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

*Correspondence: catherine.do@hmh-cdi.org or benjamin.tycko@hmh-cdi.org

Abstract

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

Results: We performed whole genome methyl-seq on diverse normal human cells and tissues from multiple individuals, plus a group of cancers (multiple myeloma, lymphoma, and glioblastoma multiforme). After excluding imprinted regions, the data pinpointed 11,233 high-confidence ASM differentially methylated regions (DMRs), of which 821 contained SNPs in strong linkage disequilibrium or precisely coinciding with GWAS peaks. ASM was increased 5-fold in the cancers, due to widespread allele-specific hypomethylation and focal allele-specific hypermethylation in regions of poised chromatin. Allele-switching at ASM loci was increased in the cancers, but destructive SNPs in specific classes of CTCF and transcription factor (TF) binding motifs correlated strongly with ASM in both normal and cancer cells. Allele-specific binding site occupancies from ChIP-seq data were enriched among ASM loci, but most ASM DMRs lacked such annotations, and some were found in otherwise uninformative "chromatin deserts".

Conclusions: ASM is increased in cancers but it is produced by shared underlying mechanisms in normal and neoplastic cells. Dense maps of ASM in normal plus cancer samples, provided here as genome browser tracks, uncover mechanistically informative rSNPs that are difficult to find by other approaches. We show examples of TF binding sites disrupted by these rSNPs that point to altered transcriptional pathways in autoimmune, 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, $p<5x10^{-8}$ for a GWAS, and this level can lead to the rejection of biological true-positives with sub-threshold 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 between the two alleles (i.e. ASM) 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 general 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 general importance of polymorphic CTCF and TF binding sites. 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 mechanistic involvement in ASM and limited the yield of candidate rSNPs in disease-associated chromosomal regions. Further, the unique strengths and potential non-redundancy of dense ASM mapping compared to other post-GWAS methods have not been assessed, and while a few studies have been done using targeted methyl-seq [15-18], the genome-wide features and mechanisms of ASM in human cancers 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 tissues and cell types from multiple individuals, plus three types of human cancers. 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

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

The biological samples in this study are listed in **Table S1**. Our experimental approach for identifying ASM DMRs, and our analytical pipeline for testing ASM mechanisms and nominating disease-associated rSNPs are diagrammed in **Figure S1**. The sample set included diverse tissues and purified cell types from multiple individuals, with an emphasis on immune system, 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) (**Table S1** and **Figure S2**). 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 included 9 brain (cerebral cortex), 7 T cell (CD3+), 3 whole peripheral blood leukocyte (PBL), 2 adult liver, 2 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 9 brain samples) that consisted of one frontal cortex (Brodmann area BA9) and two temporal cortex samples (BA37 and BA38) from the same autopsy brain.

To further increase the number of samples and cell types, and to obtain complete genomic coverage, we performed WGBS on 61 non-cancer and 14 cancer samples, including 16 T cell preparations (10 CD3+, 4 CD4+, and 2 CD8+), 9 B cell samples (CD19+), 4 monocyte (CD14+) and 2 monocyte-derived macrophage samples, 2 PBL, 1 reactive lymph node, 2 placental samples (whole tissue and purified villous cytotrophoblast from the same term placenta), 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, 4 frontal cerebral cortex grey matter samples, 6 NeuN+ FANS-purified cerebral cortex neuron preparations, 4 NeuN- FANS-purified cerebral cortex glial cell preparations, 1 ENCODE LCL (GM12878), 3 B cell lymphomas (1 follicular and 2 diffuse large B cell type), 6 multiple myeloma cases (CD138+ cells from bone marrow aspirates), and 5 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 (**Table S1**). While the two series were mostly distinct, 5 samples were assessed by both SureSelect and WGBS (**Table S1**). Numbers of mapped reads and depth of sequencing are in **Table S1**, and numbers of informative (heterozygous) SNPs are in **Figure S2**. For quality control we performed Principle Component Analysis (PCA) using net methylation values for single CpGs informative in both SureSelect and WGBS. This procedure revealed the expected segregation of samples according to cell and tissue type and cancer or non-cancer status. It also revealed some expected findings for cell lineages, particularly highlighting both similarities and differences in methylation patterns in the brain cells (whole cerebral cortex, glia, neurons) and the GBMs (**Fig. S3**).

Our analytical pipeline (**Figure S1**) 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 testing mechanistic hypotheses for ASM in normal cells and tissues and in cancers. 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. As shown in **Figure S4**, using these cut-offs we found a good yield of recurrent ASM regions, but also many more loci with ASM seen in only one sample. We utilized such rare or "private" ASM loci for analyzing per-sample ASM frequencies, but for most of our downstream analyses, focused on testing mechanisms and disease associations, we required ASM in at least two samples. Using these stringent criteria, in the combined SureSelect and WGBS dataset we found 11,233 recurrent ASM DMRs, tagged by 13,210 index SNPs, representing 0.57% 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 **Table S2**, with parameter definitions in **Table S3**.

ASM in imprinted chromosomal regions

While this study focuses mainly on non-imprinted SNP- or haplotype-dependent ASM, genomic imprinting also produces ASM, due to parent-of-origin dependent DNA methylation in a small number of imprinted chromosomal domains (~100 imprinted loci). Therefore, we used the GeneImprint database [19, 20] 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 (**Table S4).** Since a hallmark of ASM due to parental imprinting is 50/50 allele switching between individuals in unselected populations, 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 (Methods). The number of ASM DMRs decreases steeply when they are required to be found in many individuals since identifying such loci requires both a high number of informative individuals and highly recurrent ASM (**Fig. S4**). Accordingly, among the non-cancer samples 126 ASM DMRs outside of imprinted regions were identified as showing significant ASM in more than 10 individuals. Only 15/126 (12%) of this informative group of DMRs showed allele switching at a frequency of greater than or equal to 20 percent of individuals. In comparison, among ASM DMRs identified in our dataset and located in or near known imprinted genes, nearly all (34/35; 97%) showed high frequency allele switching, with an approximately 50:50 ratio, as expected for parental imprinting. These results suggest that most of the ASM loci identified by our genome-wide analysis reflect non-imprinted ASM, not ASM due to imprinting. Interestingly, even among the 15 very highly recurrent ASM loci with frequent allele switching in normal cells and tissues and located outside of validated imprinted domains, some (e.g. *IGF2R*, *IGF1R*) have been reported as imprinted in humans with inconsistent findings or variability. This small group of loci (**Table S5)** are not pursued further here but will be of interest for future testing of parent-of-origin dependent behavior using samples from families.

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 **Figure S5** for a DNA sample analyzed by both methods. Within the fraction of the genome that was adequately covered by both methods and contained informative SNPs, we found 1,590 (42.4%) shared ASM “hits” (**Fig. S3**). This substantial but partial overlap is expected, given that most ASM loci show a significant allelic methylation bias in some but not all individuals (**Table S2**). In addition, some ASM DMRs passed our statistical cutoffs in SureSelect but not in WGBS due to the greater sequencing depth of SureSelect in some regions. The majority (79%) of these loci were sub-threshold (i.e. showing at least 1 CpG with ASM) and showed an allelic methylation bias in the predicted direction in at least one individual in the WGBS data. Conversely, some of the genomic regions (44%) that were covered by both methods but revealed ASM DMRs only in the WGBS data were sub-threshold in SureSelect, with the smaller overlap partly due to inter-individual variation in ASM and inclusion of more individuals, and hence more informative heterozygotes, in the larger WGBS series. Based on these results, the current dataset provides dense maps of ASM but is still non-saturating.

To assess the true-positive rate for ASM calling more directly, we selected 18 ASM DMRs, spanning a range from high to low ASM strength and confidence scores, for targeted bisulfite sequencing (bis-seq). As summarized in **Table S6**, this procedure validated the presence of ASM in two or more independent samples, 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 83% (15/18) of the DMRs assayed (examples in **Figures S6-S9**). The three remaining loci were one very low-ranked DMR validated in the single available index case (**Fig. S9**), also with a concordant direction of the ASM, and two middle-ranked DMRs validated in only one out of two available index cases, with concordant direction of the ASM in the positive cases. This high overall validation rate by targeted bis-seq suggests a high true-positive rate of the genome-wide data.

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 **Figure 1**, when the numbers of ASM DMRs per sample were normalized to the numbers of informative SNPs and then graphed with samples grouped by normal and cancer status it became obvious that the number of ASM DMRs in cancers (multiple myeloma, lymphoma, GBM) is on average 5-fold greater than in non-neoplastic samples (Wilcoxon p=6.9x10^{-9}; **Fig. 2A**). These differences in the frequency of ASM between cancer and non-cancer are particularly convincing since our series included lineage-matched normal cell types for each of the three cancer types: non-neoplastic B cells for comparing to the diffuse large B cell lymphomas (DLBCL), follicular lymphoma (FL), and multiple myelomas and non-neoplastic glial cells for comparing to the GBMs. The increase in per-sample ASM loci was stronger for the multiple myelomas and lymphomas and weaker for the GBMs. 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 (5-fold greater than the average of the non-neoplastic samples, **Fig. 1**), 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 [21, 22] to evaluate mechanisms that could account for the gain of ASM in the cancers we first asked whether there might be an inverse correlation between global methylation levels and frequencies of ASM. Global genomic hypomethylation was observed in the LCL and the three types of primary cancers in our series (**Fig. 1** and **Fig. S10**). 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 obvious hypomethylation in the GBMs compared to normal glia (**Fig. S10**). In primary surgical specimens, GBM cells are nearly always 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 [23], which revealed characteristic GBM-associated chromosomal gains and losses. Across the entire series of cancer and non-cancer samples, we found a non-linear but strongly significant anti-correlation (i.e. inverse correlation) between per-sample ASM frequencies and global CpG methylation levels (Spearman's rho=-0.6 and p-value= $2.3\text{x}10^{-8}$). Arguing for global hypomethylation, not the malignant phenotype per se, as a main driving factor for increased ASM, the immortalized but euploid GM12878 LCL showed global hypomethylation and a high frequency of ASM, and even among the non-neoplastic and non-immortalized samples, those that showed modestly reduced global methylation (e.g. placenta and primary epithelial cell lines that had been expanded in tissue culture) showed slightly higher per-sample frequencies of ASM.

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 informative 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 **Figure 2** and **Figure S11,** this approach revealed that the average configuration was a relative loss of methylation (LOM) on one allele in the cancers. In 70% of cancer-only ASM occurrences in myelomas, 73% in lymphomas and 41% in GBMs, a strongly "hypermethylated/hypermethylated" configuration of the two alleles ("black/black") in non-cancer became a "hypomethylated/hypermethylated" ("white-grey/black") configuration in cancer (**Figure 2**). The terminology here is a practical shorthand to describe the analytical approach: "LOM" does not mean to imply that the normal cell types evolve into cancers; it is 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 in this series is unequivocal and driven mostly by the cancer and LCL samples, a multivariate regression analysis suggested that additional mechanisms might also play roles. This analysis showed that the anti-correlation between global methylation and per-sample ASM frequencies is partly independent of neoplastic status ($p=3.9x10^{-17}$ after controlling for neoplastic status), and conversely, that the higher ASM frequencies in the cancers are only partly explained by global methylation levels ($p=2x10^{-05}$ 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 32% 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 **Fig. S12**).

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 strong over-representation of the poised "bivalent" promoter state among the ASM DMRs with allele-specific GOM in the cancers, compared to ASM loci overall and to ASM loci with allele-specific LOM in the cancers (**Fig. 3** and **Table 1**). Poised promoters, as annotated by ENCODE chromatin state segmentation, 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-neoplastic stem cells [24] and can transition to a CpG-hypermethylated repressed state in cancer cells that acquire de-differentiated or stem cell-like phenotypes [25].

Finally, 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 the cancer and non-cancer shared ASM loci (**Fig S13**). This finding suggests there might be similarities in the underlying mechanisms in the two classes of ASM, which are investigated by more specific tests in the next sections.

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 [26-31]. Among the ASM DMRs found in the normal samples in the current dataset, 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 and heterochromatin states are depleted, relative to the background of adequately covered genomic regions (**Table 1**). This over-representation of promoter/enhancer elements among ASM DMRs in normal cells and tissues 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**).

To assess similarities and differences in the characteristics of ASM in non-cancer vs. 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 and second, we compared the percentage of ASM loci overlapping each chromatin state among cancer-only ASM loci versus ASM loci found in at least one non-cancer sample. The results of both approaches, using bivariate logistic regressions, showed that ASM DMRs in cancer and non-cancer show a parallel enrichment in all the strongly enriched chromatin features (**Table 1**), albeit with some differences among the less strongly enriched features (**Table 1** and **Table S7**). These findings suggest that the mechanisms leading to ASM are at least partly similar in non-neoplastic and neoplastic cells **-** a conclusion that is further supported by analysis of correlations of ASM with polymorphisms in recognition motifs for DNA binding proteins, 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 mapping, 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 [32-34]. Previously we showed that ASM DMRs can overlap with strong CTCF ChIP-seq peaks and polymorphic CTCF binding sites [8, 35]. 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 significantly affects the predicted binding affinity, requiring 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 CTCF occupancy on at least one allele). We identified 2,302 ASM SNPs (17%) that significantly disrupted at least one of the canonical or ENCODE-discovery CTCF motifs [21, 33]. 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 8.6% of these non-ASM informative SNPs significantly disrupted a CTCF motif, corresponding to a significant enrichment for disrupted CTCF motif among ASM SNPs (OR=2.6; p-value=8x10^{-232}). Importantly from a mechanistic standpoint, the enrichment persists, albeit slightly weaker, when considering only non-CpG-containing polymorphic CTCF motif instances (OR=1.97; p=5.9x10^{-40}).

When testing enrichment for the 14 distinct ENCODE/JASPAR-defined CTCF motifs, we found significant enrichment for 13 of them (**Table S7**). Moreover, as shown in **Figure 4, Figure S14,** and **Table S8**, the difference in binding affinity score between alleles is significantly anti-correlated (inversely correlated) with the difference in methylation for 4 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 motifs, facilitated by the larger number of ASM occurrences in the expanded dataset, show that while the presence of a methylatable CpG in the binding site increases the likelihood of producing or propagating ASM, this feature is not required; rather, the essential feature is allele specific CTCF binding.

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 [36] we found 11,633 polymorphic binding site occurrences overlapping with an informative index SNP for high-confidence recurrent ASM. Of these, 9,043 overlapped at least one ENCODE DNase I hypersensitive site and 2,384 at least one ENCODE cognate ChIP-seq TF peak. From a panel of 1,493 TF motifs with at least 10 occurrences, we found 860 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 (**Table S7**). Next, using linear regression of allele-specific affinity score differences on allele-specific CpG methylation differences, we found 179 TF binding motifs, corresponding to 114 cognate TFs, where DNA methylation appears to be shaped by binding site occupancies (**Fig. 4, Fig. S14,** and **Tables S8** and **S9**). Among these motifs, 116 also showed significant enrichment among ASM loci (**Table S9**). Using stringent statistical criteria (FDR<0.05 and $R^2 \geq 0.4$), all but one of these TF motifs that were both correlated and enriched show inversely correlated behavior, such that a relatively higher binding likelihood correlates with CpG hypomethylation (examples in **Figure 4** and **Figures S14** and **S15)**. Multivariate linear regression of the 160 (out of 179) 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 loci that were detected only in the cancers (including the GM12878 EBV-transformed cell line) vs those found in 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 **Figs. S14** and **S15**), which indicates sharing of this fundamental mechanism of ASM in normal and cancer cells. 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 **Fig. S15**).

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 1,898 high-confidence ASM index SNPs from our GM12878 data 1,555 overlapped at least one ChIP-seq peak in this cell line and had enough ChIP-seq reads (≥10X) to assess allele-specific binding of at least one ENCODE-queried TF. We found that 8% (156) 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 129 (83%) 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, $p=2.2e10^{-16}$). Confirming this pooled analysis, among 14 TFs with more than 10 ASB occurrences associated with ASM, 11, 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: 86%-100%, FDR <0.05).

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

For post-GWAS mapping of rSNPs that underlie GWAS signals much attention has been appropriately focused on cataloguing 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 [37-39], and RegulomeDB [40, 41], 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 (>90%) in our expanded dataset currently lack ASB annotations (**Table S2**). In addition, index SNPs for strong ASM DMRs sometimes have weak RegulomeDB scores (**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 56% 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 (**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 in this study are in chromatin deserts (**Table 1** and **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, particularly if some desert regions contain cryptic binding motifs that were active (occupied) at some point in the history of the cell.

To test 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 analysis, which was performed on the subset of 62 TF motifs that had at least 3 occurrences per ASM class. The correlations were significant and in the same direction (inverse correlation of predicted binding affinity with allelic methylation) in all 4 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. **Figure S16** shows examples of ASM DMRs in desert regions that contain disruptive SNPs in ASM-correlated 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 (**Table S2**), with variability not only between cell types and cancer status but also within a single cell type. As diagrammed in **Figure S17A**, given the connection between TF binding site occupancies and ASM, one hypothesis to explain this variability invokes differences in intracellular levels of TFs. Alternatively, genetic differences (i.e. haplotype effects due to the influence of other SNPs near the ASM index SNP) could also play a role (**Fig. S17B)**. 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 (10%), while the rate of allele switching is strikingly higher (43%) among cancer-only ASM loci (**Fig. 5A, B** and **Fig. S17C**). This finding suggests that biological states, here neoplastic vs non-neoplastic, can influence the stability of ASM, with greater epigenetic variability or instability in cancers manifesting as increased allele switching.

To investigate this variability, we compared the features of ASM DMRs that showed allele switching versus those that did not. As shown in **Figure 5C**, 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 (**Fig. S18**). This finding suggests a working model that postulates two classes of binding motifs: one group of motifs in which destructive SNPs show strong correlations with ASM and stably bind their cognate factors, independently of the neoplastic cellular phenotype, thereby mitigating against allele switching; and another group of motifs with more labile TF binding, which are sensitive to changes in the intracellular levels of their cognate factors and that can participate in allele switching via "TF competition". According to this model (**Fig. S17C**), in situations with adequate chromatin accessibility, there could 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 experimental findings indicating that global DNA hypomethylation in transformed cells is associated with increased chromatin accessibility at regulatory elements [42, 43].

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

For assessing the value of ASM as a signpost for rSNPs in disease-associated chromosomal regions, we defined lenient and stringent haplotype blocks by applying the algorithm of Gabriel et al [44], 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 (**Figs. S19 and S20**). We also calculated R^2 between each ASM and GWAS 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 821 ASM DMRs that contained SNPs in strong LD ($R^2>0.8$) or precisely coinciding with GWAS peak SNPs. A much larger group of 5,988 ASM DMRs were in leniently defined haplotype blocks that contained GWAS peaks. Highlighting mechanistic information from these ASM loci, among the ASM index SNPs in strong LD or precisely coinciding with GWAS peak SNPs, 627 were in ASM-enriched classes of CTCF or TF binding motifs and 174 were in significantly ASM-correlated CTCF or TF binding motifs.

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

We found 207 ASM DMRs containing 231 index SNPs in strong LD (R^2>.8) with GWAS peak SNPs for autoimmune and inflammatory diseases (**Table S10**)**,** plus a larger number in the leniently defined blocks containing such peaks (**Table S2**). Among the loci in strong LD, about half showed ASM in immune system cell types (T cells, B cells, PBL, monocyte/macrophages; **Table S10**). In these DMRs, 51 ASM index SNPs precisely coincided with GWAS peak SNPs, supporting the candidacy of these statistically identified SNPs as biologically functional rSNPs. Another partly overlapping group of 51 ASM index SNPs altered strongly correlated CTCF or TF binding motifs, providing mechanistic leads to disease-associated transcriptional pathways (**Tables S2** and **S10**). Some interesting ASM index SNPs in the stringent blocks, some precisely coinciding with GWAS peak SNPs and others in strong LD with these peaks include rs2145623, coinciding with a GWAS peak SNP for ulcerative colitis, sclerosing cholangitis, ankylosing spondylitis, psoriasis and Crohn's disease (nearest genes *PSMA6*, *NFKBIA*), rs10411630 linked to multiple sclerosis (MS) via LD with GWAS peak SNP rs2303759 (nearest genes *TEAD2*, *DKKL1*, and *CCDC155*; **Fig. S6**), 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**), rs6603785 coinciding with a GWAS peak for SLE and hypothyroidism (nearest genes *TFNRSF4*, *SDF4*, *B3GALT6*, *FAM132A*, *UBE2J2*; **Fig. S21**), and rs11516512 in LD with a GWAS peak SNP for multiple sclerosis, rheumatoid arthritis, and celiac disease (nearest genes *MMEL1*, *TTC34*). Among these examples, index SNPs rs2145623, rs10411630, rs2664280, and rs6603785 each disrupt one or more strongly ASM-correlated TF binding motifs, implicating multiple ETS-family TFs, EGR1, AP1 (JUNB), and MYC, respectively, as candidate transcriptional pathways in these diseases.

ASM index SNPs in LD with GWAS peaks for cancer susceptibility

We found 189 ASM DMRs containing 206 index SNPs in strong LD (R^2>.8) with GWAS peak SNPs for cancer susceptibility or response to treatment (**Table S11**), plus a larger number of ASM loci in the more leniently defined blocks containing such peaks (**Table S2**). Among the DMRs that contained index SNPs in strong LD with the GWAS peaks, 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, etc.) and/or in T cells, which are relevant to cancer via immune surveillance. In these DMRs, 43 of the ASM index SNPs precisely coincided with the GWAS peak SNPs, suggesting that these peak SNPs might be bona fide rSNPs, and another partly overlapping group of 37 index SNPs altered strongly ASM-correlated binding motifs, providing mechanistic information about disease-associated transcriptional pathways (**Table S11**). Some interesting ASM index SNPs in the stringently defined haplotype blocks, some of which precisely coincide with GWAS peak SNPs, include rs416981 (nearest genes *FAM3B*, *MX2*, *MX1*) associated with cutaneous melanoma and nevus counts, rs4487645 (nearest genes *SP4*, *DNAH11*, *CDCA7L*; **Fig. 7**) associated with multiple myeloma and immunoglobulin light chain amyloidosis, rs2853677 linked to lung cancer, gliomas, and other malignancies, as well as benign prostatic hyperplasia via LD with several GWAS peak SNPs (genes *SLC6A18*, *TERT*, *MIR4457*, *CLPTM1L*; **Fig. 7**), rs3806624 linked to multiple myeloma and lymphoma (nearest gene *EOMES*; **Fig. S21**), and rs2754412 linked to breast cancer via LD with GWAS peak SNP rs2754412 (nearest genes *HSD17B7P2*, *SEPT7P9*, *LINC00999*; **Fig. S22**). Potentially informative examples in the lenient blocks include rs2427290 linked to colorectal cancer via GWAS peak SNP rs4925386 (nearest genes *OSBPL2*, *ADRM1*, *MIR4758*, *LAMA5*, *RPS21*, *CABLES2*; **Fig. S7**), and rs2283639 linked to non-small cell lung cancer via GWAS peak SNP rs1209950 (nearest genes *LINC00114*, *ETS2*, *LOC101928398*; **Fig. S8**), plus many others. Among these examples, index SNPs rs416981, rs4487645, rs2427290, rs2283639, rs3806624 and rs61837215 each disrupt a strongly ASM-correlated TF binding motif, implicating TATA_disc1 (a discovery motif that has been suggested to bind

YY1), PAX5, CCNT2, SMC3, BATF, and ELF1_2, respectively, as candidate transcriptional or chromatin organizing pathways in susceptibility to these cancers.

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

We found 152 ASM DMRs containing 164 index SNPs in strong LD (R^2>.8) with GWAS peak SNPs for neurodegenerative, neuropsychiatric, or behavioral phenotypes (**Table S12**), plus a larger number in the more leniently defined haplotype blocks encompassing the GWAS peaks (**Table S2)**. Among the loci in the stringently defined blocks, 21 showed ASM in brain cells and tissues (grey matter, neurons, glia), and a somewhat larger number showed ASM in immune system cell types. Both can be phenotypically relevant, since studies have linked brain disorders not only to neuronal and glial cell processes but also to the immune system [45]. In addition, many loci in this list showed ASM in GBMs, which have partial glial and neuronal differentiation. In these DMRs, 36 of the ASM index SNPs precisely coincided with the GWAS peak SNPs, supporting a functional regulatory role for these variants, and another partly overlapping group of 34 SNPs altered strongly ASM-correlated binding motifs, providing mechanistic information about disease-associated transcriptional pathways. Some interesting examples in the stringent blocks (**Table S12**) include rs1150668 linked to risk tolerance/smoking behavior and wellbeing spectrum via GWAS peak SNPs rs1150668 (precisely 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*; **Fig. S21**), 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*), rs667897 linked to Alzheimer's disease via GWAS peak SNP rs610932 (nearest genes *MS4A2*, *MS4A6A*) and rs13294100 that precisely coincides with a GWAS peak SNP for Parkinson's disease (nearest gene *SH3GL2*), plus others. Among these examples, index SNPs rs2710323 (super enhancer with multiple TF binding sites), rs4976977 (TAL1_5 motif), rs667897 (NFE2, NFE2L2 motifs), and rs13294100 (STAT6_2 motif) each disrupt strongly ASM-correlated TF binding motifs, implicating specific candidate transcriptional pathways in these traits and diseases.

