7 DNA methylation

ENCODE analysis identifies dynamic DNA methylation patterns and relationships to regulatory elements

Methylation of cytosine, usually at CpG dinucleotides, is involved in epigenetic regulation of gene expression. Promoter methylation is typically associated with repression, whereas genic methylation correlates with transcriptional activity[42]. We used reduced representation bisulfite sequencing (RRBS) to quantitatively profile DNA methylation for an average of 1.2 million CpGs in each of 82 cell lines and tissues (8.6% of non-repetitive genomic CpGs), including CpGs in intergenic regions, proximal promoters, and in intragenic regions (gene bodies)[43], although it should be noted that the RRBS method preferentially targets CpG rich islands. We found 96% of CpGs exhibited differential methylation in at least one cell type or tissue assayed (Varley *et al.* Personal Communication), and levels of DNA methylation correlated with chromatin accessibility. The most variably methylated CpGs are found more often in gene bodies and intergenic regions, rather than in promoters and upstream regulatory regions. In addition, we identified an unexpected correspondence between unmethylated genic CpG islands and binding by P300, a histone acetyltransferase linked to enhancer activity[44].

Because RRBS is a sequence-based assay with single-base resolution, we were able to identify CpGs with allele-specific methylation consistent with genomic imprinting, and determined that these loci exhibit aberrant methylation in cancer cell lines (Varley *et al.* Personal Communication). Furthermore, we detected reproducible cytosine methylation outside CpG dinucleotides in adult tissues[45], providing further support that this non-canonical methylation event may play important roles in human biology (Varley *et al.* Personal Communication).

Similarly, DNA methylation shows marked distinctions between segments, recapitulating the known biology of predominantly unmethylated active promoters (TSS states) followed by methylated gene bodies[42] (T state, Figure 5D). The two enhancer-enriched states show distinct patterns of DNA methylation, with the less active enhancer state (by H3K27ac/H3K4me1 levels) showing higher methylation.

BProtein-DNA interactions are also sensitive to cytosine methylation[15,16]. Comparing DNaseI footprints and whole-genome bisulphite sequencing methylation data from pulmonary fibroblasts (IMR90), we found that CpG dinucleotides contained within DNaseI footprints were significantly less methylated than CpGs in non-footprinted regions of the same DHS (Mann-Whitney *U*-test; $P < 2.2 \times 10\text{-}16$; Fig. 2b). Footprints therefore seem to be selectively sheltered from DNA methylation, indicating a widespread connection between regulatory factor occupancy and nucleotide-level patterning of epigenetic modifications.

CpG methylation has been closely linked with gene regulation, based chiefly on its association with transcriptional silencing[25]. However, the relationship between DNA methylation and chromatin structure has not been clearly defined. We analysed ENCODE reduced-representation bisulphite sequencing (RRBS) data, which provide quantitative methylation measurements for several million CpGs[26]. We focused on 243,037 CpGs falling within DHSs in 19 cell types for which both data types were available from the same sample. We observed two broad classes of sites: those with a strong inverse correlation across cell types between DNA methylation and chromatin accessibility (Fig. 4a and Supplementary Fig. 11a), and those with variable chromatin accessibility but constitutive hypomethylation (Fig. 4a, right). To quantify these trends globally, we performed a linear regression analysis between chromatin accessibility and DNA methylation at the 34,376 CpG-containing DHSs (see Supplementary Methods). Of these sites, 6,987 (20%) showed a significant association (1% FDR) between methylation and accessibility (Supplementary Fig. 11b). Increased methylation was almost uniformly negatively associated with chromatin accessibility (>97% of cases). The magnitude of the association between methylation

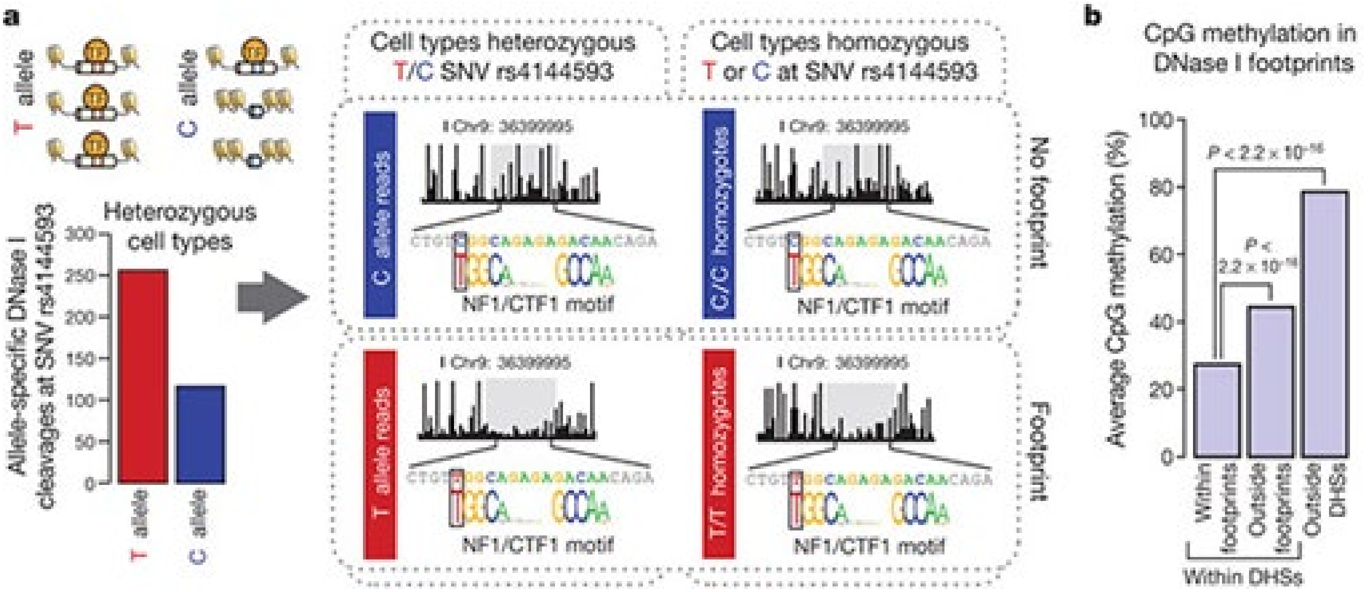


Figure 2 | DNase I footprints mark sites of *in vivo protein* occupancy. (**a**) Schematic and plots showing the effect of T/C SNV rs4144593 on protein occupancy and chromatin accessibility. The *y* axis of the bar graph shows the number of DNase I cleavage events containing either the T or C allele. Middle plots show T or C allele-specific DNase I cleavage profiles from ten cell lines heterozygous for the T/C alleles at rs4144593. Right plots show DNase I cleavage profiles from 18 cell lines homozygous for the C allele at rs4144593 and one cell line homozygous for the T allele at rs4144593. Cleavage plots are cut off at 60% cleavage height. (**b**) The average CpG methylation within IMR90 DNase I footprints, IMR90 DHSs (but not in footprints) and non-hypersensitive genomic regions in IMR90 cells. CpG methylation is significantly depleted in DNase I footprints ($P<2.2\times10^{-16}$, Mann-Whitney *U*-test).

and accessibility was strong, with the latter on average 95% lower in cell types with coinciding methylation versus cell types lacking coinciding methylation (Supplementary Fig. 11c). Fully 40% of variable methylation was associated with a concomitant effect on accessibility.

