Methylation Calling

To call methylated sites, we summed the number of reads that supported methylation at a site and the number of reads that did not. We used these counts to perform a binomial test with a probability of success equal to the non-conversion rate, which was determined by computing the fraction of methylated reads in the lambda genome (spiked in during library construction). The false discovery rate (FDR) for a given p-value cutoff was computed by calculating the fraction of sites in the lambda genome that had a p-value less than or equal to the cutoff and then dividing that quantity by the fraction of sites that had a p-value less than or equal to the cutoff across all other chromosomes. Because the p-value distributions for each methylation context are different, this procedure was applied to each three nucleotide context independently (e.g., a p-value cutoff was calculated for CAT cytosines).

DMR Finding

We defined a stochastic model of our methylation data sets in which the observed number of reads supporting methylated and unmethylated cytosines at each position in each sample is drawn from a binomial distribution. In each sample at each cytosine in the CG context there is a single parameter, x_n^i, corresponding to the true fraction of methylated alleles in the population, or the methylation level, where *i* denotes the position of cytosine and *n* denotes the sample. Our null hypothesis is that the methylation level at this position is equal in all samples ($x_n^i = x^i$ for all *n*). Our procedure is designed to test whether the observed data are consistent with the null hypothesis, or alternatively if there is a significant deviation from equal methylation levels.

To do this, first we compute a goodness-of-fit statistic, *s*, which was introduced and validated by Perkins et al.[5]. Specifically, we arrange the observed data in an Nx2 table, with one row for each of N samples and a column for reads supporting methylated and unmethylated cytosines respectively. The number of observed reads in sample *n* at position *i* is o_{nj}^i, where *j* =

1 for methylated reads and *j* = 2 for unmethylated reads. The expected number of reads in sample *n* with methylation state *j* under the null hypothesis is e_{nj}^{i}:

$$e_{nj}^{i} = \left(\sum_{m=1}^{N} o_{mj}^{i}\right)\left(\sum_{k=1}^{2} o_{nk}^{i}\right)/M^{i}$$

where $M^{i} = \sum_{n=1}^{N}\sum_{k=1}^{2} o_{nk}^{i}$ is the total number of reads in all samples. The statistic for the goodness of fit is

$$s^{i} = \sqrt{\frac{1}{2N}\sum_{n=1}^{N}\sum_{j=1}^{2}\left(o_{nj}^{i} - e_{nj}^{i}\right)^{2}}$$

Next, we simulated read count data under our stochastic model assuming the null hypothesis in the following way:

- Set all cell counts in the table to zero
- Randomly select a cell in the table with probability equal to the expected counts divided by the total number of counts in the table ($\frac{e_{nj}^{i}}{M^{i}}$). Increment the value in this cell by one.
- Repeat this procedure M^{i} times.
- Finally, calculate the value of the statistic, s_{shuff}^{i}, for the randomly generated table.

This randomization procedure was repeated until we observed 100 iterations with a value of s_{shuff}^{i}that was at least as extreme as that of the observed data, *s*, up to a maximum of 3,000 iterations. The p-value at position i was then computed as:

$$p^{i} = \frac{R^{i} + 1}{T^{i}}$$

Where R^{i} is the number of randomized tables with a statistic greater than or equal to the original table's statistic and T^{i} is the total number of randomized tables that were computed. Our adaptive permutation procedure ensures that any sites which we may potentially identify as significantly differentially methylated with $p^{i} < 0.01$ will be sampled 3,000 times. At other sites,

we have observed an appreciable number (100) of permutations more extreme than our original test statistic ($s \geq s_{shuff}$) and the p-value for these sites will be $p \geq (100+1)/3000 = 0.034$; these sites will therefore not be called as differentially methylated.

To control the false discovery rate (FDR) at our desired rate of 1%, we used a computationally efficient procedure designed for comparing multiple sequential permutation-derived p-values[6]. This procedure is designed to account for the effect of our adaptive permutation procedure on the form of the distribution of p-values. First we generated a histogram of the p-values across all cytosines in CG context. We also calculated the expected number of p-values to fall in a particular bin under the null hypothesis. This expected count is computed by multiplying the width of the bin by the current estimate for the number of true null hypotheses (m_0), which is initialized to the number of tests performed. We then identified the first bin (starting from the most significant bin) where the expected number of p-values is greater than or equal to the observed value. The differences between the expected and observed counts in all the bins up to this point are summed, and a new estimate of m_0 is generated by subtracting this sum from the current total number of tests. This procedure was iterated until convergence, which we defined as a change in the m_0 estimate less than or equal to 0.01. With this m_0 estimate, we were able to estimate the FDR corresponding to a given p-value cutoff by multiplying the p-value by the m_0 estimate (the expected number of positives at that cutoff under the null hypothesis) and dividing that product by the total number of significant tests we detected at that p-value cutoff. We chose the largest p-value cutoff that still satisfied our FDR requirement.

In the next stage of analysis, we combined significant sites (DMSs) into blocks if they were within 250 bases of one another and had methylation changes in the same direction (e.g., sample A was hypermethylated and sample B was hypomethylated at both sites). A sample was considered hypo or hyper methylated if the deviation of observed counts from the expected counts was in the top or bottom 1% of deviations. These residuals were calculated for a position

i using the following formula for a given cell in row *n* and column *j* of the table:

$$\frac{o_{nj}^{i} - e_{nj}^{i}}{\sqrt{e_{nj}^{i} * (1 - \sum_{m=1}^{N} \frac{e_{mj}^{i}}{M^{i}}) * (1 - \sum_{k=1}^{2} \frac{e_{nk}^{i}}{M^{i}})}}$$

The distinction between hypermethylation and hypomethylation was made based on the sign of the residuals. For example, if the residual for the methylated read count of sample A was positive, it was counted as hypermethylation. Furthermore, blocks that contained fewer than 10 differentially methylated sites were discarded.