In addition to the disease categories detailed above, we observed several hundred high confidence ASM index SNPs in strong LD with GWAS peaks for cardiometabolic diseases and traits, for example rs2664280 linked to Type 2 diabetes mellitus as well as psoriasis (**Fig. 7**), or with other medically important phenotypes including pharmacogenetic profiles.

Visualization of the ASM mapping data as annotated genome browser tracks

The final set of high-confidence recurrent ASM loci averaged 25 ASM DMRs per Mb of DNA genome wide. We provide the data both in tabular format (**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 is shown in **Figure 7**. These tracks (see Availability of Data section) can be displayed, together with other relevant tracks, including chromatin structure for mechanistic studies and the GWAS catalogue track for potential disease associations, in UCSC Genome Browser sessions [46].

Discussion

These data from dense mapping of ASM in normal human cell types and tissues, plus a group of cancers, identify 13,210 index SNPs in 11,233 DMRs that show strong and recurrent 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 contiguous 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 DNA 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). Overall, 50% of our informative SNPs were also informative in the HEP dataset and 31% of our ASM index SNPs passed criteria for ASM in the HEP data. Given the differences in analytical methods, and more importantly, 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. 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 percent of assessed genomic sites [38], and the data provided here bolster and refine previous results from us and others [8, 9, 11, 27, 29, 30, 47] supporting a dominant role for binding site occupancies in shaping both net and allele-specific DNA methylation patterns in normal 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.

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, an interesting 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 one of the 116 motifs that are both enriched and significantly ASM-correlated (**Table S9**) show a strongly inversely correlated direction of the relationship, such that higher predicted binding affinity (greater predicted binding site occupancy) tracks with relative CpG hypomethylation, which can be heuristically understood as protection of the occupied binding site from methylation. When we only require ASM correlation, without enrichment as a criterion (**Table S8**), we find 174 motifs with this inversely correlated behavior, and only 5 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 5 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 [48].

A crucial qualitative 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, and in the GM12878 LCL, compared to normal cells and tissues and non-transformed low-passage primary epithelial cell explants. Special aspects of ASM detected in the cancers, and in the immortalized but euploid LCL, included allele-specific hypomethylation genome-wide and allele-specific hypermethylation at loci in poised chromatin, as well as 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.

Based on this shared mechanism, 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. Among the ASM index SNPs that we find in strong LD or precisely coinciding with GWAS peak SNPs, 627 are in ASM-enriched classes of CTCF or TF binding motifs and 174 are in significantly ASM-correlated CTCF or TF binding motifs.

This informative situation is highlighted by our findings for ASM index SNP rs4487645, which coincides with a GWAS peak for AL amyloidosis and multiple myeloma and disrupts a TF motif (PAX5) that is significantly correlated with ASM. Since the PAX5 TF is known to function as a master regulator of B cell development [49], these ASM mapping data are post-GWAS evidence for involvement of a relevant biological pathway in susceptibility to multiple myeloma, an important type of 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 studies. A similarly useful example, in a non-neoplastic disease, is provided by ASM index SNP rs2664280, which disrupts a JUNB motif (a binding site for the AP1 TF complex) and is in strong LD with a GWAS peak SNPs for psoriasis. 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 [50].

Lastly, regarding the usefulness and 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 percent 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 many individuals, plus three types of cancers. The data reveal 11,233 high-confidence ASM regions, of which 821 contain SNPs in strong LD or precisely coinciding with GWAS peaks for common 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. Despite these differences, enrichment and correlation analyses indicated 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 been non-informative. We provide our dense ASM maps as genome browser tracks and show examples of destructive variants in TF binding sites that nominate altered transcriptional pathways in susceptibility to autoimmune, neuropsychiatric, and neoplastic diseases.

Materials and methods

Human cells and tissues

Human tissues and cell types analyzed in this study are listed in **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 one week as described [51]. 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 [52] and documented for purity of cell types by immunostaining of cytospin slides, as shown previously [23]. Biopsy samples of human cancers were obtained with I.R.B. approval in a de-identified manner from the Tissue Biorepository of the John Theurer Cancer Center. The GM12878 lymphoblastoid cell line was purchased from Coriell, primary cultures of non-neoplastic human urinary bladder epithelial cells were purchased from A.T.C.C., and primary cultures of non-neoplastic human mammary epithelial cells 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 (**Table S1**). In this protocol, targeted regions (total of 3.7M 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 60 primary non-neoplastic and 14 primary neoplastic 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 **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 **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 [21, 53] 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 [54] 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 [55] 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 **Table S1**, with metrics calculated using Picard tools. SNP calling was performed with BisSNP [56] using default settings, except for the maximum coverage filter set at 200 to encompass deep sequencing, and quality score recalibration. SNP calling was carried out using human genome GRCh37 and dbSNP147 as references. 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. 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 were 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 150Kb 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 [35] and which showed frequent allele switching in normal samples in the current dataset, consistent with imprinting (**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 [57], 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 [58, 59].

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 [55], 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 **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 (GRCh37 assembly) we downloaded RefSeq annotations, DNase hypersensitive sites, TF peaks by ChIP-seq, and chromatin state segmentation by HMM in ENCODE cell lines[60]. 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 [41]. For each relevant feature, enrichment among ASM index SNPs compared to the genome-wide set of nformative 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 [53] and ASB SNPs [38, 61] 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 [21, 33], as well as PWMs from the JASPAR database [36, 53]. We scored allele specific binding affinity at each index SNP using the atSNP R package [39], 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[60]. For each motif, we used data from Kheradpour and Kellis [21, 33] 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 O.R.s for each disrupted polymorphic motif. Enrichment was defined as an O.R.>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 3,394 ASM SNPs for the 2 ASM-class analysis and a total of 62 TF motifs with 10,709 motif occurrences disrupted by 1,967 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. 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 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 [41]. We defined haplotype blocks using 1000 Genomes phase 3 data [62] based on the method of Gabriel et al. for scoring linkage disequilibrium (LD) with emphasis on D-prime values [44] in PLINK [63]. 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 [44]. 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 200kb. 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.

Author Contributions

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

Ethics approval and consent to participate

Peripheral blood samples were obtained with informed consent. Primary tumor samples were collected by the JTCC Tissue Biorepository and transferred to the laboratory in a de-identified manner under I.R.B.-approved protocols.

Acknowledgments

The authors thank Nicholas Ilsey and Sonia DaSilva-Arnold for two of the placenta samples.

Funding

This work was supported by NIH grants R21 AI133140-01, R01 MH092580, 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

The Agilent SureSelect and WGBS data have been submitted to NCBI/GEO in two series (GSE137287, GSE137879) linked by superseries GSE137880. Examples of custom genome browser tracks with annotated ASM loci can be viewed at a UCSC browser session hosted by our laboratory (https://genome.ucsc.edu/s/TyckoLab/High%20Confidence%20ASM).

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, Stockholm SE-171 77, 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

[6]Department of Pathology & Cell Biology, Columbia University Medical Center, New York, NY 10032

[7]Division of Gastroenterology and Celiac Center, Department of Medicine, Columbia University Medical Center, New York NY 10032, USA

[8]Departments of Dermatology and Genetics and Development, Columbia University Medical Center, New York NY 10032, USA

[9]Lombardi Comprehensive Cancer Center of Georgetown University, Washington DC 20057, USA

[10]MNG Laboratories, Atlanta GA 30342, USA

[11]Taub Institute for Research on Alzheimer’s disease and the Aging Brain, Columbia University Medical Center, New York NY 10032, USA

[12]Department of Neurology, Columbia University Medical Center, New York NY 10032, USA

[13]Departments of Psychiatry and Behavioral Medicine and Obstetrics and Gynecology, Columbia University Medical Center, New York, NY 10032, USA

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

Parameters are listed in descending order of enrichment odds ratio among ASM loci found in normal samples. N refers to the numbers of ASM index SNPs.

Parameter	Normal ASM [a] (N=9,667) O.R. (p-value)	Cancer ASM [b] (N=3,543) O.R. (p-value)	Enrichment/depletion in same direction in cancer vs normal ASM	Enrichment strength in cancer/normal (p-value)
ASM SNP is ASB SNP	16.5 (< 10^{-999})	5.3 ($3x10^{-31}$)	YES: enriched	0.34 ($7x10^{-13}$)
ASB SNP within 1kb	8.2 (< 10^{-999})	3.3 ($9x10^{-46}$)	YES: enriched	0.42 ($2x10^{-21}$)
enriched poly. motif OR >4[c]	5.7 (< 10^{-999})	1.7 ($2x10^{-42}$)	YES: enriched	0.46 ($2x10^{-39}$)
poised promoter	5.4 (< 10^{-999})	5.8 (< 10^{-999})	YES: enriched	1.1 ($3.7x10^{-02}$)
weak promoter	4.9 (< 10^{-999})	3.7 ($4x10^{-272}$)	YES: enriched	0.77 ($4x10^{-10}$)
active promoter	4.3 (< 10^{-999})	4.2 ($2x10^{-228}$)	YES: enriched	0.97 ($6x10^{-01}$)
correlated poly. motif [c]	4.0 (< 10^{-999})	0.87 (2x 10^{-02})	NO	0.35 ($1x10^{-59}$)
enriched poly. motif OR >2[c]	3.9 (< 10^{-999})	1.5 ($3x10^{-27}$)	YES: enriched	0.47 ($5x10^{-68}$)
weak/poised enhancer	3.0 (< 10^{-999})	1.6 ($6x10^{-43}$)	YES: enriched	0.53 ($3x10^{-55}$)
any polymorphic motif	2.7 ($3x10^{-92}$)	0.87 ($2x10^{-03}$)	NO	0.48 ($2x10^{-39}$)
repetitive	2.3 ($2x10^{-74}$)	1.8 ($6x10^{-13}$)	YES: enriched	0.79 ($2x10^{-02}$)
strong enhancer	2.2 ($2x10^{-296}$)	1.4 ($1.3x10^{-16}$)	YES: enriched	0.63 ($4x10^{-21}$)
insulator	2.2 ($7x10^{-195}$)	1.4 ($6x10^{-14}$)	YES: enriched	0.67 ($3x10^{-13}$)
ASM SNP is eQTL SNP	2.2 ($4x10^{-78}$)	1.9 ($1x10^{-17}$)	YES: enriched	0.86 ($7x10^{-02}$)
polycomb repressed	2.0 ($1x10^{-263}$)	1.9 ($3x10^{-77}$)	YES: enriched	0.92 ($4x10^{-02}$)
GWAS peak precise overlap	2.0 ($5x10^{-22}$)	2.1 ($5x10^{-11}$)	YES: enriched	1.1 ($7x10^{-01}$)
GWAS peak rsq>0.8	1.5 ($2x10^{-29}$)	1.4 ($1x10^{-09}$)	YES: enriched	0.96 ($6x10^{-01}$)
Txn state	1.3 ($1x10^{-41}$)	1.0 ($8x10^{-01}$)	Weak	0.74 ($3x10^{-13}$)
Heterochromatin [d]	0.82 ($1x10^{-12}$)	1.0 ($5x10^{-01}$)	Weak	1.2 ($4x10^{-04}$)
quiescent chromatin	0.23 (2x10-^{224})	0.53 ($5x10^{-33}$)	YES: depleted	2.2 (4x10-^{31})
chromatin desert [e]	0.12 (<10^{-999})	0.48 ($1x10^{-71}$)	YES: depleted	3.6 (3x10-^{114})

[a] "Normal ASM" refers to ASM in at least one non-cancer sample, allowing ASM in cancer samples.

[b] "Cancer ASM" refers to ASM present in one or more cancer samples (including the GM12878 LCL), but not in any normal sample.

[c] Enriched and correlated motifs determined on the complete set of ASM SNPs; includes ENCODE discovery motifs (_disc)

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

[e] Chromatin desert defined in Methods.

Figure Legends

Figure 1. ASM is increased in cancers and the increase correlates with global DNA hypomethylation.

A, Violin plots showing the percentage of index SNPs showing ASM among the sets of informative SNPs (heterozygous) in each sample. The frequency of ASM is significantly higher in the cancers and lymphoblastoid cell line (LCL) compared to non-neoplastic cells/tissues. N indicates the number of samples. For uniformity in this analysis, only WGBS samples (not SureSelect) are shown. *The GM12878 LCL is grouped with the cancer samples. **B**, Relationship between global DNA methylation and the percentage of informative SNPs that reveal ASM in each sample, showing a significant inverse correlation between per sample ASM frequencies and global methylation levels. Cancer samples are color-coded in red scale and non-cancer samples in blue scale. **mammary epithelial cell lines (N=3) and epithelium-rich normal breast tissue (N=2).

Figure 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 were compared to the relevant non-cancer cell types (B cells for myeloma and lymphoma; glia for glioblastoma). Average fractional methylation was estimated using a linear mixed model with random intercept and random slope (Methods). For each sample type, the squares represent the model estimate of the average fractional methylation in the low and high methylated alleles. **B**, Specific 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. Rows are bisulfite sequence reads separated by REF and ALT allele. Methylated CpGs are black and unmethylated CpGs are white. **C**, Bar graphs showing distribution of net methylation in normal B-cells (left) and glia (right) grouped into 3 classes (low, intermediate and high methylation) at all informative CpGs (which reflects the random expectation) and at CpGs where ASM was observed in the cancer samples but not in the cell lineage-matched non-cancer samples. While allele-specific LOM in cancer numerically 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 in fact 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) represent a significant enrichment over random expectation, given the globally hypomethylated genomic background of the cancers.

Figure 3. Gains of ASM in cancers due to allele-specific GOM at loci in poised chromatin

A, Bar 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 non-neoplastic 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. Absence of circulating myeloma cells in the paired B cell sample was verified by cytopathology and by the absence of DNA copy number aberrations that were seen in the multiple myeloma.

Figure 4. ASM is driven by allele specific CTCF and TF binding in both normal and neoplastic cells

A, XY plots showing examples of TF motifs with strong correlations between predicted allele-specific binding site affinities (estimated by PWM scores) and methylation differences across all occurrences showing ASM. These examples are among 179 significantly correlated motifs, listed in **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,** A large majority of the polymorphic motifs with significant correlations between allelic methylation and predicted binding affinities are also statistically enriched among ASM regions (**Table S9**). The heatmap shows the enrichment or depletion, in $\log^2$(O.R.), for the top 20 enriched TF binding motifs among cancer or non-cancer ASM loci in regions defined as chromatin desert or non-desert (Methods). **C,** Graphs showing significant correlations between allelic TF binding affinity scores and ASM in each of the 4 classes of ASM loci. The left panel 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 small differences in the slopes for each class. **D,** Pairwise comparisons of the correlations in each of the 4 classes of ASM loci, Bonferroni-adjusted for multiple testing. While all the slopes are in a similar range, the correlations in the mixed model are weakest for cancer-only ASM loci, with a modest but statistically significant difference between the cancer vs non-cancer ASM classes, but not between desert and non-desert ASM loci. N: number of occurrences included in the mixed model.

Figure 5. Increased allele switching at ASM loci in cancers

A, Map of the ASM region tagged by index SNP rs11864188 in the *PKD1L3* gene. The ASM shows 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. These motifs include 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. **B**, Frequency of switching, scored for ASM index SNPs that were heterozygous in at least 2 samples, is increased among cancer-only ASM loci (43%) compared to normal ASM loci (10%). As an internal control, informative SNPs mapping to known imprinted regions showed 57% switching, approximating the expected 50% based on parent-of-origin dependent ASM. **C**, Enrichment analyses of polymorphic CTCF and TF binding motifs among switching vs non-switching ASM loci: in the left panel, polymorphic motifs with very 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 dotted vertical lines show the average values for depletion for each set of motifs.

Figure 6. Examples of ASM index SNPs in strong LD or coinciding with GWAS peaks

A, Map of the ASM DMR tagged by index SNP rs4487645, which coincides with a GWAS peak SNP for multiple myeloma ($p=3.0x10^{-14}$; O.R.=1.38) and AL amyloidosis ($p=2.0x10^{-9}$; O.R.=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 TF binding motif on the ALT allele. The REF allele, with intact high-affinity motif, is relatively hypomethylated, as predicted by the TF binding site occupancy model. **B**, Map of the ASM region tagged by index SNP rs2664280, which is in strong LD with GWAS peak SNPs rs2675662 for psoriasis ($p=3.0x10^{-8}$; O.R.=1.14) and rs2633310 for T2D ($p=2.0x10^{-8}$; Beta=-.044). The ASM index SNP is in an enhancer region (yellow-coded chromatin state; GM12878) in the *CAMK2G* gene on chromosome 10. This SNP disrupts several AP1 binding motifs (JUNB shown) on the ALT allele, with higher binding affinity on the REF allele, which is relatively hypomethylated as predicted. **C,** Map of the ASM DMRs tagged by the rs2853677 and rs6420020 index SNPs in the *TERT* gene on chromosome 5. The DMRs are in chromatin that is quiescent/repressed in most ENCODE samples (light and dark gray coded chromatin state; K562 track), but this region is transcribed in undifferentiated H1-hESC. ASM for these DMRs was found only in cancer samples (GBMs). The index SNP rs2853677 is a GWAS peak SNP for non-small cell lung cancer and benign prostatic hyperplasia ($p<10^{-999}$; O.R.=1.41 and $p=2.0x10^{-22}$; O.R.=1.12 respectively). The other ASM index SNP, rs6420020

(**Table S2**) is in LD with GWAS peak SNPs for GBM (p=6.0x10^{-24}; O.R.=1.68), breast carcinoma (p=3.0x10^{-8}; O.R.=1.07), and chronic lymphocytic leukemia (p=6.0x10^{-10}; O.R.=1.18). ASM allele switching is seen at rs6420020; candidate polymorphic TF binding motifs are in **Table S2**. **D,** Map of the ASM DMR tagged by a closely spaced group of SNPs, rs114627468, rs9357065, rs1225618 and rs1150668, in the promoter region of the *ZNF192P1* pseudogene, closely flanked by coding genes in the *ZSCAN* family, on chromosome 6. ASM index SNP rs1150668 is a GWAS peak SNP for body height (p=2.0x10^{-7}; Beta=-.060), smoking status (p=6.0x10^{-15}; Beta=-.0086), smoking behavior (p=3.0x10^{-8}; Beta=+.011), myopia (p=1.0x10^{-11}; Beta=+6.78), and schizophrenia with autism spectrum disorder (p=8.0x10^{-11}; O.R.=1.07). In addition, the 4 ASM index SNPs are in a stringently defined haplotype block containing GWAS peak SNP rs62620225, for multiple phenotypes including wellbeing spectrum (p=6x10^{-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 **Table S12**. Additional examples of disease-linked ASM loci are in **Figures S6-S8**, **S21**, and **S22**.

Figure 7. ASM loci displayed as annotated genome browser tracks

Track of high confidence ASM are provided in UCSC browser format (see Availability of Data). The detailed bed file is provided as supplemental data and can be uploaded to the UCSC Genome Browser. ASM is color coded in blue scale for negative direction ASM (hypomethylation of ALT allele compared to REF allele on average across all ASM samples) and positive direction ASM in red scale (hypermethylation of ALT allele compared to REF allele). Information about the index ASM SNP is displayed by clicking on the SNP (box). Reported information includes sample-aggregated information on the ASM-DMR and the index SNP, sample-specific information on ASM strength (p-value and methylation difference), the two classes of polymorphic motifs disrupted by the index SNP (i.e. enriched among ASM and/or with binding affinity-methylation correlation). Motif logo and sequences of the 2 alleles at the motif occurrences, generated using atSNP, are displayed by clicking on the motif name. Additional annotations of ASM index SNPs are in **Table S2**.

Supplemental Figure Legends

Figure S1. Flow charts of computational and analytical approaches in this study.

A, Steps for ASM calling and ranking, including ASM definition and criteria (See Methods). Our ASM definition combines both individual CpG and DMR-wide (multiple CpG) criteria. **B**, Analytical pipeline for post-calling annotation and analyses to test ASM mechanisms, comparing ASM sub-classes (cancer vs non-cancer; desert vs non-desert), and overlaying that information with public GWAS data to nominate disease associated rSNPs.

Figure S2. Summary of sample types and numbers and yield of informative SNPs

A, Summary of samples sequenced by Agilent and WGBS. Additional information is provided in **Table S1**. Since our high confidence ASM set required ASM in at least 2 samples, the final informative SNP set used for downstream analyses corresponds to the 2,263,262 SNPs that were informative in at least 2 samples. The Venn diagrams are diagrammatic, not drawn to scale. **B**, Map of a region of chromosome 20, showing an increased yield of ASM SNPs in WGBS compared to SureSelect, as expected based on genomic coverage, with consistent findings in regions covered by both methods.

Figure S3. PCA of the WGBS and SureSelect methyl-seq data and overlap between ASM loci detected by the two methods

A, PCA performed using net methylation values of CpGs on chromosome 20. Only CpGs informative (>10X) in both Agilent SureSelect and WGBS were used. The PCA shows clear clustering by cell/tissue and cancer type. Similar results were found using methylation data from other autosomes. **B**, Pie chart showing the proportion of high confidence ASM SNPs found in more than two biological samples, identified by WGBS and by SureSelect methyl-seq. Numbers of ASM SNPs are in parenthesis. **C**, Venn diagram showing a cross-platform comparison, with the percentage of high confidence ASM index SNPs that were identified in both assays. Only SNPs informative in both assays (adequate sequence coverage and heterozygous genotype calls) were considered for this comparison.

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.

A, Most of the ASM calls were found only in 1 sample. While many might be genuine ASM associated with rare SNPs or with inter-individual variability, all downstream analyses in the current study are focused on recurrent ASM detected in at least 2 samples (13,210 ASM index SNPs). **B** and **C**, Approximately one third of the ASM DMRs were identified only in cancer samples (referred to here as “cancer-only ASM”). Given that our study included more non-cancer than cancer samples, this high proportion of ASM SNPs found only in the cancers is significantly increased compared to random expectation.

Figure S5. Example of a chromosome region illustrating consistency between SureSelect methyl-seq, WGBS, and targeted bisulfite sequencing

A, Map of the region on chromosome 20 containing the ASM index SNP rs2427290. When covered in both SureSelect and WGBS, the net methylation is consistent between both assays, and shows low methylation “wells” at CpG islands, as expected. ASM dictated by index SNP rs2427290 is detected in both assays, with additional ASM SNPs found by WGBS, as expected. **B**, Primary sequencing data from WGBS, Agilent SureSelect methyl-seq, and targeted bis-seq, showing consistent findings of ASM in T cell samples. Rows represent sequence reads, and columns CpG sites in these reads. All samples are heterozygous, and the reads are separated by allele. Methylated CpGs are black circles, and unmethylated CpGs are white circles.

Figure S6. Validations of ASM DMRs in disease-associated chromosomal regions: rs1041163 and multiple sclerosis

A, Map showing the ASM region associated with rs1041163 in a putative tissue-specific promoter/enhancer region of the *CCDC155* gene, downstream of the *DKKL1* gene on chromosome 19. ENCODE chromatin state tracks suggest dynamic regulation, with active or quiescent marks depending on the cell type. Bisulfite PCR amplicons were designed to overlap the ASM and flanking SNPs, and to include at least 3 CpGs. The SNP is in high LD and R^2 (R^2 =0.98) with the rs2303759 GWAS peak SNP associated with multiple sclerosis, and another amplicon was designed to assess possible ASM at this position (which did not show ASM in our genome-wide data). The ASM index SNP disrupts an EGR1 TF binding motif and a weak EGR1 ChIP-seq peak found in K562 cells, supporting rs1041163 as a bona fide regulatory SNP. **B**, Targeted bis-seq reads, validating a discrete ASM regions (amplicon 2 and part of amplicon 3) spanning ~700 bp in T cells and brain. Number of ASM samples, informative tissue types and additional annotations are in **Table S2**. The targeted bis-seq showed no evidence of ASM at the GWAS peak SNP, as expected. Rows represent sequence reads, and columns CpG sites in these reads. All samples are heterozygous, and the reads are separated by allele. Methylated CpGs are black circles, and unmethylated CpGs are white circles.

Figure S7. Validation of ASM DMRs in disease-associated chromosomal regions: rs2427290 and colorectal cancer

A, Map showing the ASM region tagged by SNP rs2427290 in the *LAMA5* gene on chromosome 20. The same region is shown, for a different purpose, in **Figure S5**. This region has an active promoter state (color-coded red) in the GM12878 LCL but is in a Txn state (coded green) without promoter characteristics in other ENCODE cell lines. The ASM index SNP is in moderate LD (lenient haplotype block; D'=0.8) with GWAS peak SNP rs4925386 associated with colorectal cancer. The ASM index SNP is in a region of open chromatin (DNAse hypersensitivity) and disrupts an ENCODE discovery motif for CCNT2 TF binding. The relatively hypomethylated allele is the one with the higher predicted CCNT2 binding affinity. **B**, Targeted bis-seq reads validating a discrete ASM regions (amplicon 2) spanning ~400 bp in T cells and colonic mucosa. ASM is not found at the GWAS peak SNP. Numbers of ASM samples in each cell and tissue type, and additional annotations, are in **Table S2**. Rows represent sequence reads, and columns CpG sites in these reads. All samples are heterozygous, and the reads are separated by allele. Methylated CpGs are black circles, and unmethylated CpGs are white circles.