The role of DNA methylation in causation of gene silencing is presently unclear. Does methylation reduce chromatin accessibility by evicting transcription factors? Or does DNA methylation passively 'fill in' the voids left by vacating transcription factors? Transcription factor expression is closely linked with the occupancy of its binding sites[27]. If the former of the two above hypotheses is correct, methylation of individual binding site sequences should be independent of transcription factor gene expression. If the latter, methylation at transcription factor recognition sequences should be negatively correlated with transcription factor abundance (Fig. 4b).

Comparing transcription factor transcript levels to average methylation at cognate recognition sites within DHSs revealed significant negative correlations between transcription factor expression and binding site methylation for most (70%) transcription factors with a significant association ($P > 0.05$). Representative examples are shown in Fig. 4c and Supplementary Fig. 12a. These data argue strongly that methylation patterning paralleling cell-selective chromatin accessibility results from passive deposition after the vacation of transcription factors from regulatory DNA, confirming and extending other recent reports[28].

Interestingly, a small number of factors showed positive correlations between expression and binding site methylation (Supplementary Fig. 12b), including MYB and LUN-1 (also known as TOPORS). Both of these transcription factors showed increased transcription and binding site methylation specifically within acute promyelocytic leukaemia cells (NB4), and both interact with promyelocytic leukaemia (PML) bodies[29,30], a sub-nuclear structure disrupted in PML cells. The anomalous behaviour of these two transcription factors with respect to chromatin structure and DNA methylation may thus be related to a specialized mechanism seen only in pathologically altered cells.

FIGURE 3

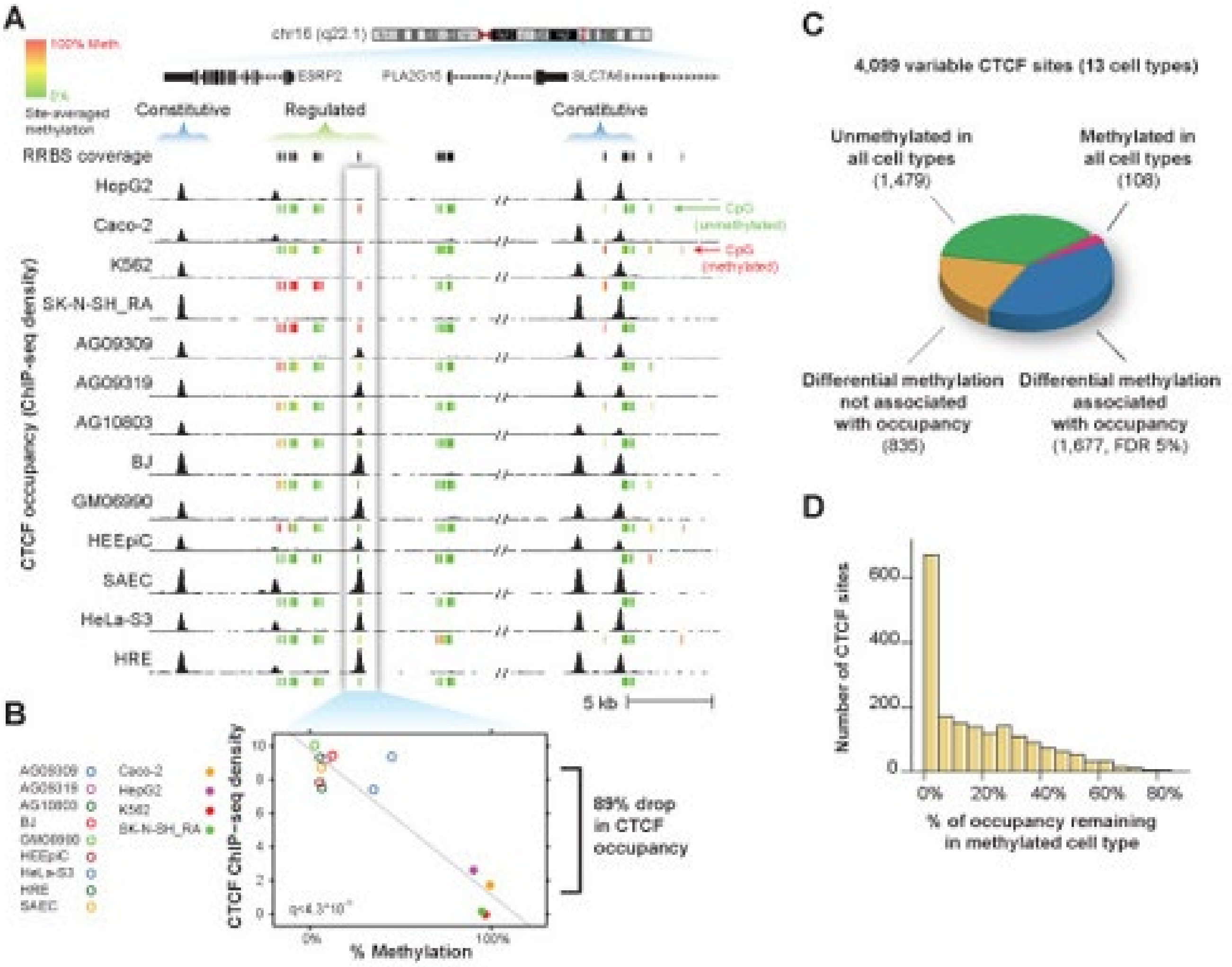


Figure 3 | Impact of DNA methylation on cell-selective CTCF binding. (a) Example CTCF binding sites, where occupancy (*above*) quantitatively increases as local CpG methylation decreases (*below*). Green indicates CpG is 0% methylated; yellow, 50%; and red, 100%. (b) Quantitative analysis of methylation at the boxed CTCF binding site in a. (c) Global impact of methylation at variable CTCF sites monitored by RRBS. Sixty-five percent of sites with cell-type selective patterns of methylation also exhibited differences in occupancy. (d) At methylated binding sites, occupancy was reduced on average by 87% compared with cell lines without methylation at the same site. Shown are sites where increased methylation was associated with decreased occupancy (98% of all significant sites).

Pre-existing methylation can antagonize CTCF binding *in vitro* (Bell and Felsenfeld 2000; Hark *et al.* 2000; Kanduri *et al.* 2000). Therefore we asked whether differential methylation was associated with variable sites *in vivo*. To study this, we compared CTCF occupancy and reduced representation bisulfite sequencing (RRBS) data (Fig. 3A). We studied a subset of CTCF sites in 13 cell types (n=6,707) for which RRBS data was available from the ENCODE project (Varley). We obtained methylation status of 44,048 CpGs dinucleotides in the region centered on these sites (see Methods), with each CpG monitored in an average of 12 out of 13 cell types (Supplemental Fig. S6).

First, we assessed overall methylation status at the 6,707 CTCF sites with RRBS data. We found that methylation was substantially more variable at variable CTCF sites than at constitutive ones (Supplementary Fig. S5). Only 10% of these sites tested showed intermediate methylation status (between 25% and 75% methylation) (Supplemental Fig. S6). Overall, 98% of CTCF sites were unmethylated (defined as <50% methylation) in at least one of the cell types tested, confirming an inverse relationship between methylation and CTCF occupancy. However, 47% of CTCF sites were methylated (>50% methylation) in at least one cell type, suggesting a widespread potential link between methylation and CTCF occupancy.

