#### **Role of methylation haplotype blocks in formation of specifically methylated regions**

To identify specifically methylated regions from a diverse set of samples, we incorporated several ideas that were previously applied to DNA methylation and genes expression analyses. These approaches include local linear smoothing of methylation frequencies [PMID 23034175], dispersion modeling [PMID 26780092], Hidden Markov Model segmentation of methylomes [PMID 23995138], and categorizing methylation specificity with Shannon entropy [PMID: 23925113, PMID 23995138, PMC2759137]. Briefly, segmentation was performed on the methylome using a five states Hidden Markov Model on each individual site’s estimate of dispersion across all samples. Each resulting segment belongs to a dispersion state, with higher dispersion states tending to have higher average dispersion scores and lower entropy scores for the segment-wide methylation levels for samples with minimum 20X depth of coverage (Supp. Figure 1a-b).

A normalized Shannon entropy cutoff of 0.87 bit was applied for high specificity regions and a robust AIC-based outlier detection method [PMC2759137] identified 249,264 segments as loss of methylation DMRs (- methyl DMRs) and 27,957 segments as gain of methylation DMRs (+methyl DMRs). Segments with less than 0.1 maximum absolute difference (Supp. Figure 1c) were classified as either low methylation segments (n = 67,928, average methylation < 0.50) or high methylation segments (n = 1,949, average methylation ≥ 0.50). Segments which did not meet the entropy cutoff but have greater than 0.1 maximum absolute difference were considered a fuzzy methylation segment (n=1,908,956, median average methylation = 0.82). In total, ~ 3.2 million autosomal CpGs were part of a gain or loss DMR (~13 % of covered CpGs). The distributions of these segments in previously characterized genomic regions are shown in Figure 2a. Both sets of DMRs are neither fully methylated nor unmethylated in the outlying sample, but tend to be either fully methylated or unmethylated in the majority of samples (Figure 1b-c). An example for each of the two categories of DMRs is shown for the ZNF236 locus and the SETBP1 locus (Figure 1d-e).

We compared these segment classifications with the MethylMark segments that were identified using a method described in a recent publication [PMID: 26635396]. MethylMark segments were specifically methylated in a minority of cell types and hypomethylated marks were found to be frequently co-localizing with cell type specific transcription factor binding sites and super-enhancers. While 80% of the MethylMark CpGs were considered fuzzily methylated in our analysis, 84% of CpGs within – methyl DMRs and 26% of CpGs within + methyl DMRs were also considered a MethylMark. Some of the causes for why +/- methyl DMRs CpGs would be missed by MethylMark are the low methylation differences observed for those segments while MethylMark requires 0.3 minimum difference and the differences in how segment boundaries were defined.

Another interesting class of regions, partially methylated domains (PMDs), are large regions of hypomethylation which are conserved. We obtained a published catalogue of tissues PMDs [PMID: 26030523] and found that 26% of the fuzzily methylated CpGs contributes to nearly all of CpGs within PMDs (92.8% of PMDs are classified as fuzzily methylated). Fuzzy methylation segments are 82% of covered CpGs and the 61% which are not due to PMDs have a median maximum absolute difference of 0.36. Previous dynamic DMR finding algorithms which require a minimum 0.3 difference would have missed these segments. The actual percentage of dynamic DMRs might be larger than the 15%-25% previously reported [PMID: 26030523, PMID: 26635396, PMID: 23925113 ]. For chr10, we utilized a method based on a Beta-Binomial hierarchical model (MOABS) to detect differentially methylation between pairs of samples at the individual site level with minimum 5X depth of coverage [PMID: 24565500] and to determine dynamic CpGs. We compared the overlap of dynamic CpGs with fuzzy methylation segments and estimated that 91.3% of fuzzy methylation segment contains a dynamic CpG.

Strikingly, the ratio of –methyl DMRs to + methyl DMR is 1 to 7.46 respectively in MHBs. This suggests that cells tend to gain methylation in a spreading process. Furthermore, methylation haplotype blocks (MHBs), DNA methylation valleys (DMVs), and CpG islands have substantial representation of +methyl DMRs even though these segments only comprise of 1.8% of CpGs. DMVs were discovered in a study examining DNA methylation dynamics during early embryonic development. These ultra-long (>5 kbp) regions of low methylation overlaps highly conserved sequences and are enriched in transcription factor and developmental genes [PMID 23664764]. These regions are also most likely to observe gain of methylation through development. We found that 3,927 of the 4,481 (87.6%) DMVs overlap with at least one +methyl DMR, and 23% of +methyl DMR CpGs overlapping MHBs also overlaps with a DMVs. MHBs and DMVs only share 6.2% and 10.4% of CpGs overlapping MHBs for fuzzy methylation and low methylation segments respectively.

CpG islands and MHBs share 81% and 87% of CpGs overlapping MHBs for + methyl DMRs and low methylation segments respectively, but they only share 35% of CpGs overlapping MHBs for fuzzy methylation segments. Low methylation regions and + methyl DMRs are similar in that they both are cases where methylated haplotypes are rare, but in the case of + methyl DMRs those methylated haplotypes are found only in a few tissues or cell types. Fuzzy methylation regions tend to have many methylated haplotypes, and these regions may include allele specific methylation, imprinting regions, and regions with cellular heterogeneity. The observations argues for a haplotype level analysis of DNA methylation variability which would be more powerful in detecting rare methylation differences and differences due to cellular heterogeneity.

FIGURE LEGEND:

#### **Figure 2.** Characterization of differentially methylated regions. (a) Distribution overlapping CpG sites within the five categories of segments for (1) DMRs identified by MethylMarks software, (2) a catalog of all partially methylated domains (PMDs) found in adult tissues, (3) methylation haplotype blocks (MHBs), (4) DNA methylation valleys (DMVs) found in cultured human progenitor cell types, and (5) CpG islands (CGIs). (b-c) cumulative distribution of individual sample methylation levels when they are classified as an outlier (hypermethylated or hypomethylated) versus when they are classified as methylated with the majority of samples. (d) Methylation levels across the ZNF236 locus for a representative set of human tissue samples. Purple box indicates a –methyl DMR region with methylation loss observed for CD14 and colon tissues. (e) Methylation levels across the SETBP1 locus for a representative set of human tissue samples. Green box indicates a +methyl DMR region with methylation gain observed across multiple methylation haplotype blocks for placenta tissue.

#### **Supplementary Figure 1.** Identification of differentially methylated regions. (a) Distribution of average dispersion score for individual segments for the five HMM states. (b) Cumulative distribution of normalized entropy scores for each segmented state. The entropy cutoff of 0.87 determines the level at which a DMR can be defined which limits the number of DMR from S1 and maximizes the number of DMR from S5. (c) Distribution of maximum absolute difference between two samples for each segment, the cutoff of 0.1 determines all of the regions with uniform low and uniform high methylation. (d) Distribution of the five HMM states for each of the five categories. Low methylation segments have average methylation levels less than 0.5, and maximum absolute difference of 0.1. High methylation segments have average methylation levels greater or equal to 0.5 and maximum absolute difference of 0.1. Fuzzy methylation segments have maximum absolute difference greater than 0.1 and entropy scores above 0.87 cutoff, thus not being able to be classified as a DMR. + methyl DMR segments have entropy scores below 0.87 and contain a sample with gain of methylation compared with majority of samples. – methyl DMR segments have entropy scores below 0.87 and contain a sample with loss of methylation compared with majority of samples.(e-i) Distribution of the average methylation level across samples for each segment category described in (d).