Figure S8. Validation of ASM DMRs in disease-associated chromosomal regions: rs2283639 and non-small cell lung carcinoma

A, Map showing the ASM region tagged by index SNP rs2283639, located in an enhancer region (color-coded in yellow) located immediately upstream of the promoter of the *ETS2* gene on chromosome 21. The SNP is in partial LD (lenient haplotype block; D'=.96) with GWAS peak SNP rs1209950, associated with survival after treatment of non-small cell lung carcinoma. The ASM index SNP disrupts an ENCODE-discovery motif for SMC3 (cohesion complex component), and it co-localizes with a CTCF ChIP-seq peak and a and weak SMC3 ChIP-seq peak. Three amplicons were designed for targeted bis-seq of the ASM region. **B**, Graphical representation of the targeted bis-seq results, validating a discrete ASM regions (amplicon 2) spanning ~600 bp in T cells and lung. The relatively hypomethylated allele is the one with higher predicted SMC3 binding affinity. Numbers of ASM samples in each tissue and cell type, and additional annotations, are in **Table S2**. Rows represent sequence reads, and columns CpG sites in these reads. All samples are heterozygous, and the reads are separated by allele. Methylated CpGs are black circles, and unmethylated CpGs are white circles.

Figure S9. Validations of ASM DMRs spanning a range of ASM ranks

Targeted bis-seq showing validation of additional ASM regions (others in **Figures S6-S8**), with ASM index SNPs that have high, middle or low overall ranks. The results of all validations are summarized in **Table S6**. Rows represent sequence reads, and columns CpG sites in these reads. All samples are heterozygous, and the reads are separated by allele. Methylated CpGs are black circles, and unmethylated CpGs are white circles. In each illustrated case, the relatively hypermethylated allele (REF or ALT) in the targeted bis-seq data is consistent with the relatively hypermethylated allele detected in the primary SureSelect or WGBS data (**Table S2**).

Figure S10. Kernel density plots of methylation levels showing global hypomethylation and decrease in the percentage of high methylated CpGs in cancers

Distribution of the averaged percentage of net methylation genome wide for all informative CpGs in cancer and lineage matched normal samples, by Kernel density estimation. CpG methylation has a bimodal distribution with a large major mode at the high methylated CpGs (>80%) and a weak minor mode at the low methylated CpGs (<5% methylation) in the 2 non-neoplastic cell type (B cells and glia). In multiple myeloma and lymphoma, a strong global hypomethylation with the loss of the high methylated CpGs peak is observed. Hypomethylation is present, but milder, in GBMs.

Figure S11. Allele-specific losses of methylation leading to ASM in cancers

Graphs showing the fitted values of the percentage methylation in cancer (red) for myeloma, lymphoma and glioblastoma versus the lineage-matched non-neoplastic cell types (B cells for myeloma and lymphoma and glia for glioblastoma) for regions where ASM was found only in cancer. In non-neoplastic cells, on average, the methylation levels in these regions were high or intermediate on both alleles and ASM in cancer reflects losses of methylation on one of the alleles. The average fractional methylation was estimated using a linear mixed model with random intercept and random slope (Methods). The light lines represent the fitted values for each locus and the bold line the average fit. The slope between low and high methylation estimates the ASM magnitude. The non-significant and small slope in non-neoplastic cells reflects the absence of significant ASM in these regions.

Figure S12. Kernel density plots of methylation level distributions showing statistically enriched instances of allele-specific gains of methylation leading to ASM in cancers.

These Kernel density plots show the distribution of methylation values in non-neoplastic cells comparing methylation at loci where ASM was found neither in the non-neoplastic cells nor in the matched cancer samples vs loci where ASM was found only in the matched cancer. These graphs show that allele-specific loss of methylation (LOM), which represents the most common scenario for cancer-only ASM, is under-represented compared to random expectation in the globally hypomethylated genomic background, while the less frequent allelic-specific gains of methylation (GOM) are over-represented relative to this background. As shown in **Figure 3**, these instances of GOM in the cancers often map to regions of poised chromatin.

Figure S13. Shared ASM loci in cancer and non-cancer have similar ASM magnitude

Graphs and diagrams showing the fitted values of the average percent methylation of the low and high methylated alleles in cancer (RED) for multiple myeloma (MM), lymphoma, and glioblastoma multiforme (GBM) vs cell lineage-matched non-neoplastic cell types (BLUE), namely B cells for MM and lymphoma and glia for GBM, for DMRs where ASM was found both in cancer and non-cancer. The average fractional methylation of each allele and in each cancer or normal sample class (middle panels) was estimated using a linear mixed model with random intercept and random slope (Methods).On the left panels, the light lines represent the fitted values for each locus and the bold line the average fit. The slopes between low and high methylated alleles estimate the ASM magnitude and are similar (parallel) in cancer and non-cancer samples, with a non-significant statistical interaction between cancer vs normal status and ASM magnitude. The right panels show primary WGBS data for representative examples, with sequence reads separated by allele. Methylated CpGs are black circles, and unmethylated CpGs are white circles.

Figure S14. Correlations between ASM magnitude and disruption of CTCF motifs by ASM index SNPs does not depend on the presence of CpG dinucleotides in the binding motifs

X-Y plots showing the allelic methylation to binding affinity relationship for the 4 CTCF motifs with a significant correlation. Motif occurrences containing at least one CpG site are in green and non-CpG containing motifs in orange. No significant difference in the correlation was observed between the 2 classes of motifs.

Figure S15. Correlations between allelic TF binding affinity scores and ASM cancer versus non cancer and specific examples of TF binding motifs, showing significant correlations between predicted allele-specific binding site affinities and ASM amplitude in both normal and cancer samples.

A, Graphs showing significant correlations between allelic TF binding affinity scores and ASM in each of the 2 classes of ASM loci. The left panel 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. The slopes of the fitted lines 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 compared to non-cancer ASM, with slightly weaker slope in cancer. The right panel shows the pairwise comparison of the correlations in each of the 2 classes of ASM loci with a significant difference between the cancer vs non-cancer ASM classes. N: number of occurrences included in the mixed model. **B**, These examples were selected requiring at least 3 occurrences per ASM classes. The X-Y plots show ASM magnitude vs differences in predicted allele-specific binding affinities (PWM scores) for the EHF_1, SPI1_3, SPIB_2 and ETV6_1 motifs. All 4 classes of ASM loci show similar anti-correlations, but there is a slight decrease in the slope for cancer compared to normal ASM. Desert vs non-desert classes of ASM loci show essentially identical slopes. Regression lines were not plotted if there were less than 3 occurrences within the ASM class (non-cancer/desert for SPIB_2 and cancer/desert for ETV6_1)

Figure S16. Examples of ASM DMRs in chromatin deserts

A, Map showing the ASM DMR tagged by index SNP rs2272697 in the body of the *MANBA* gene. ENCODE chromatin state data show that this region is marked only by a non-regulatory chromatin state (Txn, color-coded green), and the index SNP has a weak regulomeDB score (5), this SNP in fact disrupts an ETS1 motif, suggesting that it could have a regulatory role via the ETS1 transcriptional pathway at some stage of cellular differentiation. This SNP is in high LD (R2>0.9) with multiple GWAS peak SNPs (rs5026472, rs1054037 and rs7665090) associated lymphocyte count, liver cirrhosis, and multiple sclerosis. **B,** Map showing the ASM DMR tagged by index SNP rs13097644 in the intergenic region upstream of the *SETMAR* gene. This region is flagged as quiescent by ENCODE chromatin state (color-coded light gray), consistent with the low regulomeDB score (6) for this SNP. However, the index ASM SNP disrupts an ASM-correlated Erg TF binding motif, suggesting that rs13097644 might act as a regulatory genetic variant at some stage of cell differentiation.

Figure S17. Models for inter-individual variability and allele-switching at ASM loci

A, ASM is not present, with high methylation on both alleles, when either the TFBS is not accessible (closed chromatin) or the TF is not sufficiently expressed (left panel). For accessible TFBS, the magnitude of ASM depends on the level of free TFs (middle panels). When the TF level is low, binding occurs on allele A (with high binding affinity) but stochastically in only a subset of DNA molecules. The overall proportion of low methylated reads (bound TFBS) reflects the steady state between dissociation and binding rates, defined by the concentration of the TF. At the other end of the concentration curve (right panel), strongly overexpressed TFs can bind both high and low binding affinity sites, leading to protection of both alleles against methylation and a loss of ASM. **B**, Inter-individual variability and allele-switching at ASM loci can be explained by a haplotype effects, in which multiple SNPs rather than a single SNP, or a dominant SNP in weak LD with the scored index SNP, dictate the ASM. This situation is "pseudo-switching". **C,** Since most ASM SNPs found in this study can potentially disrupt multiple TF motifs, a TF competition model can explain bona fide allele-switching. This model appears to apply more often in cancer cells, which show a high frequency of ASM allele-switching in this study and are known to frequently over-express oncogenic TFs (e.g. c-MYC; **Fig. 5**).

Figure S18. 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

Graph showing the percentage of switching ASM loci in cancer and non-cancer samples as a function of the number of non-cancer samples where ASM is seen. For ASM loci in cancer, the x=0 data point corresponds to the percentage of switching among cancer-only ASM loci, while the subsequent data points show a decrease in switching among ASM loci found in both cancer and normal as the number of non-cancer samples (in addition to the cancer samples) showing ASM increases. As a comparison, the percentage of switching among ASM loci found in non-cancer samples is low and independent of the total number of samples showing ASM. This finding supports a working model that postulates two classes of binding motifs: one group in which destructive SNPs show strong correlations with ASM, independently of the neoplastic cell phenotype, and stably bind their cognate factors, mitigating against allele switching; and another group of motifs with more labile TF binding, which are sensitive to global increases in chromatin accessibility and changes in intracellular levels of their cognate factors, leading to allele switching via “TF competition”.

Figure S19. Examples of haplotype blocks defined by stringent and lenient parameters

Example of haplotype blocks on chromosome 5 using the Gabriel et al. approach based on confidence interval of D’ values, with stringent (top) and lenient parameter (bottom). The lenient parameters, with relaxed D-prime confidence intervals and historical recombination rate (Methods), lead to haplotype blocks with larger sizes. Graphs were generated using Haploview using 1000 Genome data.

Figure S20. Utility of D’ and R-square parameters for assessing candidate disease-associated rSNPs

Example of D’ (left) and R^2 (right) values between GWAS SNP rs710987 and all SNPs within 200 kb. The GWAS SNP is in red and ASM SNPs are in blue. The lenient haplotype block borders are shown in dashed green. The D’ graph confirms that most of the SNPs within the block (including the ASM SNPs) exhibit high D’ with the GWAS SNP and in this regard are in strong LD with it. The R^2 graph of the same window and SNPs shows that only a small subset of the SNPs in LD also exhibits high R^2 values, because even among SNPs in perfect LD only those with similar allele frequencies are expected to have high R^2 values. A complete understanding of disease associations, including possible effects of more than one rSNP in the same haplotype block, requires extending the identification of rSNPs to those in strong LD with the GWAS peak SNP based on D’, even without high R^2 values.

Figure S21. Additional examples of mechanistically informative disease associated ASM index SNPs: autoimmune and neuropsychiatric disorders

A, Map of the ASM region tagged by index SNP rs6603785 and located in an active enhancer region (color coded dark yellow) downstream of the *UBE2J2* gene on chromosome 1. ASM was observed in multiple blood cell types, including B cells. The ASM index SNP coincides with a GWAS peak SNP, associated with SLE ($p=9.0x10^{-6}$; O.R.=1.11) and hypothyroidism ($p=2.0x10^{-9}$; O.R. not listed). The SNP disrupts a MYC motif, with lower binding affinity and hypermethylation on the ALT allele, as predicted by the TF binding site occupancy model. **B**, Map of the ASM region tagged by index SNP rs2710323 and located in an active enhancer region (color coded in dark yellow) in the gene body of *ITIH1* on chromosome 3. ASM was observed in multiple blood cells, including T cells. The ASM index SNP coincides with a supra-threshold GWAS peak SNP for BMI measurements and multiple neuropsychiatric phenotypes including feeling nervous measurement, anxiety measurement, schizoaffective disorder, schizophrenia, and bipolar disorder (p-values and O.R. or Beta values in **Table S12**). The SNP disrupts an ELF1 motif, with lower binding affinity and higher methylation, as predicted, on the REF allele.

Figure S22. Additional examples of mechanistically informative disease associated ASM index SNPs: breast cancer and lymphoma

A, Map of the ASM region tagged by index SNP rs61837215 and located in the active promoter region (color coded red) of the *SEPT7P9* pseudogene (nearest coding gene, *ZNF37A*) on chromosome 10. ASM is observed in multiple myeloma cells and in normal B cells. The index SNP is in strong LD with GWAS peak SNP rs2754412 associated with breast cancer ($p=6.0x10^{-7}$; Beta=+.031). The ASM index SNP disrupts an ELF1_2 motif, with lower binding affinity and higher methylation on the REF allele, as predicted by the TF binding site occupancy model. **B**, Map of the ASM region tagged by SNP rs3806624 and located in a poised promoter region (color coded purple) of the *EOMES* gene on chromosome 3. ASM was observed in DLBCL and in GBM, with allele switching between the two cancer types. The ASM index SNP coincides with a GWAS peak SNP for Hodgkin lymphoma ($p=1.0x10^{-12}$; O.R.=1.26) and is in strong LD with GWAS peak SNP rs9880772 associated with chronic lymphocytic leukemia ($p=3.0x10^{-11}$; O.R.=1.19), as well as with multiple myeloma (**Table S11**). The SNP disrupts multiple motifs, including a BATF motif (lower binding affinity and higher methylation on the ALT allele) and a MAZ motif with opposite disruption of the binding affinity (lower binding affinity and higher methylation on the REF allele).

Supplemental Table Legends

Table S1. Biological samples analyzed in this study

List of genomic DNA samples sequenced by Agilent SureSelect methyl-seq and by WGBS. Cell/tissue type, normal or cancer status, and diagnoses of the cancer patients from which the samples were collected are listed. Multiple biological samples from the same subjects can be identified by subject ID. Sequencing depth and QC information are also reported.

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

ASM loci, excluding known imprinted chromosomal regions (Methods), are listed using index SNPs as unique identifiers. Thus, some ASM DMRs are listed more than once, when multiple ASM index SNPs lie in the same DMR. Information relative to samples with and without ASM are aggregated, with the samples listed as concatenated entries. However, information at the single sample level can be retrieved from the UCSC-format detailed bed file: (https://genome.ucsc.edu/s/TyckoLab/High%20Confidence%20ASM). Relevant annotations were selected to characterize ASM SNPs based on their potential regulatory functions and disease associations. Briefly, the ASM index SNPs and DMRs are ranked for ASM strength and confidence (Methods), annotated using information about chromatin states, TF binding motifs, and allele-specific marks (ASB, eQTLs) from public databases, with some parameters being calculated using analytical procedures described in the Methods. ASM index SNPs are also annotated for their haplotype block locations and LD with GWAS peak SNPs.

Table S3. Definitions of the terms in Table S2

Description and definition of the columns in table S2. The order by row corresponds to the column order.

Table S4. Imprinted regions with known ASM detected in this study

ASM DMRs detected within 75 kb of known and validated imprinted genes (i.e. 150 kb windows) was considered as likely due to imprinting and was therefore excluded from our downstream analyses but listed in this table. The detection of these instances of ASM, with the expected high rate of allele switching (due to parent of origin dependence) in imprinted regions serves as a positive internal control for the initial steps of the ASM calling pipeline.

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

Table S6. ASM loci tested for validations by targeted bisulfite sequencing

Primers for bisulfite PCR were designed using MethPrimer, with the resulting amplicons spanning the indicated genomic coordinates.

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

Results are from testing for enrichment of motif occurrences in which the motif is disrupted by the ASM index SNP, with a significant difference in affinity score between the two alleles. Significant enrichment among ASM loci was defined as FDR<0.05 and OR>2 (no depletion was observed). Background number of polymorphic occurrences for each motif (random expectation) was computed by screening a random sample of 40,000 occurrences from the list of non-ASM heterozygous (informative) SNPs in our study.

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

Significance was defined as FDR<0.05 and model R^2 >0.4. Results from the linear regressions with and without controlling for the motif CpG content (when the number of occurrences was >3 in each group) are reported.

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

Subset of enriched motifs with significant correlation of predicted allele-specific binding affinities with ASM magnitude. Results of the enrichment analysis by ASM classes (non-desert-non-neoplastic, desert-non-neoplastic, non-desert-cancer and non-desert-cancer) are reported.

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

R^2 cutoff was set at 0.8. GWAS peak SNPs associated with immune related diseases were identified using EFO parent-terms mapped to the GWAS reported traits (provided by the GWAS catalog) with additional manual curation.

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

R^2 cutoff was set at 0.8. GWAS peak SNPs associated with cancer susceptibility were identified using EFO parent-terms mapped to the GWAS reported traits (provided by the GWAS catalog) with additional manual curation.

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

R^2 cutoff was set at 0.8. GWAS peak SNPs associated with neuropsychiatric disorders and traits, or with neurodegenerative diseases, were identified using EFO parent-terms mapped to the GWAS reported traits (provided by the GWAS catalog) with additional manual curation.

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:**904-908.

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:**109-112.

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:**411-419.

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:**142-147.

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:**934-955.

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 Biology* 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**.

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 in health and disease* 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:**3700-3707.

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:**472-484.

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.

19. Das R, Hampton DD, Jirtle RL: **Imprinting evolution and human health.** *Mamm Genome* 2009, **20:**563-572.

20. http://www.geneimprint.com/site/home.

21. 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:**6883-6894.

22. Feinberg AP, Gehrke CW, Kuo KC, Ehrlich M: **Reduced genomic 5-methylcytosine content in human colonic neoplasia.** *Cancer Res* 1988, **48:**1159-1161.

23. 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.

24. 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:**315-326.

25. 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:**837-849.

26. 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:**1091-1097.

27. 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:**490-495.

28. 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:**1128-1138.

29. 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.

30. 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.

31. Wen B, Wu H, Bjornsson H, Green RD, Irizarry R, Feinberg AP: **Overlapping euchromatin/heterochromatin- associated marks are enriched in imprinted gene regions and predict allele-specific modification.** *Genome Res* 2008, **18:**1806-1813.

32. 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:**1231-1245.

33. Kheradpour P, Kellis M: **Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments.** *Nucleic Acids Res* 2014, **42:**2976-2987.

34. Schmitt AD, Hu M, Ren B: **Genome-wide mapping and analysis of chromosome architecture.** *Nat Rev Mol Cell Biol* 2016, **17:**743-755.

35. 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.

36. 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-D266.

37. 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.

38. 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.

39. Zuo C, Shin S, Keles S: **atSNP: transcription factor binding affinity testing for regulatory SNP detection.** *Bioinformatics* 2015, **31:**3353-3355.

40. 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:**1790-1797.

41. 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-D1012.

42. 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**.

43. 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:**317-330.

44. 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:**2225-2229.

45. Debnath M, Berk M: **Functional Implications of the IL-23/IL-17 Immune Axis in Schizophrenia.** *Mol Neurobiol* 2016.

46. 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-D858.

47. 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:**3476-3487.

48. 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:**30919-30932.

49. Medvedovic J, Ebert A, Tagoh H, Busslinger M: **Pax5: a master regulator of B cell development and leukemogenesis.** *Adv Immunol* 2011, **111:**179-206.

50. Uluckan O, Guinea-Viniegra J, Jimenez M, Wagner EF: **Signalling in inflammatory skin disease by AP-1 (Fos/Jun).** *Clin Exp Rheumatol* 2015, **33:**S44-49.

51. 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:**7303-7311.

52. Matevossian A, Akbarian S: **Neuronal nuclei isolation from human postmortem brain tissue.** *J Vis Exp* 2008.

53. 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:**i185-194.

54. Krueger F, Andrews SR: **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.** *Bioinformatics* 2011, **27:**1571-1572.

55. Li LC, Dahiya R: **MethPrimer: designing primers for methylation PCRs.** *Bioinformatics* 2002, **18:**1427-1431.

56. 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.

57. 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-E2890.

58. He Y, Ecker JR: **Non-CG Methylation in the Human Genome.** *Annu Rev Genomics Hum Genet* 2015, **16:**55-77.

59. 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:**212-216.

60. Gama-Sosa MA, Midgett RM, Slagel VA, Githens S, Kuo KC, Gehrke CW, Ehrlich M: **Tissue-specific differences in DNA methylation in various mammals.** *Biochim Biophys Acta* 1983, **740:**212-219.

61. 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:**485-497.

62. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human genetic variation.** *Nature* 2015, **526:**68-74.

63. 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:**559-575.

Figure 1

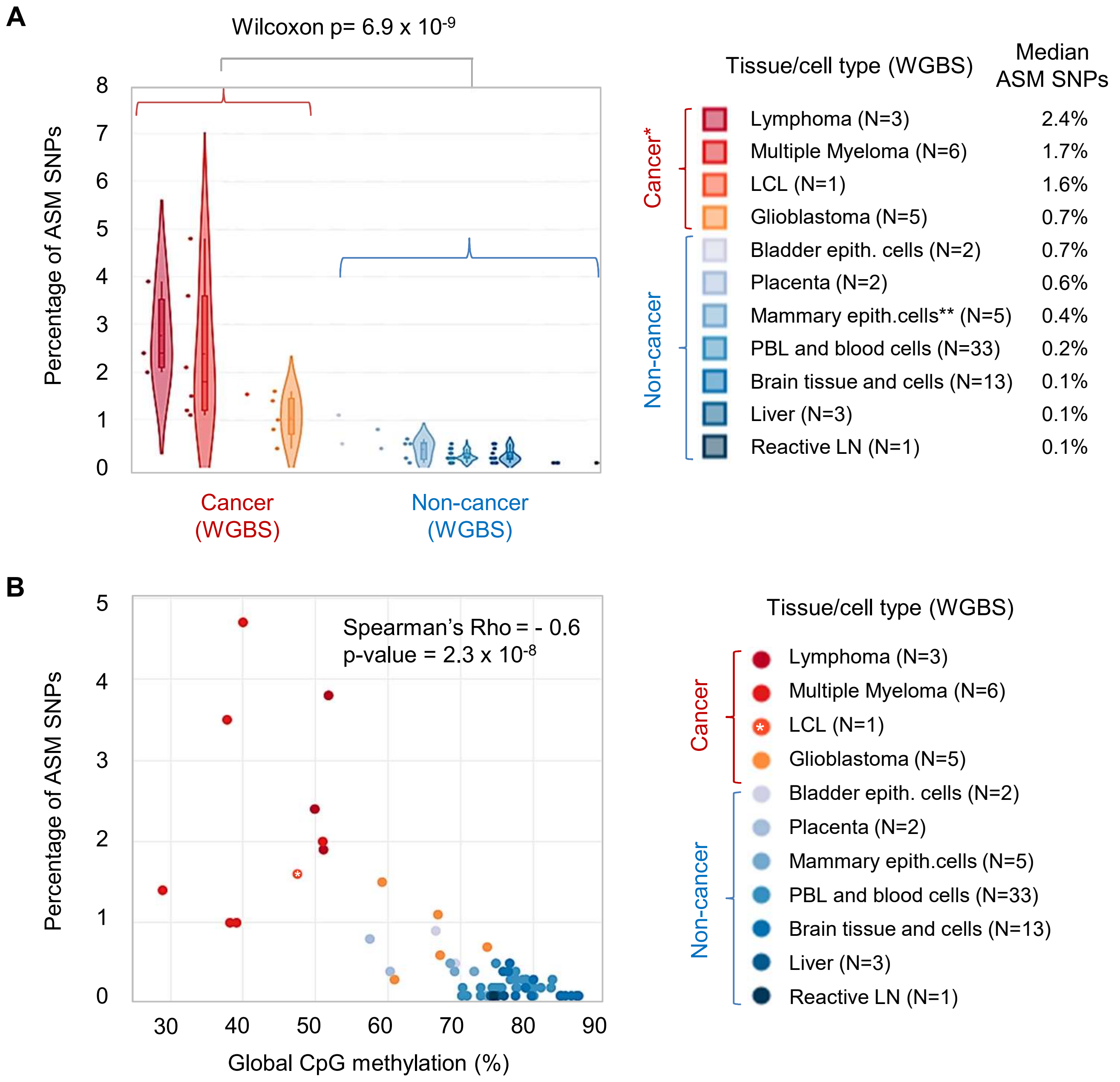

A
Wilcoxon p= 6.9 x 10-9
Percentage of ASM SNPs
0
1
2
3
4
5
6
7
8
Cancer
(WGBS)
Non-cancer
(WGBS)
Tissue/cell type (WGBS)
Median
ASM SNPs
Cancer*
Lymphoma (N=3) 2.4%
Multiple Myeloma (N=6) 1.7%
LCL (N=1) 1.6%
Glioblastoma (N=5) 0.7%
Non-cancer
Bladder epith. cells (N=2) 0.7%
Placenta (N=2) 0.6%
Mammary epith.cells** (N=5) 0.4%
PBL and blood cells (N=33) 0.2%
Brain tissue and cells (N=13) 0.1%
Liver (N=3) 0.1%
Reactive LN (N=1) 0.1%
B
Spearman's Rho = - 0.6
p-value = 2.3 x 10-8
Percentage of ASM SNPs
0
1
2
3
4
5
30
40
50
60
70
80
90
Global CpG methylation (%)
Tissue/cell type (WGBS)
Cancer
Lymphoma (N=3)
Multiple Myeloma (N=6)
LCL (N=1)
Glioblastoma (N=5)
Non-cancer
Bladder epith. cells (N=2)
Placenta (N=2)
Mammary epith.cells (N=5)
PBL and blood cells (N=33)
Brain tissue and cells (N=13)
Liver (N=3)
Reactive LN (N=1)