FIGURE 4

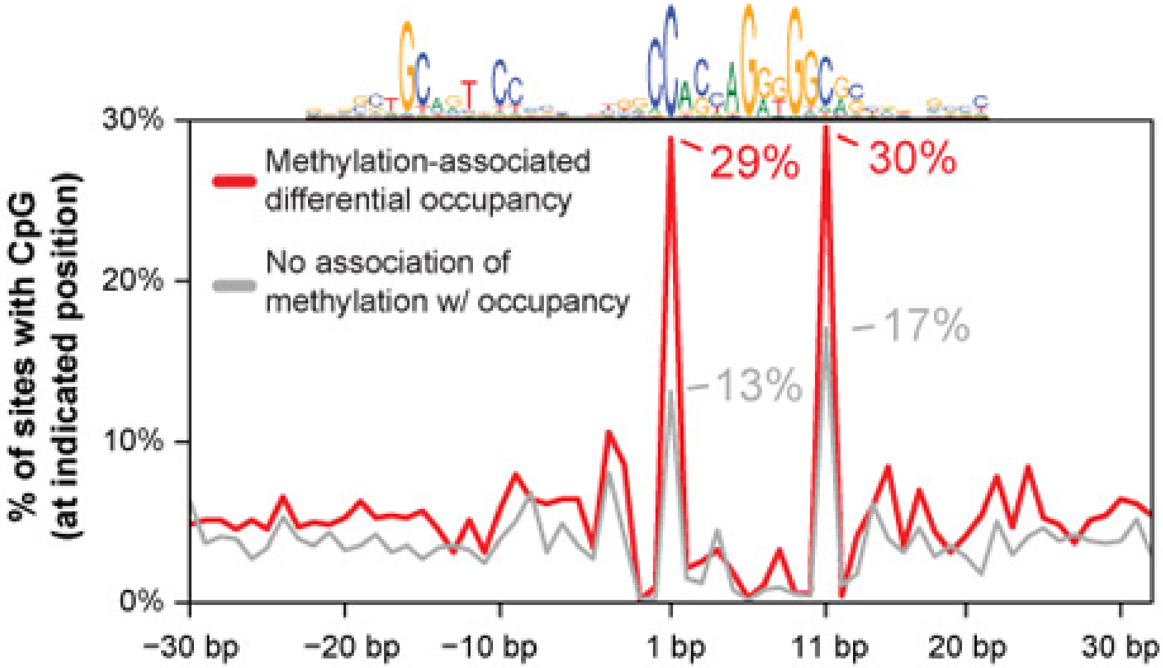


Figure 4 | Sites significantly affected by methylation are enriched for CpGs at two positions. Frequency of a CpG (*y*-axis) at positions relative to the CTCF motif (*x*-axis) is shown for sites with variable methylation that is associated (red) and is not associated (gray) with occupancy changes. Note that at positions 1 and 11, there is a 2.2- and 1.8-fold enrichment, respectively, for the presence of a CpG at sites where the variable methylation was not associated with occupancy. Twenty-nine percent of CTCF motifs genome-wide contain a CpG at one or both of these positions.

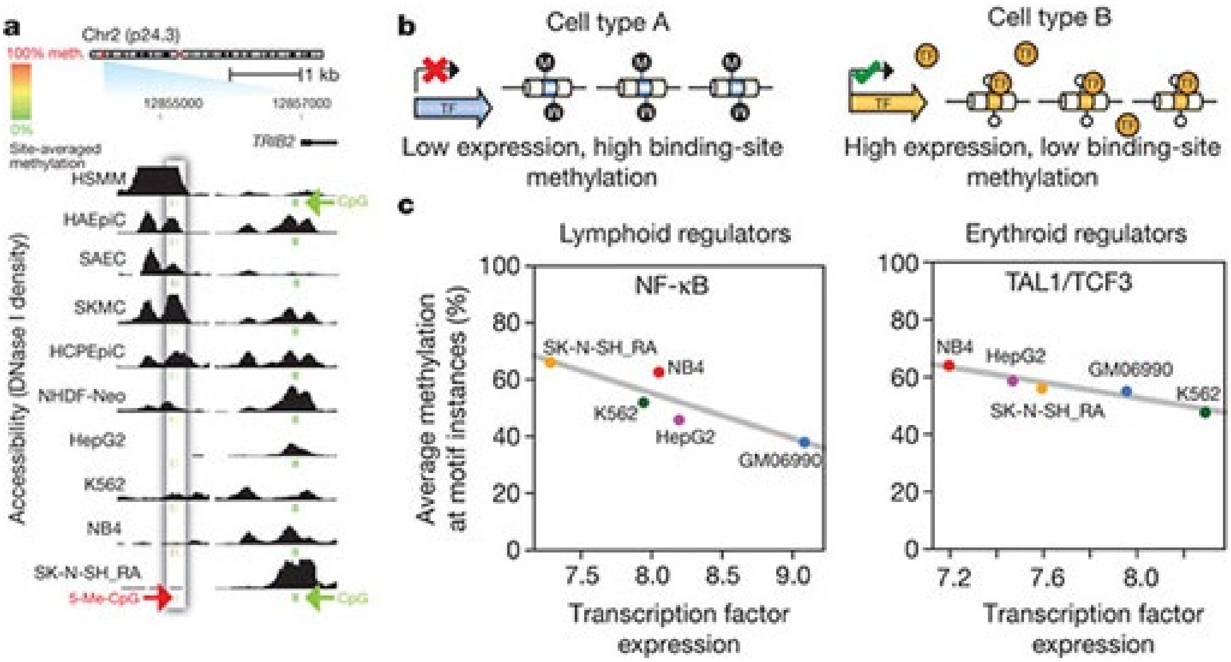


Figure 4 | Chromatin accessibility and DNA methylation patterns. (a) DNase I sensitivity in 10 cell types with ENCODE reduced representation bisulphite sequencing data. Inset box: accessibility (*y* axis) decreases quantitatively as methylation increases. Other DHSs (right) show low correlation between accessibility and methylation. CpG methylation scale: green, 0%; yellow, 50%; red, 100%. (b) Model of transcription factor (TF)-driven methylation patterns in which methylation passively mirrors transcription actor occupancy. (c) Relationship between transcription factor transcript levels and overall methylation at cognate recognition sequences of the same transcription factors. Lymphoid regulators in B-lymphoblastoid line GM06990 (left) and erythroid regulators in the erythroleukaemia line K562 (right). Negative correlation indicates that site-specific DNA methylation follows transcription factor vacation of differentially expressed transcription factors.

