ISSUES IN TESTING DNA METHYLATION USING NEXT-GENERATION SEQUENCING

Douglas Baumann

R. W. Doerge

ISSUES IN TESTING DNA METHYLATION USING NEXT-GENERATION SEQUENCING

Douglas Baumann and R.W. Doerge*
Department of Statistics, Purdue University

*Corresponding Author:
R.W. Doerge
Department of Statistics
Purdue University
250 N. University St.
West Lafayette, IN 47907
email: doerge@purdue.edu
phone: 765-494-6030
fax: 765-494-0558

ABSTRACT: DNA methylation is an epigenetic modification known to affect gene expression, cellular differentiation, as well as phenotypes. Recent advancements in next-generation sequencing technologies have provided unparalleled insight into the location and function of DNA methylation in a variety of organisms. These data require vastly different statistical procedures than data from previous genomic-based technologies. We outline the biological and chemical processes involved in several approaches for gaining DNA methylation data. The implications of the differences between the approaches are discussed relative to the statistical methodology, and the use of genome annotation is explored for the purpose of improving the statistical power when testing for differential methylation.

1 Introduction

Epigenetics is broadly defined as the study of heritable changes in physical characteristics of organisms that are not attributable to changes in the DNA sequence. There are two key areas of focus in epigenetics, namely histone modifications and DNA methylation (herein referred to as methylation). The study of histone modifications primarily addresses how DNA coils and uncoils itself in a cell's nucleus when transitioning from active and inactive transcription states. This type of epigenetic modification is not a direct modification to the DNA sequence. Methylation, on the other hand, involves the attachment of methyl groups typically to cytosine bases, which directly affects how the cell's infrastructure behaves during transcription (Figure 1). Methylation abundance has been associated with silencing gene expression (Figure 2) (Aoyama et al., 2004; Finnegan, 2010), and promoting the occurrence of cancer (Feinberg and Vogelstein, 1983; Kouzarides, 2002).

Methylation studies that examine the entire genome have become increasingly prevalent in the literature (Wang et al., 2006; Irizarry et al., 2008; Cokus et al., 2008; Lister et al., 2008, 2009; Bock et al., 2010). Using next-generation sequencing (NGS) technologies, it is now possible to interrogate the epigenome at single-base detail (or single-base 'resolution'). This is providing unparalleled insight into the role and function of DNA methylation in a variety of organisms. While the advancement of NGS has moved epigenetics forward, the diversity of the technologies has presented new challenges. Luckily, the major NGS technologies follow a similar workflow (Figure 3). Specifically, DNA is sampled, broken into fragments which are selected for size, and then sequenced. The resulting sequences are then compared to, and placed on (i.e., aligned to), a fully sequenced genome (referred to as a 'reference genome'). The occurrence of an observed nucleotide, at a specific position relative to the reference genome, is counted (this is referred to as 'sequencing depth' or simply 'depth'). The proportion of the entire genome covered by at least one base (or, at a designated depth) is referred to as 'coverage.' The most popular goal for NGS count data are applications of in genomics. They have promoted many novel applications of statistical techniques, ranging from Fisher's Exact Test (Lister et al., 2008) to Bayesian deconvolution methods (Down et al., 2008).

The cost and efficacy of current NGS technologies for whole-genome methylation analysis are discussed in detail in Harris et al. (2010) and Bock et al. (2010), where each method is applied to human embryonic stem cells in an effort to compare cost, detection levels, and concordance among the techniques. We review the biological and chemical properties of these methods, and how these properties affect the statistical procedures that both summarize and test differential methylation between samples or treatments. In an effort to improve on existing statistical approaches, we explore the concept of incorporating of genome annotation into the statistical analysis. This additional information serves to increase interpretability of testing results, and also improves the statistical power to detect small changes in methylation levels.

2 Immunoprecipitation-Based Technologies

One major class of technologies for analyzing DNA methylation using NGS is immunoprecipitation-based approaches. These methods utilize antibodies to select only DNA fragments which exhibit cytosine methylation. Two algorithms, "MBD-isolated Genome Sequencing" (MiGS) (Serre et al., 2010) and "Bayesian tool for methylation analysis" (Batman) (Down et al., 2008), are popular tools