Figure 2

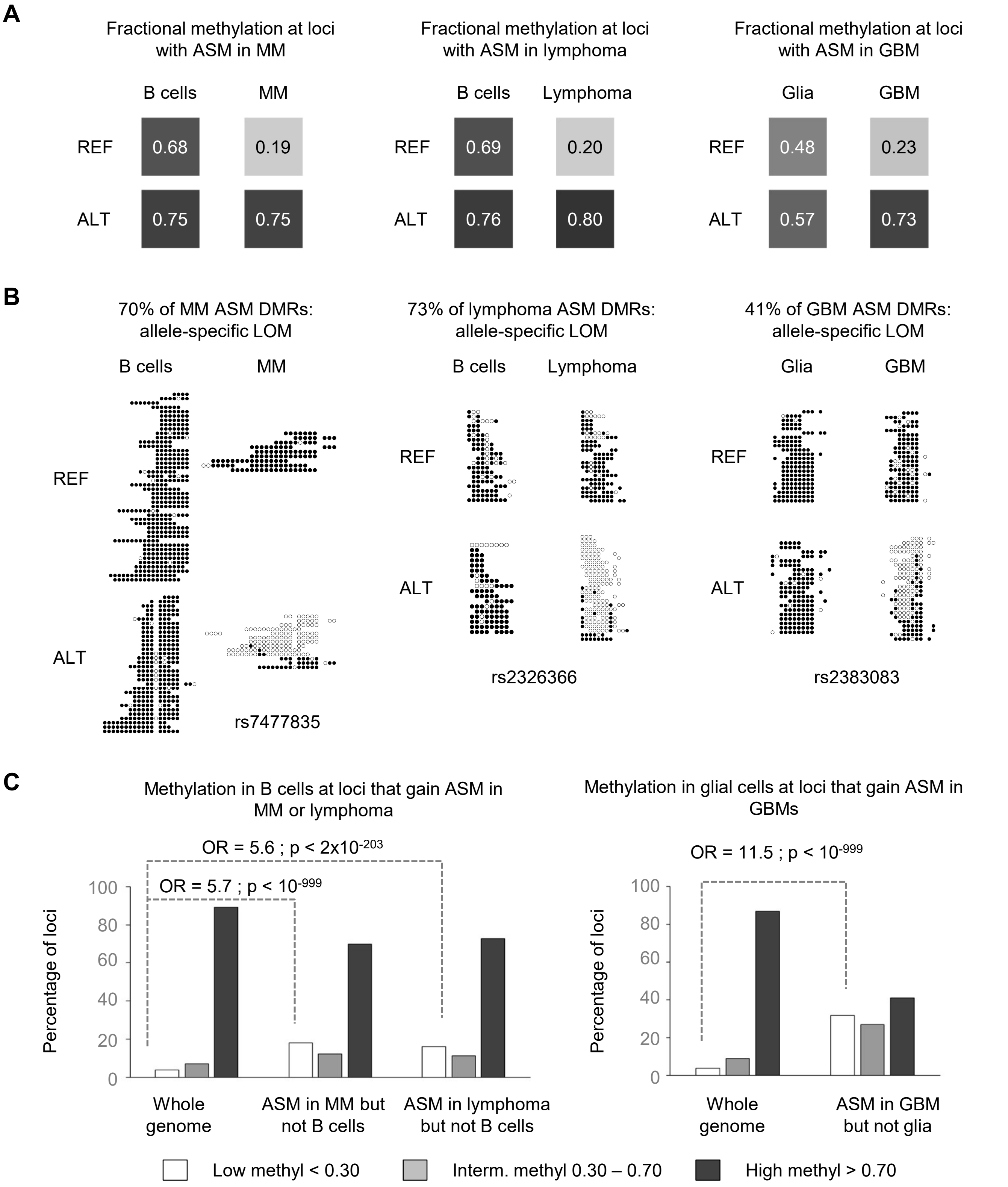

A
Fractional methylation at loci with ASM in MM
B cells
MM
REF
0.68
0.19
ALT
0.75
0.75
Fractional methylation at loci with ASM in lymphoma
B cells
Lymphoma
REF
0.69
0.20
ALT
0.76
0.80
Fractional methylation at loci with ASM in GBM
Glia
GBM
REF
0.48
0.23
ALT
0.57
0.73
B
70% of MM ASM DMRs: allele-specific LOM
B cells
MM
REF
ALT
rs7477835
73% of lymphoma ASM DMRs: allele-specific LOM
B cells
Lymphoma
REF
ALT
rs2326366
41% of GBM ASM DMRs: allele-specific LOM
Glia
GBM
REF
ALT
rs2383083
C
Methylation in B cells at loci that gain ASM in MM or lymphoma
OR = 5.6 ; $p < 2x10^{-203}$
OR = 5.7 ; $p < 10^{-999}$
Percentage of loci
100
80
60
40
20
0
Whole genome
ASM in MM but not B cells
ASM in lymphoma but not B cells
Methylation in glial cells at loci that gain ASM in GBMs
OR = 11.5 ; $p < 10^{-999}$
Percentage of loci
100
80
60
40
20
0
Whole genome
ASM in GBM but not glia
□ Low methyl < 0.30
▨ Interm. methyl 0.30 – 0.70
■ High methyl > 0.70

Figure 3

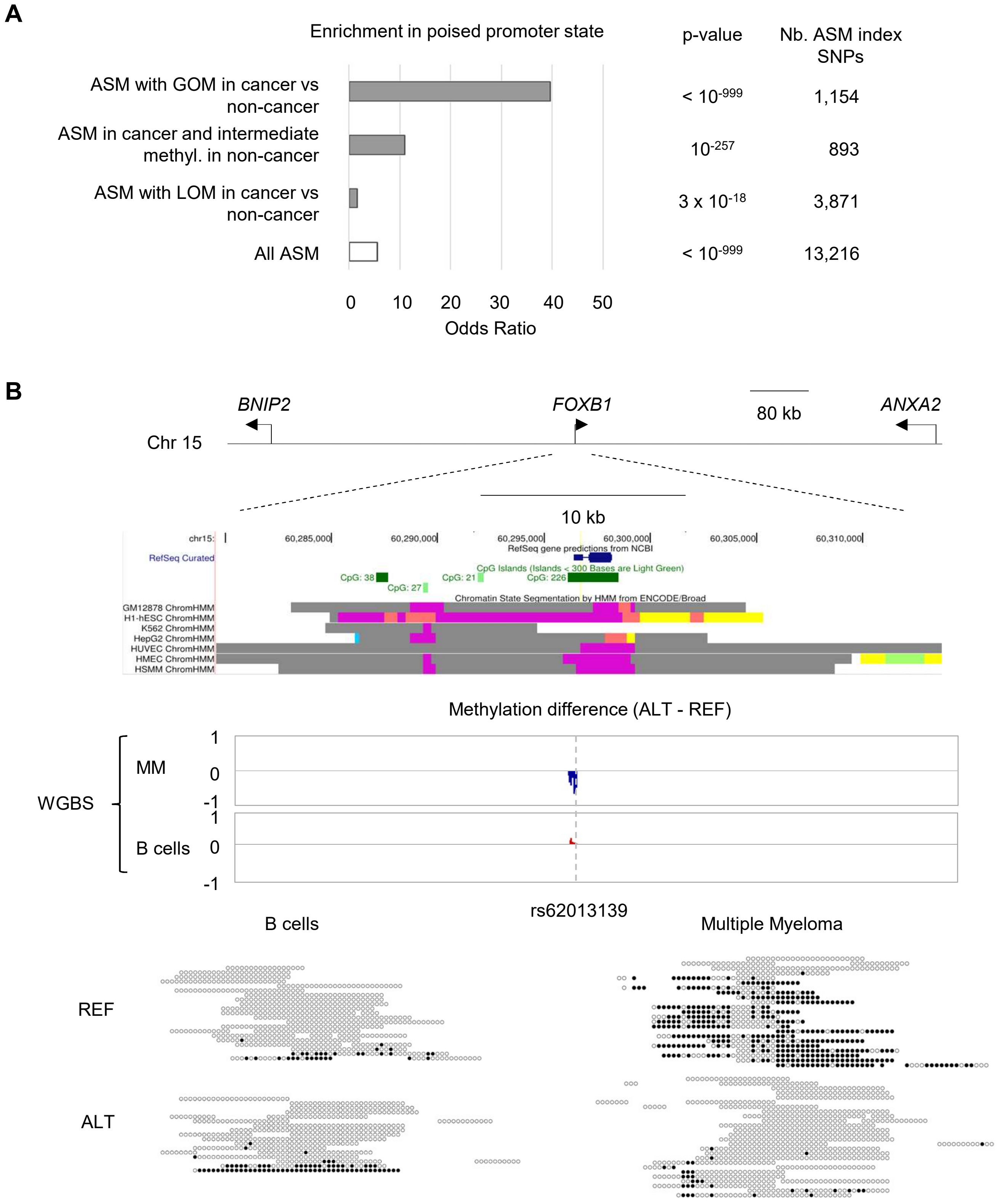

A
Enrichment in poised promoter state
p-value
Nb. ASM index SNPs
ASM with GOM in cancer vs non-cancer
< 10^-999
1,154
ASM in cancer and intermediate methyl. in non-cancer
10^-257
893
ASM with LOM in cancer vs non-cancer
3 x 10^-18
3,871
All ASM
< 10^-999
13,216
0
10
20
30
40
50
Odds Ratio
B
BNIP2
FOXB1
80 kb
ANXA2
Chr 15
10 kb
chr15:
60,285,000
60,290,000
60,295,000
60,300,000
60,305,000
60,310,000
RefSeq gene predictions from NCBI
RefSeq Curated
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 38
CpG: 27
CpG: 21
CpG: 226
Chromatin State Segmentation by HMM from ENCODE/Broad
GM12878 ChromHMM
H1-hESC ChromHMM
K562 ChromHMM
HepG2 ChromHMM
HUVEC ChromHMM
HMEC ChromHMM
HSMM ChromHMM
Methylation difference (ALT - REF)
WGBS
MM
B cells
1
0
-1
1
0
-1
rs62013139
B cells
Multiple Myeloma
REF
ALT

Figure 4

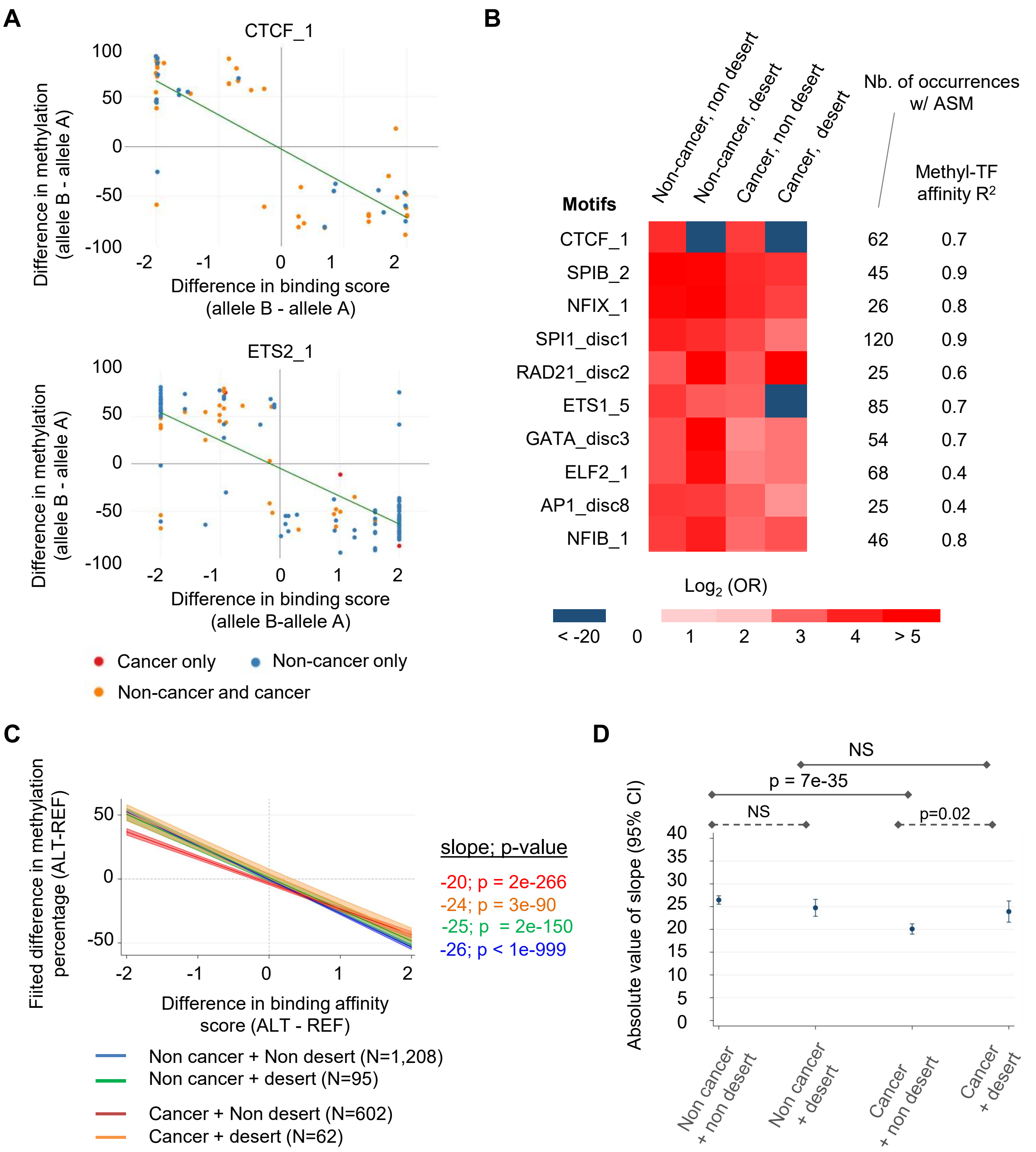

A
CTCF_1
Difference in methylation (allele B - allele A)
100
50
0
-50
-100
-2
-1
0
1
2
Difference in binding score (allele B - allele A)
ETS2_1
Difference in methylation (allele B - allele A)
Difference in binding score (allele B-allele A)
● Cancer only
● Non-cancer only
● Non-cancer and cancer
B
Motifs
Non-cancer, non desert
Non-cancer, desert
Cancer, non desert
Cancer, desert
Nb. of occurrences w/ ASM
Methyl-TF affinity R²
CTCF_1 62 0.7
SPIB_2 45 0.9
NFIX_1 26 0.8
SPI1_disc1 120 0.9
RAD21_disc2 25 0.6
ETS1_5 85 0.7
GATA_disc3 54 0.7
ELF2_1 68 0.4
AP1_disc8 25 0.4
NFIB_1 46 0.8
Log2 (OR)
< -20
0
1
2
3
4
> 5
C
Fitted difference in methylation percentage (ALT-REF)
50
0
-50
-2
-1
0
1
2
Difference in binding affinity score (ALT - REF)
slope; p-value
-20; p = 2e-266
-24; p = 3e-90
-25; p = 2e-150
-26; p < 1e-999
Non cancer + Non desert (N=1,208)
Non cancer + desert (N=95)
Cancer + Non desert (N=602)
Cancer + desert (N=62)
D
NS
p = 7e-35
NS
p=0.02
Absolute value of slope (95% CI)
40
35
30
25
20
15
10
5
0
Non cancer + non desert
Non cancer + desert
Cancer + non desert
Cancer + desert

Figure 5

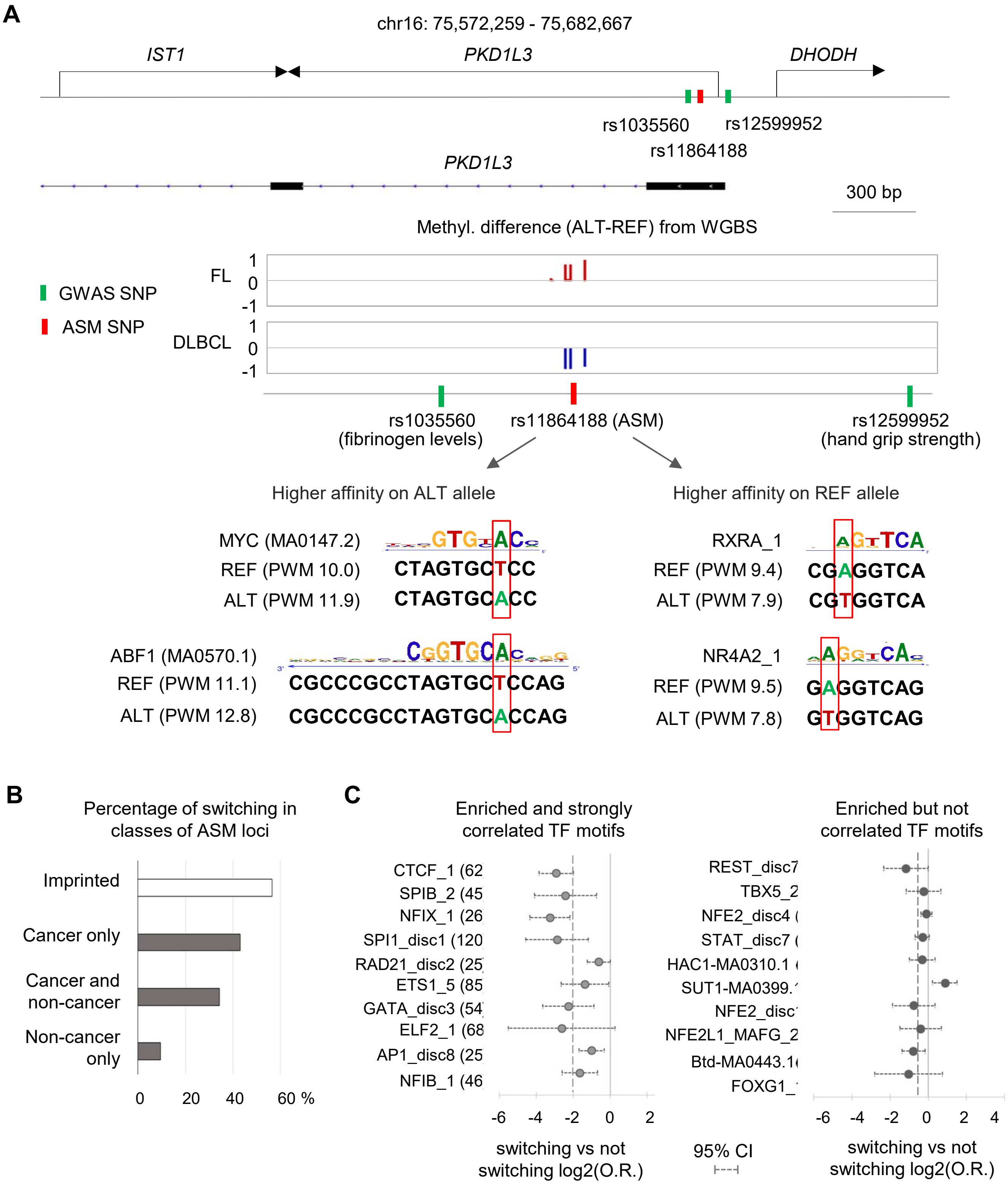

A
chr16: 75,572,259 - 75,682,667
IST1
PKD1L3
DHODH
rs1035560
rs12599952
rs11864188
PKD1L3
300 bp
Methyl. difference (ALT-REF) from WGBS
GWAS SNP
ASM SNP
FL
DLBCL
1
0
-1
rs1035560
(fibrinogen levels)
rs11864188 (ASM)
rs12599952
(hand grip strength)
Higher affinity on ALT allele
Higher affinity on REF allele
MYC (MA0147.2)
REF (PWM 10.0)
ALT (PWM 11.9)
CTAGTGCTCC
CTAGTGCACC
RXRA_1
REF (PWM 9.4)
ALT (PWM 7.9)
CGAGGTCA
CGTGGTCA
ABF1 (MA0570.1)
REF (PWM 11.1)
ALT (PWM 12.8)
CGCCCGCCTAGTGCTCCAG
CGCCCGCCTAGTGCACCAG
NR4A2_1
REF (PWM 9.5)
ALT (PWM 7.8)
GAGGTCAG
GTGGTCAG
B
Percentage of switching in classes of ASM loci
Imprinted
Cancer only
Cancer and non-cancer
Non-cancer only
0
20
40
60 %
C
Enriched and strongly correlated TF motifs
CTCF_1 (62
SPIB_2 (45
NFIX_1 (26
SPI1_disc1 (120
RAD21_disc2 (25
ETS1_5 (85
GATA_disc3 (54
ELF2_1 (68
AP1_disc8 (25
NFIB_1 (46
-6
-4
-2
0
2
switching vs not switching log2(O.R.)
95% CI
Enriched but not correlated TF motifs
REST_disc7
TBX5_2
NFE2_disc4 (
STAT_disc7 (
HAC1-MA0310.1 (
SUT1-MA0399.1
NFE2_disc1
NFE2L1_MAFG_2
Btd-MA0443.1(
FOXG1_1
-6
-4
-2
0
2
4
switching vs not switching log2(O.R.)

