Supporting Material for FadE

Introduction

This supplement is focused on providing more detailed analysis regarding interesting topics for this research and future research regarding methylation estimation. The first item of interest concerns different methods with which to facilitate unbiased SBT alignment. Additionally, the results of color space methylation estimation of IMR90 fibroblast data are explored in-depth with regard to the methylation rate at and adjacent to genes with CpG island promoter regions. Additionally, the results of a nucleotide implementation of the FadE algorithm is compared to the results from Lister et al's nucleotide space analysis on the IMR90 fibroblast dataset. The concordance of the results when a shared dataset is used further suggests of the utility of FadE and also may provide some insight as to the differences between the SOLiD and Illumina sequencing platforms. Finally, we explore the utility of the credible interval for analysis of methylation data. A large scale simulation is used to show that the credible interval is a more accurate proxy for estimation accuracy than the read depth, suggesting that use of the credible interval size to filter methylation results will provide a greater increase to accuracy of methylation results.

Alignment of Sodium Bisulfite Treated Reads

Total Cytosine Translation

In total cytosine translation all cytosine positions on the read and reference sequence are translated to thiamine nucleotides. Alignment is then carried out after which the alignment results are saved but read and reference sequences are translated back to their original sequence, and "C" "C" alignments are interpreted as resulting from methylation [4, 2]. This method is carried out on the nucleotide sequence of the reads which makes it unlikely to perform accurately on color reads where translation to the nucleotide sequence will result in vastly different nucleotide sequences downstream all single color errors. As even a moderate color error rate will result in many incorrectly translated read sequences this method is considered only for nucleotide space reads. In nucleotide space, this method does indeed eliminate bias because the translation is carried out on every cytosine position on both sequences. However, the drawback to this method is that the DNA sequence is reduced from a four to a three letter

alphabet. While this may not matter for short reference sequences, reducing a long reference sequence from four to three letters vastly increases the likelihood of ambiguous alignments. In fact if we consider a uniform random model for DNA where

$$P(x) = \frac{1}{4}, \forall x \epsilon \{A, C, G, T\}. \quad (1)$$

Then the probability than an n-basepair read will find a spurious alignment with k or fewer substitutions to a genome of size $|G|$:

$$|G| \sum_{0}^{k} Binom(n, p(x)). \quad (2)$$

Thus, for normal four letter DNA, ($p(x) = \frac{1}{4}$) and a double stranded 3 Billion base pair human genome, approximately one in fifty trillion reads will find a spurious alignment with five or fewer mismatches under a uniform normal model. In contrast, a three letter alphabet results in one in thee million reads finding spurious alignments. While both of these numbers are small in comparison to the read set, it should be noted that the human genome is largely semi-repetitive and repetitive which results in vastly more ambiguity than is estimated by a uniform random model. For this reason, the absolute number of spurious alignments predicted with a three and four letter alphabet are less important than the fact that the number of spurious alignments grows by a factor of more than 170,000. In fact this can be illustrated by considering only the totally repetitive regions on the forward strand of the first human chromosome. When the normal four letter alphabet is used for chromosome one, approximately 1.3% of 50mers have a duplicate sequence located elsewhere on the chromosome and 0.09% have more than five duplicate sequences. When the three letter alphabet resulting from total ($C \rightarrow T$) translation is considered these values grow to 1.9% and 0.18%. Such an increase is likely magnified when inexact duplicates on the entire genome are considered which may substantially affect alignment accuracy.

Index Substring Translation

Many alignment algorithms operate by indexing substrings into multiple smaller subsequences which are stored in an index table. These subsequences are then queried for matches by the reads to facilitate mismatch tolerant alignment. Such algorithms can be altered so that instead of just storing subsequences all combinations of ($C \rightarrow T$) translations of subsequences are stored and queried by the reads. If any subsequence match is found, the read and native sequence can then be compared and an alignment can be accepted or rejected. If all combinations of translations are indexed, this method will also serve to eliminate methylation bias. Additionally, because this method acts only on the reference index, it is suitable for use on color space reads and was used successfully by Hansen et al. [3]. The main drawback to such a method is the slowdown to alignment required to query all combinations of ($C \rightarrow T$) subsequence translation and the