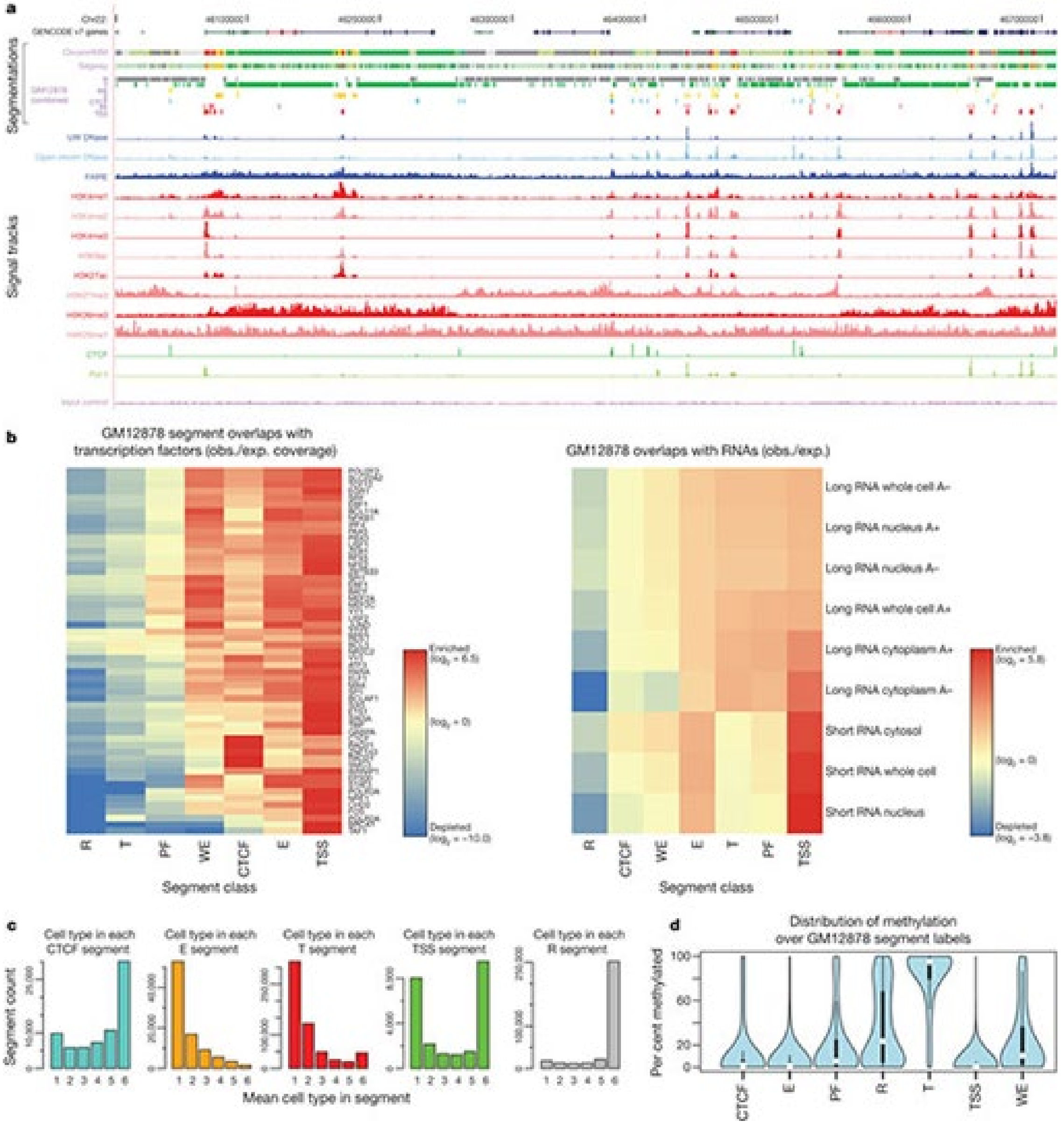


Figure 5 | Integration of ENCODE data by genome-wide segmentation. (**a**) Illustrative region with the two segmentation methods (ChromHMM and Segway) in a dense view and the combined segmentation expanded to show each state in GM12878 cells, beneath a compressed view of the GENCODE gene annotations. Note that at this level of zoom and genome browser resolution, some segments appear to overlap although they do not. Segmentation classes are named and coloured according to the scheme in Table 3. Beneath the segmentations are shown each of the normalized signals that were used as the input data for the segmentations. Open chromatin signals from DNase-seq from the University of Washington group (UW DNase) or the ENCODE open chromatin group (Openchrom DNase) and FAIRE assays are shown in blue; signal from histone modification ChIP-seq in red; and transcription factor ChIP-seq signal for Pol II and CTCF in green. The mauve ChIP-seq control signal (input control) at the bottom was also included as an input to the segmentation. (**b**) Association of selected transcription factor (left) and RNA (right) elements in the combined segmentation states (*x* axis) expressed as an observed/expected ratio (obs./exp.) for each combination of transcription factor or RNA element and segmentation class using the heat-map scale shown in the key besides each heat map. (**c**) Variability of states between cell lines, showing the distribution of occurrences of the state in the six cell lines at specific genome locations: from unique to one cell line to ubiquitous in all six cell lines for five states (CTCF, E, T, TSS and R). (**d**) Distribution of methylation level at individual sites from RRBS analysis in GM12878 cells across the different states, showing the expected hypomethylation at TSSs and hypermethylation of genes bodies (T state) and repressed (R) regions.