Figure 6

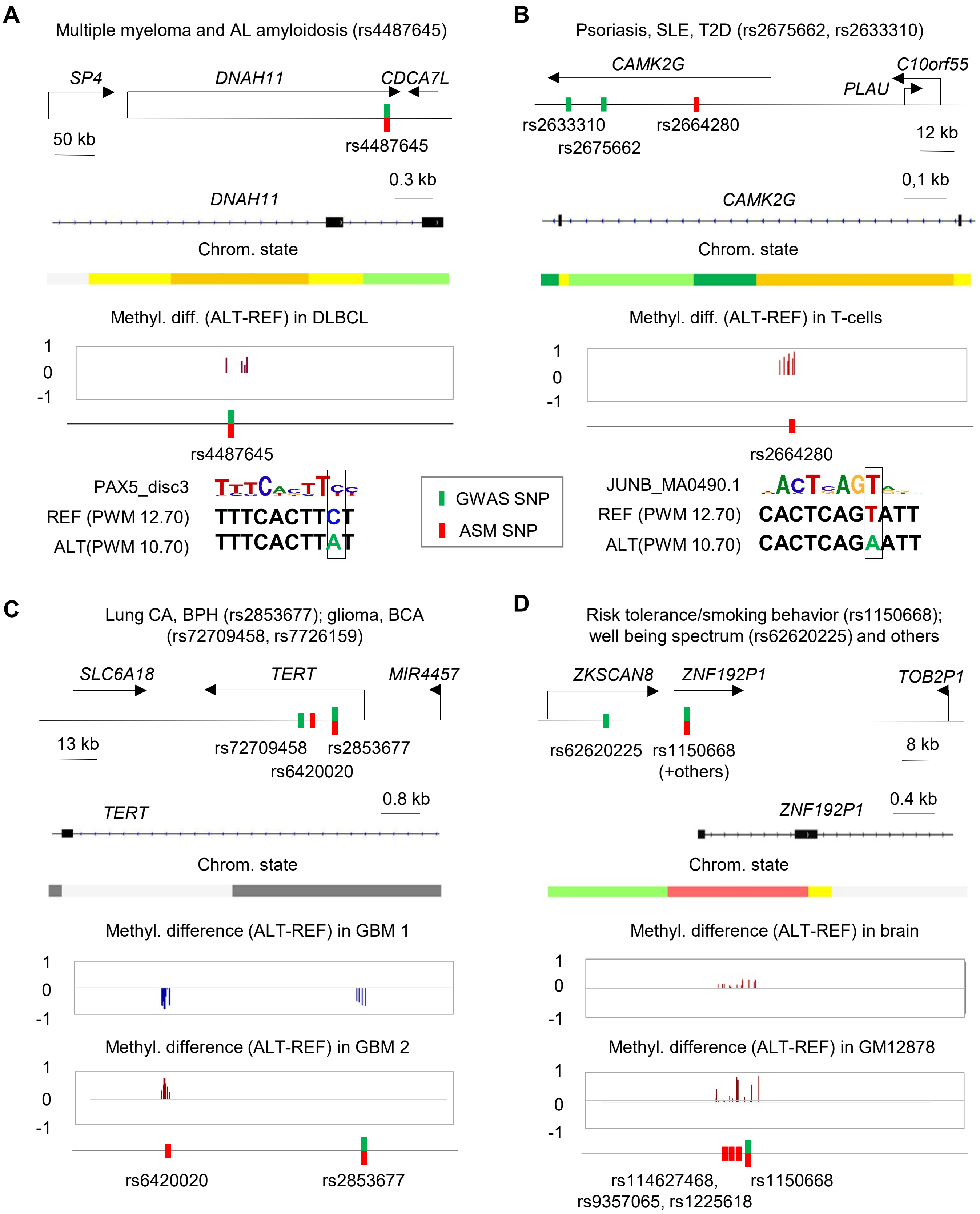
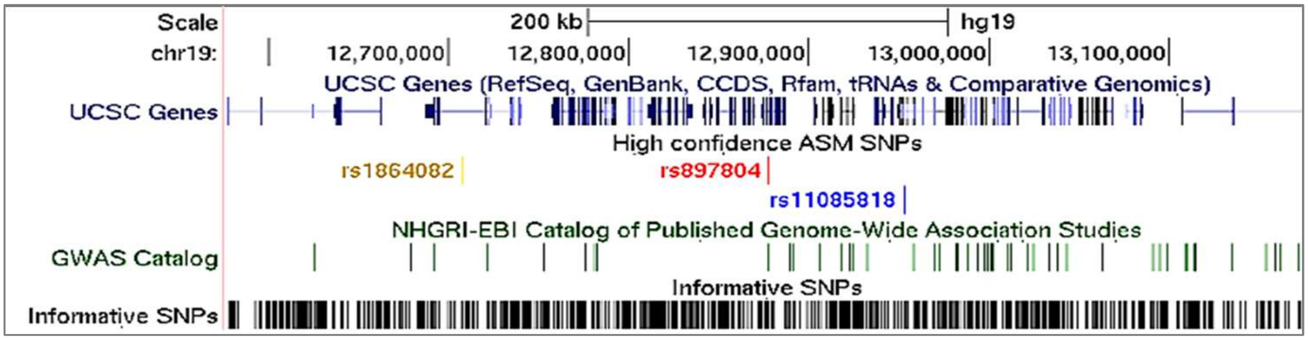
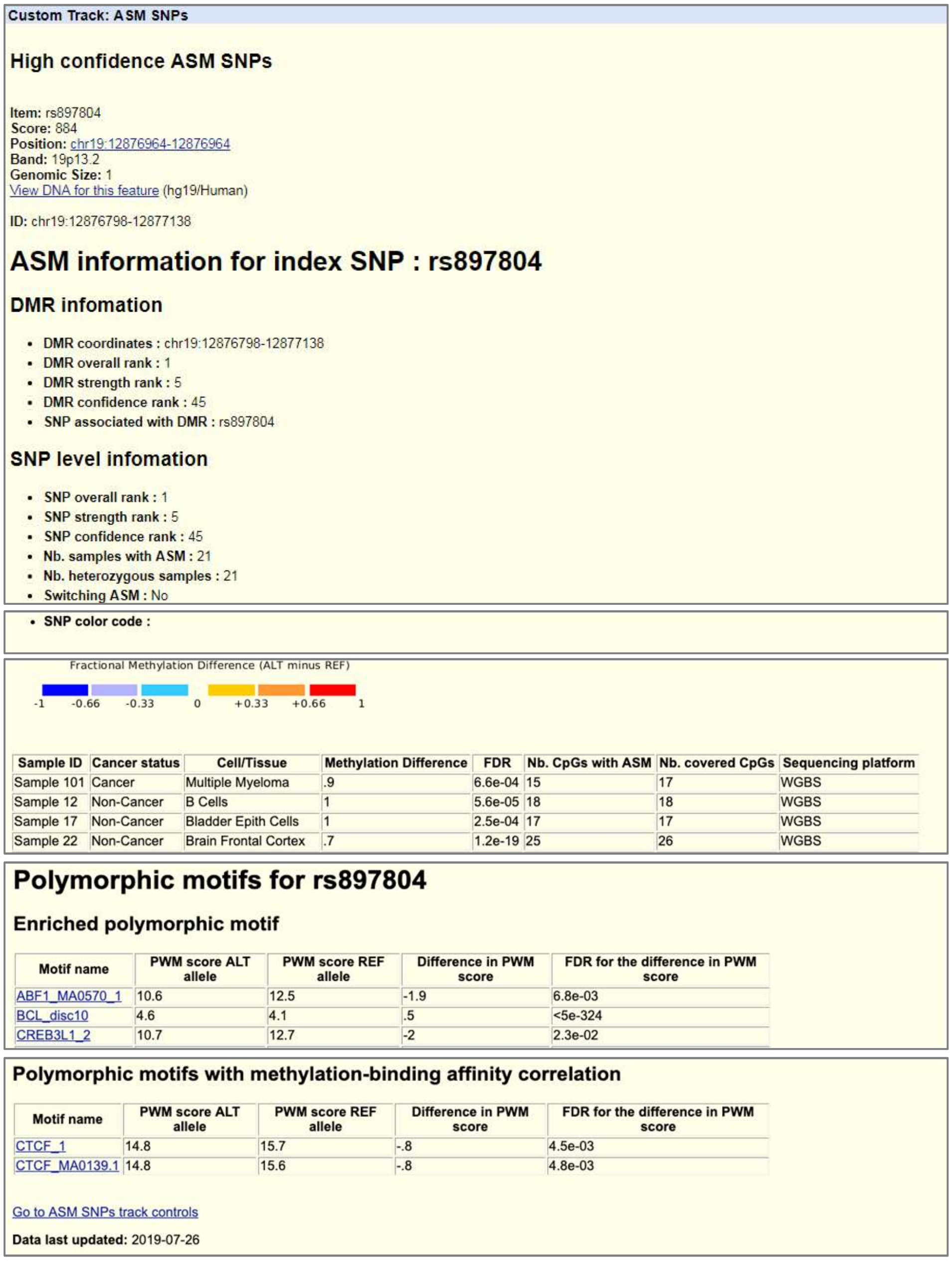

A
Multiple myeloma and AL amyloidosis (rs4487645)
SP4
DNAH11
CDCA7L
50 kb
rs4487645
0.3 kb
DNAH11
Chrom. state
Methyl. diff. (ALT-REF) in DLBCL
1
0
-1
rs4487645
PAX5_disc3
REF (PWM 12.70)
TTTCACTTCT
ALT(PWM 10.70)
TTTCACTTAT
GWAS SNP
ASM SNP
B
Psoriasis, SLE, T2D (rs2675662, rs2633310)
CAMK2G
C10orf55
PLAU
rs2633310
rs2664280
rs2675662
12 kb
0,1 kb
CAMK2G
Chrom. state
Methyl. diff. (ALT-REF) in T-cells
1
0
-1
rs2664280
JUNB_MA0490.1
REF (PWM 12.70)
CACTCAGTATT
ALT(PWM 10.70)
CACTCAGAATT
C
Lung CA, BPH (rs2853677); glioma, BCA (rs72709458, rs7726159)
SLC6A18
TERT
MIR4457
13 kb
rs72709458
rs2853677
rs6420020
0.8 kb
TERT
Chrom. state
Methyl. difference (ALT-REF) in GBM 1
1
0
-1
Methyl. difference (ALT-REF) in GBM 2
1
0
-1
rs6420020
rs2853677
D
Risk tolerance/smoking behavior (rs1150668); well being spectrum (rs62620225) and others
ZKSCAN8
ZNF192P1
TOB2P1
rs62620225
rs1150668 (+others)
8 kb
ZNF192P1
0.4 kb
Chrom. state
Methyl. difference (ALT-REF) in brain
1
0
-1
Methyl. difference (ALT-REF) in GM12878
1
0
-1
rs114627468, rs9357065, rs1225618
rs1150668

Figure 7





Custom Track: ASM SNPs

High confidence ASM SNPs

Item: rs897804
Score: 884
Position: chr19:12876964-12876964
Band: 19p13.2
Genomic Size: 1
View DNA for this feature (hg19/Human)

ID: chr19:12876798-12877138

ASM information for index SNP : rs897804

DMR infomation

- **DMR coordinates :** chr19:12876798-12877138
- **DMR overall rank :** 1
- **DMR strength rank :** 5
- **DMR confidence rank :** 45
- **SNP associated with DMR :** rs897804

SNP level infomation

- **SNP overall rank :** 1
- **SNP strength rank :** 5
- **SNP confidence rank :** 45
- **Nb. samples with ASM :** 21
- **Nb. heterozygous samples :** 21
- **Switching ASM :** No

- **SNP color code :**

Fractional Methylation Difference (ALT minus REF)

🟦	🟪	🟦	🟨	🟧	🟥
-1 / -0.66	-0.66 / -0.33	-0.33 / 0	0 / +0.33	+0.33 / +0.66	+0.66 / 1

Sample ID	Cancer status	Cell/Tissue	Methylation Difference	FDR	Nb. CpGs with ASM	Nb. covered CpGs	Sequencing platform
Sample 101	Cancer	Multiple Myeloma	.9	6.6e-04	15	17	WGBS
Sample 12	Non-Cancer	B Cells	1	5.6e-05	18	18	WGBS
Sample 17	Non-Cancer	Bladder Epith Cells	1	2.5e-04	17	17	WGBS
Sample 22	Non-Cancer	Brain Frontal Cortex	.7	1.2e-19	25	26	WGBS

Polymorphic motifs for rs897804

Enriched polymorphic motif

Motif name	PWM score ALT allele	PWM score REF allele	Difference in PWM score	FDR for the difference in PWM score
ABF1_MA0570_1	10.6	12.5	-1.9	6.8e-03
BCL_disc10	4.6	4.1	.5	<5e-324
CREB3L1_2	10.7	12.7	-2	2.3e-02

Polymorphic motifs with methylation-binding affinity correlation

Motif name	PWM score ALT allele	PWM score REF allele	Difference in PWM score	FDR for the difference in PWM score
CTCF_1	14.8	15.7	-.8	4.5e-03
CTCF_MA0139.1	14.8	15.6	-.8	4.8e-03

Go to ASM SNPs track controls

Data last updated: 2019-07-26

General information about the ASM index SNP

Information about the ASM DMR

Ranking of the ASM index SNP

Color code for fractional methylation difference

Samples with ASM

Polymorphic TB binding motifs enriched among ASM loci and disrupted by the ASM index SNP

Figure S1

A

Identifying and ranking ASM DMRs

Sequence alignment to the reference methylome

- BisMark - default settings with PE mode (WGBS, Agilent) and SE mode (Agilent unpaired reads after trimming).

Heterozygous SNP calling

- BisSNP - default settings with Quality Score recalibration
- Non-G/A SNP coverage > 5x per allele (total coverage > 10x)
- Allele B coverage > 20 and < 80 of total coverage
- Filter out false calls : SNPs with multiple alignment, > 2 alleles with AF>0.01, indels, no AF (UCSC browser annotation of dbSNP147)
- Filter out false calls: SNPs with in HW disequilibrium (exact FDR<0.05) and het. freq > expected het. freq (dbSNP147)

Identification of CpGs with ASM

- CpG coverage > 5X
- Filter out CpGs destroyed by common SNPs (> 5% MAF)
- Filter out CpGs within 10 bp of PE read 2 for Nextera ("fill-in" region) and 7 bp of both reads for TruSeq
- Fisher exact test comparing methylation on allele A vs B ($p < 0.05$)
- Check predicted differences in methylation in AA vs. BB homozygotes using an mQTL-like approach

Identification & ranking of ASM DMRs

- Estimate DMR border (first and last ASM CpG) and count the # of significant CpGs in the DMR
- Compare methylation between alleles across the DMR: avg methylation across all covered CpGs between the first and last ASM CpG of the same DMR.
- ASM calling : DMR difference > 20 and BH-corrected Wilcoxon p-value < 0.05 and at least 3 ASM CpGs including at least 2 consecutive CpGs (overlapping DMRs merged)
- Exclude DMRs in known imprinted chromosomal regions
- Rank DMRs by absolute methylation difference, number and percentage of ASM CpGs
- Independent validations by targeted bis-seq on a set of ASM loci with strong and weak ranks

B

Testing mechanisms in normal and cancer ASM; nominating disease-associated rSNPs

Functional annotation and enrichment analyses of features in ASM DMRs

- eQTLs; DNAse-hs, TF binding (ChIP-seq)
- Disrupted TFBS motif occurrences overlapping a cognate ChIP-seq peak (200 bp window)
- TF peaks for which the motif is enriched (≥10 fold) compared to background based on ENCODE ChIP-seq

Identification of TFBS motifs with disruptive SNPs that correlate with ASM

- Identify polymorphic TF motifs occurrences using ENCODE and JASPAR PWM and AtSNP software (require DNAse-hs peaks)
- Test for enrichment of disrupted vs non-disruptive polymorphic TF occurrences among ASM DMRs
- Test for correlations of ASM strength with PWM scores

Assess mechanistic similarities and differences between cancer and non-cancer ASM

- Multivariate analysis to compute odds ratios of finding ASM DMRs from cancer and non-cancer samples in specific chromatin states and associated with SNPs that disrupt specific TF binding motifs
- Assess rates of allele switching in cancer vs non-cancer ASM loci
- Assess frequencies of chromatin desert locations for cancer vs non-cancer ASM loci

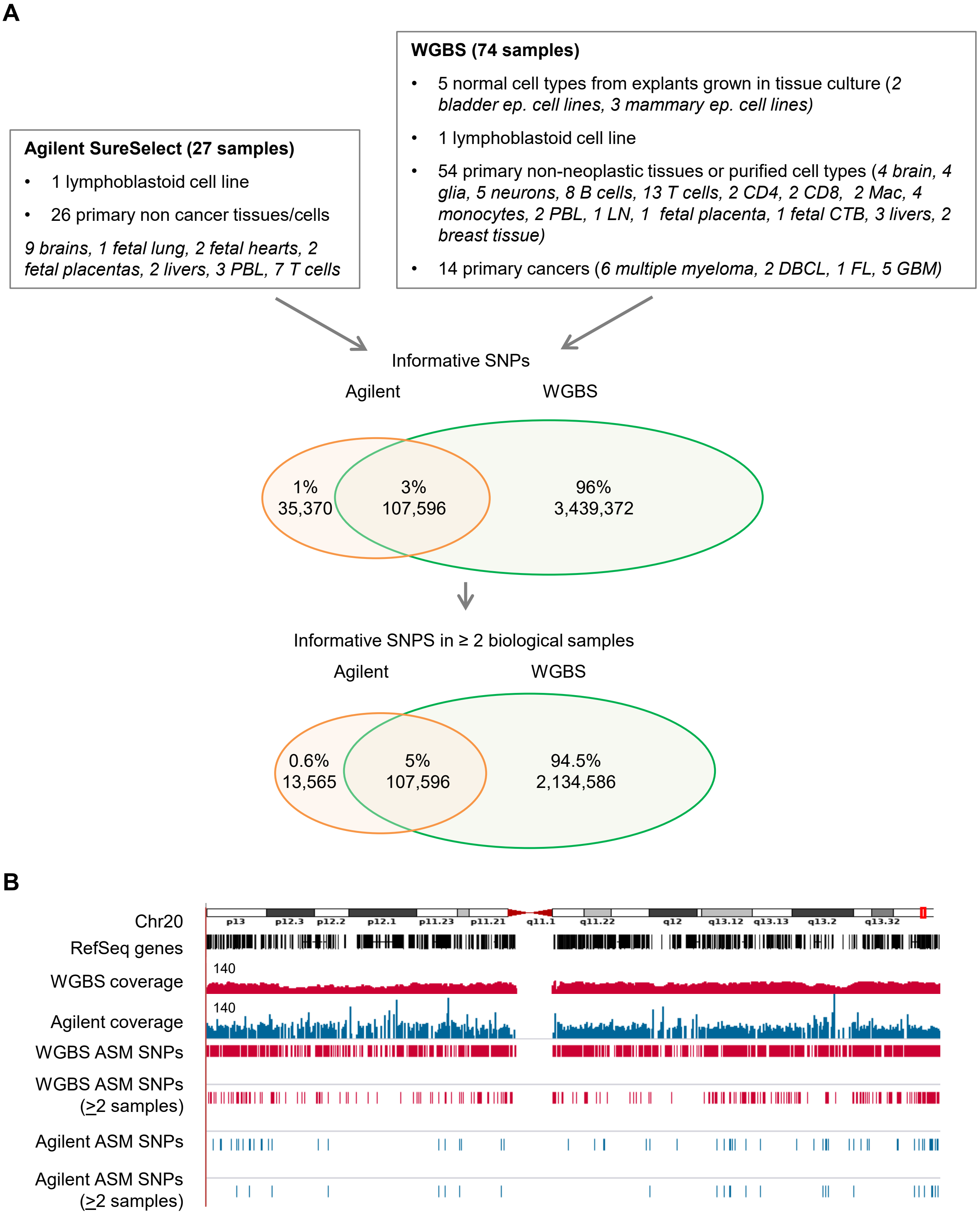
Identification of hap-ASM DMRs in haplotype blocks that contain GWAS peaks

- Determine stringent and lenient LD and haplotype blocks
- Require distance between ASM and GWAS SNP < 200 kb
- Annotate ASM index SNPs for associations with immune/inflammatory, neuropsychiatric and neurodegenerative, neoplastic, and cardiometabolic diseases and traits

Creation of genome browser tracks for visualization and prioritization of candidate disease-associated rSNPs

- Custom tracks of ASM for each chromosome in UCSC Genome Browser format; tracks provide multiple annotations of each ASM index SNP.
- Annotations include ranking based on ASM strength and mechanistically relevant features including the identities of enriched and correlated TF and CTCF binding motifs that are disrupted by each ASM index SNP

Figure S2



Agilent SureSelect (27 samples)

- 1 lymphoblastoid cell line
- 26 primary non cancer tissues/cells

9 brains, 1 fetal lung, 2 fetal hearts, 2 fetal placentas, 2 livers, 3 PBL, 7 T cells

WGBS (74 samples)

- 5 normal cell types from explants grown in tissue culture (*2 bladder ep. cell lines, 3 mammary ep. cell lines)*
- 1 lymphoblastoid cell line
- 54 primary non-neoplastic tissues or purified cell types (*4 brain, 4 glia, 5 neurons, 8 B cells, 13 T cells, 2 CD4, 2 CD8, 2 Mac, 4 monocytes, 2 PBL, 1 LN, 1 fetal placenta, 1 fetal CTB, 3 livers, 2 breast tissue)*
- 14 primary cancers (*6 multiple myeloma, 2 DBCL, 1 FL, 5 GBM)*

Figure S3

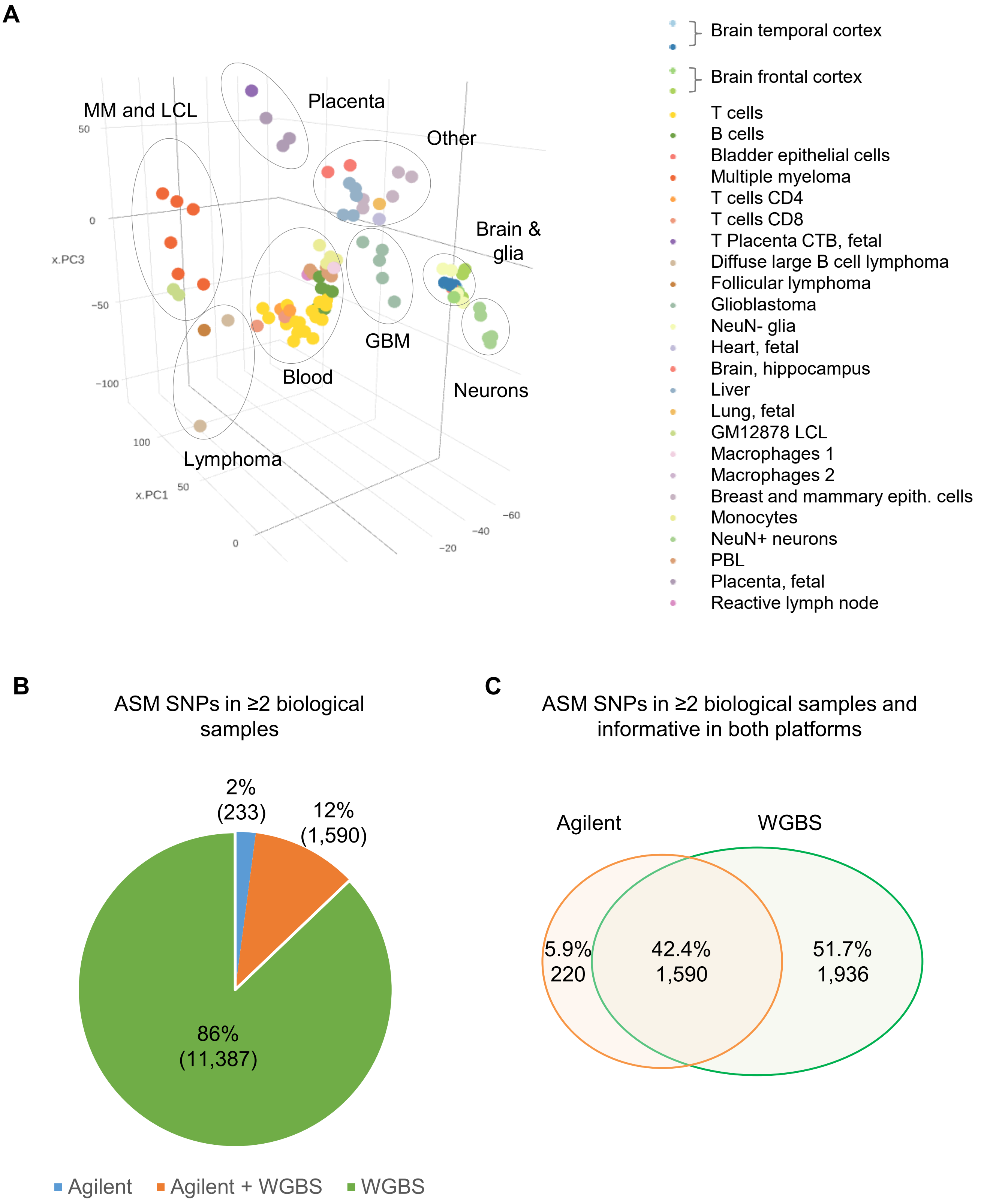

A
MM and LCL
Placenta
Other
Brain & glia
GBM
Blood
Neurons
Lymphoma
x.PC3
x.PC1
50
0
−50
−100
100
50
0
−60
−40
−20
Brain temporal cortex
Brain frontal cortex
T cells
B cells
Bladder epithelial cells
Multiple myeloma
T cells CD4
T cells CD8
T Placenta CTB, fetal
Diffuse large B cell lymphoma
Follicular lymphoma
Glioblastoma
NeuN- glia
Heart, fetal
Brain, hippocampus
Liver
Lung, fetal
GM12878 LCL
Macrophages 1
Macrophages 2
Breast and mammary epith. cells
Monocytes
NeuN+ neurons
PBL
Placenta, fetal
Reactive lymph node
B
ASM SNPs in ≥2 biological samples
2% (233)
12% (1,590)
86% (11,387)
Agilent
Agilent + WGBS
WGBS
C
ASM SNPs in ≥2 biological samples and informative in both platforms
Agilent
WGBS
5.9% 220
42.4% 1,590
51.7% 1,936