prohibitively large index table that may be required to store them. The number of subsequences combinations ($C(S)$) required for indexing each position grows exponentially with respect to the number of cytosine positions ($|c|$) in the subsequence: $C(S) = 2^{|c|}$. The length of subsequences ($|S|$) used is often a function of the read length and may be slightly larger than half the read length [5] but are rarely smaller than fourteen basepairs [6], which is effectively a computational bottleneck for efficient alignment to a reference as large as the human genome. Increasing the subsequence length from 14 to approximately 25 basepairs will have a significant increase to alignment efficiency. Without translation the number of subsequences required to index a chromosome is directly proportional to it's length and depends the mismatch tolerance required of the algorithm. Assuming a positive integer k to describe the number of subsequences required to index each position in the reference genome, $2kn$ represents the total number of subsequences required to index each strand of a reference sequence of length n. Assuming this relationship we can perform a quick back of the envelope calculation to approximate the increase in subsequences required to store all combinations of $(C \rightarrow T)$ translations. To accommodate all combinations of translation, each reference position will produce approximately $2k^{|c|}$ combinations, thus the total number of reference subsequences required is described by a sum of each of n positions on the reference sequence approximately:

$$2\sum_{i=0}^{i=n} 2k^{|c|}. \tag{3}$$

Applying this equation to chromosome 21 and calculating the factor increase in stored subsequences reveals that the number of required subsequences necessary to index all $(C \rightarrow T)$ translations increases by a factor of approximately ten to over 400,000 to for subsequence lengths from 14bp to 50bp. Such a large increases prohibits such an algorithm from being used on most computing clusters and the increase in the number of subsequences will likely slow the alignment to the degree that it is not possible to execute in a reasonable amount of time. If it is known beforehand that a sample has no methylation in a non CpG context then the requirement in memory can be cut drastically by indexing only combinations of $(CpG \rightarrow TpG)$ translations and translating all CpH $(H \neq G)$ positions to thymine. This was the strategy Feinberg et al used to perform bisulfite alignment in color space, and results in an increase to the number of subsequences indexed by factors of approximately 9% to 60% for subsequence lengths of 14bp to 50bp. Such increases are tolerable provided the alignment algorithm has a compact way to store the extra indexes required to represent the translations. One drawback to the method is that while some tissues are thought to be devoid of CpH methylation, the bisulfite conversion rate is often less than perfect and as such alignments with untranslated CpH positions will not be aligned.

Multiple Reference Alignment

To avoid the perils of the translation of color reads and to methylation to be analyzed in a both a CpG and CpH context, a multi-reference translation strategy can be used. Such a strategy allows any alignment algorithm to be used without modification and requires little to no additional computing power or memory requirement when compared to regular alignment. In this method bias is reduced by translating the reference sequence into multiple sequences which favor alignment for reads with differing amounts of methylation. The goal of such a strategy is to choose reference translations such that the combination of alignment biases results in an overall reduction of bias concerning alignment over cytosine nucleotides. Reference translations may include:

- The native reference sequence. The native reference sequence favors alignment of reads to highly methylated regions.
- A ($C \rightarrow T$) translated reference sequence which favors alignment of reads to largely unmethylated regions.
- A ($CpX \rightarrow TpH$) translated reference sequence will favor alignment of reads to methylated CpG positions. In most mammalian cells only CpG positions will have a significant level of methylation.
- Probabilistic ($C \rightarrow T$) translated reference sequence will favor regions which are not strictly unmethylated or methylated.

If an alignment algorithm is tolerant to a large number of substitutions, most reads will find an alignment to at least one of the translated references. Afterward the results can be joined into a single alignment file where each read is assigned to the positions with the fewest number of substitutions over non-cytosine reference positions. In our experiments we used a strategy which used the first three reference translations described above. To demonstrate the utility of this strategy we performed multiple simulations using human chromosome 21. In each simulation a methylated version of chromosome 21 was created wherein CpG and CpX positions were randomly assigned methylation rates between 0.30 to 1 and 0 to 0.7, respectively. 70 million 50 nt bisulfite treated color reads were simulated according to these conditions and an additional 70 million 50 reads were simulated from the native reference sequence. A uniform 1% color error rate was imposed on all reads. The simulated bisulfite treated reads were aligned to the three reference sequences described above while the untreated reads were aligned to the native sequence using PerM [5] and allowing a maximum of eight substitutions. To obtain a measure of the prevalence of incorrect alignment resulting from genomic structure and color error a likelihood ratio test was first performed for each native reference transition covered in the untreated simulations. For each transition this test compared whether a null model (the ratio of colors which should align to the position) differed significantly from an alternative model (the ratio of bases observed in the alignment). Thus, the likelihood ratio statistic calculated was as follows:

