASM with slight hypermethylation of the ALT allele

neoplastic stem cells [27] and can transition to a CpG-hypermethylated repressed state in cancer cells that acquire de-differentiated or stem cell-like phenotypes [28].

For completeness, using a similar statistical approach and mixed model for the set of ASM occurrences that were shared by cancer and non-cancer samples, we asked whether ASM might be not only more frequent in cancers, but also stronger. We found no significant differences in average ASM magnitude between cancer and non-cancer ASM loci (Additional file 2: Figure S15).

Enrichment for chromatin states suggests mechanistic similarities between cancer and non-cancer ASM

Different chromatin states, and different classes of binding sites for TFs and CTCF, can be associated with specific patterns of CpG methylation [29–33]. Among the ASM DMRs found in the non-cancer samples, enrichment of active and poised promoter regions and enrichment of the poised/bivalent enhancer state are strong, the active transcription state is slightly enriched, and quiescent chromatin is depleted, relative to the background of adequately covered genomic regions (Table 1). This over-representation of promoter/enhancer elements among ASM DMRs suggests that ASM may contribute to inter-individual differences in gene expression—a conclusion that is supported by our observation of enrichment for eQTLs in ASM DMRs (Table 1). Using chromatin state data from the Roadmap Epigenomics project [34], which is available for T cells CD3, T cells CD4, T cells CD8, B cells, monocytes, cerebral cortex, and the GM12878 cell line, we tested chromatin state enrichment among ASM index SNPs separately in each of these tissues and cell types and found that each of the major enrichments is shared across these tissues and cell types. This finding suggests that while ASM maps are partly tissue-specific (see below) the ASM is produced by shared underlying mechanisms.

To assess similarities and differences in the characteristics of ASM in cancer vs non-cancer, we took two approaches: first, we tested for enrichment of chromatin states among ASM loci that were detected only in cancers ("cancer-only" ASM; observed in at least 2 cancer samples but in none of the non-neoplastic samples) and ASM loci detected in non-cancer samples ("normal ASM"; present in at least one non-cancer sample, but allowing ASM in cancers as well), separately using bivariate logistic regression and second, testing the differential enrichment between the two groups using multivariate regressions including the interaction term between ASM and cancer status. Both approaches showed that ASM DMRs in cancer and non-cancer show a parallel enrichment in all the strongly enriched chromatin states, albeit with some differences among the less strongly enriched features (Table 1). These findings suggest that the basic mechanisms leading to ASM are similar in non-neoplastic and neoplastic cells—a conclusion that is further supported by analysis of correlations of ASM with SNPs in CTCF and TF binding sites, described below.

ASM correlates with allele-specific binding affinities of specific CTCF and TF recognition motifs in both cancer and normal samples

The hypothesis that allele-specific TF binding site occupancy (ASB) due to sequence variants in regulatory elements could be a mechanism leading to ASM has been supported by previous data from us and others [8, 10, 11]. To test this hypothesis using denser maps, and to ask whether this mechanism might underlie ASM in both normal and neoplastic cells, we analyzed the set of ASM loci for enrichment of sequence motifs recognized by classical TFs, and motifs recognized by CTCF, which defines the insulator chromatin state and regulates chromatin looping [35–37]. Previously, we showed that ASM DMRs can overlap with strong CTCF ChIP-seq peaks and polymorphic CTCF binding sites [8, 38]. In our expanded dataset, we used atSNP to identify CTCF motif occurrences where the ASM index SNP not only overlaps a CTCF motif but also

Table 1 Enrichment analysis for mechanistically relevant features reveals similarities between normal and cancer ASM

Parameter	Normal ASM[a] ($N = 13{,}069$) OR (p value)	Cancer ASM[b] ($N = 4361$) OR (p value)	Same direction in cancer vs normal ASM?	Enrichment strength in cancer vs. normal OR (p value)
ASM SNP is an ASB SNP	14.5 ($< 1 \times 10^{-999}$)	5 (3.1×10^{-32})	Yes: enriched	0.3 (2.8×10^{-13})
Poised promoter	5.1 ($< 1 \times 10^{-999}$)	5.6 ($< 1 \times 10^{-999}$)	Yes: enriched	1.1 (0.069)
Polymorphic TFBS motif[c]	4.9 (2×10^{-184})	2 (3.7×10^{-34})	Yes: enriched	0.4 (6.4×10^{-28})
Weak promoter	4.4 ($< 1 \times 10^{-999}$)	3.5 (3×10^{-290})	Yes: enriched	0.8 (5.9×10^{-10})
Active promoter	4.1 ($< 1 \times 10^{-999}$)	3.9 (2×10^{-238})	Yes: enriched	1 (0.44)
Correlated polymorphic TF binding motif[c]	3.7 ($< 1 \times 10^{-999}$)	0.8 (0.3×10^{-4})	NO	0.2 (3×10^{-106})
Weak/poised enhancer	2.7 ($< 1 \times 10^{-999}$)	1.5 (8.9×10^{-45})	Yes: enriched	0.6 (1.4×10^{-58})
Repetitive sequences	2.2 (3.8×10^{-76})	1.9 (9.6×10^{-16})	Yes: enriched	0.9 (0.083)
ASM SNP is eqtl SNP	2.2 (1.2×10^{-92})	1.9 (4.7×10^{-21})	Yes: enriched	0.9 (0.11)
Strong enhancer	2.2 ($< 1 \times 10^{-999}$)	1.3 (1.1×10^{-14})	Yes: enriched	0.6 (6.1×10^{-34})
ASM SNP is GWAS peak SNP	2 (2.2×10^{-32})	2 (9.1×10^{-12})	Yes: enriched	1 (0.97)
Insulator element	2 (5×10^{-214})	1.5 (9×10^{-19})	Yes: enriched	0.7 (5.11×10^{-11})
Polycomb repressed	2 ($< 1 \times 10^{-999}$)	1.9 (2×10^{-103})	Yes: enriched	1 (0.22)
ASM SNP in stringent block with GWAS peak SNP (autoimmune/inflammatory)	1.5 (1.7×10^{-17})	1.3 (0.0043)	Yes: enriched	0.9 (0.13)
ASM SNP in stringent block with GWAS peak SNP (cancers)	1.5 (3.8^{-15})	1.6 (9.5×10^{-9})	Yes: enriched	1.1 (0.39)
ASM SNP is GWAS peak or LD Rsq > =0.8	1.5 (3.7×10^{-40})	1.4 (2.1×10^{-9})	Yes: enriched	0.9 (0.2)
Active transcriptional state (txn)	1.3 (9.1×10^{-47})	1 (0.18)	No	0.7 (1.1×10^{-17})
ASM SNP in stringent block with GWAS peak SNP (all diseases/phenotype)	1.2 (1.7×10^{-13})	1.1 (0.0019)	Weak	1 (0.34)
Quiescent chromatin[d]	0.3 (3×10^{-274})	0.6 (9.5×10^{-35})	Yes: depleted	1.9 (2.1×10^{-30})
Chromatin desert[e]	0.2 ($< 1 \times 10^{-999}$)	0.5 (3.3×10^{-82})	Yes: depleted	3.1 (4×10^{-124})

[a] “Normal ASM” is ASM in at least one non-cancer sample, allowing ASM in cancer samples

[b] “Cancer ASM” is ASM present in cancer samples (including GM12878 LCL), but not in any normal sample

[c] Enriched and correlated motifs determined on the complete set of ASM SNPs

[d] Heterochromatin in at least one cell lines and no other states observed in the other cells

[e] Chromatin desert defined in the “Materials and methods” section

significantly affects the predicted binding affinity, as reflected in the Position Weight Matrix (PWM) score. For this analysis, we required a significant difference in binding likelihood between the two alleles (FDR < 0.05) and a significant binding likelihood ($p < 0.005$) for at least one of the alleles (reflecting potential CTCF occupancy of at least one allele). We identified 3.075 ASM SNPs (17%) that significantly disrupted at least one of the canonical or ENCODE-discovery CTCF motifs [36] (http://compbio.mit.edu/encode-motifs/). To estimate the random expectation of polymorphic CTCF motif occurrences in the genome (the background frequency), we ran atSNP on a random sample of 40,000 non-ASM informative SNPs (1:3 ASM vs non-ASM SNP ratio) and found that 8.4% of these non-ASM informative SNPs significantly disrupted a CTCF motif, corresponding to a substantial enrichment for disrupted CTCF motifs among ASM SNPs (OR = 2.3; p value = 10^{-218}).

As noted in our previous smaller study [8], in the enlarged dataset, this overall enrichment of CTCF motif-disrupting SNPs among ASM loci persists, albeit slightly weaker, when considering only non-CpG-containing polymorphic CTCF motif instances (OR = 1.8; $p = 6.3 \times 10^{-34}$). When testing enrichment separately for each of the 14 distinct ENCODE/JASPAR-defined CTCF motifs, we found significant enrichment for 13 of them (Additional file 8: Table S7). Moreover, as shown in Fig. 4, Additional file 2: Figure S16, and Additional file 9: Table S8, the difference in binding affinity score between alleles is significantly anti-correlated (i.e., inversely correlated) with the difference in methylation for three of these motifs, and these correlations persist after adjustment for the presence or absence of CpGs in the motif occurrences in a multivariate model. Thus, consistent with our previous conclusions in the smaller dataset, which required motif pooling [8], these results from individual motif classes show that the presence of a methylatable CpG in the CTCF binding motif is not required; rather, the essential feature is allele-specific binding site occupancy.

Like CTCF, classical TFs could account for instances of ASM via ASB. When we scanned each ASM SNP for all ENCODE/JASPAR defined TF motifs [39], we found 17,022 (95%) recurrent ASM SNPs disrupting at least one TF binding site occurrence, representing a significant enrichment compared to the 32,790 (82%) such disruptions in the 40,000 SNP randomized non-ASM background set. Of these ASM SNPs, 12,853 overlapped at least one ENCODE DNase I hypersensitive site and 3044 at least one ENCODE cognate ChIP-seq TF peak. From a panel of 2263 TF motifs with at least 10 occurrences, we found 856 motifs with a specific enrichment (OR $\geq$ 2 and FDR corrected q value < 0.05, compared to the random sample of 40,000 non-ASM informative SNPs) among ASM DMRs (Additional file 8: Table S7). Next, using linear regression of allele-specific binding affinity (PWM) score differences on allele-specific CpG methylation differences, we found 177 TF binding motifs, corresponding to 115 cognate TFs, where DNA methylation appears to be shaped by binding site occupancies (Fig. 4, Additional file 2: Figure S12 and S16, Additional file 9: Table S8, and Additional file 10: Table S9). Among these motifs, 144 also showed significant enrichment among ASM loci (Additional file 10: Table S9). Regarding the motif classes that are significantly enriched among ASM index SNPs but do not show significant correlations of PWM scores with ASM magnitude, it is likely that some simply have too few ASM occurrences in the current dataset to achieve significance in the correlation analysis.

Using stringent statistical criteria (FDR < 0.05 and $R^2 \geq 0.4$), all but two of the TF motifs that were correlated with ASM show inversely correlated behavior, such that a relatively higher binding likelihood (stronger PWM score) correlates with CpG hypomethylation (Additional file 9: Table S8, examples in Fig. 4 and Additional file 2: Figure S12 and S16). Multivariate linear regression of the 158 (out of 177) significantly correlated motifs with at least three CpG-containing and three non-CpG-containing occurrences revealed that these inverse correlations between binding affinity scores and methylation levels persist after adjustment for the presence or absence of CpGs in the motifs. Like the findings for CTCF sites, these results suggest that ASM regions form around polymorphic TF binding sites because of allele-specific differences in binding site occupancy (ASB), not requiring a methylatable CpG in the binding motif.