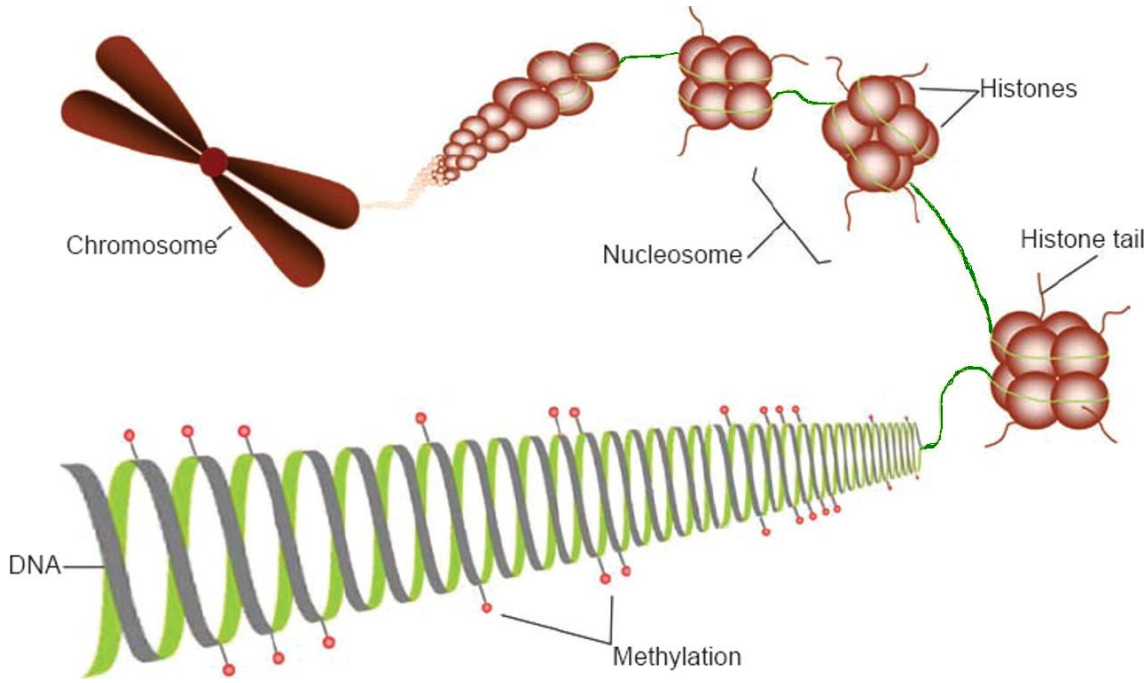

Figure 1: Primary components of epigenetics. DNA is tightly wound around nucleosomes, that are octamers of histones. Modifications to histone tails affect how DNA coils and uncoils around the nucleosomes. DNA methylation typically occurs at cytosine sites, and is known to affect transcription(Groom et al., 2010).

for the analysis of methylation data. Both methods are restricted to CG (i.e., a cytosine followed by a guanine base) methylation contexts (i.e., these methods ignore other possible contexts) and accuracy to within 100 bases. The base pair coverage across the genome for these two approaches are similar as well (∼65% coverage at 1 read threshold, ∼25% coverage at 5 reads threshold, for each approach) (Beck, 2010).

The MiGS procedure uses a methyl CG binding domain protein called MBD2 to precipitate (i.e., capture) methylated DNA fragments after they are broken into pieces. As the frequency of methylation increases for a particular fragment, the likelihood that it is captured also increases. This procedure guards against over-representation of fragments showing intermittent methylation. Once captured, the MBD2 proteins are stripped and the fragments are sequenced via a NGS platform and aligned to a reference genome (Figure 6). While the size of the fragments is platform-dependent, the average number of bases actually sequenced is generally much lower than the overall fragment length. For example, the average fragment length for the Illumina GAII platform in Serre et al. (2010) was 120 bp, while the average sequence length was 36 bp.

The discrepancy between average fragment length and average sequence length is important since only the approximately 36 bases actually sequenced are used to align the read information to a reference genome, even though methylation can occur at any CG dinucleotide in the entire fragment. To overcome this limitation, Serre (Serre et al., 2010) suggests three steps. First, the entire reference genome can be divided into non-overlapping 100-bp bins. After aligning the reads to the reference genome, the start and end position of each of the reads are recorded. As these reads represent the