Table 1: An Illustration of how Reference Bias is reduced through the union of multiple reference alignments. Shown here is the mean coverage ($\overline{cov}$, the % of positions signifying methylated alignment (% M), the % of positions signifying unmethylated alignment (% T) and the percentage of positions where the likelihood ratio test statistic for an unbiased alignment was greater than 1.0 for the alignment to three reference genomes. It can be observed that the Union of such alignments provides a reasonable unbiased alignment of SBT data.

	CpG Positions				CpH Positions			
	$\overline{cov}$	% M	% T	$\%D > 1$	$\overline{cov}$	% M	% T	$\%D > 1$
Read Composition	76.8	76.2	23.1	-	76.7	20.0	79.5	-
Native Sequence	16.0	53.7	45.8	36.3	16.7	35.9	63.5	42.2
$(C \rightarrow T)$	64.2	77.1	21.3	8.9	70.1	7.0	92.4	6.0
$(CpH \rightarrow TpH)$	67.0	81.6	17.8	6.2	67.9	7.5	91.9	5.1
Union*	69.4	77.6	21.9	1.5	73.3	16.3	83.3	2.1

*A union alignment breaks ties according to non cytosine mismatches and discards ambiguous alignments.

$$D = -2ln(\frac{(R_B)^{N_B}(R_G)^{N_G}(R_Y)^{N_Y}(R_R)^{N_R}}{(A_B)^{N_B}(A_G)^{N_G}(A_Y)^{N_Y}(A_R)^{N_R}}) \tag{4}$$

where R_X is the rate of color "X" is the reads, A_X is the rate of color "X" in the alignment and N_X is the number of times color X was observed in the alignment. Perfect alignment would result in $R_X = A_X$ and $D = 0$. In the untreated case substitutions result from sequencing errors, repetitive regions and incorrect alignment. In our untreated alignment, 98.1% of the reads aligned correctly and approximately 1% of positions spanning a cytosine nucleotide had a likelihood ratio score of greater than 1.0. This value was used as a benchmark to asses the alignment accuracy in the reference translated treated alignments.

Table 1 illustrates the how bias is reduced using a multiple reference strategy. While alignments to each of the three translated reference sequences show significant bias, the union of the alignments displays far less bias across cytosine positions. Additionally, most of the bias shown in the simulation (under representation of the number of methylated CpH comes from a few large runs of almost only CpH positions, of which only a small fraction are partially methylated. Such regions can either be ignored or additional reference translations can be made which focus on such reads and translate a moderate fraction of CpG positions for such regions to facilitate unbiased alignment.

Table 2: Methylation rates at three regions near CpG promoters in five genes

Gene Name	Loc	CpG Isl.	Upstr	CpG Isl	Dwnstr
BIVM	chr13:103455399	-750:1750	0.86	0.046	0.59
VAPA	chr18:9918000	-750:1250	0.85	0.077	0.63
UTP14c	chr13:52586534	-1250:750	0.62	0.081	0.89
EGR1	chr5:137787179	-2000:1250	0.82	0.097	0.86
TBCB	chr19:36605888	-1000:1250	0.80	0.13	0.81

Additional Results on Cell line IMR90 using Color Reads

One common pattern observed on genes with CpG island promoter regions was highly methylated CpG positions upstream of the promoter region and downstream of the transcription start site. We defined a CpG island promoter as a region between 500-2000 basepairs with greater than 5% CpG positions. For each gene with a CpG island promoter we calculated the methylation rate for three regions: upstream of the promoter, from the promoter past the transcriptional start site (TSS), and upstream of the TSS. Shown in Table 2 are the methylation rates across different regions for five of the most differentially methylated CpG island promoters. Regions were selected by choosing 250 base-pair windows such that the methylation pattern was most preserved. This Table shows size of each region and clearly shows the often observed pattern of a high methylation rate upstream and downstream of the CpG island promoter which itself is hypo-methylated. Figure 2 provides an illustration of this, in this figure one-hundred base-pair windows are binned and averaged for each of the five genes. Despite, relatively different promoter region lengths, the same pattern of hypomethylation at the transcriptional start site is observed. Finally, Figure 1 illustrates the pattern on for a single gene. Although the pattern for a single CpG promoter is noisier, the same pattern is also observed.