Lastly and importantly, we tested for enrichment of TF and CTCF binding motifs and correlations of ASM with predicted binding affinities separately in the sets of ASM

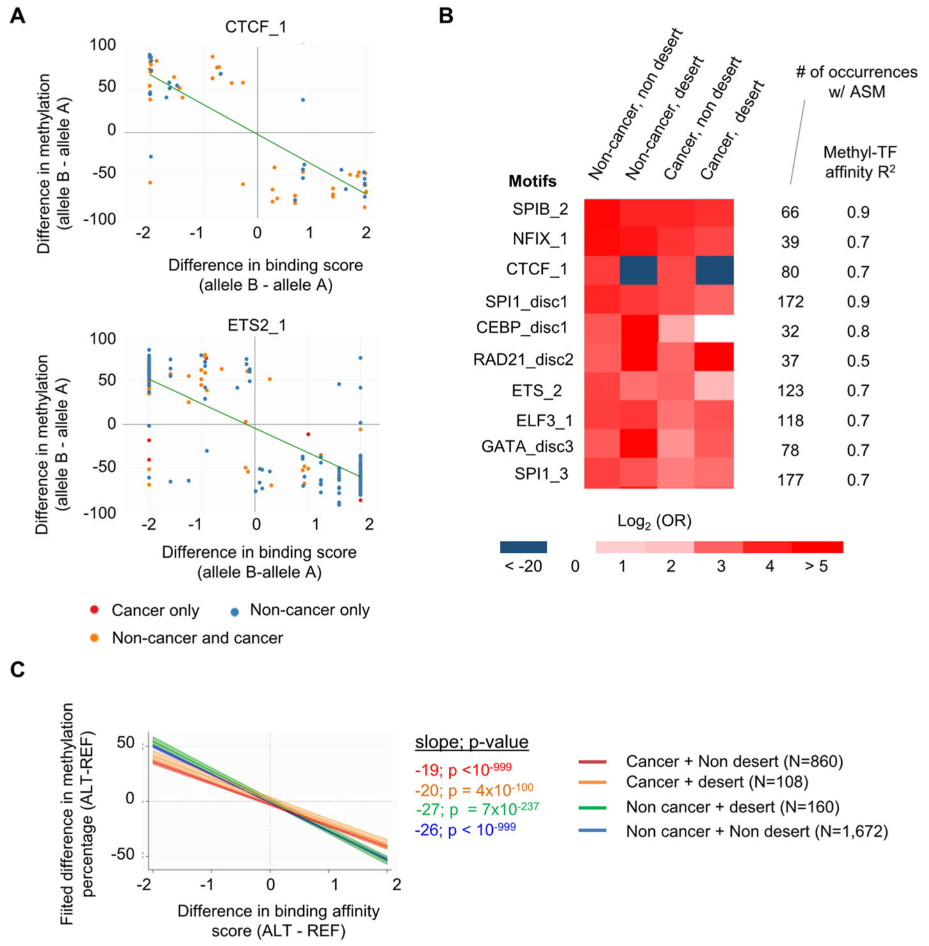

Fig. 4 ASM is driven by allele-specific CTCF and TF binding in both normal and neoplastic cells. **a** X-Y plots showing examples of TF motifs with strong correlations between predicted allele-specific binding site affinities (PWM scores) and methylation differences across all occurrences showing ASM. These examples are among 179 significantly correlated motifs (Additional file 9: Table S8). Each data point represents one occurrence of the motif overlapping an ASM index SNP in cancer (orange) or non-cancer samples (blue). For occurrences showing ASM in multiple samples, allelic methylation differences were averaged across samples by sample type. R^2 and B-H corrected *p* values (FDR) were calculated using linear regression. **b** Most of the polymorphic motifs with significant correlations between allelic methylation and predicted binding affinities are also enriched among ASM regions (Additional file 10: Table S9). The heatmap shows the enrichment or depletion, in $\log^2$(OR), for the top 10 enriched TF binding motifs among cancer or non-cancer ASM loci in regions defined as chromatin desert or non-desert (the "Materials and methods" section). **c** Significant correlations between allelic TF binding affinity scores and ASM in each of the 4 classes of ASM loci. The graph shows the fitted ASM difference on PWM score using a multivariate mixed model. The fitted line and its 95-confidence intervals (area) are shown for each ASM class; slopes were calculated by the marginal effects of the interaction term between PWM score and ASM class and were significantly different from zero. The correlations are similar in cancer ASM (in both non-desert and desert) compared to non-cancer ASM, with only small differences in the slopes for each class

loci that were detected only in the cancers (including the GM12878 LCL) vs those found in the total group of non-cancer samples. We also analyzed the full set of ASM loci using a multivariate mixed model to test for interactions of normal vs cancer status with the TF binding site affinity to ASM strength correlations. The results showed that ASM loci in cancer and non-cancer samples have similar directions of the correlations of ASM with destructive SNPs in the top-ranked classes of polymorphic TF binding motifs (Fig. 4 and Additional file 2: Figure S16), which indicates sharing of this fundamental mechanism of ASM in normal and cancer cells. This key result was confirmed when we restricted our analyses to the data from a single sequencing facility using a single library construction method (Additional file 2: Figure S12B, C). However, the correlations between predicted TF binding site affinities and ASM amplitude were slightly weaker on average (shallower slope in the X-Y plot) among the cancer-only ASM loci (Fig. 4 and Additional file 2: Figure S16), and ASM-correlated motifs were not enriched among these loci (Table 1). These findings are explained by the presence of a subgroup of cancer-only ASM loci that show allele-switching (see below).

Direct testing of the TF binding site occupancy mechanism of ASM

As a crucial validation, using our GM12878 SureSelect and WGBS data and the large number of ENCODE ChIP-seq experiments available for this cell line, we could directly ask whether ASM regions with or without polymorphic CTCF and classical TF binding sites exhibit allele-specific binding of the cognate factors. Among the 2102 high-confidence ASM index SNPs from our GM12878 data, 787 overlapped at least one ChIP-seq peak in this cell line and had enough ChIP-seq reads ($\geq$ 10×) to assess allele-specific binding of at least one ENCODE-queried TF. We found that 16.6% (131) of these ASM index SNPs showed ASB for at least one TF that could be assessed using available ENCODE data. As predicted from the binding site occupancy hypothesis for ASM, at 100 (76%) of these sites, considering both CTCF and TF motifs, the hypomethylated allele showed significantly greater occupancy. This percentage far exceeds random expectation (exact binomial test, $p = 1.2\text{e}10^{-9}$). Confirming this pooled analysis, among 9 TFs with more than 10 ASB occurrences associated with ASM, 7 examples, including the ELF1 (ETS-family) motif and others, showed a significant enrichment in ASM occurrences with an inverse correlation of predicted binding affinity with allelic CpG methylation (ASB-ASM instances with inverse correlation: 90–100%, FDR < 0.05).

Somatic mutations in TF binding sites can produce ASM in human cancers

To more completely understand the features of ASM in cancers, and to further test the hypothesis that destructive SNPs in TF binding motifs give rise to ASM, we searched for somatic mutations in the 4 multiple myeloma cases that were paired with non-neoplastic peripheral blood B cells from the same patients (analyzed using the same WGBS library protocol to ensure similar regional coverage depth) which served as germline reference sequences. We found somatic mutations at frequencies of 499 to 1023 per case, and among these mutations from 6 to 17% were associated with gains of ASM (referred to here as “de novo ASM,” examples in Fig. 5). We next filtered out mutations situated within 1 kb of known ASM index SNPs that had already been seen in other samples, since such instances might simply be uncovering normal ASM by conferring heterozygosity in regions that were non-informative in the patient’s germline sequence. Using the filtered list of 410 de novo ASM occurrences, we asked whether the somatic mutations associated with de novo ASM might be disrupting TF binding motifs at a frequency greater than random expectation. We found a significant enrichment (OR > 2 and $p < 0.05$) for 54 TF binding motifs among the de novo ASM occurrences, compared to the representation of these motifs among all the somatic mutations that were not associated with ASM. Even more convincingly, we found that a majority (71%) of the TF binding site motif classes that were enriched among instances of de novo ASM belonged to the same motif classes that were enriched among the much larger set of ASM loci that were tagged by germline SNPs (OR = 3.6, p value = 6.3×10^{-4}; examples in Fig. 5). Thus, while mutation-associated de novo ASM does not make a large numerical contribution to the overall gains of ASM in cancer vs non-cancer, this special phenomenon is informative in emphasizing the shared underlying mechanism, namely TF binding motif disruption or creation, for ASM in cancer and normal cells.

ASM DMRs are found both in active chromatin and in quiescent “chromatin deserts”

For post-GWAS mapping of rSNPs, much attention has been appropriately focused on cataloging SNPs that are expression quantitative trait loci (eQTLs) and/or lie within regions of ASB. Such efforts are aided by databases such as AlleleDB for allele-specific marks [40–42], and RegulomeDB [43, 44], which highlights potential rSNPs in non-coding regions by assigning a score to each SNP based on criteria including location in regions of DNAase hypersensitivity, binding sites for TFs, and promoter/enhancer regions that regulate transcription. Our cross-tabulations indicate that, despite a strong enrichment in ASB SNPs among ASM index SNPs (Table 1), most of the ASM index SNPs (> 95%) in our expanded dataset currently lack ASB annotations (Additional file 3: Table S2). In addition, index SNPs for strong ASM DMRs sometimes have weak RegulomeDB scores (Additional file 3: Table S2). Thus, from a practical standpoint with existing public databases, ASM mapping for identifying rSNPs appears to be largely non-redundant with other post-GWAS modalities.

To further assess the unique value of ASM mapping, we defined “chromatin desert” ASM regions as 1 kb genomic windows, centered on ASM index SNPs, that contained no DNAse peaks or only one DNAse peak among the 122 ENCODE cell lines and tissues, and no strong active promoter/enhancer, poised, or insulator chromatin state in any ENCODE sample. Less than 55% of such regions have SNPs listed in RegulomeDB, and when they are in that database, they almost always (93%) have weak scores equal to or greater than 5 (Additional file 3: Table S2). While most ASM loci map to active chromatin and are depleted in desert regions overall (Table 1), we find that 8% of ASM index SNPs in normal cells and 22% of cancer-only ASM SNPs are in chromatin deserts (Table 1 and Additional file 3: Table S2). Although deserts lack evidence of TF and CTCF binding in available databases, ASM DMRs found in these regions might be informative for localizing bona fide rSNPs if some desert regions contain cryptic binding motifs that were active (occupied) at an earlier point in the history of the cell.

To address this possibility, we asked whether correlations of ASM with destructive SNPs in TF binding motifs might also pertain to ASM in desert regions. We analyzed the full set of ASM loci using a multivariate mixed model to test for interactions of normal vs cancer status and desert vs non-desert location (i.e., 4 classes of ASM loci) with the TF binding site affinity to ASM strength correlations. Some motifs, such as CTCF binding sites, were highly depleted in deserts and therefore excluded from the

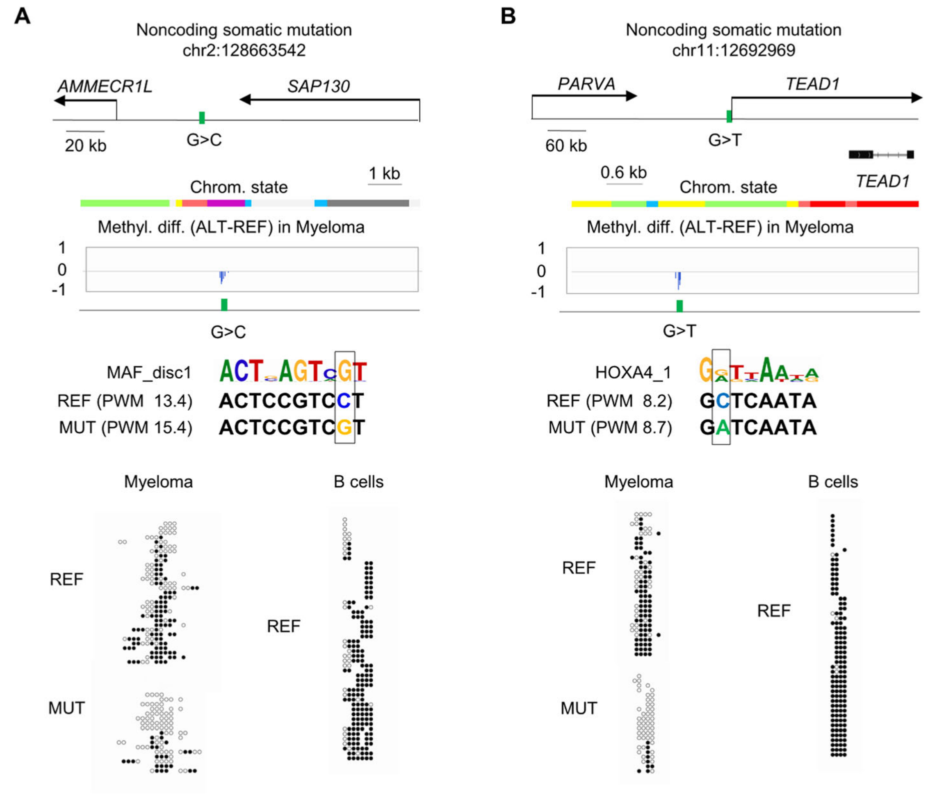

Fig. 5 Examples of de novo ASM associated with somatic mutations in cancers. **a** Map of the de novo ASM region tagged by a somatic mutation at chr2:128663542 in case SUB006. The mutation is intergenic between *AMMECR1L* and *SAP130*. It overlaps a poised chromatin region and creates a de novo binding motif for the MAF TF, with a higher binding affinity on the mutated allele than on the reference (germline) allele. The graphical representation of the WGBS data shows strong ASM with less methylation on the mutated allele. The reference allele shows similar methylation (high) in both the myeloma and paired B cells. **b** Map of the de novo ASM region tagged by the somatic mutation at chr11:12692969, also in case SUB006. The mutation is intergenic downstream of *PARVA* and upstream of *TEAD1*. It overlaps an active enhancer chromatin region and creates a de novo binding motif for HOXA4 with a higher binding affinity on the mutated allele than on the reference (germline) allele. The graphical representation of the WGBS data shows strong ASM with lower methylation on the mutated allele. The reference allele showed a similar methylation (high) in both the myeloma and paired B cells. For these two motifs, the occurrences map to the negative strand and are oriented 3′ to 5′ per atSNP convention

analysis, which was performed on the subset of 74 TF motifs that had at least three occurrences per ASM class. The correlations, when significant (FDR < 0.05), were in the same direction (inverse correlation of predicted binding affinity with allelic methylation) in all ASM classes. As expected from the findings above, we observed a slightly weaker correlation for cancer-only ASM loci compared to ASM loci in non-cancer samples. However, no differences in the strength of the correlations were found when comparing ASM occurrences in desert versus non-desert locations, both for normal and cancer-associated ASM loci. The simplest hypothesis to explain these results is that ASM DMRs in desert regions are footprints left by rSNPs that disrupt cryptic TF binding sites that were active at some stage of normal or neoplastic cell differentiation (or de-differentiation) but are no longer active in available cells or tissue types. Additional file 2: Figure S17 shows examples of ASM DMRs in desert regions that contain disruptive SNPs in ASM-correlated ETS- and ERG-family TF binding motifs.