FIGURE 5

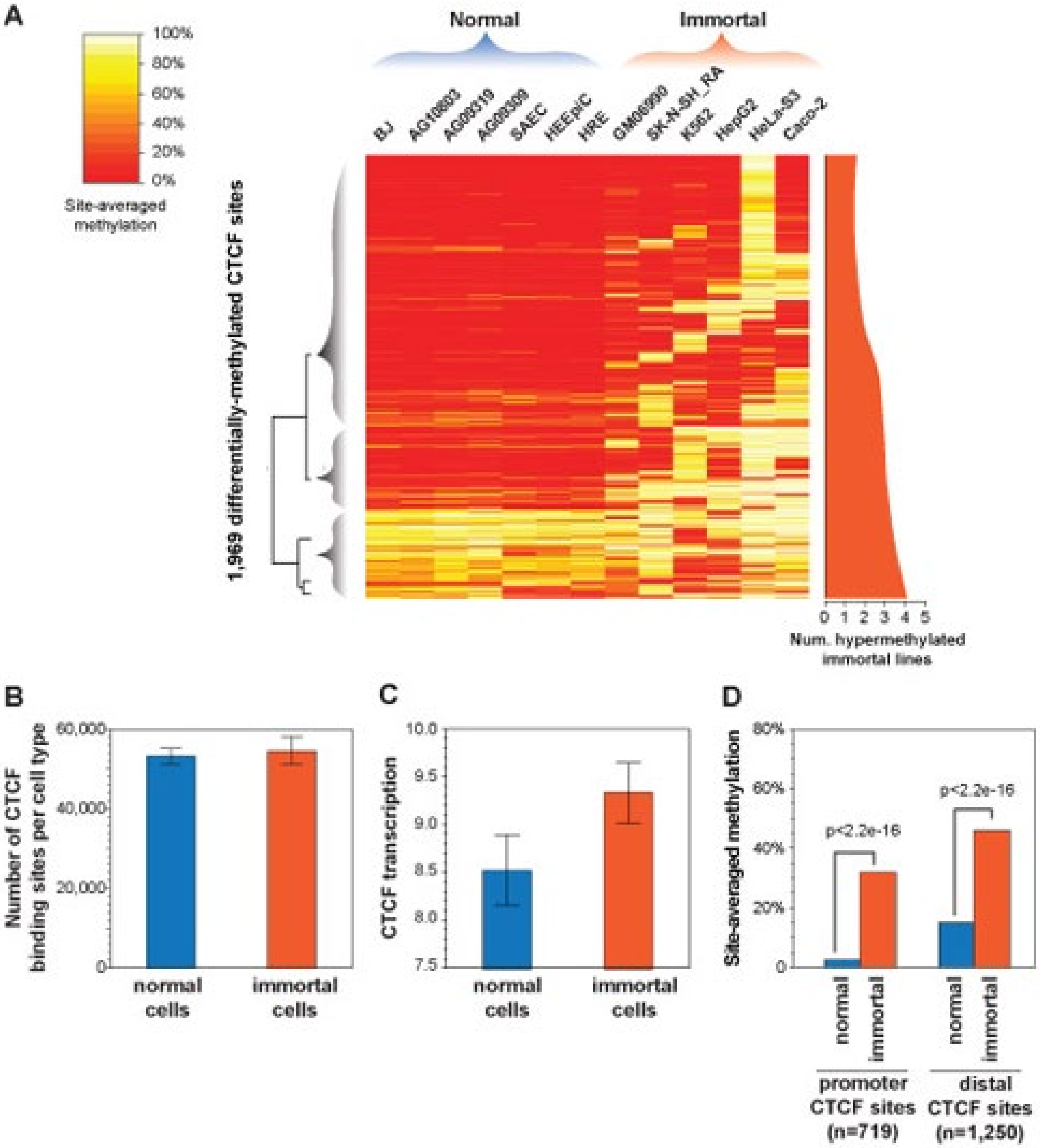
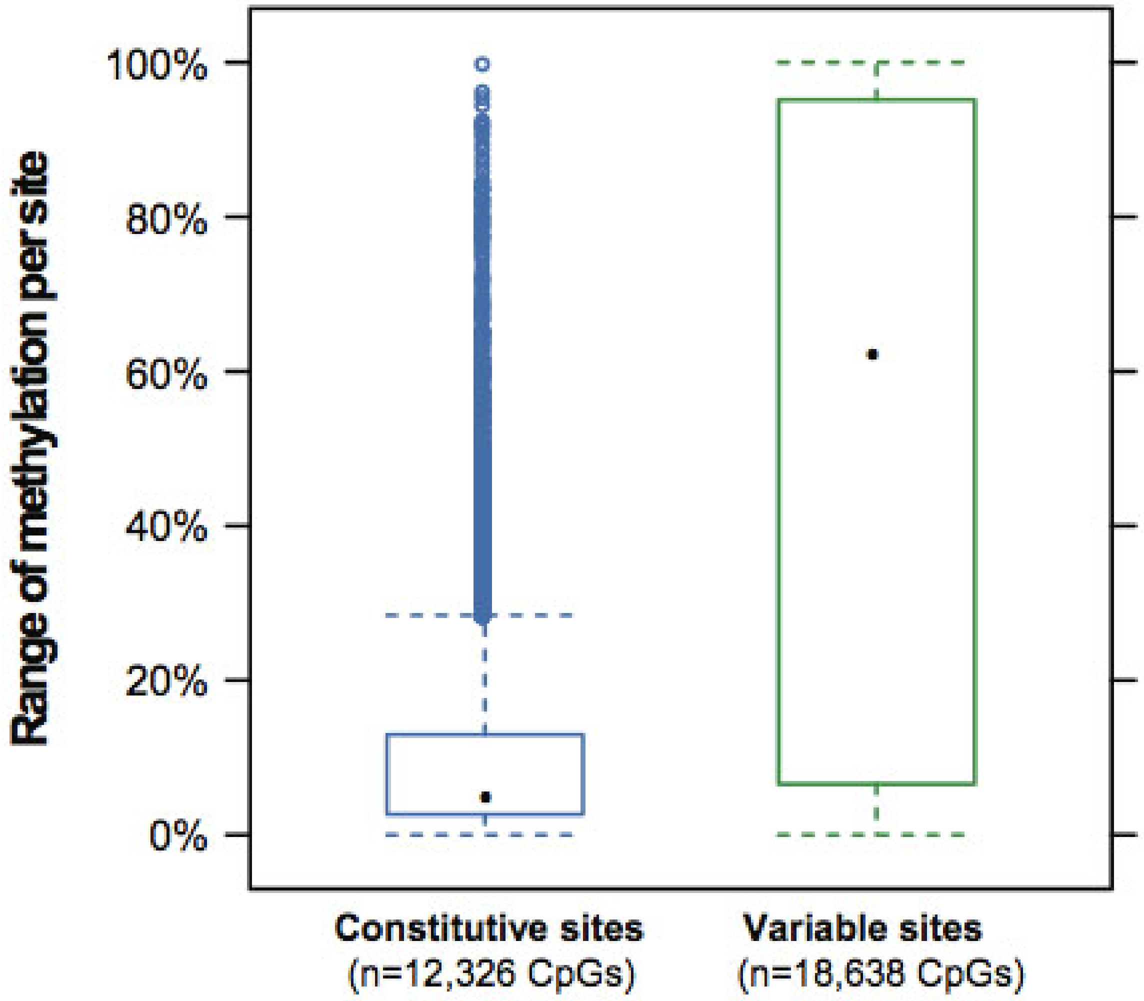


Figure 5 | Cell-selective patterns of methylation associated with occupancy differences. (a) Methylation status at 1969 CTCF sites where differential methylation is significantly associated with occupancy differences. Color corresponds to the percentage of bisulfite sequencing tags at each site overlapping methylated CpG positions. Dendrogram (*left*) highlights pattern of hypermethylation in immortal cell lines. (*Right*) Smoothed plot of number of immortal lines exhibiting hypermethylation at each site. (b) Immortal lines show no significant difference in number of occupied CTCF sites (*y*-axis, mean). Error bars, SD. (c) immortal lines demonstrate increased CTCF transcript levels (*y*-axis, mean). Error bars, SD. (d) Immortal lines exhibit increased methylation relative to the other cell types, though significant promoter methylation is rarely observed in normal lines. *y*-axis, genome-wide median of per-site methylation. *P*-values, Wilcoxon. Promoter, ±2.5 kb of RefSeq transcription start site.