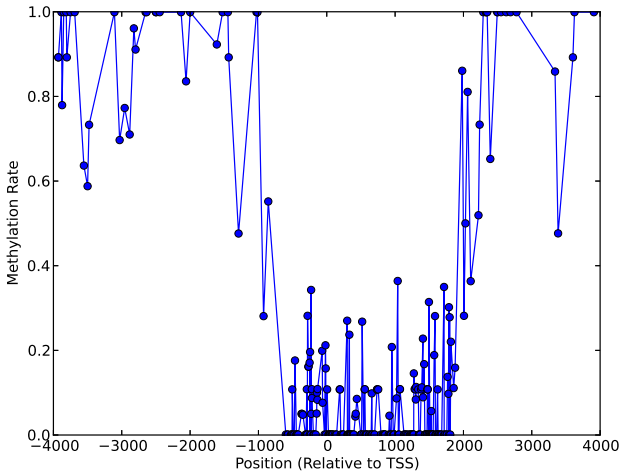


Figure 1: The methylation rate adjacent to the transcriptional start site of gene BIVM

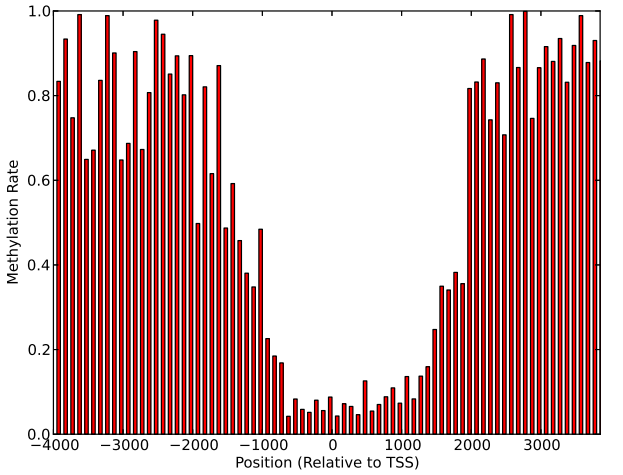


Figure 2: The average methylation rate for five genes containing CpG island promoters

Table 3: Comparison of FadE and the results obtained by Lister et al on the same nucleotide space alignment file

	FadE	Lister et al	Validation Rate*
Global Rate	62.7%	63.5%	-
Total Shared Sites	41,685,156	41,685,156	99.6%
$\hat{\rho_j} < 0.25$	7,023,184	7,023,184	99.7%
$\hat{\rho_j} > 0.75$	23,158,054	23,158,054	99.3%
$cov > 10$	122,400,411	-	99.8%*

*Validation occurs where Lister et al's estimate falls within FadE's credible interval

Nucleotide comparison of Lister et al. Alignment

To investigate further if site-by-site the differences between FadE's color space estimation of methylation and Lister et al's high coverage nucleotide space experiment were the result of multiple sources of noise or the result of differences between algorithms or errors in implementation, FadE was ran on the same dataset used by Lister et al. To eliminate differences resulting from alignment FadE was ran directly on the alignment file supplied by Lister et al. Since control SBT alignment data was not available, error rates were estimated using the alignment to non-cytosine bases. While the alignment to non-cytosine nucleotides does not provide information as to the behavior of the sequencer when sequencing bisulfite treated cytosine nucleotides, it is capable of accurately estimating the error rate with respect to read position, strand and quality score.

After estimating the emission rates FadE was ran on each of the twenty-two human autosome alignments, after which approximately 42 Million CpG methylation calls were compared between FadE and the results of Lister et al. When using Illumina data the results were very similar (Table 3), with an average difference in site specific methylation estimation between platforms of only 2.1%. This similarity suggests that the site-by-site differences observed between Lister et al's results and Fade's color-space parameter estimation result largely from sources such as biological variation or differences between the error distribution of the sequencing platforms rather than large differences between the optimization routine used by FadE and the binomial model used by Lister et al.