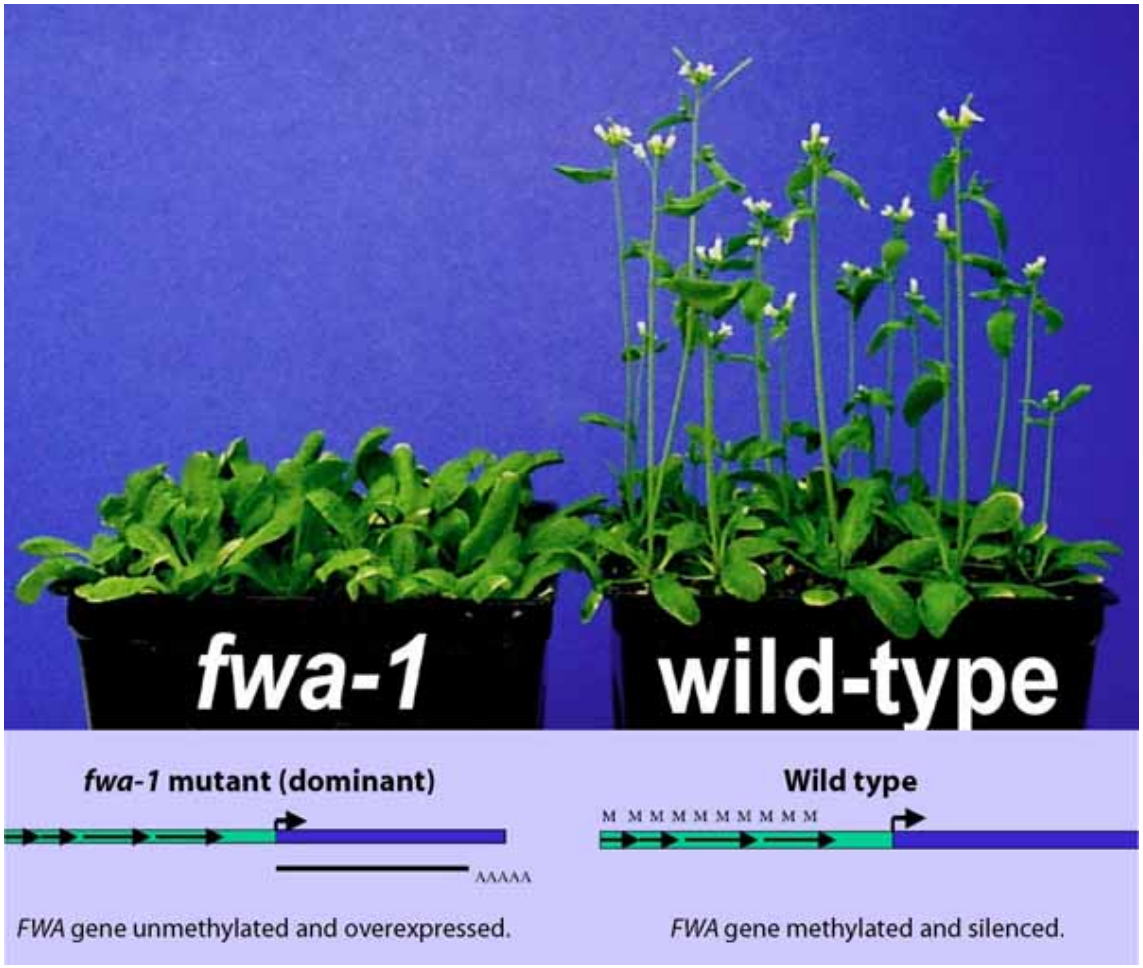

Figure 2: DNA methylation and phenotype. Genetically identical *Arabidopsis* plants, differing only in methylation levels at the *fwa-1* locus, exhibit vastly different flowering times when grown under similar conditions(Jacobsen, 2011).

position of an entire fragment, the coverage of each read is extended to the average fragment length (e.g., 120 bp). Each extended read is then assigned to the genomic bin that it most covers. In this fashion, 100-bp resolution is achieved, and non-empty genomic bins indicate putative methylation sites. After correcting for the total number of unique reads mapped in each treatment, testing for differential methylation is typically achieved via a Fisher's Exact Test (Fisher, 1922) with the observed number of reads for two treatments in each window (see Table 1).

Bin Index	Treatment 1	Treatment 2
1	n_{11}	n_{12}
2	n_{21}	n_{22}
...	...	...
N	n_{N1}	n_{N2}

Table 1: Contingency table for bin level Fisher's Exact Test used in MiGS.

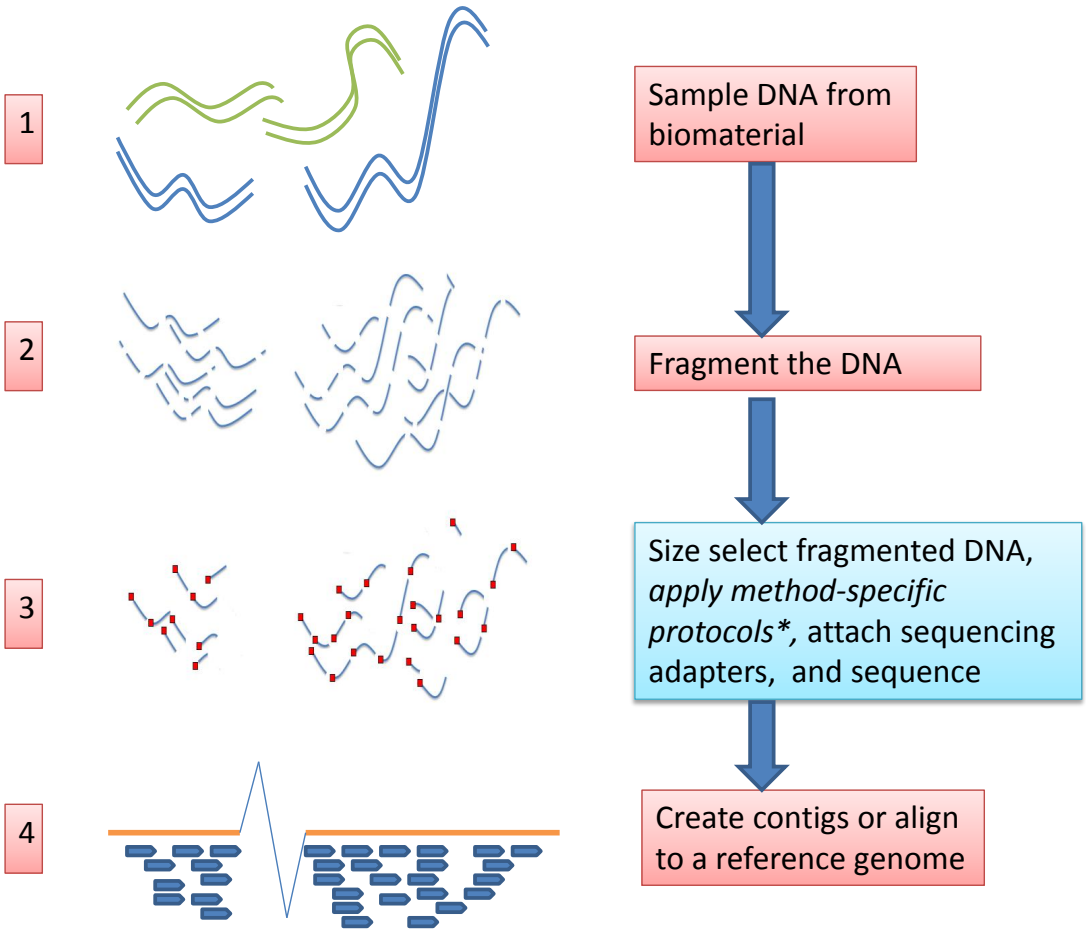

Figure 3: General workflow for next-generation sequencing whole-genome methylation technologies. The workflows differ primarily with respect to method-specific chemical compounds (blue box step). These differences are highlighted in Figure 4.