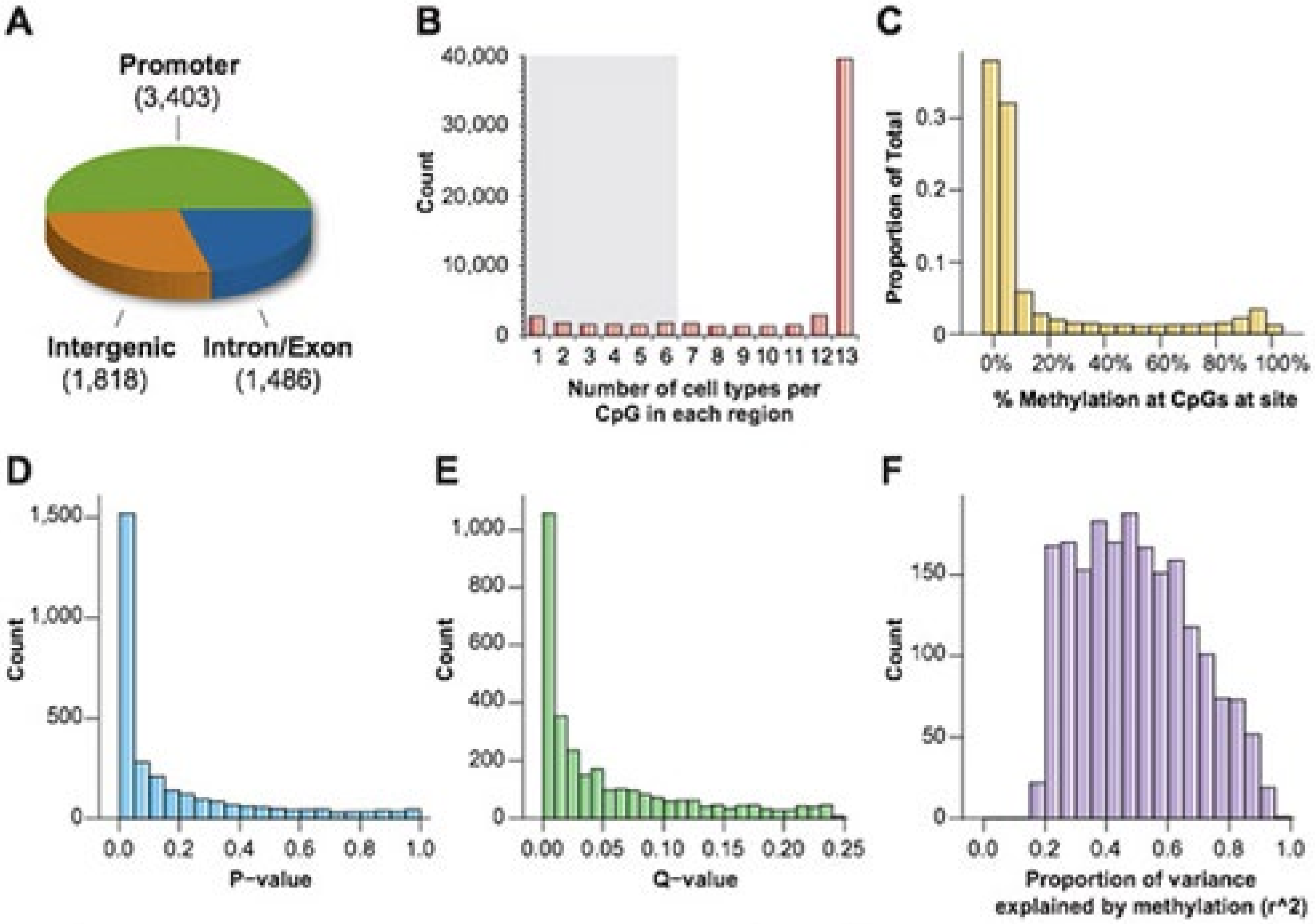
To quantify the global association of differential methylation status with variable CTCF occupancy, we performed a linear regression analysis at the 6,707 sites for which we had RRBS data (Fig. 3B; see Methods). 4,099 (61%) of these sites exhibited variable CTCF binding in the 13 cell types tested. Of the 4,099 variable sites with RRBS



Supplementary Figure S5 | Overall methylation variability across constitutive and variable CTCF sites in 13 cell types. ***Y*-axis denotes the range between the most- and least- methylated cell types at each CpG overlapping constitutive and variable CTCF sites.**

data, 1,677 (41%) showed a significant association (5% FDR) between methylation and occupancy (Fig. 3C). At significant sites, increased methylation was negatively associated with occupancy in 98% of cases. The magnitude of the association between methylation and occupancy was strong-occupancy was on average 87% lower at significant sites in the methylated cell types relative to the unmethylated cell types (Fig. 3D). Further supporting a strong link to methylation, 67% of variable methylation was associated with a concomitant affect on occupancy. The remaining 36% of sites with variable methylation that was not associated with occupancy nevertheless demonstrated an aggregate reduction in occupancy in methylated cell types (Supplemental Fig. S7), confirming the overall inverse association of methylation with CTCF occupancy, but suggesting that this relationship may be complicated by additional factors at this subset of sites.

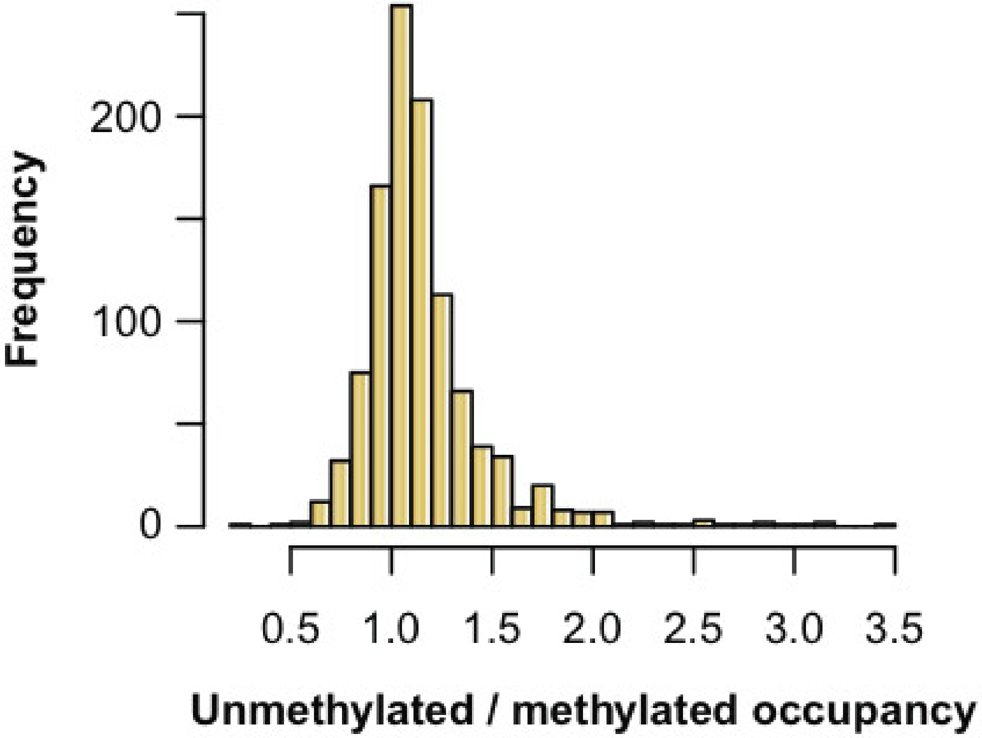
Next we asked whether the inverse relationship between methylation and CTCF occupancy is characterized by regional hypermethylation or if instead methylation is concentrated specifically at the region of protein-DNA interaction. We examined the location of all CpG dinucleotides relative to the CTCF motif at sites with variable methylation. Indeed, sites of differential methylation associated with occupancy differences showed an enrichment of CpG dinucleotides at two positions in the CTCF recognition sequence (Fig. 4). This finding is consistent with previous reports showing methylation outside the recognition sequence does not affect CTCF binding *in vitro* (Engel *et al.* 2004; Chadwick 2008). Within the recognition sequence, methylation at one of these CpGs (position 1) has been shown to inhibit binding of CTCF *in vitro* (Renda *et al.* 2007). The second