Allele-switching at ASM loci is infrequent in normal samples but increased in cancers

Most of the ASM DMRs passed statistical cutoffs for ASM in less than half of the informative samples (Additional file 3: Table S2), with variability not only between cell types and cancer status but also within a single cell type. Given the connection between TF binding site occupancies and ASM, one hypothesis to explain this variability invokes differences in intracellular levels of TFs (Additional file 2: Figure S18A). Alternatively, genetic differences (i.e., haplotype effects due to the influence of other SNPs near the ASM index SNP) could also play a role (e.g., Additional file 2: Figure S18B). A more extreme form of variation was observed at some ASM loci, namely "allele switching" [8], in which some individuals have relative hypermethylation of Allele A while others show hypermethylation of Allele B, when assessed using a single index SNP. Some instances of allele switching reflect haplotype effects [8] or parental imprinting, but other occurrences might have other explanations. In this regard, a striking finding in the current dataset is that the frequency of allele switching among ASM loci in normal samples is low (14%), while the rate of allele switching is strikingly higher (43%) among cancer-only ASM loci (Fig. 6). This finding suggests that biological states, here neoplastic vs non-neoplastic, can influence the stability of ASM, with greater epigenetic variability or instability in the cancers.

To investigate this variability, we compared the features of ASM DMRs that showed allele switching versus those that did not. As shown in Fig. 6c, the sets of ASM index SNPs for two classes of loci differed significantly in the relative representation of specific CTCF and TF binding motifs, such that the CTCF_1 motif and nearly all of the most strongly ASM-correlated classical TF binding motifs were markedly under-represented among the switching loci. Reinforcing this finding, ASM loci that were highly recurrent across multiple normal cell types and individuals showed a low frequency of switching, even when these loci had ASM in some cancers (Additional file 2: Figure S19).

These results suggest a working model that posits two classes of binding motif occurrences. One group of motif occurrences stably bind their cognate TFs when the motif sequence is optimal but are sensitive to the effects of destructive SNPs in the motif. These motif occurrences therefore show strong unidirectional correlations of PWM scores with ASM, independently of the neoplastic cellular phenotype. Another group of motifs is postulated to have more labile binding of their cognate TFs, are sensitive to changes in the intracellular levels of their cognate factors, and can participate in ASM allele switching, via "TF competition." According to this model (Additional file 2: Figure S18C), in situations with adequate chromatin accessibility, there can be replacement of one TF by another more highly expressed one that recognizes a nearby or overlapping DNA sequence motif. The credibility of this hypothesis is supported by the well-known over-expression of various oncogenic TFs in cancer cells, and by data indicating that global DNA hypomethylation in transformed cells is associated with increased chromatin accessibility at regulatory elements [34, 45].

ASM index SNPs in LD or precisely coinciding with GWAS peak SNPs

For assessing ASM as a signpost for rSNPs in disease-associated chromosomal regions, we defined lenient and stringent haplotype blocks by applying the algorithm of Gabriel

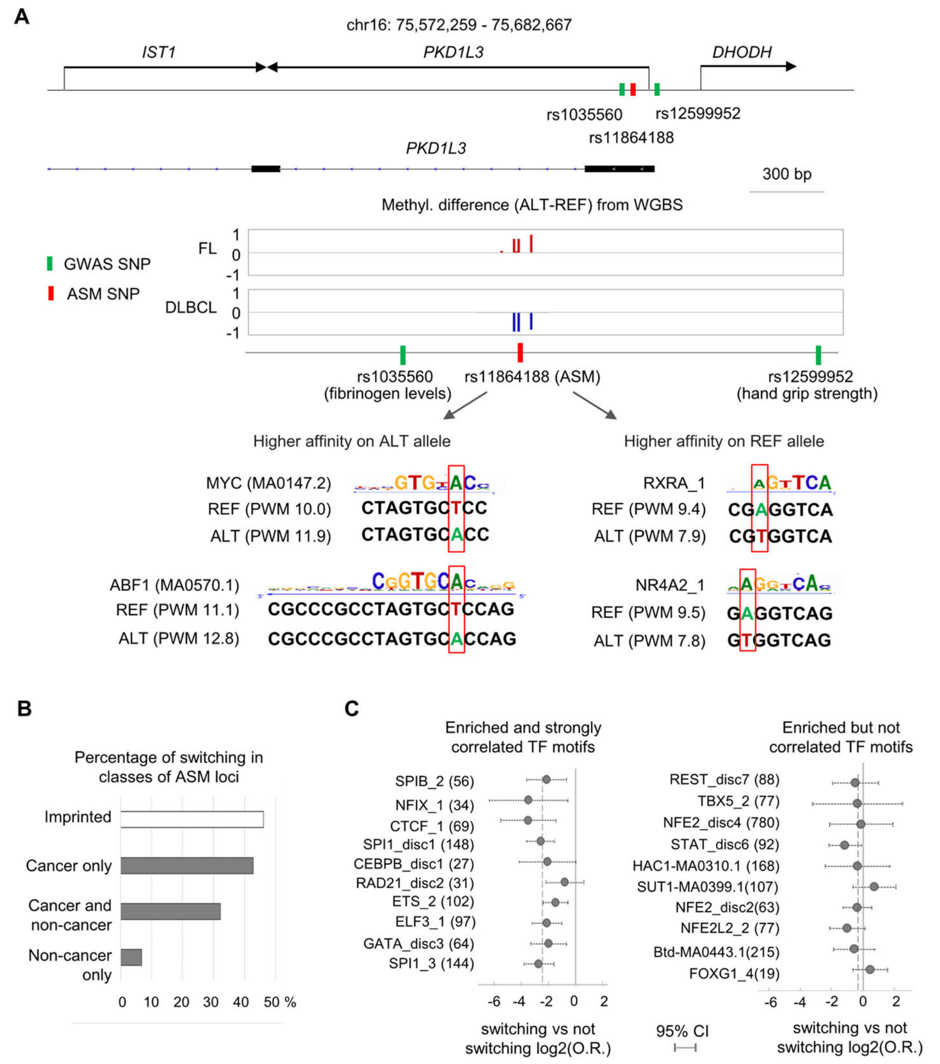

Fig. 6 Increased allele switching at ASM loci in cancers. **a** Map of the ASM region tagged by SNP rs11864188 in the *PKD1L3* gene. The ASM shows allele switching, with the ALT allele being hyper-methylated in the FL but with an opposite direction of the allelic methylation bias in the DLBCL. No ASM was detected in 23 non-cancer heterozygous samples. The rs11864188 SNP disrupts multiple TF motifs, some of which have opposite allele-specific predicted affinity differences, including MYC and ABF1 motifs with a higher affinity on the ALT allele and RXRA and NR4A2 motifs with a higher affinity on the REF allele. The MYC motif maps to the negative strand and is oriented 3′ to 5′. **b** Frequency of switching, scored for ASM index SNPs that were heterozygous in ≥ 2 samples, is increased among cancer-only ASM loci (43%) compared to ASM loci found in normal samples (10%). As a control, ASM index SNPs mapping to imprinted regions showed 46% switching, approximating the expected 50% based on parent-of-origin dependent ASM. **c** Enrichment analyses of CTCF and TF binding motifs among switching vs non-switching ASM loci: in the left panel, polymorphic motifs with strong correlations between ASM magnitude and affinity score differences are found to be depleted among switching compared to non-switching ASM loci. In the right panel, polymorphic TF motifs enriched among ASM but with little or no correlations of predicted binding affinity with ASM magnitude showed little or no depletion among switching ASM loci. The solid vertical lines represent zero (no enrichment or depletion), and the dotted lines show the average values for depletion for each set of motifs

et al. [46], using 1000 Genomes data and employing *D*-prime (*D*′) values, both with standard settings utilizing high *D*′ and *R*-squared (R^2) values to define "stringent" blocks (median size 5 kb) and with relaxed R^2 criteria to define larger "lenient" blocks with a median size of 46 kb (Additional file 2: Figure S20 and S21). We also calculated R^2 between each ASM index SNP and GWAS peak SNP to identify SNPs in the same haplotype block and with high R^2, plus SNPs in strong LD located in genomic regions that lacked a haplotype block structure. We took this two-fold approach because (i) R^2 can fail to identify SNPs in perfect LD when rare mutations have occurred over time on pre-existing common alleles in the population—a situation that can have a high D', and (ii) for some loci, the combined effect of multiple regulatory SNPs, some with weak R^2 values but high D', might be responsible for net effects on disease susceptibility. Using our complete list of ASM DMRs and GWAS data from NHGRI-EBI, including both supra-threshold and suggestive peaks ($p < 10^{-6}$), we identified 1842 ASM SNPs in strong LD ($R^2 > 0.8$) or precisely coinciding with GWAS peak SNPs (Additional file 3: Table S2). Highlighting mechanistic information from these ASM loci, among the ASM index SNPs in strong LD or precisely coinciding with GWAS peak SNPs, 1450 disrupted ASM-enriched classes of CTCF or TF binding motifs and 310 disrupted significantly ASM-correlated CTCF or TF binding motifs.

ASM index SNPs in LD with GWAS peaks for autoimmune/inflammatory diseases

We found 275 ASM DMRs containing 305 index SNPs in strong LD ($R^2 > .8$) with GWAS peak SNPs for autoimmune and inflammatory diseases, or related traits such as leukocyte counts (Additional file 11: Table S10) which corresponds to a moderate but significant enrichment (OR = 1.5, $p = 1.2 \times 10^{-18}$). Of these 305 index SNPs, 8 were in HLA genes and the remainder were in non-HLA loci. About half of these regions showed ASM in immune system cell types and/or B cell tumors (Additional file 11: Table S10). Among these ASM index SNPs, 66 precisely coincided with GWAS peak SNPs, supporting the candidacy of these statistically identified SNPs as functional rSNPs. Moreover, 61 of the 305 ASM index SNPs altered strongly ASM-correlated CTCF or TF binding motifs, and 237 disrupted enriched classes of binding sites, thus providing mechanistic leads to disease-associated transcriptional pathways. Interesting ASM index SNPs for this disease category, some precisely coinciding with GWAS peak SNPs and others in strong LD with these peaks include rs2145623, precisely coinciding with a GWAS peak SNP for ulcerative colitis, sclerosing cholangitis, ankylosing spondylitis, psoriasis and Crohn's disease (nearest genes *PSMA6*, *NFKBIA*), rs840015 in strong LD with GWAS peak SNPs for celiac disease, rheumatoid arthritis, and hypothyroidism (nearest genes *POU2F1* and *CD247*), rs10411630 linked to multiple sclerosis (MS) via LD with GWAS peak SNP rs2303759 (nearest genes *TEAD2*, *DKKL1*, and *CCDC155*; Additional file 2: Figure S7), rs2272697 linked to MS via LD with GWAS peak SNP rs7665090 (nearest genes *NFKB1*, *MANBA*), rs2664280 linked to inflammatory bowel disease, systemic lupus erythematosus (SLE) and psoriasis via GWAS SNPs rs2675662 and rs2633310 (nearest genes *CAMK2G*, *PLAU*, *C10orf55*; Fig. 7), and rs6603785 which precisely coincides with a GWAS peak for SLE and hypothyroidism (nearest genes *TFNRSF4*, *SDF4*, *B3GALT6*, *FAM132A*, *UBE2J2*; Additional file 2: Figure S22). Each of these ASM index SNPs disrupts one or more strongly ASM-enriched or ASM-correlated TF binding motifs (Additional file 11: Table S10).