The Batman algorithm was designed specifically for methylated DNA immunoprecipitation (MeDIP) based experiments. Unlike the MBD2 protein used in the MiGS procedure, the monoclonal antibodies used in MeDIP procedures (for details, see (Weber et al., 2005)) do not distinguish between regions of sparse or dense CG methylation. This leads to increased coverage at the cost of potentially inflating the influence of sparsely methylated regions. Once captured, the antibodies are removed and the fragments are sequenced.

To address the problems associated with sporadic methylation, the Batman algorithm models the effects of the density of methylated CGs. Specifically, let D be the sequencing data, summarized to coverage (i.e., sequencing depth) at each base pair. The the sequencing depth is denoted D_r at read position r. A function of the distance between r and a given CG dinucleotide c, is called the coupling factor C_{cr}. Let m_c denote the methylation state at CG c, and let the set of all methylation states be defined as m. The Batman algorithm defines the probability distribution for

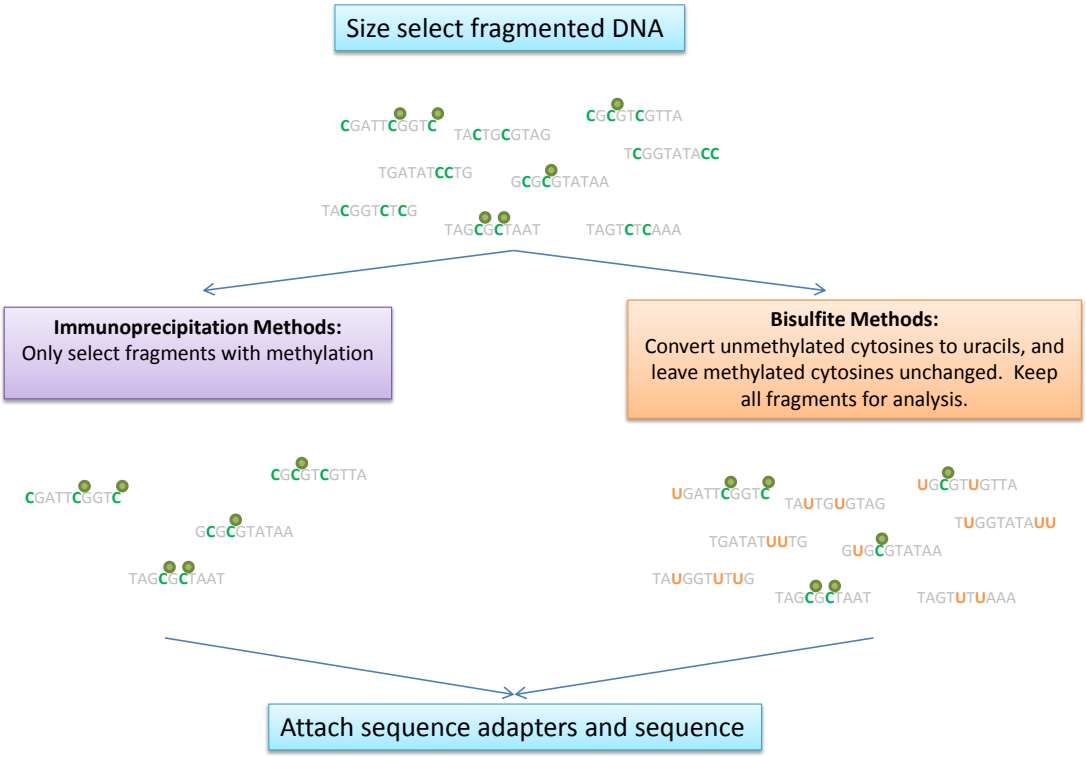

Figure 4: Primary differences in workflow for next-generation sequencing whole-genome methylation studies. Immunoprecipitation-based studies (MBD and MeDIP) only select fragments where methylation is present, without any direct modification to the base sequence. Bisulfite-based methods (RRBS and MethylC) convert unmethylated cytosines to uracils, that become thymines during PCR. These methods retain fragments indiscriminately of methylation presence, contrary to the immunoprecipitation methods.

all observations given a set of methylated states as:

$$f(D|m) = \prod_r \mathcal{G}'\big(\rho(D_r, C_{cr}, m_c), v^{-1}\big)$$

where $\mathcal{G}'$ represents a truncated (at 0) Gaussian probability density function, ρ is a polynomial function of order two, and v^{-1} is an estimate of the inverse variance of D around its mean. To ease the computational time of the algorithm, CG dinucleotides are grouped in 100-bp bins and $f(m|D)$ is subsequently estimated using nested sampling (Skilling, 2006).

Since MeDIP selects fragments showing *any* methylation, each CG pair on an observed read are typically given greater probability of methylation, regardless of whether the read was fully methylated or only had one methylated CG site. Under the Batman model, the probability of methylation for a given CG dinucleotide depends on the number of methylated CGs on the read via the coupling factor C_{cr}, or, more generally, the total coupling for the read $C_{tot} = \sum_c C_{cr}$. As the density of methylation increases around a given CG dinucleotide, this composite factor C_{tot} gives additional