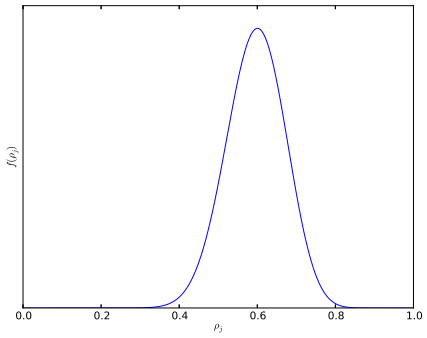
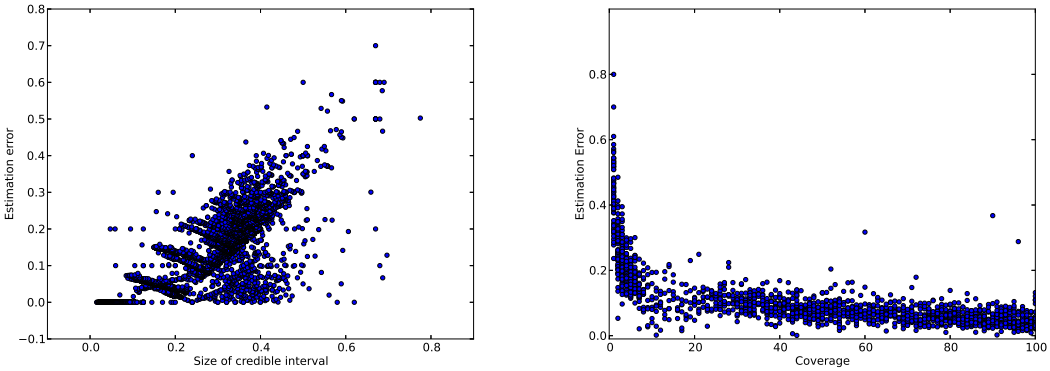


Figure 3: The probability density function $f(p_j)$ is concentrated near its maximum $\hat{\rho_j}$ when coverage is sufficient. Shown here is the pdf resulting from the alignment of thirty reads to an isolated reference cytosine with a simulated methylation rate of 0.6. This simulation imposed uniform color-error rate of 3% per position. A continuous uniform prior distribution was assumed. The credible interval for this distribution would be between approximately 0.50 and 0.70

Credible Interval Accuracy

The credible interval returned is calculated from the posterior probability distribution described in the manuscript (Page 4, *Implementation*). Returning both a credible interval as well as a point estimate for the methylation level provides the researcher with an estimate of both the likely methylation rate and the likely level of deviation from the true parameter associated with the estimate. To demonstrate the utility of the credible interval we performed multiple simulations in both color and sequence space similar to those described in the manuscript (Page 5 *Color Space Simulation*, Page 6 *Nucleotide Space Simulations*) In Figure 4, scatter plots display the relationships between the size of the credible interval and estimate accuracy and the read depth and the estimate accuracy. Both figures were generated by imposing a five percent error rate on the reads which resulted in an incorrect or ambiguous alignment rate of approximately 3%. Coverage values from 1 to 100 were analyzed and credible interval sizes were rounded to two decimals places to allow one hundred values for which to analyze. These plots illustrate that accurate credible intervals provide a more acute estimation of error than the coverage at a certain position.



(a) Credible Interval Size

(b) Read depth

Figure 4: Shown are the relationship between the total credible interval and the absolute value of the estimation error when generated using a 5% sequencing error rate with simulated nucleotide data. While low coverage is associated with larger estimate errors, performance does not continue to increase when coverage continues to improve, this is because often times high coverage is the result of multiple ambiguous, incorrect, or low quality mappings and in such cases actually serves to decrease estimation accuracy. Using the credible interval as proxy for estimate accuracy takes into account the inferred emission rate and may provide more accurate results. that the stronger correlation with reduced credible interval sizes and estimate accuracy. The plots shown were made by generating a large amount of data and repeatedly averaging the estimate accuracy of five randomly chosen positions with equal coverage or credible interval size (credible interval sizes were rounded to two decimal places).

References

[1] Holliday, R and Pugh, JE. (1975) DNA modification mechanisms and gene activity during development. *Science*, **187**, 226-232

[2] Krueger, F and Andrews, S R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571-1572

[3] Hansen et al. (2011) Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, **43** 768-775

[4] Lister et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315-322

[5] Chen, Y and Souaiaia, T and Chen, T. (2009) PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics*, **25**, 2514-2521

[6] Hen Li, Jue Ruan, Richard Durbin. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, **18** 1851-1858