ASM index SNPs in LD with GWAS peaks for cancer susceptibility

We found 247 ASM DMRs containing 268 index SNPs in strong LD ($R^2 > .8$) with GWAS peak SNPs for cancer susceptibility or response to treatment (Additional file 12:

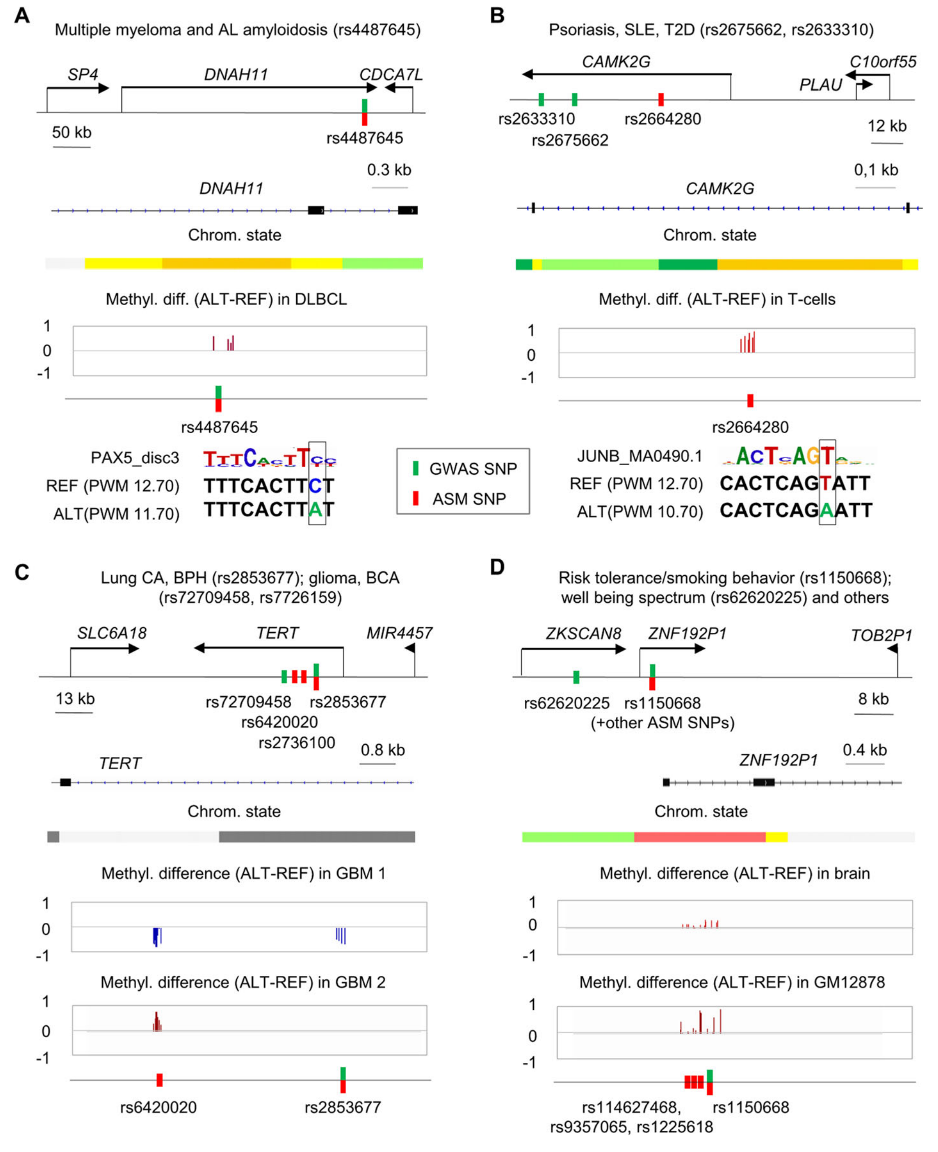

Fig. 7 (See legend on next page.)

(See figure on previous page.)
Fig. 7 Examples of ASM index SNPs in strong LD or precisely coinciding with GWAS peaks. **a** Map of the ASM DMR tagged by index SNP rs4487645, coinciding with a GWAS peak SNP for multiple myeloma ($p = 3.0 \times 10^{-14}$; OR = 1.38) and AL amyloidosis ($p = 2.0 \times 10^{-9}$; OR = 1.35). The ASM index SNP is in an enhancer region (yellow-coded chromatin state; GM12878 track) of the *DNAH11* gene on chromosome 7. This SNP disrupts a PAX5_disc3 TF binding motif on the ALT allele. The REF allele, with intact high-affinity motif, is relatively hypomethylated. Additional motifs are in Additional file 3: Table S2. **b** Map of the ASM region tagged by index SNP rs2664280, in strong LD with GWAS peak SNPs rs2675662 for psoriasis ($p = 3.0 \times 10^{-8}$; OR = 1.14) and rs2633310 for T2D ($p = 2.0 \times 10^{-8}$; beta = – 0.044). The ASM index SNP is in an enhancer region (yellow-coded; GM12878) in the *CAMK2G* gene on chromosome 10. This SNP disrupts AP1 binding motifs (JUNB shown) on the ALT allele, with higher binding affinity on the REF allele, which is relatively hypomethylated. The motif maps to the negative strand and is reported 3′ to 5′. Additional motifs are in Additional file 3: Table S2. **c** Map of the ASM regions tagged by rs2853677 and rs6420020 index SNPs in the *TERT* gene on chromosome 5. The DMRs are in quiescent/repressed chromatin in most ENCODE samples (light and dark gray; K562 track), but this region is transcribed in undifferentiated H1-hESC. ASM for these DMRs was found only in GBMs. The index SNP rs2853677 is a GWAS peak SNP for non-small cell lung cancer and benign prostatic hyperplasia ($p < 10^{-999}$; OR = 1.41 and $p = 2.0 \times 10^{-22}$; OR = 1.12, respectively). The other ASM index SNP, rs6420020 is in LD with GWAS peak SNPs for GBM ($p = 6.0 \times 10^{-24}$; OR = 1.68), breast carcinoma ($p = 3.0 \times 10^{-8}$; OR = 1.07), and chronic lymphocytic leukemia ($p = 6.0 \times 10^{-10}$; OR = 1.18). ASM allele switching is seen at rs6420020; polymorphic TF binding motifs are in Additional file 3: Table S2. **d** ASM DMR tagged by multiple SNPs (rs114627468, rs9357065, rs1225618, and rs1150668) in the promoter region of the *ZNF192P1* pseudogene, flanked by coding genes in the *ZSCAN* family, on chromosome 6. ASM index SNP rs1150668 is a GWAS peak SNP for body height ($p = 2.0 \times 10^{-7}$; beta = – 0.060), smoking status ($p = 6.0 \times 10^{-15}$; beta = – 0.0086), smoking behavior ($p = 3.0 \times 10^{-8}$; beta = + 0.011), myopia ($p = 1.0 \times 10^{-11}$; beta = + 0 6.78), and schizophrenia with autism spectrum disorder ($p = 8.0 \times 10^{-11}$; OR = 1.07). In addition, the 4 ASM index SNPs are in a stringently defined haplotype block containing GWAS peak SNP rs62620225, for psychiatric phenotypes including well-being spectrum ($p = 6 \times 10^{-12}$; beta = 0.023). ASM in this DMR was observed in multiple tissues, including brain. The ASM index SNP rs1225618 is as an ASB SNP for TAF1; other ASM-correlated motifs disrupted by the index SNPs are in Additional file 13: Table S12. Further examples of disease-linked ASM loci are in Additional files 11: Table S10, Additional files 12: Table S11-Additional files 13: Table S12, and Additional file 2: Figure S7-S9, S22, and S23.

Table S11), which represents a moderate but significant enrichment (OR = 1.5, $p = 3.2 \times 10^{-22}$). Among these loci, a large majority showed ASM in cancers or cell types that approximate cancer precursor cells (e.g., B cells for lymphoma and multiple myeloma, glia for GBM, mammary or bladder epithelial cells for carcinomas, normal liver for hepatocellular carcinoma) and/or in T cells, which are relevant to cancer via immune surveillance. In these DMRs, 60 of the ASM index SNPs precisely coincided with the GWAS peak SNPs, supporting the candidacy of these statistically identified SNPs as functional rSNPs. Among the 268 ASM index SNPs, overlapping groups of 40 and 207 index SNPs altered ASM-correlated or enriched TF binding motifs, respectively, providing mechanistic leads to disease-associated transcriptional pathways (Additional file 12: Table S11). Interesting ASM index SNPs for this disease category include rs398206 associated with cutaneous melanoma and nevus counts via strong LD with GWAS SNPs rs416981 and rs45430 (nearest genes *FAM3B*, *MX2*, *MX1*), rs4487645 precisely coinciding with a GWAS peak SNP for multiple myeloma and immunoglobulin light chain amyloidosis (nearest genes *SP4*, *DNAH11*, *CDCA7L*; Fig. 7), rs3806624 precisely coinciding with a GWAS peak SNP for B cell lymphomas and multiple myeloma (nearest gene *EOMES*; Additional file 2: Figure S23), rs2853677 linked to lung cancer and other malignancies, as well as benign prostatic hyperplasia, via strong LD with several GWAS peak SNPs (genes *SLC6A18*, *TERT*, *MIR4457*, *CLPTM1L*; Fig. 7), and rs61837215 linked to breast cancer via LD with GWAS peak SNP rs2754412 (nearest genes *HSD17B7P2*, *SEPT7P9*, *LINC00999*; Additional file 2: Figure S23). Potentially informative examples in lenient blocks include rs2427290 linked to colorectal cancer via GWAS peak SNP rs4925386 (nearest genes *OSBPL2*, *ADRM1*, *MIR4758*, *LAMA5*, *RPS21*, *CABLES2*; Additional file 2: Figure S8) and rs2283639 linked to non-small cell lung cancer via GWAS peak SNP rs1209950 (nearest genes *LINC00114*, *ETS2*, *LOC101928398*; Additional file 2: Figure S9). Each of these index SNPs disrupts one or more strongly ASM-enriched and/or correlated TF binding motifs (Additional file 12: Table S11).

ASM index SNPs in LD with GWAS peaks for neuropsychiatric traits and disorders and neurodegenerative diseases

We found 210 ASM DMRs containing 225 index SNPs in strong LD ($R^2 > 0.8$) with GWAS peak SNPs for neurodegenerative, neuropsychiatric, or behavioral phenotypes (Additional file 13: Table S12). Among these ASM DMRs, about 15% showed ASM in brain cells and tissues (cerebral gray matter, neurons, glia), and a larger percentage showed ASM in immune system cell types. Both can be disease relevant, since studies have linked brain disorders not only to neuronal and glial cell processes but also to the immune system [47]. In addition, many loci in this list showed ASM in GBMs, which have partial glial and neuronal differentiation and may be revealing genetic variants that can affect early neuronal proliferation and differentiation. In these DMRs, 52 of the ASM index SNPs precisely coincided with the GWAS peak SNPs, supporting a functional regulatory role for these genetic variants, and overlapping groups of 36 and 183 index SNPs altered ASM-correlated or enriched binding motifs, respectively, providing mechanistic leads to disease-associated transcriptional pathways. Some interesting examples in strong LD with GWAS peaks for this general disease category (Additional file 13: Table S12) include rs1150668 linked to risk tolerance/smoking behavior and well-being spectrum via GWAS peak SNPs rs1150668 (coinciding with the ASM index SNP) and rs62620225 (nearest genes *ZSCAN16*, *ZKSCAN8*, *ZNF192P1*, *TOB2P1*, *ZSCAN9*; Fig. 7); rs2710323 that coincides with a GWAS peak SNP for schizoaffective disorder, anxiety behavior, bipolar disorder, and others (nearest genes *NEK4*, *ITIH1*, *ITIH3*, *ITIH4*, *MUSTN1*, *MIR8064*, *TMEM110*; Additional file 2: Figure S22); rs4976977 linked to intelligence measurement, anxiety measurement, schizophrenia, and unipolar depression via strong LD with GWAS peak SNP rs4976976 (nearest genes *MIR4472-1*, *LINC00051*, *TSNARE1*); and rs667897 linked to Alzheimer's disease via GWAS peak SNP rs610932 (nearest genes *MS4A2*, *MS4A6A*) and rs13294100, which coincides with a GWAS peak SNP for Parkinson's disease (nearest gene *SH3GL2*). Each of these index SNPs disrupts one or more strongly ASM-enriched and/or correlated TF binding motifs (Additional file 13: Table S12).

Visualization of the ASM mapping data as annotated genome browser tracks

In addition to the three major disease categories detailed above, we found several hundred high confidence ASM index SNPs in strong LD with GWAS peaks for pharmacogenetic phenotypes or for cardiometabolic diseases and traits (e.g., rs2664280 linked to type 2 diabetes mellitus via GWAS SNP rs2633310, Fig. 7). The final set of high-confidence recurrent ASM loci averaged 5 ASM DMRs per Mb of DNA genome wide. We provide the data both in tabular format (Additional file 3: Table S2) and as annotated genome browser tracks that include the most useful and mechanistically relevant parameters for each ASM index SNP. These parameters include ASM confidence and strength ranks, cell and tissue types with ASM, cancer vs normal status of the samples with ASM, and presence or absence of enriched CTCF or TF binding motifs and/or motifs with significant correlations of ASM strength with allele-specific differences in predicted binding affinity scores. An example of a 500-kb region of chromosome 19 containing 5 ASM DMRs, with ranks ranging from strong to weak and the strongest one encompassing a CTCF-bound insulator element, is in Additional file 2: Figure S24. These tracks (see the "Availability of data and materials" section) can be displayed, together with other relevant tracks, including chromatin structure for mechanistic studies and the GWAS catalog track for potential disease associations, in UCSC Genome Browser sessions [48].

Discussion

These data from dense mapping of ASM in normal human cell types and tissues, plus a group of cancers, identify 17,931 index SNPs in 15,112 DMRs that show strong and recurrent non-imprinted ASM, of which a substantial subset map within haplotype blocks that contain GWAS peaks for common diseases and related traits. In this study, we focused on finding strong and high-confidence ASM DMRs, each containing multiple CpGs passing ASM criteria, and each detected in at least two independent samples. Thus, we sought to maximize true-positive findings, which were borne out by a high validation rate using targeted bis-seq. In addition to the value of these data for disease-focused post-GWAS studies, this high yield of stringently defined ASM DMRs, and inclusion of both cancer and normal cell types and tissues, allowed us to test mechanistic hypotheses for the creation of allele-specific CpG methylation patterns in ways that have not been feasible with prior datasets.