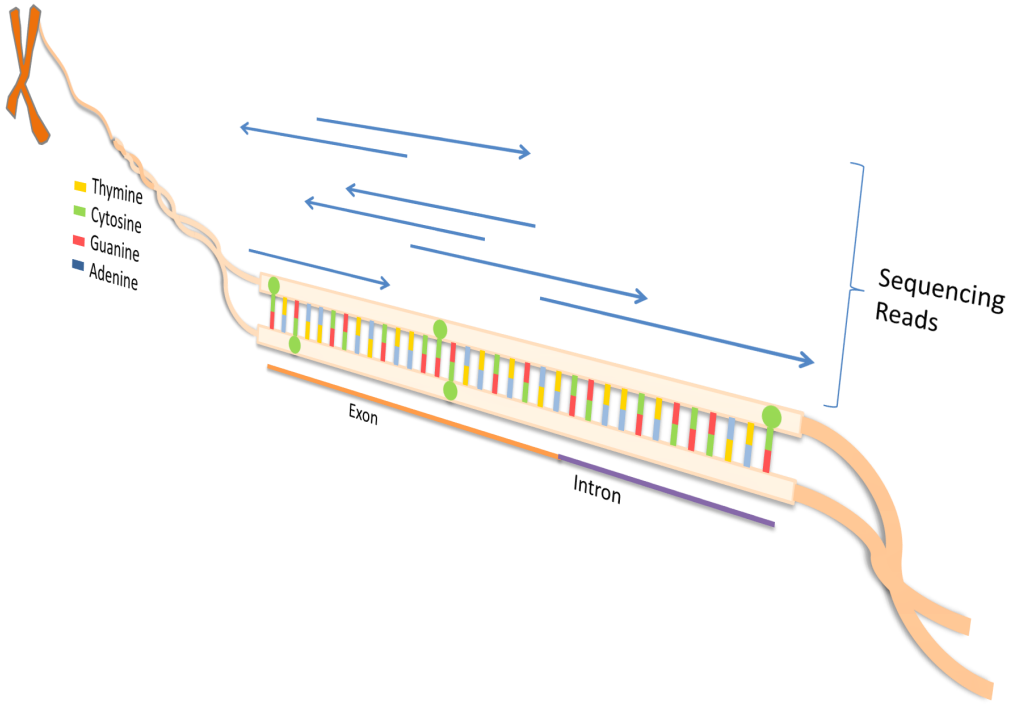

Figure 5: Illustration of data produced by MBD2 (used in MiGS) or MeDIP (used in Batman) sequencing. Sequencing reads are mapped (with direction) to a reference genome. Reads are observed only when methylation is present within the sequence fragment.

weight to the probability that the given site is methylated. In adjusting for methylation density in this way, the Batman algorithm attempts to remove some of the bias associated with the fragment selection procedures associated with MeDIP. This correction is most apparent in applications with long reads and sparse methylation.

3 Bisulfite-Based Technologies

Another major class of technologies used for examining DNA methylation in NGS employ a sodium bisulfite treatment step, which converts unmethylated cytosines to uracils (Figure 4). The two most common approaches are Reduced Representation Bisulfite Sequencing (RRBS, (Meissner et al., 2005)) and MethylC-seq (Lister et al., 2008). Unlike immunoprecipitation-based technologies, bisulfite-based technologies are able to interrogate the methylome at single base-pair resolution (Figure 6). Higher resolution comes at a price, though, since either genomic coverage (RRBS, $\sim$10% coverage genome-wide), or cost (MethylC-Seq, $\sim$5- to 10-fold increase in cost over other methods) is sacrificed.

The reduction in coverage when using RRBS is attributed to the preprocessing steps of the DNA sample. Sample DNA is digested using a restriction endonuclease enzyme, which cuts the DNA at specific locations based on the underlying nucleotide sequence. The resulting fragments are

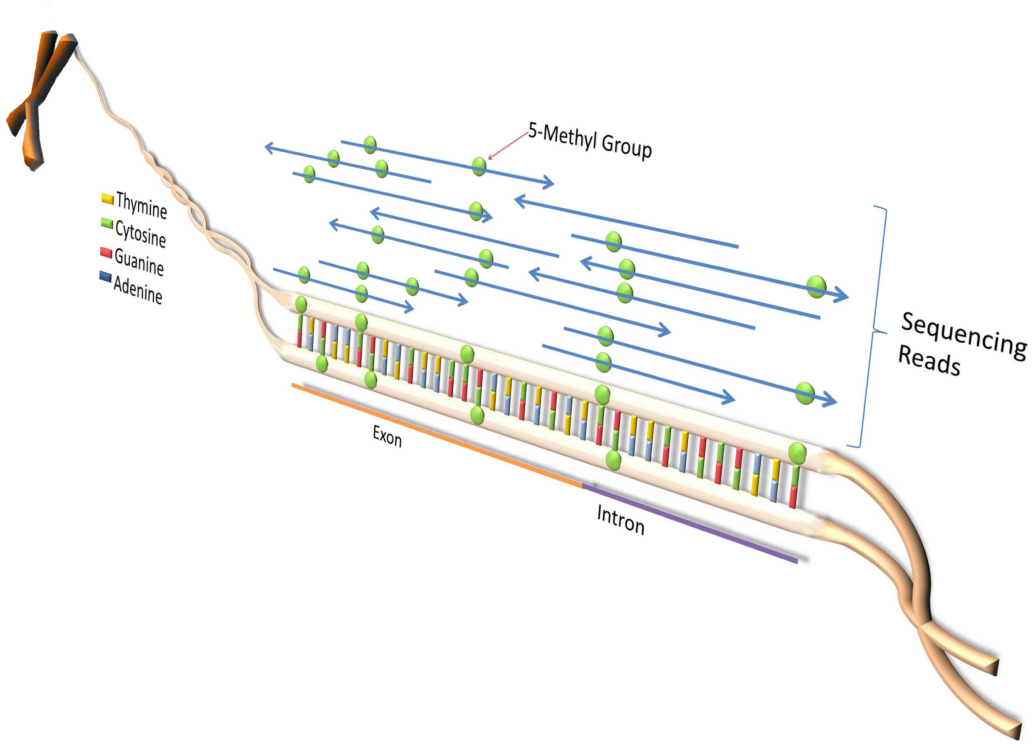

Figure 6: Illustration of data produced by Reduced Representation Bisulfite Sequencing (RRBS), or MethylC-seq. Sequencing reads are mapped (with direction) to a reference genome and methylation status is observed (green circles). Reads are observed independently of methylation status.