Figure S4

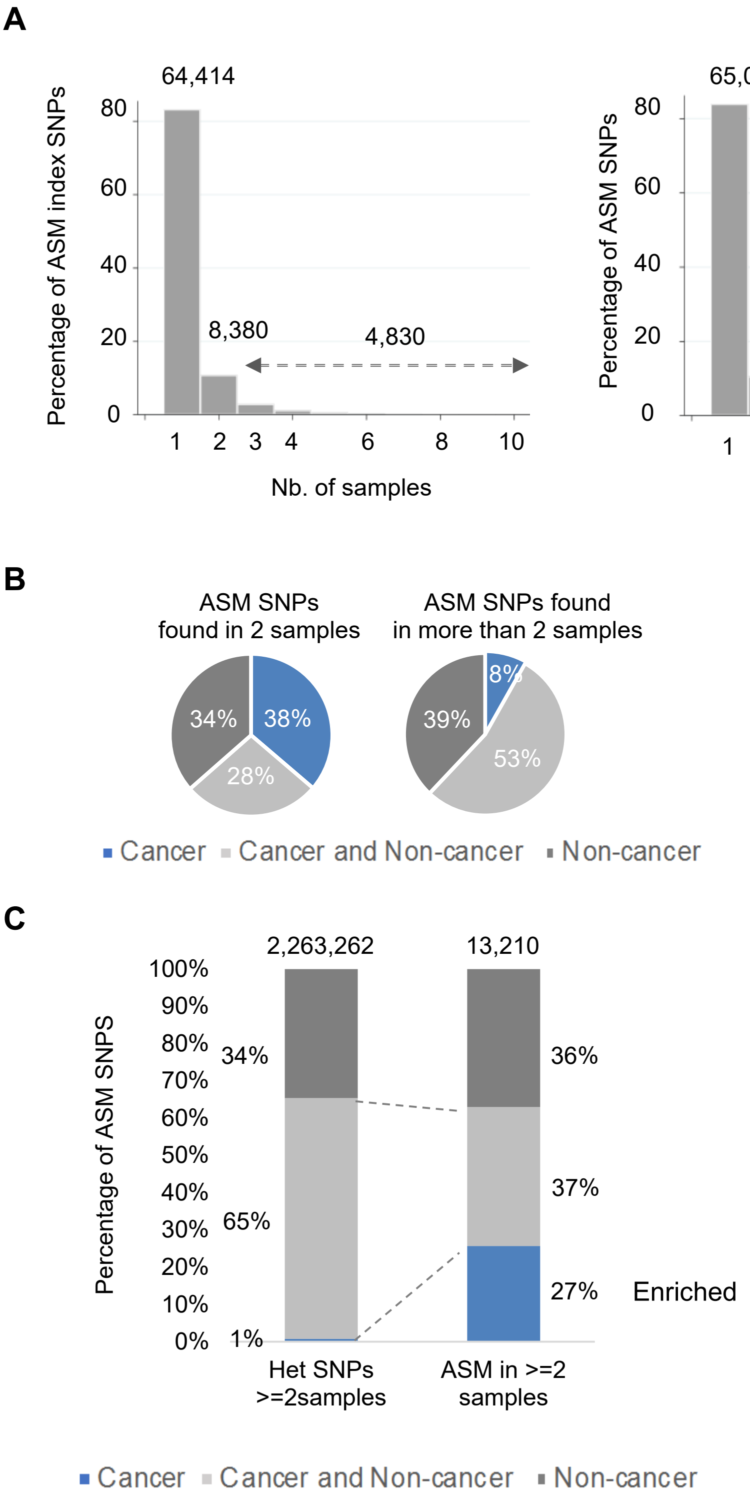
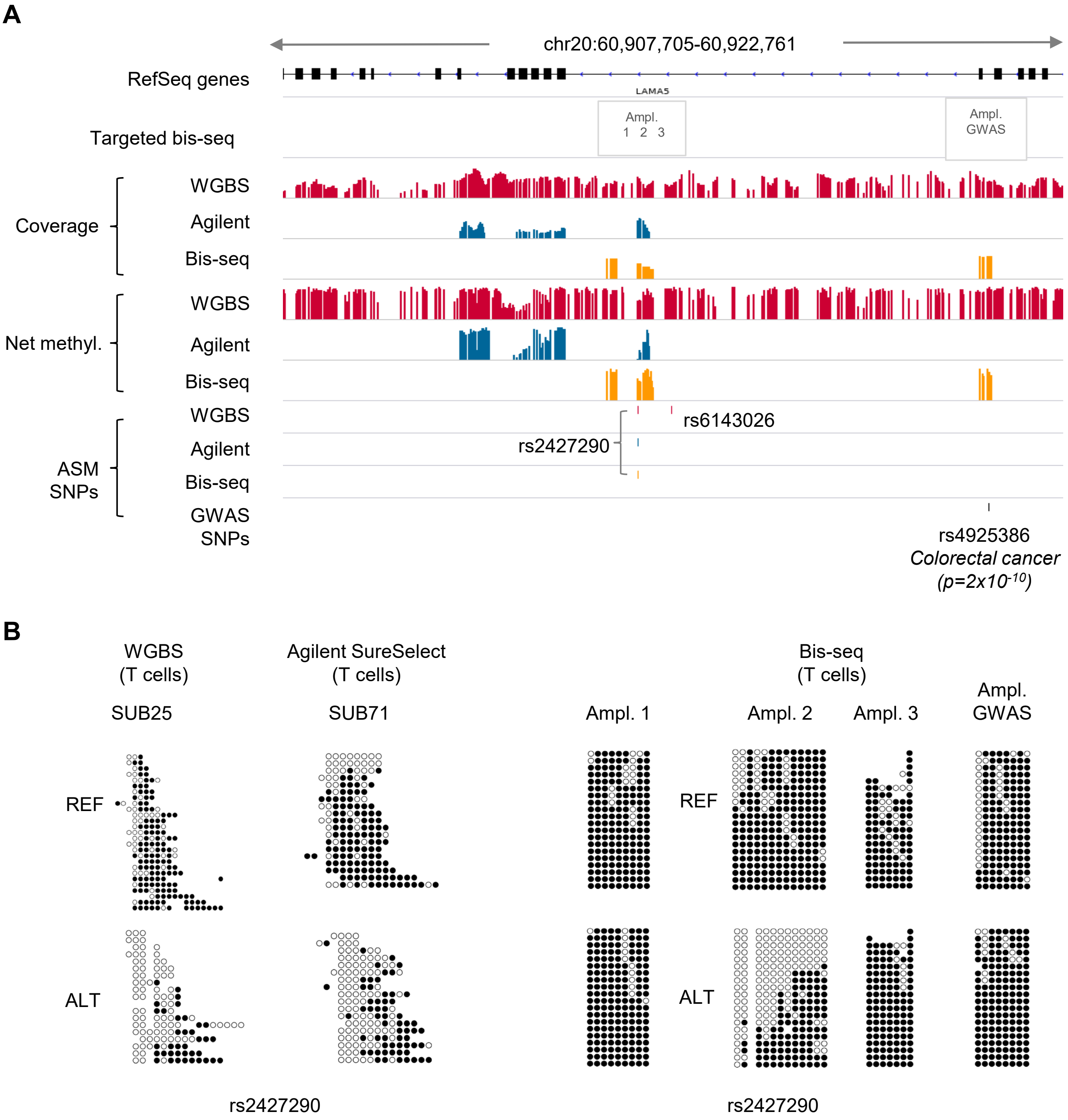

A
64,414
Percentage of ASM index SNPs
80
60
40
20
0
8,380
4,830
1
2
3
4
6
8
10
Nb. of samples
65,043
Percentage of ASM SNPs
80
60
40
20
0
8,195
4,386
1
2
3
4
6
8
10
Nb. of individuals
B
ASM SNPs found in 2 samples
34%
38%
28%
ASM SNPs found in more than 2 samples
8%
39%
53%
■ Cancer ■ Cancer and Non-cancer ■ Non-cancer
C
Percentage of ASM SNPS
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2,263,262
13,210
34%
65%
1%
36%
37%
27%
Enriched
Het SNPs >=2samples
ASM in >=2 samples
■ Cancer ■ Cancer and Non-cancer ■ Non-cancer

Figure S5



A
chr20:60,907,705-60,922,761
RefSeq genes
LAMA5
Targeted bis-seq
Ampl.
1 2 3
Ampl.
GWAS
Coverage
WGBS
Agilent
Bis-seq
Net methyl.
WGBS
Agilent
Bis-seq
ASM SNPs
WGBS
Agilent
Bis-seq
GWAS SNPs
rs6143026
rs2427290
rs4925386
Colorectal cancer
(p=2x10-10)
B
WGBS
(T cells)
Agilent SureSelect
(T cells)
Bis-seq
(T cells)
SUB25
SUB71
Ampl. 1
Ampl. 2
Ampl. 3
Ampl.
GWAS
REF
ALT
REF
ALT
rs2427290
rs2427290

Figure S6

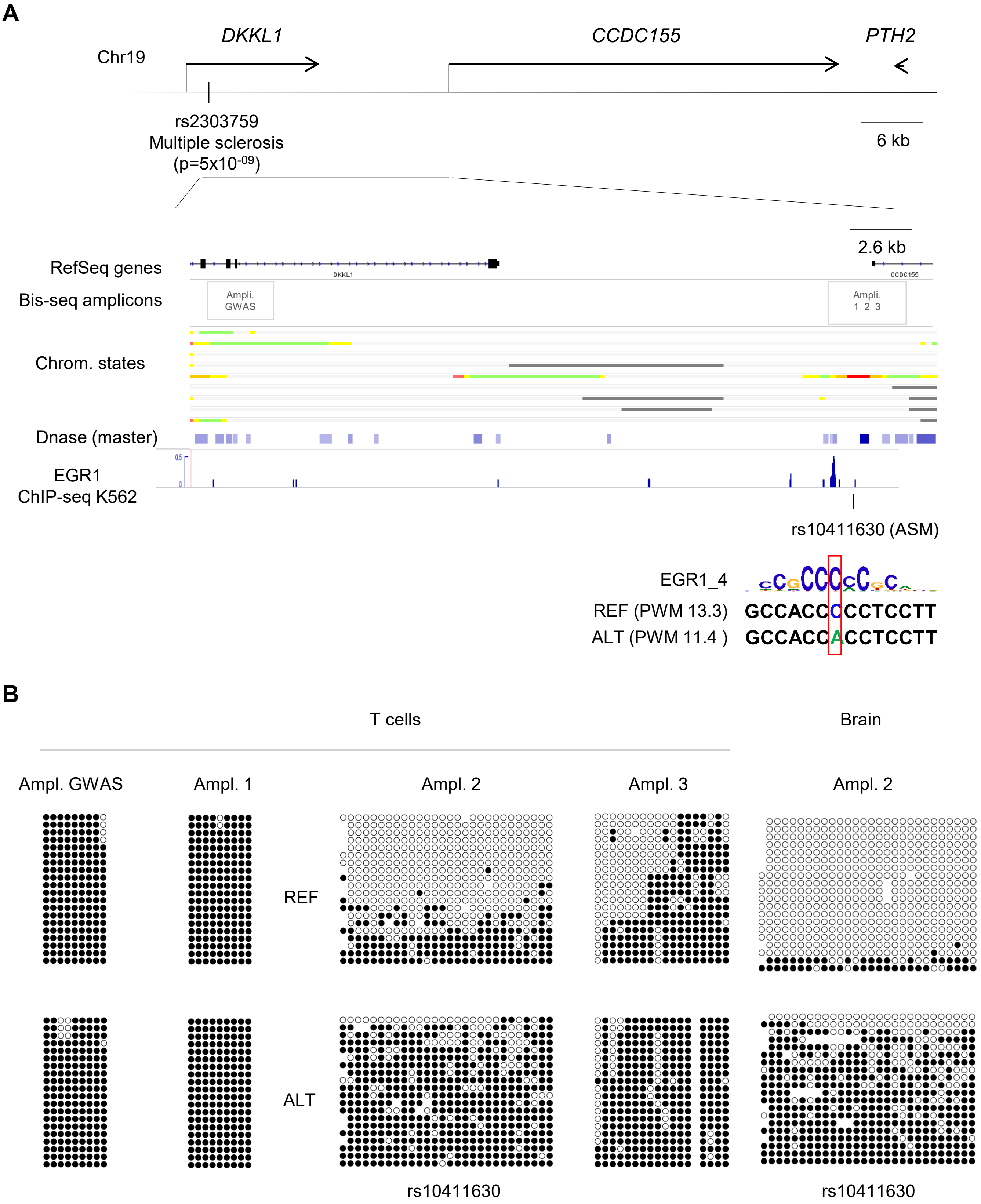
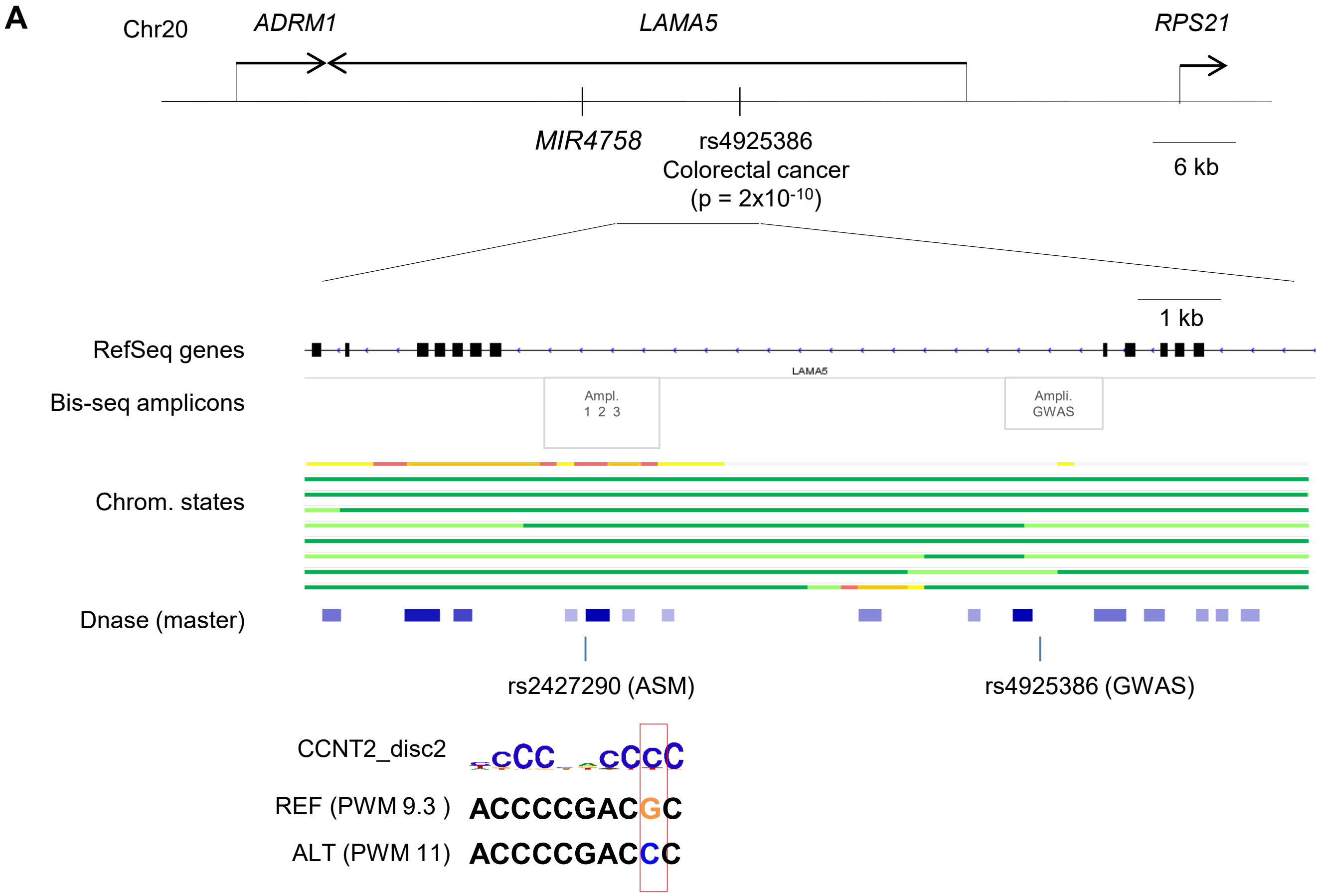

A
Chr19
DKKL1
CCDC155
PTH2
rs2303759
Multiple sclerosis
(p=5x10^{-09})
6 kb
2.6 kb
RefSeq genes
DKKL1
CCDC155
Bis-seq amplicons
Ampli. GWAS
Ampli. 1 2 3
Chrom. states
Dnase (master)
EGR1
ChIP-seq K562
0.5
0
rs10411630 (ASM)
EGR1_4
REF (PWM 13.3)
GCCACCCCCTCCTT
ALT (PWM 11.4)
GCCACCACCTCCTT
B
T cells
Brain
Ampl. GWAS
Ampl. 1
Ampl. 2
Ampl. 3
Ampl. 2
REF
ALT
rs10411630
rs10411630

Figure S7


A
Chr20
ADRM1
LAMA5
RPS21
MIR4758
rs4925386
Colorectal cancer
(p = 2x10-10)
6 kb
1 kb
RefSeq genes
LAMA5
Bis-seq amplicons
Ampl.
1 2 3
Ampli.
GWAS
Chrom. states
Dnase (master)
rs2427290 (ASM)
rs4925386 (GWAS)
CCNT2_disc2
REF (PWM 9.3)
ACCCCGACGC
ALT (PWM 11)
ACCCCGACCC

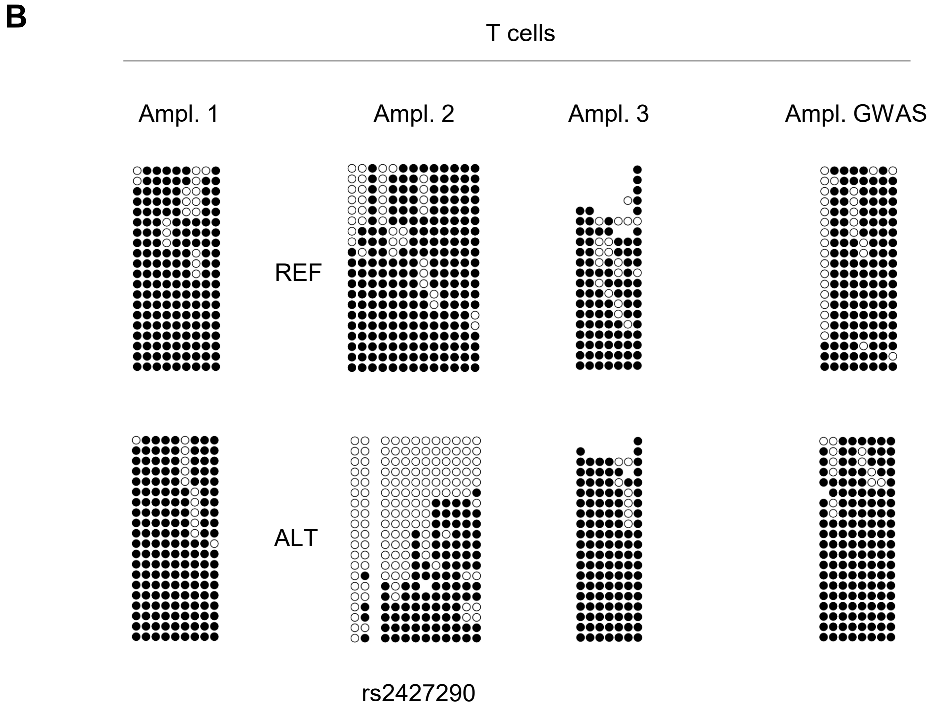

B
T cells
Ampl. 1
Ampl. 2
Ampl. 3
Ampl. GWAS
REF
ALT
rs2427290

Figure S8

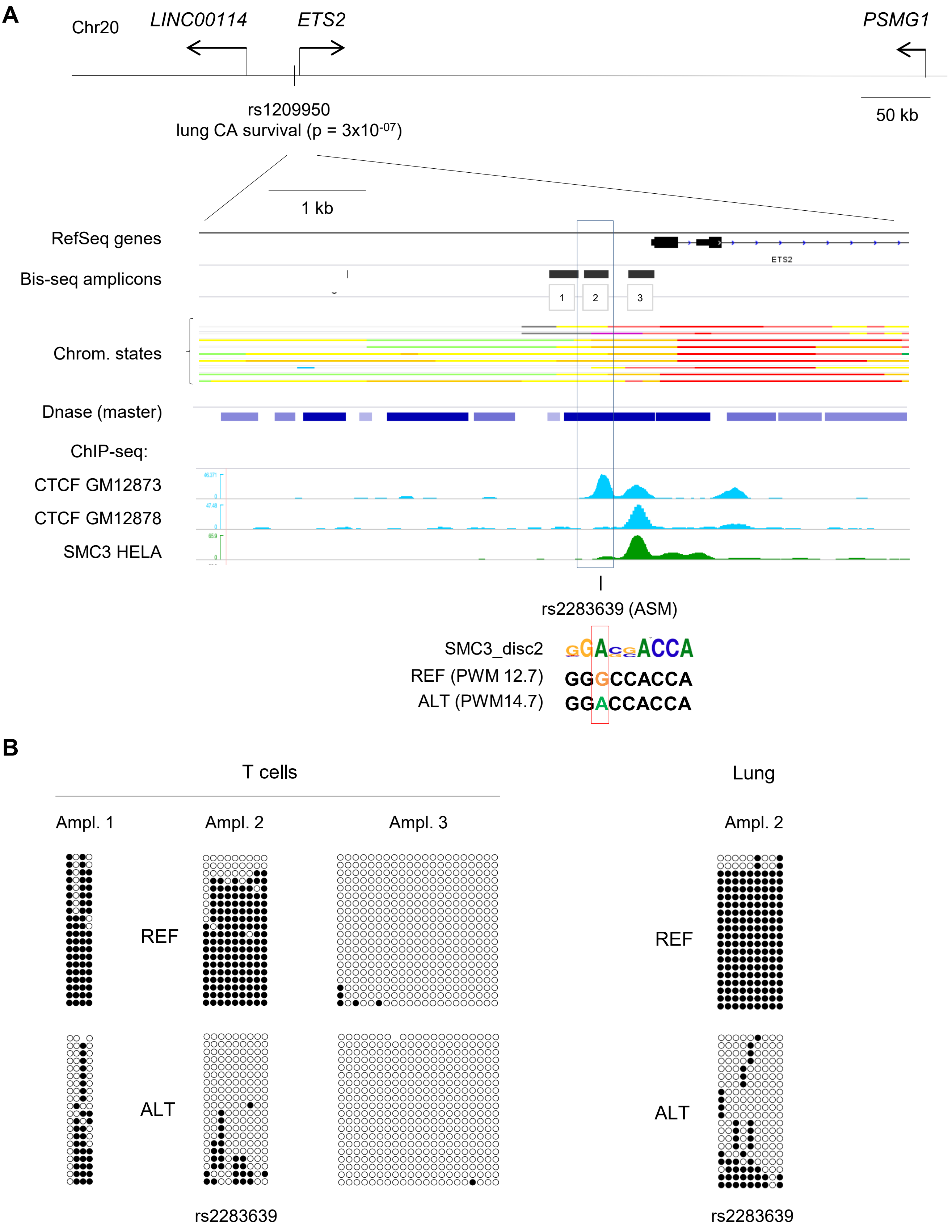

A
Chr20
LINC00114
ETS2
PSMG1
rs1209950
lung CA survival (p = 3x10-07)
50 kb
1 kb
RefSeq genes
ETS2
Bis-seq amplicons
1
2
3
Chrom. states
Dnase (master)
ChIP-seq:
CTCF GM12873
CTCF GM12878
SMC3 HELA
rs2283639 (ASM)
SMC3_disc2
REF (PWM 12.7)
GGGCCACCA
ALT (PWM14.7)
GGACCACCA
B
T cells
Lung
Ampl. 1
Ampl. 2
Ampl. 3
Ampl. 2
REF
ALT
REF
ALT
rs2283639
rs2283639