A recent study by Onuchic et al. using Human Epigenome Project (HEP) data provided a map of ASM SNPs based on 49 WGBS from 11 donors (non-cancer tissues) and 2 cell lines [11]. Using their publicly accessible processed data, we identified a set of strong ASM SNPs that pass similar effect size and p value criteria as in our analysis (> 20% methylation difference and corrected p value < 0.05). For harvesting these candidate ASM loci from their dataset, we did not require multiple CpGs in each sequence contig to show ASM, since although in our criteria this is a requirement, it was not utilized as a criterion by Onuchic et al. Overall, 50% of our informative SNPs were also informative in the HEP dataset and 31% of our ASM index SNPs passed the above criteria for ASM in the HEP WGBS data. Given the differences in analytical methods and the differences in numbers and tissue types of the individuals analyzed, this is an encouraging convergence of findings. At the same time, this comparison indicates that our dataset adds substantial new information. With even greater numbers of individuals (informative heterozygotes at more SNPs), additional cell and tissue types, and greater depth of WGBS, additional loci with ASM will be identified. Our data already reveal a large component of rare or "private" ASM with a substantial subset showing a strong ASM magnitude. Indeed, some of the ASM loci identified and validated by targeted bis-seq in our previous smaller study [8] are not included in our current list of recurrent ASM DMRs because they passed ASM criteria in only one individual. Conversely, as expected based on the requirement for multiple individuals when using a methylation QTL (mQTL) approach to detect ASM, the current ASM dataset now encompasses a larger percentage of the set of mQTLs identified in that prior study.

Allele-specific binding of TFs and CTCF has been detected at up to 5% of assessed genomic sites [41], and the data provided here bolster and refine previous results from us and others [8, 9, 11, 30, 32, 33, 49] implicating a major role for binding site occupancies in shaping both net and allele-specific DNA methylation patterns in human cells. The harvest of large numbers of strong and high-confidence ASM occurrences in this study facilitated our analysis of individual (not pooled) binding motifs, thereby producing a statistically robust list of specific ASM-correlated CTCF and TF binding motifs, nearly all of which show anti-correlated (i.e., inversely correlated) behavior in which greater predicted binding site affinity and site occupancy tracks with less methylation of CpGs on that allele—which can be heuristically understood as protection of the occupied binding site from methylation.

The set of CTCF and TF binding motifs that we find to be strongly correlated with ASM when they contain disruptive SNPs overlaps only partly with the ASM-correlated motifs identified in the HEP study [11]. Encouragingly, certain classes of motifs emerge as significantly correlated in both studies. However, in addition to some differences in the identities of the most strongly correlated and enriched motifs or motif classes, a general difference between the conclusions of the two studies concerns the numbers of motifs showing positive vs negative directions of the correlations. The HEP investigators reported a substantial minority subset (approximately 30%) of motifs for which higher predicted binding affinity was found to correlate with greater CpG methylation (i.e., positively or directly correlated behavior). In our dataset, using our ASM criteria and analytical pipeline, we find a nearly complete absence of such occurrences. All but 1 of the 144 motifs that are both enriched and significantly ASM-correlated (Additional file 10: Table S9) show an inversely correlated direction of the relationship, such that higher predicted binding affinity (greater predicted binding site occupancy) tracks with relative CpG hypomethylation. When we only require ASM correlation, without enrichment as a criterion (Additional file 9: Table S8), we find 175 motifs with this inversely correlated behavior, but only two motifs with positively correlated behavior in which greater predicted binding site occupancy tracks with CpG hypermethylation. Our combined ASM and ASB analysis, using ENCODE ChIP-seq data in the GM12878 LCL, also showed a strong enrichment of inverse correlations between binding and methylation levels. Interestingly, however, in our small set of two positively correlated motifs, we find the YY1 binding motif, which was also found by the HEP investigators in their positively correlated subset. This finding makes biological sense since the YY1 TF, acting as a component of the PRC2 polycomb repressive complex, can attract CpG methylation, at least partly through recruitment of DNA methyltransferases [50].

An advance in the current study is our ability to test and compare mechanisms of ASM in normal and neoplastic cells. We observed a dramatic increase in per sample ASM frequencies, on average, in the primary cancers compared to cell lineage-matched normal cells and to non-cancer samples overall. This increase was paralleled by a more modest but still significant increase in ASM frequency in whole placental tissue and in purified trophoblast, which, as shown here and in other studies [8, 21, 23], have global CpG hypomethylation similar to cancers. Special aspects of ASM detected in the cancers included allele-specific hypomethylation genome-wide and allele-specific hypermethylation at loci in poised chromatin, as well as relatively increased ASM in chromatin desert regions and increased allele-switching at ASM loci. Despite these differences, our findings from testing for enrichment of TF and CTCF binding motifs and correlations of ASM with destructive SNPs in these motifs clearly indicate that the same binding site occupancy mechanism pertains in both normal and cancer-associated ASM. A striking additional result that supports this shared mechanism, and which may have important implications for cancer biology, is our finding of de novo ASM affecting CpGs clustered around somatic point mutations in cancer cells. The key mechanistically informative feature of this de novo ASM is that it preferentially occurs around mutations that disrupt the same classes of TF binding motifs that are linked to ASM in normal cells. While this topic will need future work, we can speculate that some of these non-coding mutations, which are declaring their functionality by producing the observed de novo ASM, might play roles in cancer biology through effects on gene expression. A possible example is the *TEAD1* gene, which is known to be over-expressed in aggressive and treatment-refractory cases of multiple myeloma [51] and which showed de novo ASM in its upstream enhancer region in a multiple myeloma case in our series, via gain of a new TF binding motif on the mutated allele (Fig. 5).

Based on the shared general mechanism of ASM in cancer and normal cells, an important practical conclusion is that analyzing combined series of cancer cases plus non-cancer samples increases the power of ASM mapping for finding mechanistically informative rSNPs. In conjunction with GWAS data, these rSNPs can point to genetically regulated transcriptional pathways that underlie inter-individual differences in susceptibility not only to cancers but also to nearly all common human non-neoplastic diseases. Due to the LD structure of the genome, GWAS peaks by themselves can only point to disease-associated haplotype blocks, with all SNPs in strong LD with the causal SNP(s) showing similar correlations to the phenotype. Therefore, additional types of evidence are needed before causal roles can be attributed to GWAS peak SNPs or to other SNPs in strong LD with them. ASM mapping can pinpoint candidate rSNPs that declare their presence by conferring the observed physical asymmetry in CpG methylation between the two alleles. The key finding that supports such mapping for biologically meaningful rSNP discovery is the one above, namely that ASM is caused by disruptive SNPs in TF and CTCF binding sites.

This situation is highlighted by our findings for ASM index SNP rs4487645 (Fig. 7), which coincides with a GWAS peak for AL amyloidosis and multiple myeloma and disrupts an ENCODE PAX5 discovery motif (PAX5_disc3) that is significantly enriched among ASM loci. Since the PAX5 TF is a master regulator of B cell development [52], these ASM mapping data are post-GWAS evidence suggesting involvement of a relevant biological pathway in susceptibility to multiple myeloma, a B cell malignancy. That the ASM at this locus was specifically found in a sample of DLBCL (another type of B cell cancer) highlights the usefulness of including primary tumor samples in ASM mapping. Another example is the ASM index SNP rs2283639 is linked to lung cancer GWAS peak SNP rs1209950. This ASM index SNP is situated in the promoter/enhancer region of the *ETS2* gene, where it disrupts an ASM-enriched ETS1_3 TF binding motif (Additional file 2: Figure S9). A promising example in a non-neoplastic disease is provided by ASM index SNP rs2664280, which disrupts multiple ASM-enriched and ASM-correlated JUNB and AP1 binding motifs (all with greater predicted binding affinity on the REF allele) and is in strong LD with a GWAS peak SNP for psoriasis (Fig. 7). For this example, the ASM was found in T cells, which are relevant for psoriasis, and the candidacy of the JUNB motif disruption as a biological explanation for the disease association is supported by other evidence for involvement of AP1-dependent transcriptional changes in this disease [53]. These situations can be tested further by functional experiments such as CRISPR/Cas9-mediated DNA deletions in ASM DMRs and mutations of ASM index SNPs in appropriate cell types.

Lastly, regarding the non-redundancy of ASM mapping as a post-GWAS approach, while SNPs with experimental evidence for ASB are strongly enriched among the ASM loci reported here, more than 90% of the ASM index SNPs harvested in this study lack currently available ASB annotations. Thus, maps of ASM, which are readily generated from large archival collections of DNA samples, can provide information about rSNPs that has not emerged from other types of mapping data, such as ChIP-seq for ASB, which require whole cells or tissue samples and are more technically difficult to obtain. That ASM data are largely non-redundant with other post-GWAS modalities (ASB, chromatin states and chromatin accessibility, eQTLs) is further highlighted by our observation of ASM DMRs in chromatin deserts. Our finding of similar correlations of ASM with destructive SNPs in specific TF binding motifs in both non-desert and desert regions suggests that mapping ASM in deserts can pinpoint candidate rSNPs in cryptic TF binding sites, which were presumably active at earlier stages of cell differentiation and have left "methylation footprints" that can be detected as ASM but cannot be found using other mapping methods.

Conclusions

We mapped ASM genome-wide in DNA samples including diverse normal tissues and cell types from multiple individuals, plus three types of cancers. The data reveal 15,115 high-confidence ASM regions, of which 1842 contain SNPs in strong LD or precisely coinciding with GWAS peaks for human diseases and traits. We find that ASM is increased in cancers, due to widespread allele-specific hypomethylation and focal allele-specific hypermethylation in regions of poised chromatin, with cancer-associated epigenetic variability manifesting as increased allele switching. We also report rare but informative de novo ASM due to somatic mutations in TF binding sites in cancers. Despite these cancer-specific phenomena, enrichment and correlation analyses indicate that destructive SNPs in specific classes of CTCF and TF binding motifs are a shared mechanism of ASM in normal and cancer cells and that this mechanism also underlies ASM in "chromatin deserts," where other post-GWAS mapping methods have not been informative. We provide our dense ASM maps as genome browser tracks and show examples of ASM index SNPs that are in LD with GWAS peaks and disrupt TF binding motifs, thereby nominating specific transcriptional pathways in the pathogenesis of autoimmune and cardiometabolic diseases, neuropsychiatric disorders, and cancers.

Materials and methods

Human cells and tissues

Human tissues and cell types analyzed in this study are listed in Additional file 1: Table S1. The Agilent SureSelect series included 9 brain (cerebral cortex), 6 T cell (CD3+), 3 whole peripheral blood leukocyte (PBL), 2 adult liver, 1 term placenta, 2 fetal heart, 1 fetal lung, and one ENCODE lymphoblastoid cell line (LCL; GM12878). All samples were from different individuals, except for a trio among the brain samples consisting of one frontal cortex (Brodmann area BA9) and two temporal cortex samples (BA37 and BA38) from the same autopsy brain. We performed WGBS on 16 normal T cell preparations (10 CD3+, 4 CD4+, and 2 CD8+), 10 B cell samples (CD19+), 7 monocyte (CD14+) and 2 monocyte-derived macrophage samples, 2 PBL, 1 reactive lymph node, 4 fractionated samples from a term placenta (whole tissue from the chorionic plate, purified villous cytotrophoblast from chorionic plate and basal plate, and extravillous trophoblast from basal plate), 3 adult liver, 2 primary bladder epithelial cell cultures, 2 epithelium-rich non-cancer tissue samples from breast biopsies, 3 primary mammary epithelial cell cultures, 3 frontal cerebral cortex gray matter samples, 6 NeuN+ FANS-purified cerebral cortex neuron preparations, 4 NeuN− FANS-purified cerebral cortex glial cell preparations, 1 LCL (GM12878), 3 B cell lymphomas (1 follicular and 2 diffuse large B cell type), 7 multiple myeloma cases (CD138+ cells from bone marrow aspirates), and 6 cases of glioblastoma multiforme (GBM). The glia samples were paired with neuron preparations from the same autopsy brains, and several of the B cell, PBL, monocyte/macrophage, and T cell samples were from the same individuals (Additional file 1: Table S1). In the combined series, 5 samples were assessed by both SureSelect and WGBS (Additional file 1: Table S1). Peripheral blood samples were obtained with informed consent, and CD3+ T lymphocytes, CD19+ B lymphocytes, and CD14+ monocytes were isolated by negative selection using RosetteSep kits (Sigma). Macrophages were produced from monocytes by culturing in RPMI with 20% fetal calf serum with 50 ng/ml M-CSF for 1 week as described [54]. Fractionation of villous cytotrophoblast and extra-villous trophoblast from a term placenta was carried out as previously described [55]. All other non-neoplastic primary human tissues were obtained from autopsies. Neuronal and glial cell nuclei were prepared from autopsy brains using tissue homogenization, sucrose gradient centrifugation, and fluorescence-activated nuclear sorting (FANS) with a monoclonal anti-NeuN antibody [56] and documented for purity of cell types by immunostaining of cytospin slides, as shown previously [57]. Biopsy samples of human lymphomas and GBMs, and CD138+ multiple myeloma cells isolated from bone marrow biopsies by positive selection on antibody-conjugated magnetic beads (Miltenyi Biotec), were obtained with I.R.B. approval in a de-identified manner from the Tissue Biorepository of the John Theurer Cancer Center. Absence of circulating myeloma cells in the paired B cell samples was verified by cytopathology and by the absence of DNA copy number aberrations that were seen in the multiple myeloma cells. Among the 6 GBMs, we did not detect cases with a strong CpG island hypermethylator phenotype (CIMP) as defined by Noushmehr et al. [58], which is expected given that CIMP is more frequent in low-grade gliomas than in high-grade GBMs. In surgical specimens, GBM cells are mixed with non-neoplastic glial and vascular cells, but the presence of malignant cells in each GBM sample was confirmed by histopathology on sections of the tissue blocks and was verified by assessing DNA copy number using normalized WGBS read counts [57], which revealed characteristic GBM-associated chromosomal gains and losses. The GM12878 lymphoblastoid cell line DNA was purchased from Coriell, which performs cell line authentication using STR assays, primary cultures of non-neoplastic human urinary bladder epithelial cells (cytokeratin

18^+ and TE-7^-) were purchased from A.T.C.C. and Cell Applications, Inc., and primary cultures of non-neoplastic human mammary epithelial cells (cytokeratin 18^+) were purchased from Cell Applications, Inc. and ScienCell Research Laboratories.