then size-selected and treated with sodium bisulfite. The process of digestion and size selection generates predictable fragments based on prior knowledge of the DNA sequence being studied, and as such, a "reduced representation" of the genome (generally 10-12%) is generated. The bisulfite treatment converts unmethylated cytosines to uracils, but is quite damaging to the sample product. The lengthy digestion process degrades almost the entire sample (>90% degradation within the first hour of bisulfite treatments (Grunau et al., 2001)). With this in mind, the bisulfite-treated samples are amplified through polymerase chain reaction (PCR) in excess of 8 times before sequencing. Sequenced fragments are then compared against reference fragments which have been computationally derived given the genome sequence and restriction enzyme sequence digestion proclivity. This process tends to bias toward CG-rich regions in the genome, limiting the scope of RRBS studies (Beck, 2010).

Rather than using a restriction enzyme to fragment the genome, the current gold-standard for NGS methylation studies, MethylC-seq, subjects the sample DNA to high-frequency sound waves, called sonication, producing fragments that cover the entire genome with no sequence bias. These fragments are then size-selected, amplified through PCR, and sequenced. Since fragments can theoretically originate from any location in the genome, sequenced reads are aligned to the entire genome sequence (rather than the reduced genome used in RRBS). This feature requires considerably more sample DNA than does RRBS (5μg vs. 0.03-0.05μg for MethylC-seq and RRBS, respectively (Beck, 2010)).

		1	2	···	K
	Cytosine Index				
Treatment 1	Rep. 3	17/19 (M)	3/17 (UM)	···	13/16 (M)
	Rep. 2	7/20 (UM)	13/13 (M)	···	11/15 (M)
	Rep. 1	3/5 (M)	1/14 (UM)	···	14/19 (M)
Treatment 2	Rep. 1	1/15 (UM)	13/14 (M)	···	8/16 (M)
	Rep. 2	2/17 (UM)	1/17 (UM)	···	7/11 (M)
	Rep. 3	3/11 (UM)	2/15 (UM)	···	13/17 (M)
		⇓	⇓	⇓	⇓
	Base Pair Decision	D	ND	···	ND

Table 2: For each cytosine (in CG context) and replicate, the number of reads indicating methylation out of the sequencing depth (total number of reads mapped to the cytosine) are recorded. Tests for differential methylation can be performed with base-pair level testing (in blue, D="Differentially Methylated" and ND="Not Differentially Methylated").

Both methods, RRBS and MethylC-seq, produce several (5-15, on average) reads at each genomic cytosine at which the methylation status is also recorded. These data (Table 2) can be summarized between treatments in a number of ways. Meissner et al. (2005) summarized RRBS data by calculating the correlation between the read profile for wild-type and two methylation maintenance deprived mutant lines. Lister et al. (2008), on the other hand, took a testing approach. Specifically, for each cytosine to which unique reads aligned, a Fisher's Exact Test was performed (Table 3). Once completed for each cytosine, these results were summarized over 1000bp windows. If the proportion of differentially methylated cytosines exceeded a pre-specified threshold (say, 20%), which is determined in an ad hoc manner, the window was labeled differentially methylated (DM), and neighboring DM windows were combined to form a differentially methylated region (DMR).

4 Incorporating Annotation

While NGS technologies have allowed researchers to gain valuable insight into epigenomics, the inclusion of other information that is available in well-annotated databases has not been investigated. In addition to the sequence of bases which make up DNA, and the methylation information, the genomes of many species are well-documented or well-annotated, meaning that there are data for the genomic location and function of genes, as well as other intergenic regions. The annotation information not only provides information as to what the base sequence is, but it also includes the genomic context of the sequence (e.g., exon, intron, untranslated region, etc.). Annotation information is generally only used after an analysis is performed, and for the purpose of bringing scientific context to the investigation. Specifically, once the single bin (for immunoprecipitation-based technologies) or single base (for bisulfite-based technologies) tests have been performed, and statistically significant results determined, only then is the annotation information brought to the biological interpretation of the results. Interestingly, Zhang et al. (2006) showed that methylation patterns can vary greatly between annotated regions, information that indeed may be useful. In fact, testing within annotation regions has several advantages when compared to the traditional sliding window approaches (Lister et al., 2008, 2009). First, the individual base or bin tests are

Cytosine	Treatment 1	Treatment 2
Unmethylated reads	n_{11}	n_{12}
Methylated reads	n_{21}	n_{22}

Table 3: 2×2 contingency table for base-pair level Fisher's Exact Test.

likely to be more homogeneous in a single annotated region than in a window which may include multiple annotated regions. Secondly, conclusions drawn from annotated regions are generally more interpretable than those gained from windows. Testing over annotated regions offers an attractive way to incorporate preexisting biological knowledge. It has potential for a significant impact on both the analysis and the interpretation of results.

We incorporate annotation early in the analysis pipeline using a subset-level test. The read information at each cytosine is summarized into a binary representation (i.e., methylated or unmethylated) for each sample using a predetermined threshold (typically 50%). These summaries are then compared using a Fisher's Exact Test (in the case of single replicates within each treatment) or a logistic regression (multiple replicates). Summarizing and testing in this fashion generally utilizes longer gene lengths (in comparison to sequencing depth) when performing the statistical test, giving rise to increased information and increased statistical power at equal or lower depths.