Figure S9

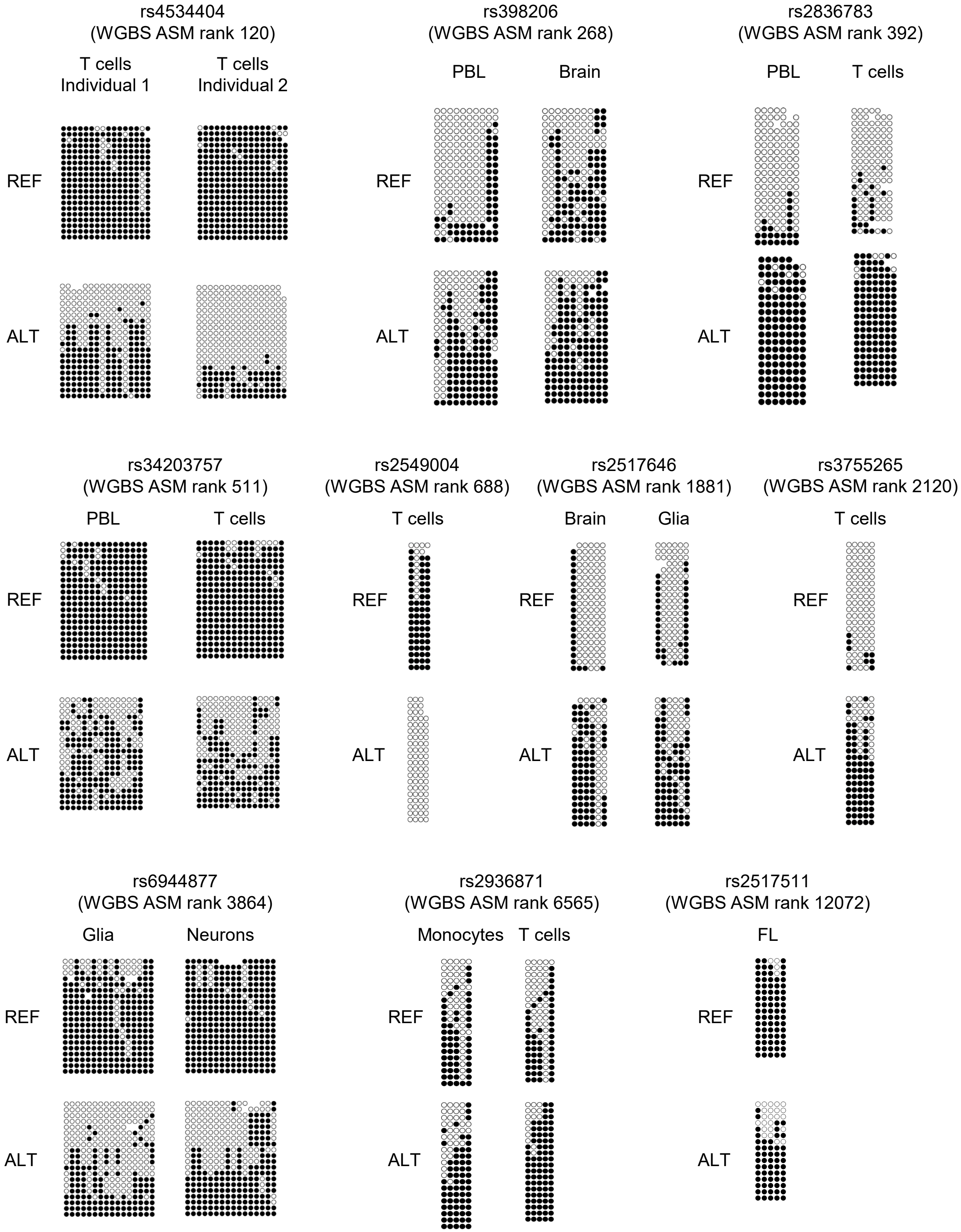
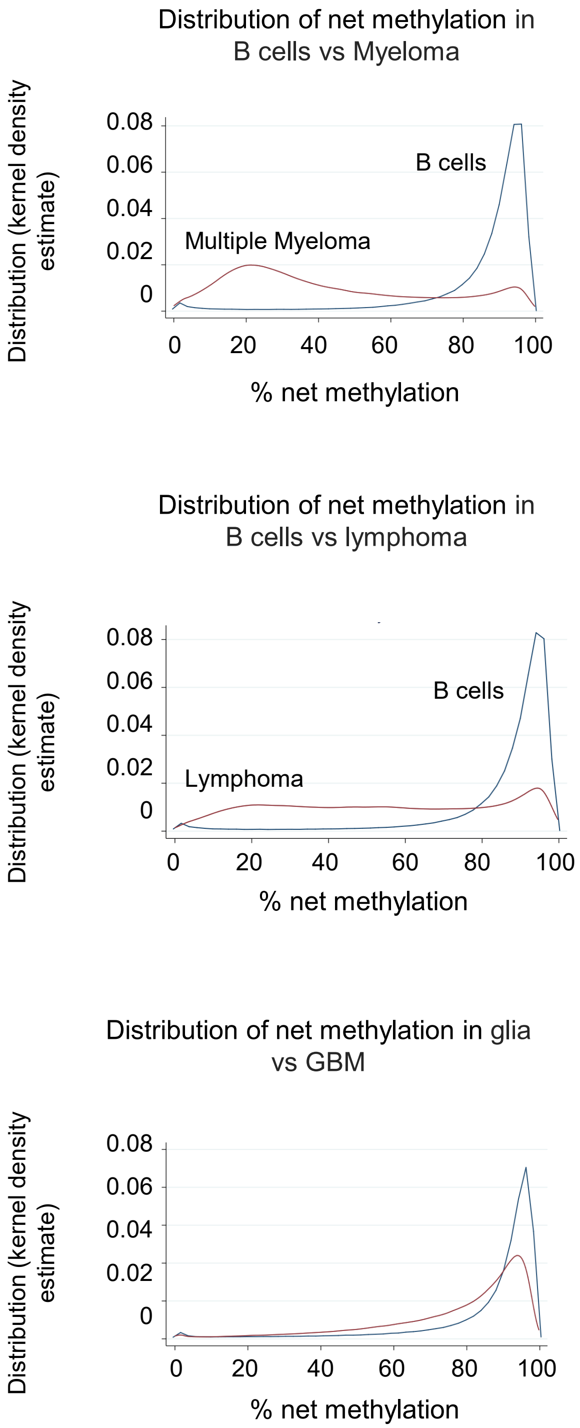

rs4534404
(WGBS ASM rank 120)
T cells
Individual 1
T cells
Individual 2
REF
ALT
rs398206
(WGBS ASM rank 268)
PBL
Brain
REF
ALT
rs2836783
(WGBS ASM rank 392)
PBL
T cells
REF
ALT
rs34203757
(WGBS ASM rank 511)
PBL
T cells
REF
ALT
rs2549004
(WGBS ASM rank 688)
T cells
REF
ALT
rs2517646
(WGBS ASM rank 1881)
Brain
Glia
REF
ALT
rs3755265
(WGBS ASM rank 2120)
T cells
REF
ALT
rs6944877
(WGBS ASM rank 3864)
Glia
Neurons
REF
ALT
rs2936871
(WGBS ASM rank 6565)
Monocytes
T cells
REF
ALT
rs2517511
(WGBS ASM rank 12072)
FL
REF
ALT

Figure S10

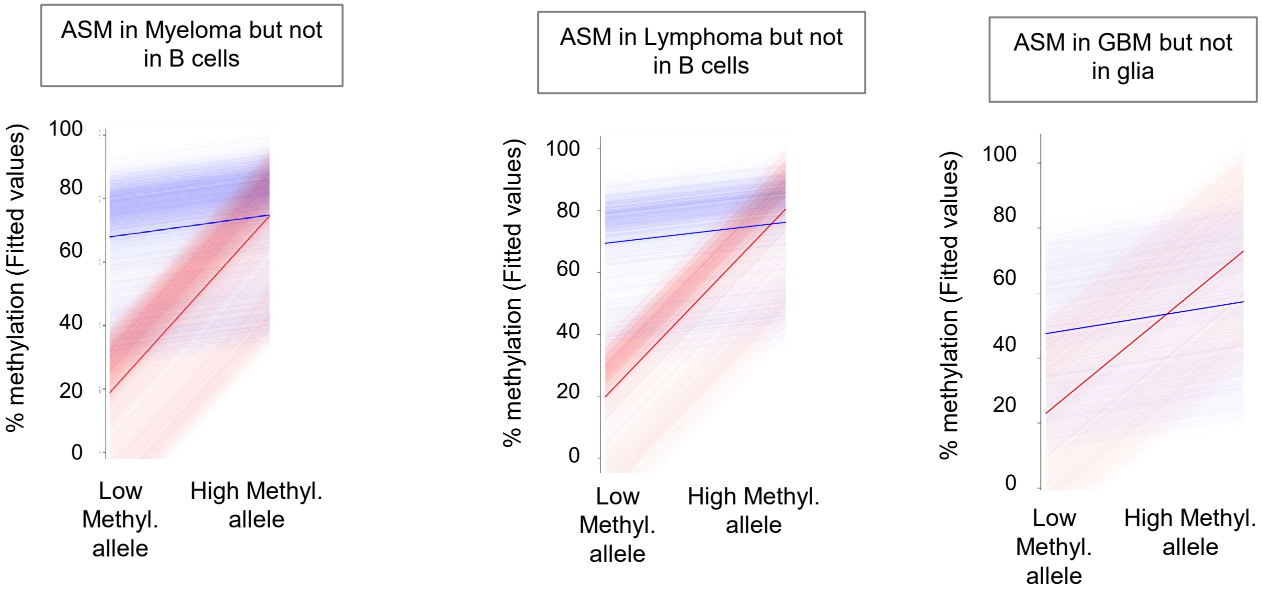


Distribution of net methylation in B cells vs Myeloma

Distribution of net methylation in B cells vs lymphoma

Distribution of net methylation in glia vs GBM

Figure S11



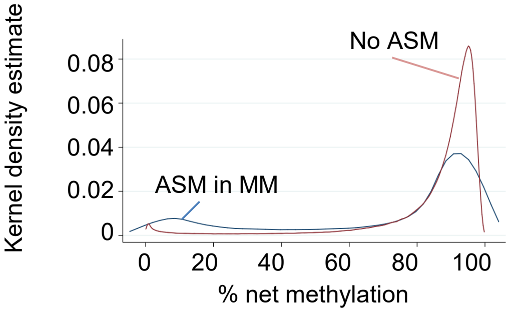
ASM in Myeloma but not in B cells

ASM in Lymphoma but not in B cells

ASM in GBM but not in glia

Figure S12

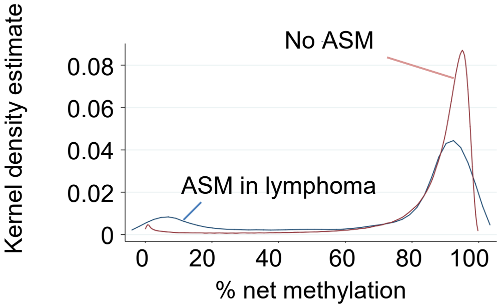
Distribution of net methylation in B cells - specifically for loci that show ASM in myeloma (BLUE) vs loci without ASM (RED)



Over-representation of low methylation (OR: 5.7, $p< 10^{-999}$)

Under-representation of high methylation (OR: 0.3, $p=6x10^{-297}$)

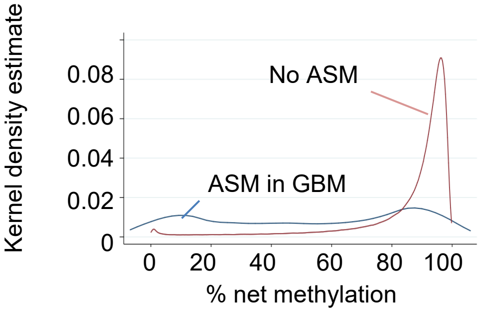
Distribution of net methylation in B cells - specifically for loci that show ASM in lymphoma (BLUE) vs loci without ASM (RED)



Over-representation of low methylation (OR: 5.6, $p=2x10^{-203}$)

Under-representation of high methylation (OR: 0.3, $p= 10^{-149}$)

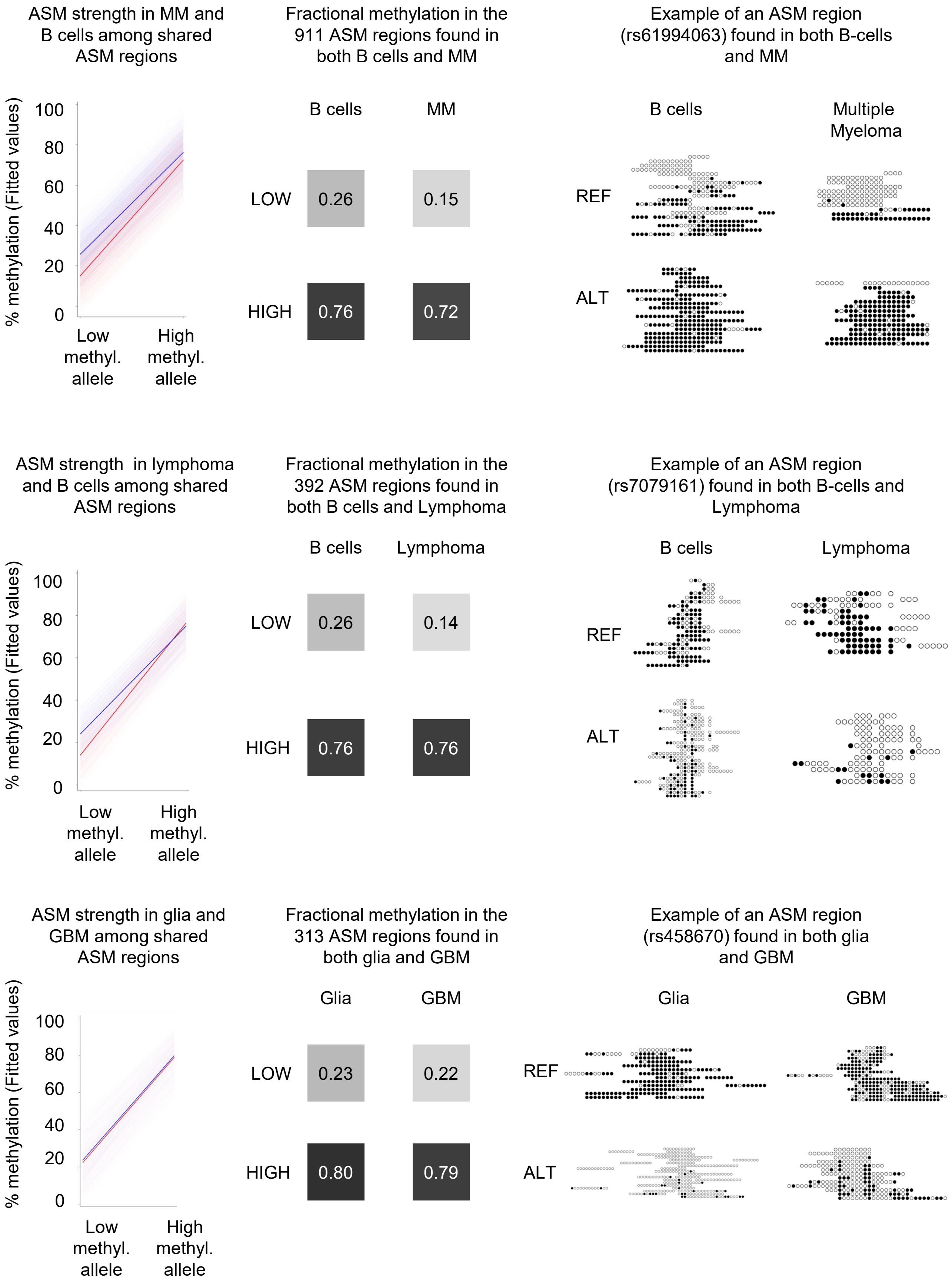
Distribution of net methylation in glia - specifically for loci that show ASM in GBM (BLUE) vs loci without ASM (RED)



Over-representation of low methylation (OR: 12, $p<10^{-999}$)

Under-representation of high methylation (OR: 0.1, $p<10^{-999}$)

Figure S13



ASM strength in MM and B cells among shared ASM regions

Fractional methylation in the 911 ASM regions found in both B cells and MM

Example of an ASM region (rs61994063) found in both B-cells and MM

ASM strength in lymphoma and B cells among shared ASM regions

Fractional methylation in the 392 ASM regions found in both B cells and Lymphoma

Example of an ASM region (rs7079161) found in both B-cells and Lymphoma

ASM strength in glia and GBM among shared ASM regions

Fractional methylation in the 313 ASM regions found in both glia and GBM

Example of an ASM region (rs458670) found in both glia and GBM

Figure S14

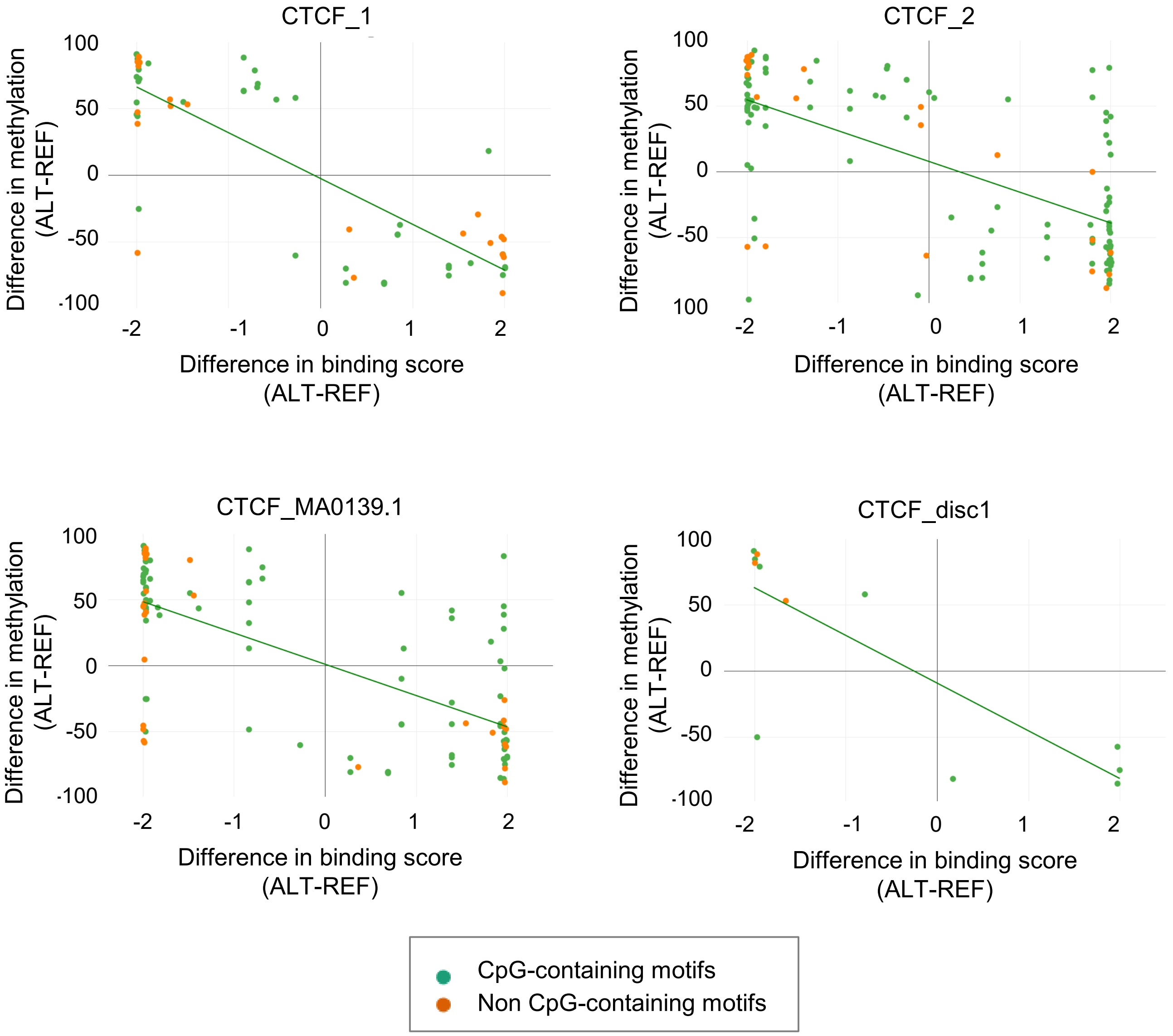

CTCF_1
CTCF_2
CTCF_MA0139.1
CTCF_disc1
Difference in methylation
(ALT-REF)
100
50
0
-50
-100
Difference in binding score
(ALT-REF)
-2
-1
0
1
2
CpG-containing motifs
Non CpG-containing motifs

Figure S15

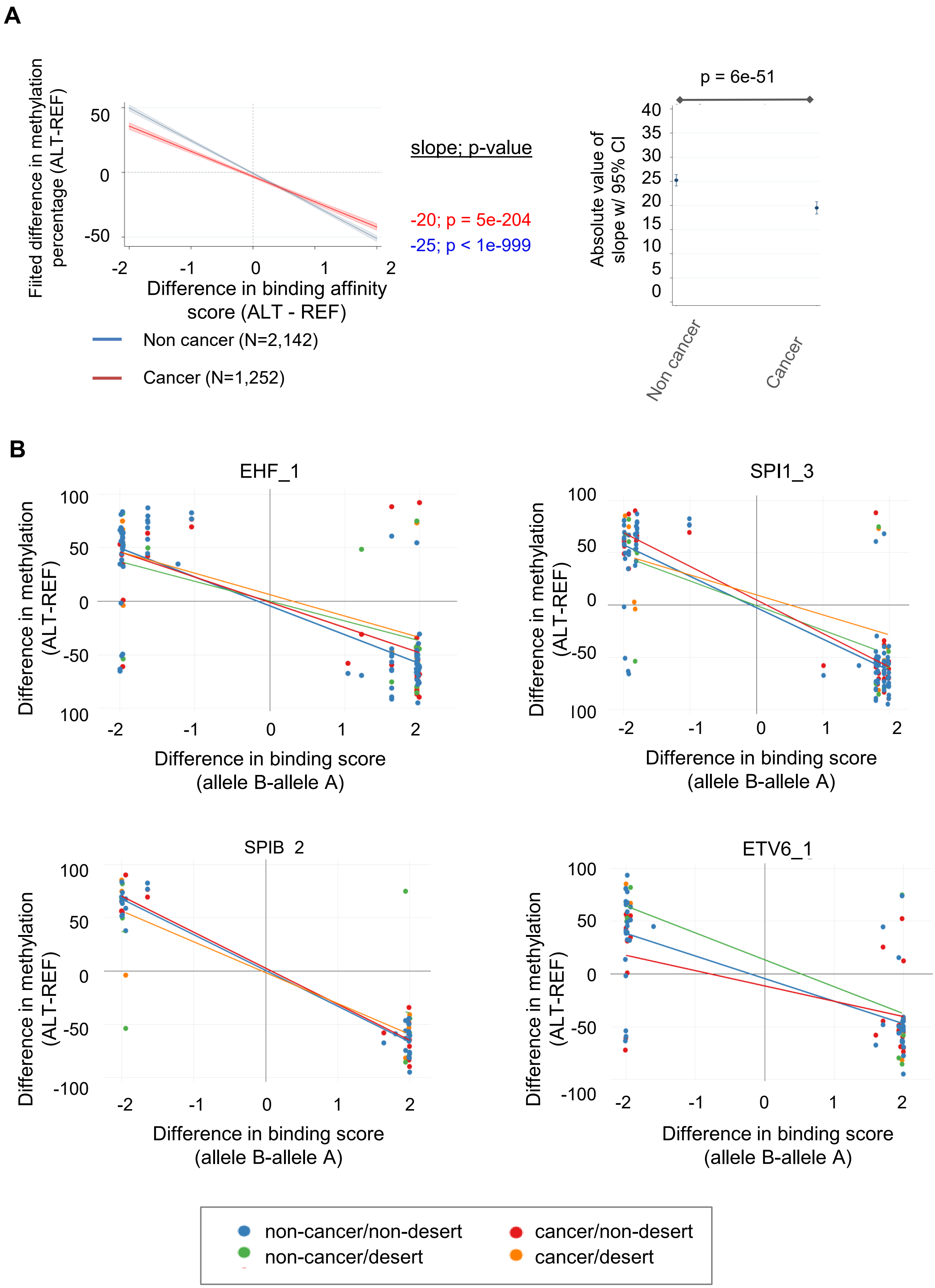

A
Fiited difference in methylation percentage (ALT-REF)
50
0
-50
-2
-1
0
1
2
Difference in binding affinity score (ALT - REF)
slope; p-value
-20; p = 5e-204
-25; p < 1e-999
Non cancer (N=2,142)
Cancer (N=1,252)
p = 6e-51
Absolute value of slope w/ 95% CI
40
35
30
25
20
15
10
5
0
Non cancer
Cancer
B
EHF_1
SPI1_3
SPIB_2
ETV6_1
Difference in methylation (ALT-REF)
100
50
0
-50
100
-100
-2
-1
0
1
2
Difference in binding score (allele B-allele A)
non-cancer/non-desert
non-cancer/desert
cancer/non-desert
cancer/desert

Figure S16

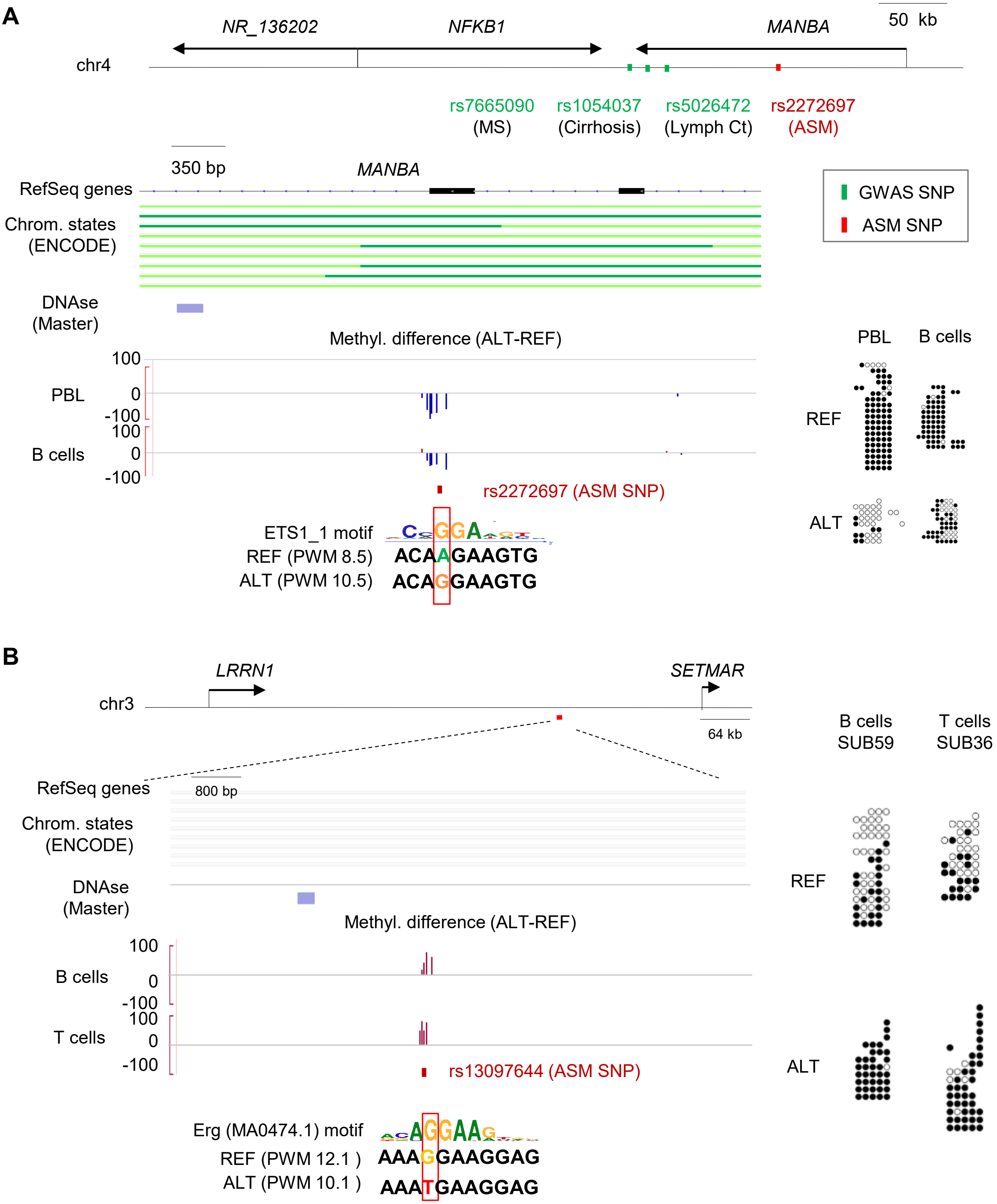

A
NR_136202
NFKB1
MANBA
50 kb
chr4
rs7665090
(MS)
rs1054037
(Cirrhosis)
rs5026472
(Lymph Ct)
rs2272697
(ASM)
350 bp
MANBA
RefSeq genes
Chrom. states
(ENCODE)
🟩 GWAS SNP
🟥 ASM SNP
DNAse
(Master)
Methyl. difference (ALT-REF)
PBL
B cells
100
0
-100
100
0
-100
PBL
B cells
REF
ALT
rs2272697 (ASM SNP)
ETS1_1 motif
REF (PWM 8.5)
ALT (PWM 10.5)
ACAAGAAGTG
ACAGGAAGTG
B
LRRN1
SETMAR
chr3
64 kb
B cells
SUB59
T cells
SUB36
RefSeq genes
800 bp
Chrom. states
(ENCODE)
DNAse
(Master)
Methyl. difference (ALT-REF)
B cells
T cells
100
0
-100
100
0
-100
REF
ALT
rs13097644 (ASM SNP)
Erg (MA0474.1) motif
REF (PWM 12.1)
ALT (PWM 10.1)
AAAGGAAGGAG
AAATGAAGGAG