Agilent SureSelect methyl-seq and WGBS

We used the Agilent SureSelect methyl-seq DNA hybrid capture kit according to the manufacturer's protocol to analyze methylomes in a total of 27 non-neoplastic cell and tissue samples (Additional file 1: Table S1). In this protocol, targeted regions (total of 3.7 M CpGs) including RefGenes, promoter regions, CpG islands, CpG island shores, shelves, and DNAse I hypersensitive sites are sequenced to high depth. DNA was sheared to an average size of 200 bp and bisulfite converted with the EZ DNA methylation kit (Zymo). Paired-end reads (100, 150, or 250 bp) were generated at the Genomics Shared Resource of the Herbert Irving Comprehensive Cancer Center of Columbia University, with an Illumina HiSeq2500 sequencer.

For analyzing complete methylomes in the normal and tumor samples, plus the GM12878 LCL, WGBS was performed at the New York Genome Center (NYGC), MNG Genetics (MNG), and the Genomics Shared Resource of the Roswell Park Cancer Institute (RPCI), as indicated in Additional file 1: Table S1. The NYGC used a modified Nextera transposase-based library approach. Briefly, genomic DNA was first tagmented using Nextera XT transposome and end repair was performed using 5mC. After bisulfite conversion, Illumina adapters and custom bisulfite converted adapters are attached by limited cycle PCR. Two separate libraries were prepared and pooled for each sample to limit the duplication rate and sequenced using Illumina X system (150 bp paired-end). WGBS performed at MNG used the Illumina TruSeq DNA Methylation Kit for library construction according to the manufacturer's instructions and generated 150 bp paired-end reads on an Illumina NovaSeq machine. WGBS performed at RPCI utilized the ACCEL-NGS Methyl-Seq DNA Library kit for library construction (Swift Biosciences) and generated 150 bp paired-end reads on an Illumina NovaSeq.

Read mapping, SNP calling, and identification of ASM DMRs

Our analytical pipeline is diagrammed in Additional file 2: Figure S1. Compared with our previous study [8], updates included improvements in sequence processing, updated database utilization and increased stringency for SNP quality control, assignment of both strength and confidence scores to ASM index SNPs, use of updated ENCODE and JASPAR databases (http://compbio.mit.edu/encode-motifs/, [60]) for scoring the effects of the ASM index SNPs on predicted TF binding affinities, and utilization of haplotype blocks and LD criteria, instead of simple distance criteria around GWAS peaks for nominating disease-associated rSNPs in ASM DMRs. 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 [59] with paired-end mode and default setting allowing about 3 mismatches in a 150 bp read. For the SureSelect methyl-seq samples, unpaired reads after trimming were aligned separately using single end-mode and the same settings. Duplicate reads were removed using Picard tools [60] and reads with more than 10% unconverted CHG or CHH cytosines (interquartile range: 0.1–2.2% of mapped reads; median 0.14%) were filtered out. Depth of sequencing for each sample in Additional file 1: Table S1, with metrics calculated using Picard tools. SNP calling was performed with BisSNP [61] using default settings, except for the maximum coverage filter set at 200 to encompass deep sequencing while avoiding highly repetitive sequences, and quality score recalibration. SNP calling was carried out using human genome GRCh37 and dbSNP147 as references (ADD PMID: 21478889 and PMID: 11125122). For ASM calling, only heterozygous SNPs are informative. We filtered out heterozygous SNPs with less than 5 reads per allele. In addition, SNP with multiple mapping positions were filtered out, as well as SNPs with more than one minor allele with allele frequency > 0.05. Informative SNPs were defined as heterozygous, bi-allelic, and uniquely mapped SNPs that did not deviate significantly from Hardy-Weinberg equilibrium based on exact tests corrected for multiple tests (FDR < 0.05 by HardyWeinberg R package) and were covered by more than 5 reads per allele. In addition, we filtered out any informative regions mapping ENCODE defined "blacklisted" regions [62]. Informative regions were defined as regions with overlapping reads covering at least one informative SNP. Bisulfite sequencing converts unmethylated C residues to T, while methylated C residues are not converted. Therefore, for C/T and G/A SNPs, the distinction between the alternate allele and bisulfite conversion is possible only on the non-C/T strand. For SureSelect methyl-seq, since only negative-stranded DNA fragments are captured, G/A SNPs were filtered out; for WGBS, C/T and G/A SNPs were assessed after filtering out reads mapping to the C/T strand.

ASM calling was performed after separating the SNP-containing reads by allele. For each heterozygous SNP, all reads overlapping the 2 kb window centered on the SNP were extracted using Samtools. Given the median insert size of our libraries (~ 200 bp), the use of a 2 kb window instead of the SNP coordinate allows extraction, in most cases, of both paired ends even if the SNP is only covered at one of the ends. SNP calling is performed on each paired read and read IDs are separated into two files as reference (REF) and alternate (ALT) alleles using R. After Bismark methylation extractor is applied, CpG methylation calls by allele are retrieved using allele tagged read IDs. Paired reads with ambiguous SNP calling (i.e., called as REF allele on one paired end and ALT allele on the other) were discarded. For Nextera WGBS, due to the fill-in reaction using 5mC following DNA tagmentation which affects the 10 first base pairs (bp) on 5′ of read 2, methylation calling for Cs mapping to these bp was not considered. In addition, a slight methylation bias due to random priming and specific to each library kit was observed within the last 2 bp on 3′ of both paired ends for Nextera WGBS, within the first 10 bp on 5′ of both paired ends and the last 2 bp on 3′ of read 2 for TruSeq WGBS, and within the first 10 bp on 5′ of read 2 for ACCEL-NGS WGBS. Therefore, methylation calls in these windows were ignored.

To further increase the stringency and accuracy of ASM calling, only regions with at least 3 CpGs covered by more than 5 reads per allele were considered. ASM CpGs were then defined as CpGs with Fisher's exact test *p* value < 0.05 and ASM DMRs were defined as regions with $\geq$ 20% methylation difference after averaging all CpGs covered between the first and last CpGs showing ASM in the region, a Wilcoxon *p* value corrected for multiple testing by the B-H method < 0.05 (FDR at 5%), and more than 3 ASM CpGs including at least 2 consecutive ASM CpGs. CpGs destroyed by common SNPs (maf > 0.05) were filtered out from both CpG and DMR level analyses. Very close or overlapping DMRs (< 250 intervening bp) were merged into one unique DMR.

We ranked the ASM SNPs using two approaches, one based on confidence/recurrence criteria and the other on percent difference in methylation of the two alleles (ASM strength). For the confidence rank, we used the geometric mean of the average coverage of each allele, the number of samples showing ASM, and the percentage of these samples among all heterozygous (informative) samples. For the strength rank, we used the geometric mean of the methylation difference, number of ASM CpGs, and percentage of ASM CpGs among all covered CpGs. An overall rank was calculated using the geometric mean of these two ranks. ASM DMRs dictated by multiple index SNPs were ranked by the top-scoring SNP. ASM calling and ranking were performed using R and Stata 15. We used the GeneImprint database to flag and exclude from downstream analyses all ASM DMRs that mapped within 150 Kb windows centered on the transcription starting site of all known high confidence imprinted genes, including in this list the *VTRNA2-1* gene, which we have previously shown to be subject to parental imprinting in trio samples [38] and which showed frequent allele switching in normal samples in the current dataset, consistent with imprinting (Additional file 5: Table S4).

Lastly, although varying levels of non-CpG methylation (mCH) have been observed in human and mouse tissues, and this non-canonical methylation appears to have unique sub-chromosomal distributions and biological functions [63], for clarity, the current report is focused only on ASM affecting classical CpG methylation. Nonetheless, giving confidence in our dataset, we found mCH to be higher in the purified cerebral cortical neurons, (2.4% +/− 0.9%, $N = 16$) than in the non-neuronal samples (0.47% +/− 0.54%, $N = 43$), which is consistent with findings from another laboratory [64, 65].

Somatic mutation calling

Somatic mutation calling was performed on the 4 multiple myeloma samples for which paired normal peripheral blood B cells from the same individuals had been bis-sequenced using the same library preparation (ACCEL-NGS WGBS). We used BisSNP (with the same setting as for SNPs but without providing reference SNP dataset) to call all heterozygous variants for both myeloma and normal B cell samples. We then filtered out any variants reported as germline SNPs by DbSNP147. Variants mapping to ENCODE blacklisted regions were removed, and we next filtered out any variants that were present in the paired B cell samples. We used a sequencing coverage requirement for candidate mutations in the myeloma cases of 10× per allele (wild-type, mutant).

Targeted bisulfite sequencing (bis-seq) for validations of ASM

Targeted bis-seq was utilized for validation of ASM regions. PCR primers were designed in MethPrimer [60], and PCR products from bisulfite-converted DNA samples were generated on a Fluidigm AccessArray system as described previously [8], followed by sequencing on an Illumina MiSeq. PCRs were performed in triplicate and pooled to ensure sequence complexity. ASM was assessed when the depth of coverage was at least 100 reads per allele. While the absolute differences between methylation of the two alleles are not exaggerated by deep sequencing, the p values for these differences tend to zero as the number of reads increases. Therefore, to avoid artificially low p values, we carried out bootstrapping (1000 random samplings, 50 reads per allele), followed by Wilcoxon tests for significance. Samplings and bootstrapping were performed using R. The tested ASM loci and amplicon coordinates are in Additional file 7: Table S6.

Annotation and enrichment analysis of ASM loci

To annotate ASM and informative SNPs, we defined small (1000 bp) and large (150 kb) windows centered on each index SNP. The small windows were used to assess mechanistic hypotheses involving local sequence elements and chromatin states and the large windows were used for functional annotation (genes and GWAS associated SNPs). We used BedTools to intersect the genomic coordinates of ASM windows to the coordinates of the annotation sets. From the UCSC Genome Browser (PMID: 12045153) (GRCh37 assembly), we downloaded RefSeq annotations, DNase hypersensitive sites, TF peaks by ChIP-seq, and chromatin state segmentation by HMM in ENCODE cell lines (https://www.encodeproject.org/). Chromatin state segmentation for relevant human primary cells and tissues (T cells CD3, T cells CD4, T cells CD8, B cells, monocytes, and cerebral cortex) were downloaded from the Roadmap Epigenomics project (PMID: 25693563). We allowed multiple chromatin states at a single location when different states were present in different cell lines. Distances between ASM loci and genes were calculated from the transcription start sites. Regulome scores were downloaded from RegulomeDB [44]. For each relevant feature, enrichment among ASM index SNPs compared to the genome-wide set of informative SNPs (SNPs that were adequately covered and heterozygous in at least 2 samples) was tested using bivariate logistic regressions. To compare characteristics of ASM observed only in cancer samples ("cancer-only ASM") vs ASM observed in at least one non-cancer sample ("normal ASM"), these analyses were stratified by cancer status. To assess enrichment for chromatin states among ASM loci that were found only in cancers or only in non-cancer samples, with the occurrences divided into subsets according to the direction of the change in methylation in the cancers compared to cell lineage-matched normal samples, we used the same approach but considering only the sets of heterozygous SNPs informative in both myelomas and B cells, or lymphomas and B cells, or GBMs and glia. To compare the regulatory features of ASM to those of other allele-specific marks, we performed similar analyses for enrichment of ASM index SNPs in the sets of publicly available eQTLs [66] and ASB SNPs [41, 67] that were informative in our dataset.

Tests for correlations of ASM with SNPs in TF and CTCF binding sites

To test for correlations of ASM with destructive SNPs in TF binding motifs, we used position weight matrices (PWMs) of TF motifs from ENCODE ChIP-seq data [36], (http://compbio.mit.edu/encode-motifs/), as well as PWMs from the JASPAR database [39, 66]. We scored allele-specific binding affinity at each index SNP using the atSNP R package [42], which computes the B-H corrected p values (i.e., q values) of the affinity scores for each allele and q value of the affinity score differences between alleles. Motifs that contained SNPs affecting allele-specific TF binding affinity were defined as motifs with a significant difference in binding affinity scores of the two alleles (q value < 0.05) and a significant binding affinity score in at least one allele (p value < 0.005). For each TF occurrence, the binding scores per allele were estimated using PWM scores calculated as described in our earlier study [8]. In addition, among the ASM index SNPs, we specifically annotated TF binding motifs that overlapped with cognate TF ChIP-seq peaks based on ENCODE data (https://www.encodeproject.org/). For each motif, we used data from Kheradpour and Kellis [36] (http://compbio.mit.edu/encode-motifs/) to define the cognate TF peaks, required a 10-fold enrichment of the motif among ASM loci compared to background, and filtered out TF peaks with less than 10 occurrences of the tested motif among ASM loci.