5 Simulation Study

Although NGS technologies have ignited DNA methylation research, it is difficult to understand the efficacy of each statistical method that has been used to analyze the data without extensive, and costly, biological confirmation studies. As an alternative, we rely on simulations to explore the strengths of each statistical approach. To expand upon the idea of subset-level testing, we simulated 1000 exons under two treatments and explored the differences between two testing strategies: single-base tests with post-hoc exon summaries (called 'base-pair level tests'), and single-base binary summaries with subsequent single tests over exons (called 'subset-level tests'). Both differentially methylated and non-differentially methylated exons were simulated such that, for each exon, the known or true methylation status was generated independently for both treatments using a $Bernoulli$ distribution with $p = .5$. Exon lengths (in CG pairs) were simulated via a $Poisson$ distribution, with mean $20, 35$, or 50; lengths are equal between treatments for each exon. Exon lengths were chosen to represent a spectrum of overall gene lengths, with emphasis on shorter genes. Within each exon, cytosine methylation was generated via a Hidden Markov Model, with transition matrices (Table 4a). Sequencing depth was simulated using a $Poisson$ (with mean 15, the current average sequencing depth attained in literature) distribution. Finally, read methylation was simulated through a $Binomial(n = depth, p)$ distribution, where p is as described in Table 4b. These parameters were chosen to reflect the averages and ranges of parameters commonly seen in *Arabidopsis* and human DNA methylation studies (Lister et al., 2008, 2009).

Base-pair level and subset-level tests were performed using either a Fisher's Exact Test (with one replicate) or with a logistic regression(with three replicates). Currently next-generation sequencing procedures are costly when compared to older technologies (i.e., microarrays); therefore, the number of biological samples are typically limited (i.e., fewer than three). As such, the sample sizes (n=1

	UMC	MC
UMC	0.50	0.50
MC	0.15	0.85

(a)

$Setting$	$P(MR\|UMC)$	$P(MR\|MC)$
1	0.20	0.50
2	0.15	0.60
3	0.15	0.70
4	0.10	0.80

(b)

Table 4: Setup for Hidden Markov Model (HMM) simulation of exons. (a) HMM transition matrix for methylated exons. This process is weighted toward the methylated cytosine (MC) state over the unmethylated cytosine (UMC) to accurately reflect the underlying biological processes. The transition matrix for unmethylated exons is constructed similarly. (b) Binomial probabilities of a methylated read (MR) given cytosine methylation status (unmethylated (UMC) and methylated (MC)).

and 3) chosen in this simulation study represent current practices. Multiple testing correction was performed by controlling the false discovery rate (FDR)(Benjamini and Hochberg, 1995) at $\alpha = 0.05$. The simulation process was repeated 1000 times, and the statistical power of each method was recorded (Figure 7).

As expected, increasing the number of replicates from one to three tends to improve the power of the tests, with this trend being more evident in the single base-pair (BP) approach. When restricted to a single replicate, the characteristics of individual cytosine methylation behavior (Figure 7 columns) have little effect on the power of the subset-level testing approach. That is, as the consistency of read methylation status given cytosine methylation status improves, little to no increase in statistical power is observed. This may imply that for single-replicate exploratory experiments, the subset-level approach will retain power more consistently as the levels of noise increase simply because the subset summaries provide more information. Specifically, subsetting based on annotation allows for testing even when the number of replicates is extremely low. The simulations presented here can be extended for use with the immunoprecipitation-based technologies, replacing cytosines with bins (with appropriate simulation modifications).

6 Conclusion

We have discussed several of the most commonly used approaches for exploring methylomes using next-generation sequencing technologies. Since each experimental protocol differs slightly, different statistical approaches are needed to model and test for differential methylation. While the immunoprecipitation-based approaches (using the MiGS, or Batman algorithms) offer researchers an inexpensive way to explore the methylome, the bisulfite-based methods (RRBS and MethylC-seq) offer single-base resolution that proves to be beneficial for many epigenomic applications. Each of these methods, however, require careful specification of the regions on which statistical tests for differential methylation are applied. Defining these windows using annotation information can lead to more homogeneous, powerful tests, even in unreplicated settings. We anticipate as the technologies for sequencing advance into single-molecule frameworks that the strategies presented here for informing tests through annotation will easily translate and will be directly applicable to the new data.