Figure S17

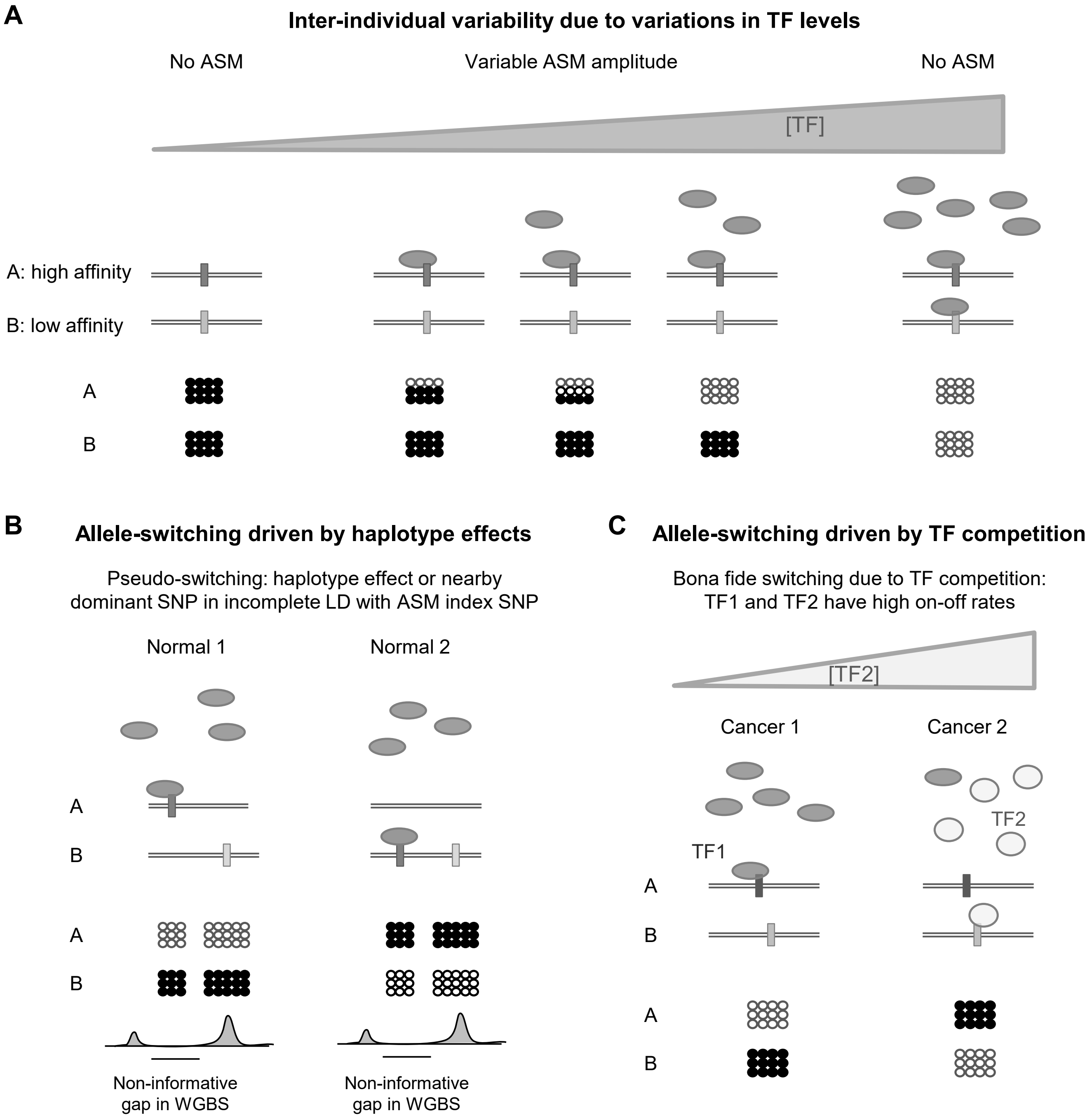

A
Inter-individual variability due to variations in TF levels
No ASM
Variable ASM amplitude
No ASM
[TF]
A: high affinity
B: low affinity
A
B
B
Allele-switching driven by haplotype effects
Pseudo-switching: haplotype effect or nearby dominant SNP in incomplete LD with ASM index SNP
Normal 1
Normal 2
A
B
A
B
Non-informative gap in WGBS
Non-informative gap in WGBS
C
Allele-switching driven by TF competition
Bona fide switching due to TF competition: TF1 and TF2 have high on-off rates
[TF2]
Cancer 1
Cancer 2
TF2
TF1
A
B
A
B

Figure S18

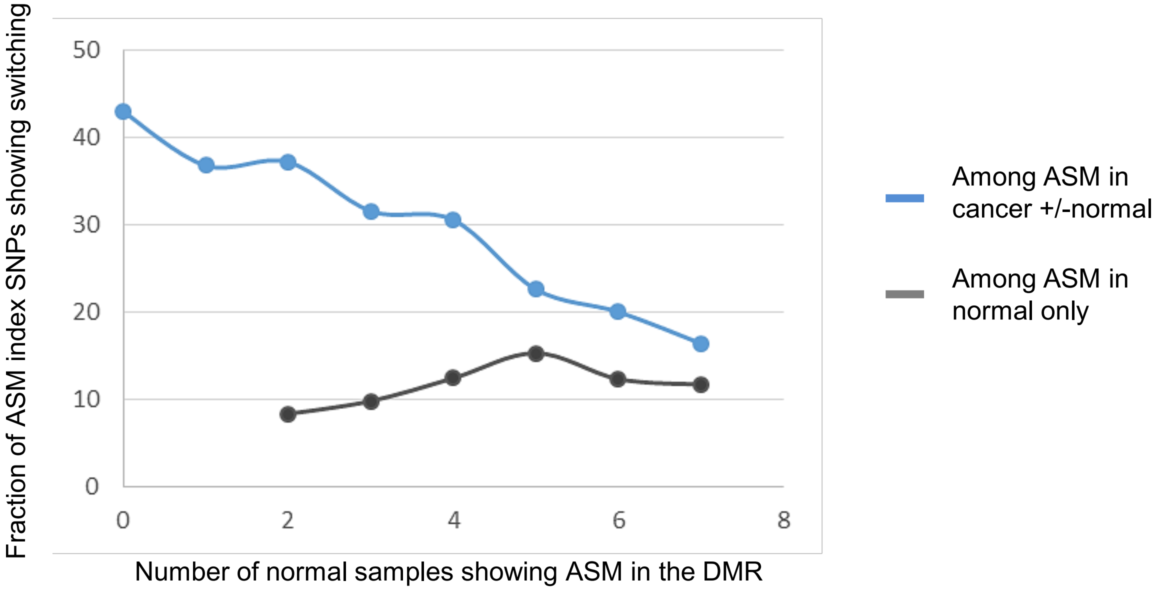
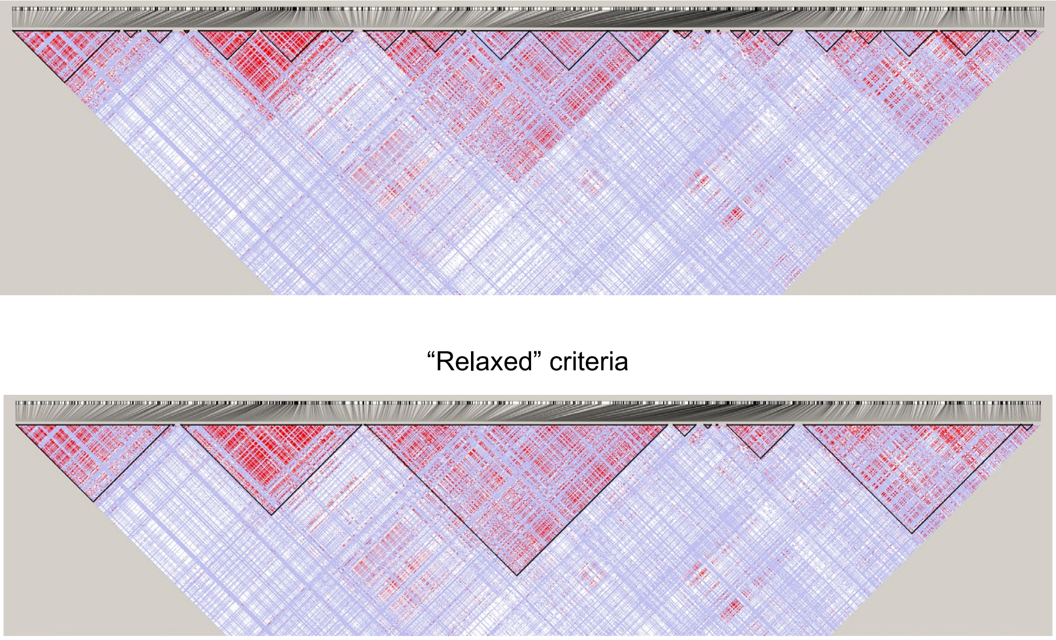

Fraction of ASM index SNPs showing switching
50
40
30
20
10
0
0
2
4
6
8
Number of normal samples showing ASM in the DMR
Among ASM in cancer +/-normal
Among ASM in normal only

Figure S19



Gabriel et al. (stringent) criteria

“Relaxed” criteria

Figure S20

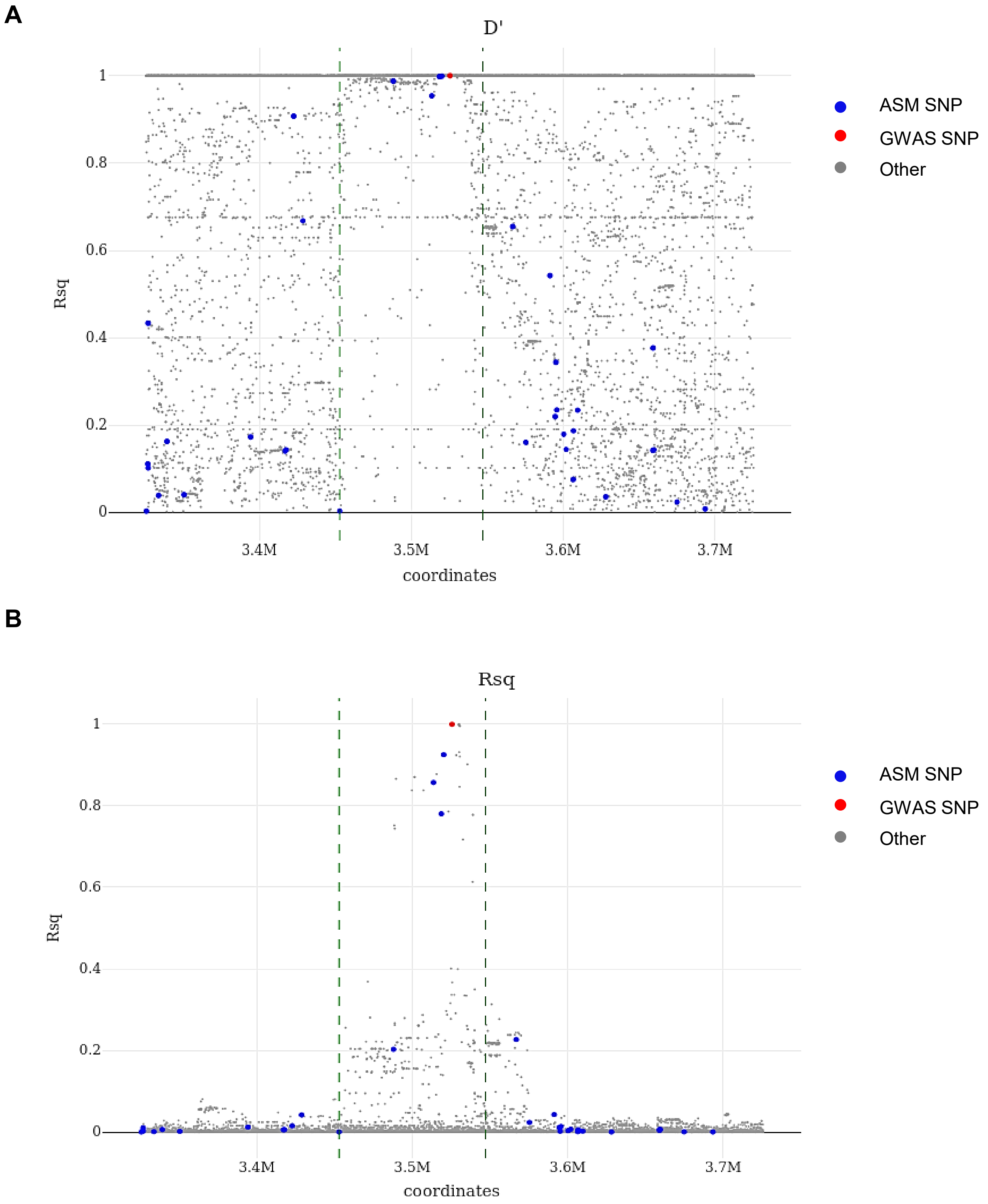

A
D'
1
0.8
0.6
0.4
0.2
0
Rsq
3.4M
3.5M
3.6M
3.7M
coordinates
● ASM SNP
● GWAS SNP
● Other
B
Rsq
1
0.8
0.6
0.4
0.2
0
Rsq
3.4M
3.5M
3.6M
3.7M
coordinates
● ASM SNP
● GWAS SNP
● Other

Figure S21

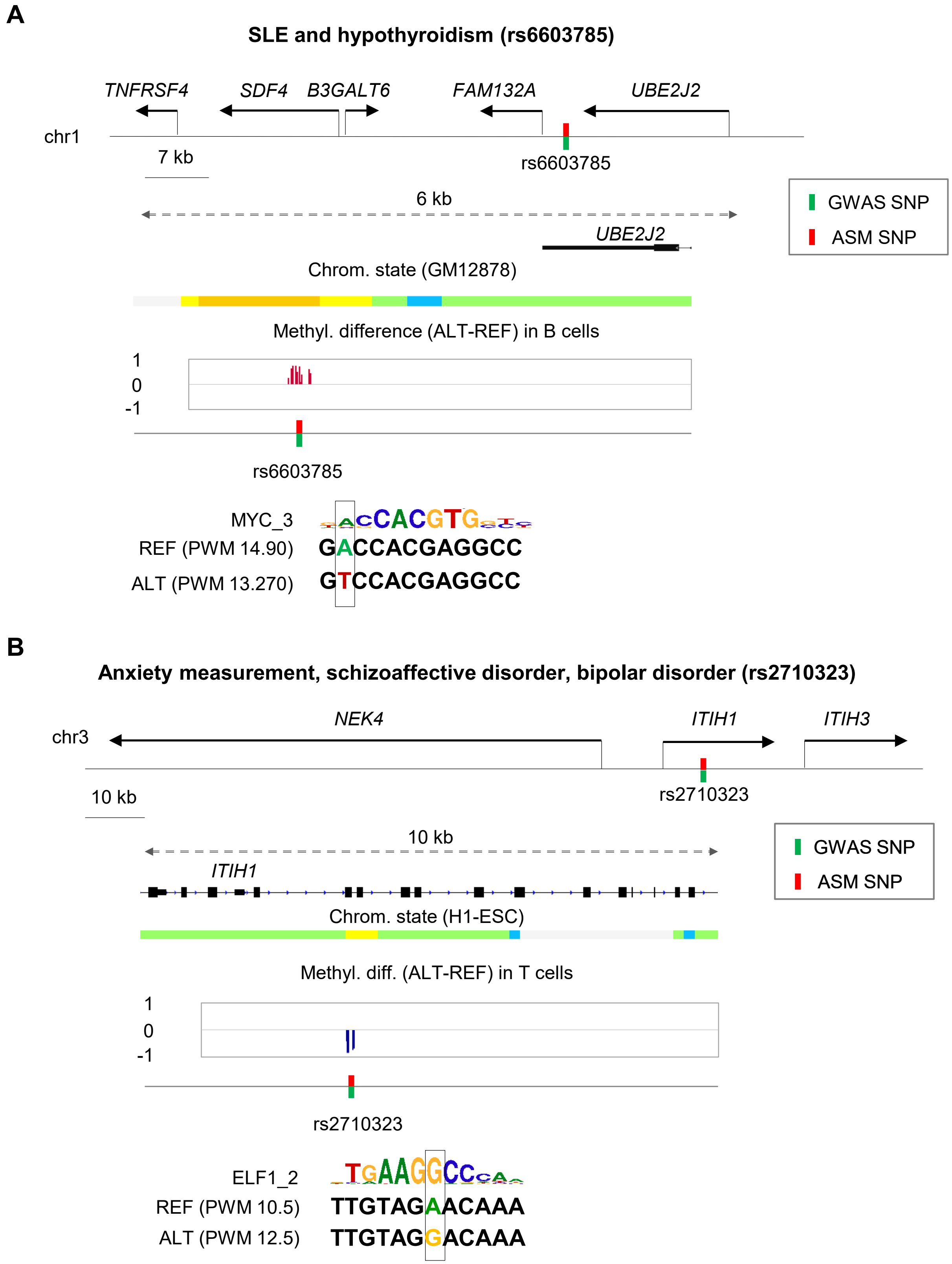
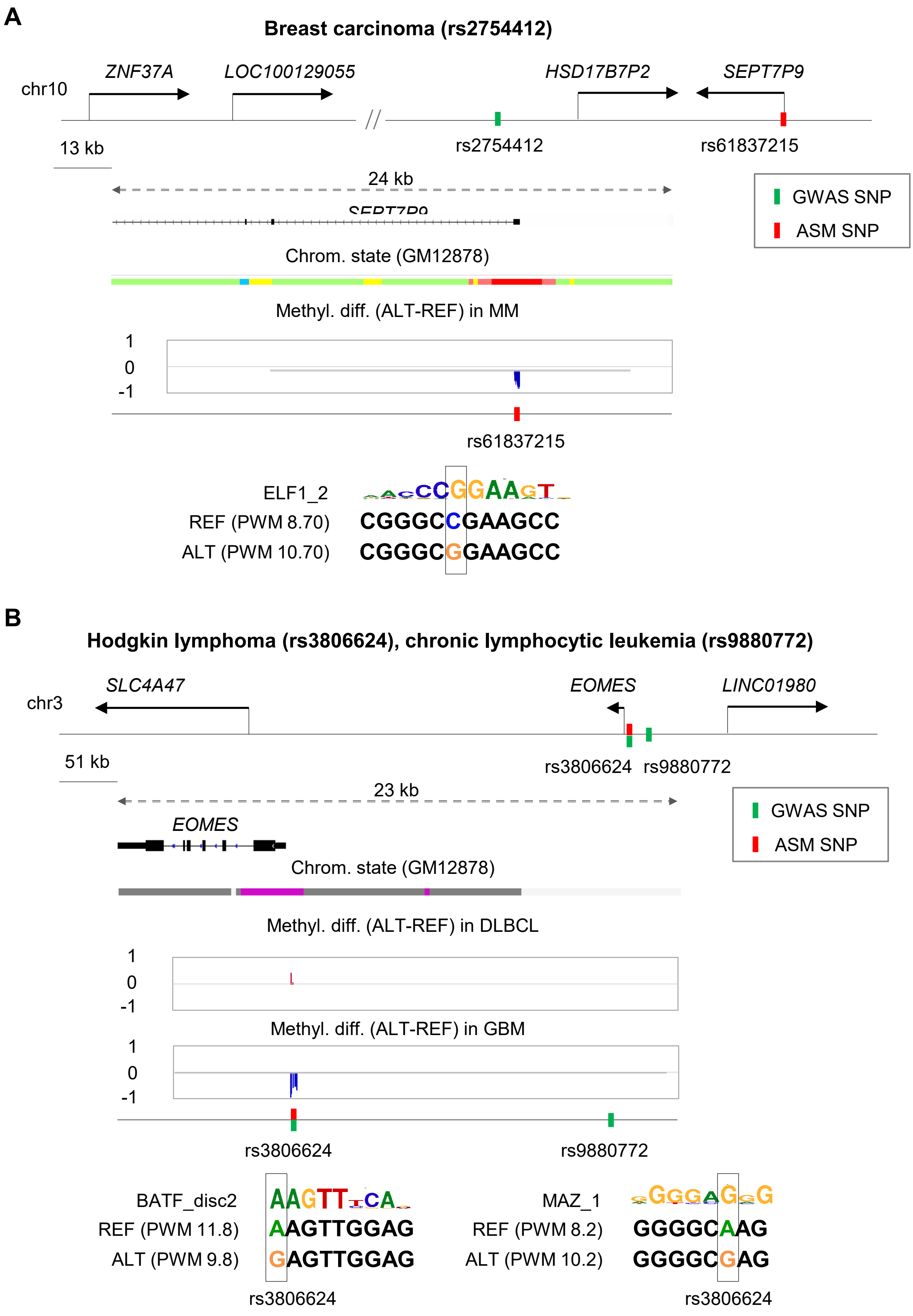

A
SLE and hypothyroidism (rs6603785)
TNFRSF4
SDF4
B3GALT6
FAM132A
UBE2J2
chr1
7 kb
rs6603785
6 kb
UBE2J2
GWAS SNP
ASM SNP
Chrom. state (GM12878)
Methyl. difference (ALT-REF) in B cells
1
0
-1
rs6603785
MYC_3
REF (PWM 14.90)
GACCACGAGGCC
ALT (PWM 13.270)
GTCCACGAGGCC
B
Anxiety measurement, schizoaffective disorder, bipolar disorder (rs2710323)
NEK4
ITIH1
ITIH3
chr3
10 kb
rs2710323
10 kb
ITIH1
GWAS SNP
ASM SNP
Chrom. state (H1-ESC)
Methyl. diff. (ALT-REF) in T cells
1
0
-1
rs2710323
ELF1_2
REF (PWM 10.5)
TTGTAGAACAAA
ALT (PWM 12.5)
TTGTAGGACAAA

Figure S22


A
Breast carcinoma (rs2754412)
chr10
ZNF37A
LOC100129055
HSD17B7P2
SEPT7P9
13 kb
rs2754412
rs61837215
24 kb
SEPT7P9
GWAS SNP
ASM SNP
Chrom. state (GM12878)
Methyl. diff. (ALT-REF) in MM
1
0
-1
rs61837215
ELF1_2
REF (PWM 8.70)
CGGGCCGAAGCC
ALT (PWM 10.70)
CGGGCGGAAGCC
B
Hodgkin lymphoma (rs3806624), chronic lymphocytic leukemia (rs9880772)
chr3
SLC4A47
EOMES
LINC01980
51 kb
rs3806624
rs9880772
23 kb
GWAS SNP
ASM SNP
EOMES
Chrom. state (GM12878)
Methyl. diff. (ALT-REF) in DLBCL
1
0
-1
Methyl. diff. (ALT-REF) in GBM
1
0
-1
rs3806624
rs9880772
BATF_disc2
REF (PWM 11.8)
AAGTTGGAG
ALT (PWM 9.8)
GAGTTGGAG
rs3806624
MAZ_1
REF (PWM 8.2)
GGGGCAAG
ALT (PWM 10.2)
GGGGCGAG
rs3806624