To test whether ASM index SNPs are enriched in variants that disrupt polymorphic TF binding motifs, we used logistic regressions to calculate ORs for each disrupted polymorphic motif. Enrichment was defined as an OR > 1.5 and B-H corrected p value < 0.05. Since computing resources required to run atSNP for > 2 million SNPs and > 2000 TF motifs are extremely large, we random sampled 40,000 non-ASM informative SNPs (1,3 ASM vs non-ASM SNP ratio) to estimate the random expectation of each TF motif occurrence. To test whether the disruption of TF binding sites could be a mechanism of ASM, we correlated the difference in PWM scores between alleles of each occurrence of a given TF motif disrupted by an ASM index SNP to the differences in methylation levels between the two alleles, using linear regression. Only TF motifs with more than 10 disrupted occurrences in ASM regions were analyzed. For index SNPs showing ASM in multiple samples, we used the average methylation difference between the two alleles. For each TF motif, a significant correlation of ASM with predicted TF binding affinity differences between the two alleles was defined as FDR < 0.05 and $R^2 > 0.4$.

To ask whether the correlations between ASM and predicted TF binding affinity differences between alleles might be similar for ASM loci found only in cancers compared to ASM loci that were observed in at least one normal sample, and to ask this same question for chromatin desert ASM vs non-desert ASM regions, we used a multivariate mixed model with random slope and intercept, with pooling of TF motifs to reach sufficient power (number of occurrences used for the regression). TF motifs with less than 10 occurrences total, or less than 3 occurrences in any ASM class, were filtered out. TF motifs included in the final mixed models for the four classes of ASM loci were pre-selected from the bivariate model (performed without distinction of ASM class; requiring FDR < 0.05 and $R^2 > 0.4$). To not bias the analysis toward TF motifs without any ASM class effect (which might be overrepresented in the set of significant TF motifs identified in the bivariate analyses), we also screened each TF motif, including CTCF motifs, using separated multivariate linear fixed models to include any motifs showing no correlation overall but a correlation trend only in one of the ASM classes (FDR < 0.05 for at least one of the ASM classes, multivariate model adjusted $R^2 > 0.4$).

We defined chromatin deserts as 1 kb genomic windows, centered on ASM index SNPs, which contained no DNAse peaks or only one DNAse peak among the 122 ENCODE cell lines and tissues, and no strong active promoter/enhancer, poised, or insulator chromatin state in any ENCODE sample. The multivariate mixed model accounts for both intra- and inter-TF motif error terms and includes the predicted TF binding affinity difference, either for two classes of ASM loci (non-cancer ASM and cancer ASM) or 4 classes of ASM loci (non-cancer ASM in non-desert regions, non-cancer ASM in desert regions, cancer ASM in non-desert regions, and cancer ASM in desert regions), the interaction between ASM class and binding affinity as fixed explanatory covariates for the methylation difference, and the TF motif as a random covariate. Marginal effects from predictions of the mixed model and Bonferroni-corrected *p* values were then computed to compare the correlation between ASM classes. The variation due to the TF motif was considered as a random effect, under the assumption that each TF motif might have a different intercept and slope. The interaction terms reflect the difference in the methylation to binding affinity correlation between each ASM class compared to the reference class, which we defined as non-cancer ASM for the 2-calss analysis and non-cancer ASM in non-desert region for the 4-class analysis. Analysis after excluding ASM loci that showed switching behavior gave similar results. TF motifs with significant correlations of disruptive SNPs with ASM for at least one of the 2 or 4 ASM classes (FDR < 0.05 and $R^2 > 0.4$) were then pooled to be tested in the final mixed model, such that the model was run using a total of 178 TF motifs with 16,609 motif occurrences disrupted by 3394 ASM SNPs for the 2 ASM-class analysis and a total of 62 TF motifs with 10,709 motif occurrences disrupted by 1967 ASM SNPs for the 4 ASM-class analysis. To assess ASB in the GM12878 cell line, ChiP-seq data for 154 TFs available for this cell line were downloaded from ENCODE (PMID: 29126249). For each TF, SNP genotyping and allele-specific read count were performed using the ChiP-seq alignment data for the set of high confidence ASM SNPs found in our GM12878 data and compared to data from WGBS. ASB SNPs were defined as SNPs showing homozygous genotype in the ChiP-seq data (but heterozygous in WGBS) with a significant allele-specific occupancy bias (FDR < 0.05, Fisher's exact test). All analyses were performed using R and STATA statistical software.

Associations of ASM with GWAS peaks

GWAS traits and associated supra and subthreshold SNPs ($p < 10^{-6}$) were downloaded from the NHGRI GWAS catalog [44, 68]. We defined haplotype blocks using 1000 Genomes phase 3 data [68] based on the method of Gabriel et al. for scoring linkage disequilibrium (LD) with emphasis on *D*-prime values [46] in PLINK [69]. To identify GWAS peaks in moderate LD with ASM index SNPs, we used relaxed criteria of *D*-prime confidence intervals (0.60–0.84) and historical recombination (0.55) but set the maximum haplotype block size at 200 kb to minimize large block calling in genomic regions lacking haplotype block structure. The blocks so defined have a median size of 46 kb. To identify ASM SNPs in strong LD with GWAS peak SNPs, we utilized the default parameters of Gabriel et al. for haplotype block calling [46]. The blocks so defined have a much smaller median size of 5 kb. Finally, we computed pairwise R^2 between our ASM SNPs and all GWAS SNPs within 200 kb. SNPs with high R^2 represent a subtype of SNPs in high LD where not only a non-random association (high D') is observed but where these SNPs can essentially be considered as proxies of each other. Statistical association between a GWAS SNP and trait can be directly imputed to any SNPs with very high $R^{2,}$ so such SNPs are obvious candidates for post-GWAS analyses. However, SNPs showing high D' but low R^2 with the GWAS SNP (which occurs when a rare SNP is in high LD with a more frequent SNP) might also contribute biologically to disease associations. We annotated each ASM index SNP for localization within these haplotype blocks, and for precise co-localization with a GWAS peak SNP or high R^2 (> 0.8), and tested for enrichment of ASM SNPs within these blocks, as well as among GWAS peak SNPs, using the same approach as described above for other genomic features.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-02059-3.

Additional file 1: Table S1. Biological samples analyzed in this study.

Additional file 2: Figure S1. Flow charts and diagram of computational and analytical approaches in this study. **Figure S2.** Summary of sample types and numbers and yield of informative SNPs. **Figure S3**. PCA of the combined WGBS and SureSelect methyl-seq data and series-wide overlap between ASM loci detected by the two methods. **Figure S4.** Distribution of ASM shows a high proportion of rare or private ASM in both cancer and normal samples and a significant increase in per-sample ASM in the cancers. **Figure S5.** Comparison of results from individual DNA samples analyzed by two WGBS library construction kits at two sequencing facilities, or by SureSelect and WGBS. **Figure S6.** Example of a chromosome region illustrating consistency between SureSelect methyl-seq, WGBS, and targeted bisulfite sequencing. **Figure S7.** Validations of ASM DMRs in disease-associated chromosomal regions: rs10411630 and multiple sclerosis. **Figure S8**. Validation of ASM DMRs in disease-associated chromosomal regions: rs2427290 and colorectal cancer. **Figure S9.** Validation of ASM DMRs in disease-associated chromosomal regions: rs2283639 and non-small cell lung carcinoma. **Figure S10.** Validations of ASM DMRs spanning a range of ASM ranks. **Figure S11.** Kernel density plots of methylation levels showing global hypomethylation and decrease in the percentage of highly methylated CpGs in cancers. **Figure S12.** Replication of the findings using WGBS from a single facility. **Figure S13.** Allele-specific losses of methylation leading to ASM in cancers. **Figure S14.** Kernel density plots of methylation level distributions showing statistically enriched instances of allele-specific gains of methylation in cancers. **Figure S15.** Shared ASM loci in cancer and non-cancer have similar ASM magnitude. **Figure S16.** Correlations between allelic TF binding affinity scores and ASM magnitude in the 4 classes of ASM loci. **Figure S17.** Examples of ASM DMRs in chromatin deserts. **Figure S18.** Models for inter-individual variability and allele-switching at ASM loci. **Figure S19.** The percentage of ASM loci that show switching behavior in cancers is smaller when considering only loci for which ASM is also detected in non-cancer samples. **Figure S20.** Examples of haplotype blocks defined by stringent and lenient parameters. **Figure S21.** Utility of D′ and R-square parameters for assessing candidate disease-associated rSNPs. **Figure S22.** Additional examples of mechanistically informative disease associated ASM index SNPs: autoimmune and neuropsychiatric. **Figure S23.** Additional examples of mechanistically informative disease associated ASM index SNPs: breast cancer and lymphoma. **Figure S24.** ASM loci displayed as annotated genome browser tracks.

Additional file 3: Table S2. ASM index SNPs and DMRs identified in this study and annotated for multiple relevant parameters.

Additional file 4: Table S3. Definitions of the terms in Table S2.

Additional file 5: Table S4. Known imprinted regions with ASM detected in this study.

Additional file 6: Table S5. New candidate imprinted regions and previously provisional imprinted loci with ASM detected in this study.

Additional file 7: Table S6. ASM loci tested for validations by targeted bisulfite sequencing.

Additional file 8: Table S7. Complete list of polymorphic CTCF and TF binding motifs found to be significantly enriched among ASM loci, requiring that the motif be disrupted by the ASM index SNP.

Additional file 9: Table S8. Complete list of CTCF and TF binding motifs that show significant correlations between allelic PWM scores and magnitude of ASM.

Additional file 10: Table S9. CTCF and TF binding motifs that show strong correlations of PWM scores with ASM and are also significantly enriched among ASM loci.

Additional file 11: Table S10. ASM index SNPs in strong LD or precisely coinciding with GWAS peak SNPs for immune-related diseases and phenotypes.

Additional file 12: Table S11. ASM index SNPs in strong LD or precisely coinciding with GWAS peak SNPs for cancer susceptibility.

Additional file 13: Table S12. ASM index SNPs in strong LD or precisely coinciding with GWAS peak SNPs for brain-related diseases and phenotypes.

Additional file 14. Review history.

Review history

The review history is available as Additional file 14.

Peer review information

Anahita Bishop was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

CD and BT designed the research. CD, ELPD, MS, AC, HM, LM, AJ, PLN, SCD, and KC generated the data. CD, ELPD, and BT carried out data analysis. CD, ELPD, and SM set up the cloud computing platform. NPI, AMC, SL, CT, KM, LH, CM, SG, GK, DS, and BT obtained the research funding. CD, ELPD, and BT wrote the manuscript, with suggestions from AC, HM, LM, AJ, SCD, NPI, GB, SL, AMC, SM, PHRG, RF, CT, KM, LH, CM, AG, KC, SG, GK, and DS. All authors read and approved the final manuscript.

Funding

This work was supported by NIH grants R01 MH092580, R01 AG036040, R01 AG035020, DP3 DK094400, P50 CA098252, and P30 CA051008. Funding was also provided by a Sanofi Innovations Award Grant, the AVHero Fund, the Caroline Fund, and research infrastructure funding to the Hackensack-Meridian Health Center for Discovery and Innovation from Celgene.

Availability of data and materials

The Agilent SureSelect and WGBS data are available in NCBI/GEO (GSE137880 and GSE79148 [70, 71]). Custom genome browser tracks with annotated ASM loci can be searched and viewed at a UCSC browser session hosted by our laboratory (https://bit.ly/tycko-asm). The Human reference genome (GRCh37) was downloaded from the GATK Bundle (ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/) [72]. DbSNP147 annotation, ENCODE ChIP-seq peaks, DNAse peaks, and chromatin state segmentation were downloaded from UCSC human genome browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/) [73]. Chromatin state segmentation data for human primary cells and tissues were downloaded from the Roadmap Epigenomics project (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state) [34].
ENCODE ChIP-seq aligned data for GM12878 cell line were downloaded from https://www.encodeproject.org/ [74]. The imprinting gene list was downloaded from GeneImprint database https://www.geneimprint.com/site/genes-by-species [75]. RegulomeDB scores were downloaded from https://www.regulomedb.org [43]. AlleleDB datasets were downloaded from http://alleledb.gersteinlab.org/download/ [41]. The GRASP dataset was downloaded from https://grasp.nhlbi.nih.gov/Overview.aspx [66]. JASPAR and ENCODE motifs were downloaded through atSNP R packages [42]. The NHGRI GWAS catalog was downloaded from https://www.ebi.ac.uk/gwas/docs/file-downloads [44]. Processed ASM data from Onuchic et al. were downloaded from ftp://ftp.genboree.org/allelic-epigenome/ [11].

Ethics approval and consent to participate

Peripheral blood samples were obtained with informed consent under Columbia University I.R.B.-approved protocols AAAI0706 and 7302R. Primary tumor samples were collected by the John Theurer Cancer Center Tissue Biorepository and transferred to the laboratory with final pathological diagnoses in a de-identified manner under Hackensack University Medical Center I.R.B.-approved protocols Pro2108-0589 and Pro2018-0020. All experimental methods comply with the Helsinki Declaration.

Competing interests

The authors have no competing interests.