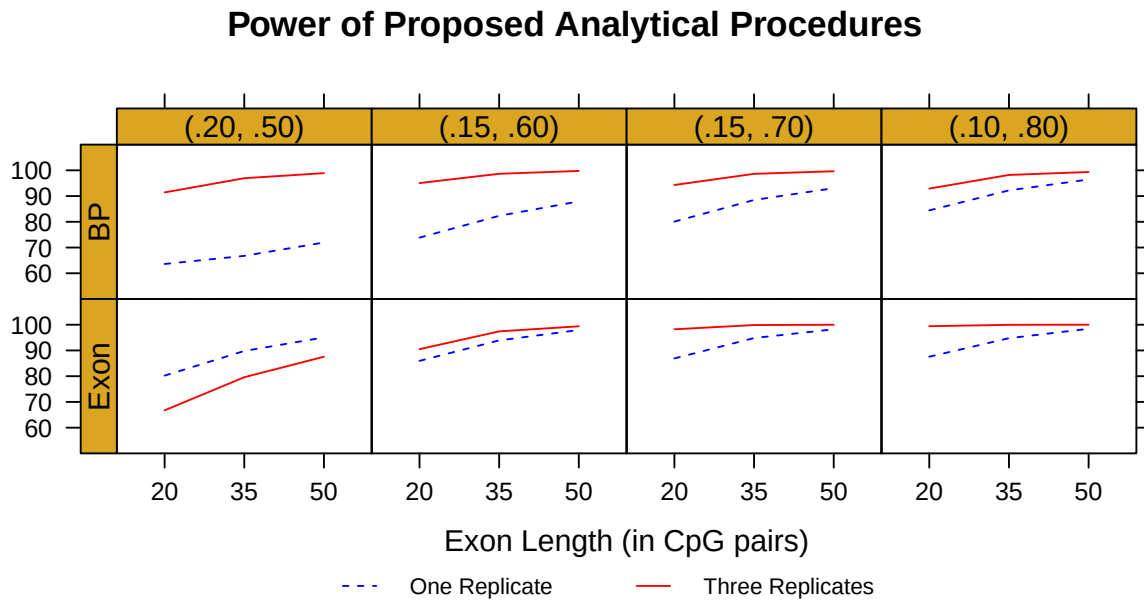

Figure 7: Power analysis results for base-pair and subset focused testing procedures with one (blue) or three (red) replicates with constant 5% false discovery rate. Power increases as exon length increases and as binomial probabilities separate. The base-pair focused method benefits more from replication than does the subset focused method, and has high performance in each of the settings examined where replication was available. Each method performs well when the binomial probabilities are well-separated. Performance of each method decreases as the HMM transition matrices become less stable (not shown).

References

Aoyama, T., T. Okamoto, S. Nagayama, K. Nishijo, T. Ishibe, K. Yasura, T. Nakayama, T. Nakamura, and J. Toguchida (2004). Methylation in the core-promoter region of the chondromodulin-i gene determines the cell-specific expression by regulating the binding of transcriptional activator sp3. *Journal of Biological Chemistry 279*, 28789–28797.

Beck, S. (2010). Taking the measure of the methylome. *Nature Biotechnology 28*, 1026–1028.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological) 57*, 289–300.

Bock, C., E. Tomazou, A. Brinkman, F. Muller, F. Simmer, H. Gu, N. Jager, A. Gnirke, H. Stunnenberg, and A. Meissner (2010). Quantitative comparison of genome-wide dna methylation mapping technologies. *Nature Biotechnology 28*, 1106–1114.

Cokus, S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, and S. E. Jacobsen (2008, March). Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature 452*, 215–219.

Down, T., V. Rakyan, D. Turner, P. Flicek, H. Li, E. Kulesha, S. Grf, N. Johnson, J. Herrero, E. Tomazou, N. Thorne, L. Bckdahl, M. Herberth, K. Howe, D. Jackson, M. Miretti, J. Marioni, E. Birney, T. Hubbard, R. Durbin, S. Tavar, and S. Beck (2008). A bayesian deconvolution strategy for immunoprecipitation-based dna methylome analysis. *Nature Biotechnology 26*, 779–785.

Feinberg, A. and B. Vogelstein (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature 301*, 89–92.

Finnegan, E. (2010). *Plant Developmental Biology - Biotechnological Perspectives*, Chapter DNA methylation: a dynamic regulator of genome organization and gene expression in plants. Springer Berlin Heidelberg.

Fisher, R. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society 85*, 87–94.

Groom, A., H. R. Elliott, N. D. Embleton, and C. L. Relton (2010). Epigenetics and child health: basic principles. *Archives of Disease in Childhood*.

Grunau, C., S. Clark, and A. Rosenthal (2001). Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Research 29*, E65–65.

Irizarry, R., C. Ladd-Acosta, B. Carvalho, H. Wu, S. Brandenburg, J. Jeddeloh, B. Wen, and A. Feinberg (2008). Comprehensive high-throughput arrays for relative methylation (charm). *Genome Research 18*, 780–790.

Jacobsen, S. (2011). Research program. *http://www.mcdb.ucla.edu/research/jacobsen/labwebsite*.

Kouzarides, T. (2002, February). Chromatin modifications and their function. *Cell 128*(4), 693–705.

Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker (2008). Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell 133*, 523–536.

Lister, R., M. Pelizzola, R. Dowen, R. Hawkins, G. Hon, J. Tonti-Filippini, J. Nery, L. Lee, Y. Zhen, Q.-M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. Millar, J. Thomson, B. Ren, and J. Ecker (2009). Human dna methylomes at base resolution show widespread epigenomic differences. *Nature 462*, 315–322.

Meissner, A., A. Gnirke, G. Bell, B. Ramsahoye, E. Lander, and R. Jaenisch (2005). Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic Acids Research 33*, 5868–5877.

Serre, D., B. Lee, and A. Ting (2010). Mbd-isolated genome sequencing provides a high-throughput and comprehensive survey of dna methylation in the human genome. *Nucleic Acids Research 38*, 391–399.

Skilling, J. (2006). Nested sampling for general bayesian computation. *Bayesian Analysis 1*, 833–860.

Wang, X., C. Zhang, L. Zhang, X. Wang, and S. Xu (2006). High-throughput assay of dna methylation based on methylation-specific primer and sage. *Biochemical and biophysical research communications 341*, 749–754.

Weber, M., J. Davies, D. Wittig, E. Oakeley, M. Haase, W. Lam, and D. Schubler (2005). Chromosome-wide and promoter-specific analyses identify sites of differential dna methylation in normal and transformed human cells. *Nature Genetics 37*, 853–862.