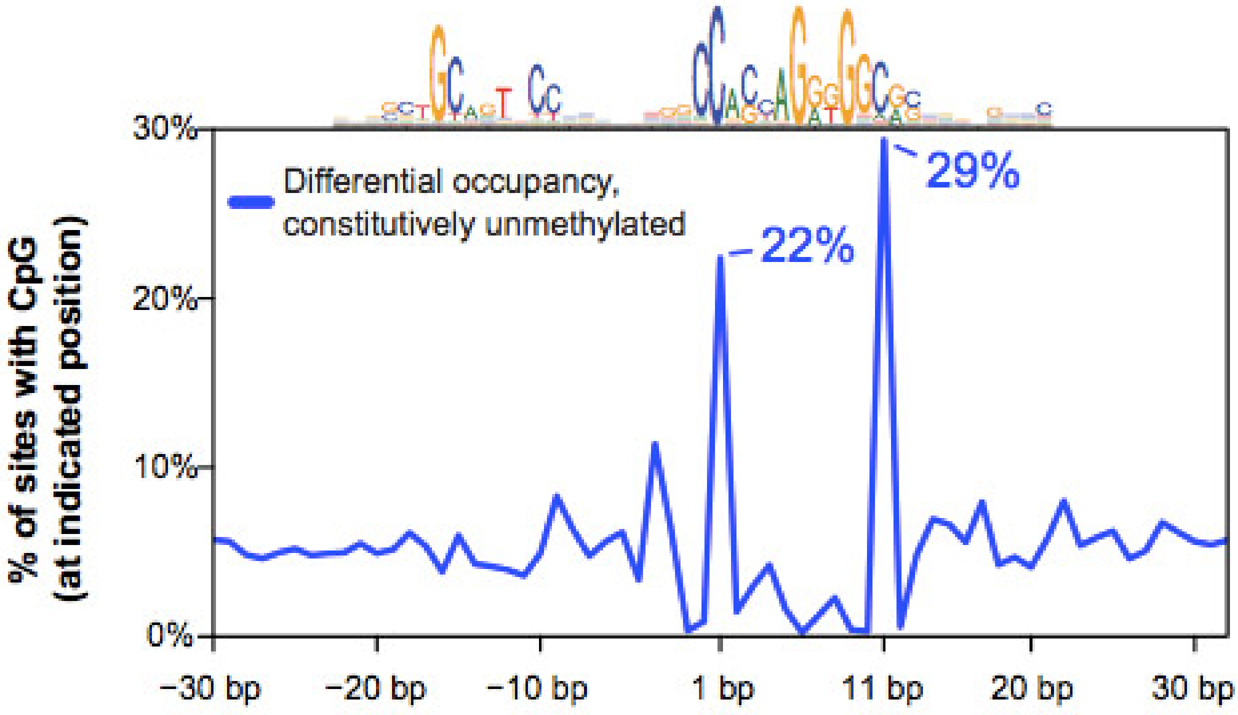
Supplementary Figure S6 | Statistical association of variable methylation with differential occupancy. (a) Location relative to genes of sites surveyed by reduced representation bisulfite sequencing, which enriches for genic and CpG-island regions, including promoters. (b) Most CpGs were surveyed in all 13 samples. Sites without with at least one surveyed CpG for at least 6 samples (gray shading) were excluded from association analysis. (c) Methylation levels (*X*-axis) observed across all CTCF sites in all cell types; most CTCF sites are unmethylated. (d-e) Association of methylation with ChIP-seq occupancy identifies 905, 1,677 and 2,046 significant sites at FDR levels 1%, 5% and 10%, respectively. Histogram of p-values (d) and FDR-adjusted q-values (e) for all tested binding sites. (f) Effect size of methylation differences associated with differences in occupancy (FDR 5%), measured by r^2 of a linear regression.

(position 11) is the predominant CpG in the motif, which has been shown to have a higher rate of C-T transitions at vertebrate-conserved binding sites (Kim *et al.* 2007), consistent with germline methylation. Interestingly, constitutively unmethylated CTCF sites also showed an enrichment of CpGs at these two positions compared to differentially methylated sites without an association to occupancy (Supplemental Fig. S8). Given that the latter sites nevertheless exhibit substantial methylation variability, this suggests that the absence of CpGs at these positions may decouple CTCF occupancy from differential methylation at these sites. Overall, 29% of CTCF recognition sequences genome-wide contain a CpG at positions 1 and/or 11, and 52% of recognition sequences contain a CpG anywhere in the sequence. The genome-wide prevalence of "susceptible" CTCF sites suggests a widespread potential for interaction between CTCF and methylation.

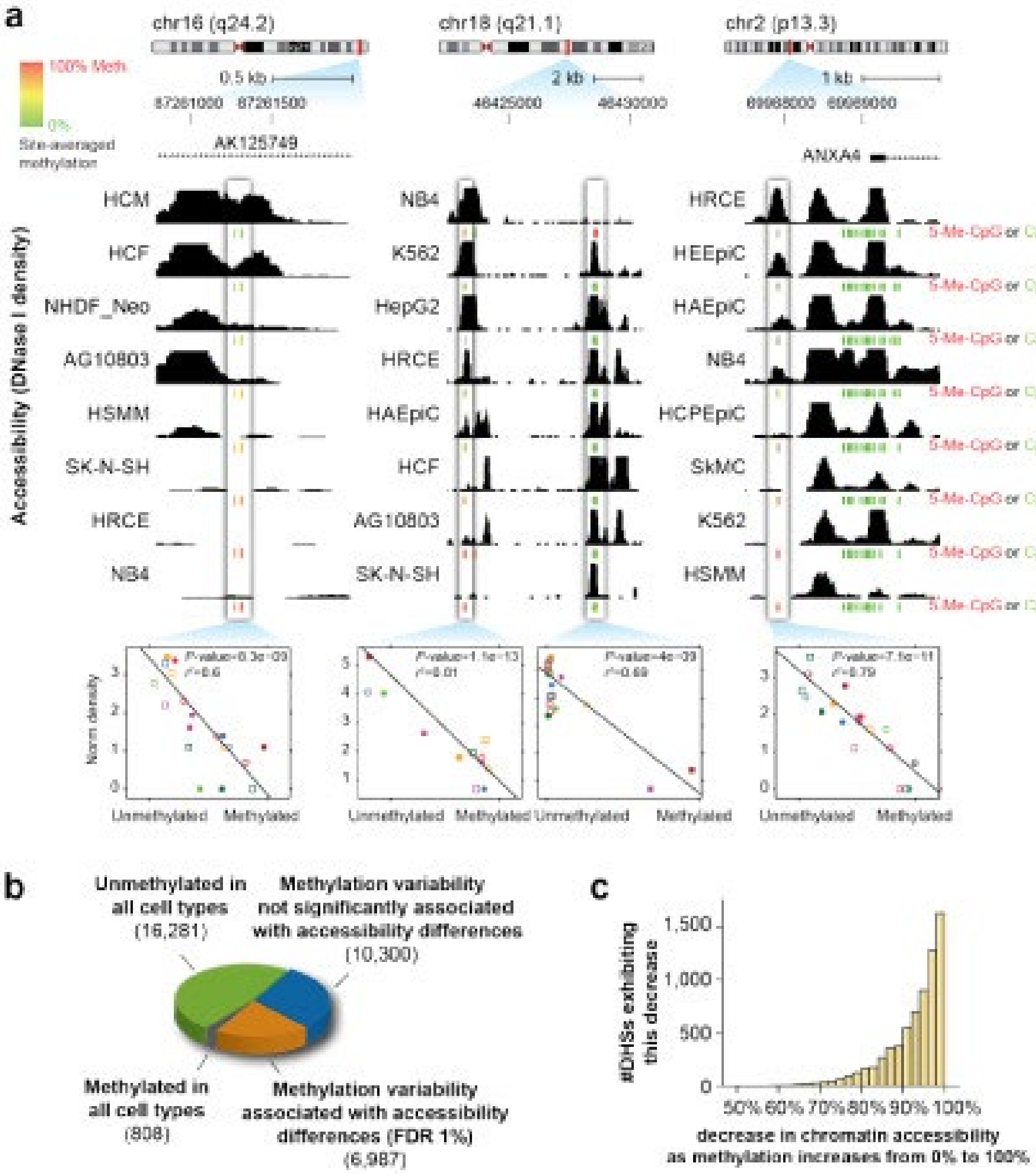
Paralleling prior reports of widespread hypermethylation in cancer (Jones and Baylin 2007; Varley), we observed a bimodal pattern of methylation at CTCF sites distinguishing normal and immortal cell types (Fig. 5A). At 31% of the sites where differential methylation was associated with CTCF occupancy, methylation was observed throughout the thirteen normal and immortal cell types (average number of methylated cell types, 7.3). In contrast, the remaining 69% of sites were characterized by cell-specific hypermethylation constrained to the 6 immortal lines (average number of methylated cell lines, 2.1; Fig. 5A, strip at right). Notably, although the neuroblastoma line SKNSH_RA clusters with epithelial cell types based purely on CTCF binding (Fig. 2A), it



Supplementary Figure S7 | Overall relationship of CTCF occupancy with methylation at 1,076 sites without a significant association. *X*-axis denotes ratio of average occupancy in unmethylated cell types divided by the average in methylated cell types.