Author details

[1]Hackensack-Meridian Health Center for Discovery and Innovation, Nutley, NJ 07110, USA. [2]John Theurer Cancer Center, Hackensack University Medical Center, Hackensack, NJ 07601, USA. [3]Department of Medicine, Huddinge, Karolinska Institutet, SE-171 77 Stockholm, Sweden. [4]Department of Gynecology and Obstetrics, Johns Hopkins Medical Institutions, Baltimore, MD 21287, USA. [5]Division of Gastroenterology, Hepatology and Nutrition, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [6]Department of Obstetrics and Gynecology, Hackensack University Medical Center, Hackensack, NJ 07601, USA. [7]Department of Pathology & Cell Biology, Columbia University Medical Center, New York, NY 10032, USA. [8]Division of Gastroenterology and Celiac Center, Department of Medicine, Columbia University Medical Center, New York, NY 10032, USA. [9]Departments of Dermatology and Genetics and Development, Columbia University Medical Center, New York, NY 10032, USA. [10]Lombardi Comprehensive Cancer Center of Georgetown University, Washington, DC 20057, USA. [11]MNG Laboratories, Atlanta, GA 30342, USA. [12]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University Medical Center, New York, NY 10032, USA. [13]Department of Neurology, Columbia University Medical Center, New York, NY 10032, USA. [14]Departments of Psychiatry and Behavioral Medicine and Obstetrics and Gynecology, Columbia University Medical Center, New York, NY 10032, USA.

Received: 16 September 2019 Accepted: 27 May 2020
Published online: 29 June 2020

References

1. Barsh GS, Copenhaver GP, Gibson G, Williams SM. Guidelines for genome-wide association studies. PLoS Genet. 2012;8:e1002812.
2. Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, Hod E, Li K, Murty VV, Schupf N, Vilain E, et al. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. Nat Genet. 2008;40:904–8.
3. 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. 2010;86:196–212.
4. Tycko B. Mapping allele-specific DNA methylation: a new tool for maximizing information from GWAS. Am J Hum Genet. 2010;86:109–12.
5. Zhang D, Cheng L, Badner JA, Chen C, Chen Q, Luo W, Craig DW, Redman M, Gershon ES, Liu C. Genetic control of individual differences in gene-specific methylation in human brain. Am J Hum Genet. 2010;86:411–9.

6. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31:142–7.
7. 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. 2014;9:e98464.
8. Do C, Lang CF, Lin J, Darbary H, Krupska I, Gaba A, Petukhova L, Vonsattel JP, Gallagher MP, Goland RS, et al. Mechanisms and disease associations of haplotype-dependent allele-specific DNA methylation. Am J Hum Genet. 2016; 98:934–55.
9. Do C, Shearer A, Suzuki M, Terry MB, Gelernter J, Greally JM, Tycko B. Genetic-epigenetic interactions in cis: a major focus in the post-GWAS era. Genome Biol. 2017;18:120.
10. Cheung WA, Shao X, Morin A, Siroux V, Kwan T, Ge B, Aissi D, Chen L, Vasquez L, Allum F, et al. Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome. Genome Biol. 2017;18:50.
11. Onuchic V, Lurie E, Carrero I, Pawliczek P, Patel RY, Rozowsky J, Galeev T, Huang Z, Altshuler RC, Zhang Z, et al. Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. Science. 2018;361:6409.
12. Shi J, Marconett CN, Duan J, Hyland PL, Li P, Wang Z, Wheeler W, Zhou B, Campan M, Lee DS, et al. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nat Commun. 2014;5:3365.
13. Kadota M, Yang HH, Hu N, Wang C, Hu Y, Taylor PR, Buetow KH, Lee MP. Allele-specific chromatin immunoprecipitation studies show genetic influence on chromatin state in human genome. PLoS Genet. 2007;3:e81.
14. Cavalli M, Pan G, Nord H, Wadelius C. Looking beyond GWAS: allele-specific transcription factor binding drives the association of GALNT2 to HDL-C plasma levels. Lipids Health Dis. 2016;15:18.
15. Boumber YA, Kondo Y, Chen X, Shen L, Guo Y, Tellez C, Estecio MR, Ahmed S, Issa JP. An Sp1/Sp3 binding polymorphism confers methylation protection. PLoS Genet. 2008;4:e1000162.
16. Stern JL, Paucek RD, Huang FW, Ghandi M, Nwumeh R, Costello JC, Cech TR. Allele-specific DNA methylation and its interplay with repressive histone marks at promoter-mutant TERT genes. Cell Rep. 2017;21:3700–7.
17. Zhou B, Ho SS, Greer SU, Zhu X, Bell JM, Arthur JG, Spies N, Zhang X, Byeon S, Pattni R, et al. Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. Genome Res. 2019;29:472–84.
18. Zhou B, Ho SS, Greer SU, Spies N, Bell JM, Zhang X, Zhu X, Arthur JG, Byeon S, Pattni R, et al. Haplotype-resolved and integrated genome analysis of the cancer cell line HepG2. Nucleic Acids Res. 2019;47:3846-61.
19. Biadasiewicz K, Fock V, Dekan S, Proestling K, Velicky P, Haider S, Knofler M, Frohlich C, Pollheimer J. Extravillous trophoblast-associated ADAM12 exerts pro-invasive properties, including induction of integrin beta 1-mediated cellular spreading. Biol Reprod. 2014;90:101.
20. DaSilva-Arnold S, James JL, Al-Khan A, Zamudio S, Illsley NP. Differentiation of first trimester cytotrophoblast to extravillous trophoblast involves an epithelial-mesenchymal transition. Placenta. 2015;36:1412–8.
21. Gamage T, Schierding W, Hurley D, Tsai P, Ludgate JL, Bhoothpur C, Chamley LW, Weeks RJ, Macaulay EC, James JL. The role of DNA methylation in human trophoblast differentiation. Epigenetics. 2018;13:1154–73.
22. Kasak L, Rull K, Vaas P, Teesalu P, Laan M. Extensive load of somatic CNVs in the human placenta. Sci Rep. 2015;5:8342.
23. Nordor AV, Nehar-Belaid D, Richon S, Klatzmann D, Bellet D, Dangles-Marie V, Fournier T, Aryee MJ. The early pregnancy placenta foreshadows DNA methylation alterations of solid tumors. Epigenetics. 2017;12:793–803.
24. Skaar DA, Jirtle RL. Analysis of imprinted gene regulation. Methods Mol Biol. 2017;1589:161–83.
25. Gama-Sosa MA, Slagel VA, Trewyn RW, Oxenhandler R, Kuo KC, Gehrke CW, Ehrlich M. The 5-methylcytosine content of DNA from human tumors. Nucleic Acids Res. 1983;11:6883–94.
26. Feinberg AP, Gehrke CW, Kuo KC, Ehrlich M. Reduced genomic 5-methylcytosine content in human colonic neoplasia. Cancer Res. 1988;48:1159–61.
27. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006;125:315–26.
28. Easwaran H, Johnstone SE, Van Neste L, Ohm J, Mosbruger T, Wang Q, Aryee MJ, Joyce P, Ahuja N, Weisenberger D, et al. A DNA hypermethylation module for the stem/progenitor cell signature of cancer. Genome Res. 2012;22:837–49.
29. Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schubeler D. Identification of genetic elements that autonomously determine DNA methylation states. Nat Genet. 2011;43:1091–7.
30. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011;480:490–5.
31. Brinkman AB, Gu H, Bartels SJ, Zhang Y, Matarese F, Simmer F, Marks H, Bock C, Gnirke A, Meissner A, Stunnenberg HG. Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. Genome Res. 2012;22:1128–38.
32. Feldmann A, Ivanek R, Murr R, Gaidatzis D, Burger L, Schubeler D. Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. PLoS Genet. 2013;9:e1003994.
33. 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. 2014;10:e1004663.
34. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
35. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell. 2007;128:1231–45.
36. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. 2014;42:2976–87.
37. Schmitt AD, Hu M, Ren B. Genome-wide mapping and analysis of chromosome architecture. Nat Rev Mol Cell Biol. 2016;17:743–55.
38. Paliwal A, Temkin AM, Kerkel K, Yale A, Yotova I, Drost N, Lax S, Nhan-Chang CL, Powell C, Borczuk A, et al. Comparative anatomy of chromosomal domains with imprinted and non-imprinted allele-specific DNA methylation. PLoS Genet. 2013;9:e1003622.

39. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, Bessy A, Cheneby J, Kulkarni SR, Tan G, et al. JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018;46:D260–6.
40. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011;7:522.
41. Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M. A uniform survey of allele-specific binding and expression over 1000-genomes-project individuals. Nat Commun. 2016;7:11101.
42. Zuo C, Shin S, Keles S. atSNP: transcription factor binding affinity testing for regulatory SNP detection. Bioinformatics. 2015;31:3353–5.
43. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22:1790–7.
44. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47:D1005–12.
45. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, Silva TC, Groeneveld C, Wong CK, Cho SW, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018;362:6413.
46. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, et al. The structure of haplotype blocks in the human genome. Science. 2002;296:2225–9.
47. Debnath M, Berk M. Functional implications of the IL-23/IL-17 immune axis in schizophrenia. Mol Neurobiol. 2016;54: 8170–78.
48. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN, et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. 2019;47:D853–8.
49. Lipka DB, Wang Q, Cabezas-Wallscheid N, Klimmeck D, Weichenhan D, Herrmann C, Lier A, Brocks D, von Paleske L, Renders S, et al. Identification of DNA methylation changes at cis-regulatory elements during early steps of HSC differentiation using tagmentation-based whole genome bisulfite sequencing. Cell Cycle. 2014;13:3476–87.
50. Ko CY, Hsu HC, Shen MR, Chang WC, Wang JM. Epigenetic silencing of CCAAT/enhancer-binding protein delta activity by YY1/polycomb group/DNA methyltransferase complex. J Biol Chem. 2008;283:30919–32.
51. Riz I, Hawley RG. Increased expression of the tight junction protein TJP1/ZO-1 is associated with upregulation of TAZ-TEAD activity and an adult tissue stem cell signature in carfilzomib-resistant multiple myeloma cells and high-risk multiple myeloma patients. Oncoscience. 2017;4:79–94.
52. Medvedovic J, Ebert A, Tagoh H, Busslinger M. Pax5: a master regulator of B cell development and leukemogenesis. Adv Immunol. 2011;111:179–206.
53. Uluckan O, Guinea-Viniegra J, Jimenez M, Wagner EF. Signalling in inflammatory skin disease by AP-1 (Fos/Jun). Clin Exp Rheumatol. 2015;33:S44–9.
54. Martinez FO, Gordon S, Locati M, Mantovani A. Transcriptional profiling of the human monocyte-to-macrophage differentiation and polarization: new molecules and patterns of gene expression. J Immunol. 2006;177:7303–11.
55. DaSilva-Arnold SC, Zamudio S, Al-Khan A, Alvarez-Perez J, Mannion C, Koenig C, Luke D, Perez AM, Petroff M, Alvarez M, Illsley NP. Human trophoblast epithelial-mesenchymal transition in abnormally invasive placenta. Biol Reprod. 2018;99:409–21.
56. Matevossian A, Akbarian S. Neuronal nuclei isolation from human postmortem brain tissue. J Vis Exp. 2008;20:914.
57. Mendioroz M, Do C, Jiang X, Liu C, Darbary HK, Lang CF, Lin J, Thomas A, Abu-Amero S, Stanier P, et al. Trans effects of chromosome aneuploidies on DNA methylation patterns in human Down syndrome and mouse models. Genome Biol. 2015;16:263.
58. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Pan F, Pelloski CE, Sulman EP, Bhat KP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010;17:510–22.
59. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for bisulfite-Seq applications. Bioinformatics. 2011;27:1571–2.
60. Li LC, Dahiya R. MethPrimer: designing primers for methylation PCRs. Bioinformatics. 2002;18:1427–31.
61. Liu Y, Siegmund KD, Laird PW, Berman BP. Bis-SNP: combined DNA methylation and SNP calling for bisulfite-seq data. Genome Biol. 2012;13:R61.
62. Amemiya HM, Kundaje A, Boyle AP. The ENCODE blacklist: identification of problematic regions of the genome. Sci Rep. 2019;9:9354.
63. Keown CL, Berletch JB, Castanon R, Nery JR, Disteche CM, Ecker JR, Mukamel EA. Allele-specific non-CG DNA methylation marks domains of active chromatin in female mouse brain. Proc Natl Acad Sci U S A. 2017;114:E2882–90.
64. He Y, Ecker JR. Non-CG methylation in the human genome. Annu Rev Genomics Hum Genet. 2015;16:55–77.
65. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015;523:212–6.
66. Leslie R, O'Donnell CJ, Johnson AD. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics. 2014;30:i185–94.
67. Cavalli M, Pan G, Nord H, Wallerman O, Wallen Arzt E, Berggren O, Elvers I, Eloranta ML, Ronnblom L, Lindblad Toh K, Wadelius C. Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. Hum Genet. 2016;135:485–97.
68. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, GA MV, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526:68–74.
69. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
70. Do C, Dumont E, Salas M, Castano A, Mujahed H, Maldonado L, Singh A, DaSilva-Arnold S, Bhagat G, Lehman S, et al. Whole genome bisulfite sequencing and Genome-wide targeted methyl-seq: Allele-specific DNA methylation is increased in cancers and its dense mapping in normal plus neoplastic cells increases the yield of disease-associated regulatory SNPs. GSE137880. Gene Expression Omnibus. 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE137880.

71. Do C, Lang CF, Lin J, Darbary H, Krupska I, Gaba A, Petukhova L, Vonsattel JP, Gallagher MP, Goland RS, et al. Mechanisms and disease associations of haplotype-dependent allele specific DNA methylation: Methyl-seq data for the identification of hap-ASM. GSE79148. Gene Expression Omnibus. 2016. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79148. Accessed 1 Apr 2020.
72. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
73. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.
74. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, et al. The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46:D794–801.
75. The GeneImprint Database. https://www.geneimprint.com/site/genes-by-species. Accessed 1 Apr 2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.