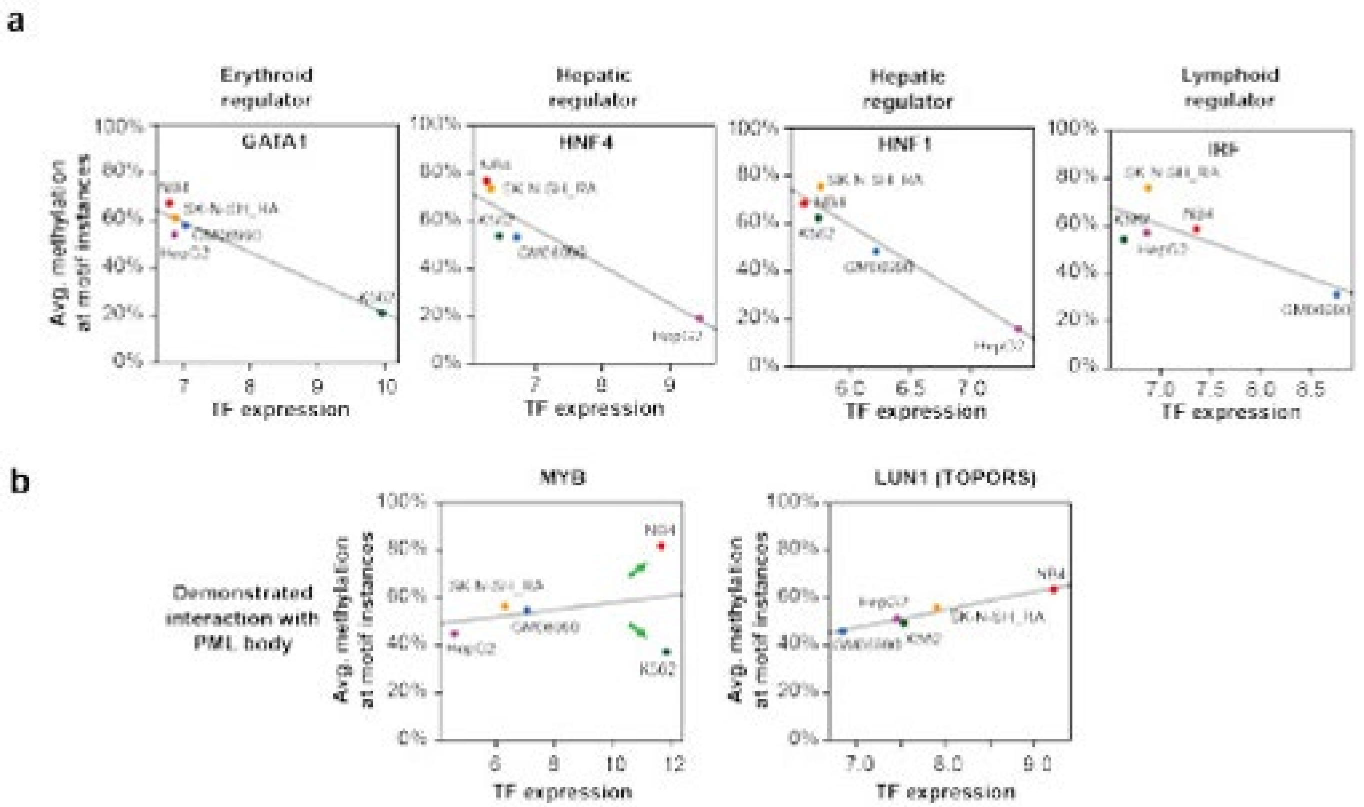


Supplementary Figure S8 | Variable CTCF sites without methylation differences. Frequency of a CpG (*Y*-axis) at positions relative to the CTCF motif (*X*-axis) shown for sites with differential binding but no differential methylation (blue). Note that the presence of a CpG at positions 1 and 11 is similar to that at sites where the variable methylation was associated with occupancy (Figure 4, red), but sites with variable methylation that is not associated with occupancy (Figure 4, gray) are depleted for these CpGs.



Supplementary Figure 11 | Further examples of association between methylation and accessibility. (a) Further examples of association between methylation and accessibility. Data tracks show DNase I sensitivity in selected cell types. Green bars, CpG is 0% methylated; yellow, 50% methylated; red, 100% methylated. Association is quantified in the plots below the tracks. Each point in the graph represents one of 19 cell-types (a susbset of which is represented in the tracks). *X*-axis is the percent methylation of the site in that cell-type; *y*-axis is the normalised DNaseI tag density at the site in that cell type. In each example, accessibility (*y*-axis) quantitatively decreases as methylation increases (left to right). (b) Global characterisation of the effect of methylation on chromatin accessibility, surveyed at 34,376 DHSs with RRBS data. 40% of sites with variable methylation across cell-types were associated with differences in chromatin accessibility. (c), In cell lines with methylated DHSs, site accessibility was reduced on average by 95%. Shown are sites where increased methylation was significantly associated with decreased accessibility (= 97% of all sites in the orange slice shown in (b)).

exhibits the hypermethylation characteristic of the other immortal lines. Surprisingly, the increased methylation in immortal lines does not correspond to a decrease in the total number of bound CTCF sites (Fig. 5B). Strikingly, we also observed that CTCF transcript levels are significantly higher in the immortal cell lines (Fig. 5C). This disruption of CTCF binding in immortal cell lines is further distinguished by an unique association between CTCF occupancy and methylation at promoter sites. Of the promoter CTCF sites where methylation was significantly associated with occupancy, 98% (281 of 288) of these sites were characterized by hypermethylation in the



Supplementary Figure 12 | Genome-wide Influence of methylation on chromatin accessibility. (**a**) Relationship between TF transcript levels and overall methylation at cognate recognition sequences of the same TFs. Negative correlation indicates that site-specific DNA methylation follows TF vacation of differentially expressed TFs. Left, erythroid regulator in the erythroleukemia line K562; centre, hepatic regulators in the liver carcinoma HepG2; and right, lymphoid regulator in the B lymphoblast line GM06990. (**b**), MYB and LUN-1 (also called TOPORS) have both been demonstrated to interact with promyelocytic leukemia (PML) bodies, and show increased transcription and binding site methylation in the acute promyelocytic leukemia (APL) line NB4. Although Myb expression is upregulated in both erythroid K562 and the APL line NB4 (green arrows), its putative binding sites exhibit altered methylation only in the APL line NB4.

immortal lines (Fig. 5D). These results suggest a widespread methylation-associated remodeling of the CTCF binding landscape in immortal cell lines.