Epigenetics 9:5, 698–703; May 2014; © 2014 Landes Bioscience

MAGI
Methylation analysis using genome information

Douglas D Baumann[1] and RW Doerge[2,*]

[1]Department of Mathematics; University of Wisconsin, La Crosse; La Crosse, WI USA; [2]Department of Statistics; Purdue University; West Lafayette, IN USA

Keywords: annotation informed, differential methylation, epigenetics, epigenomics, statistical bioinformatics, testing methylation

By incorporating annotation information into the analysis of next-generation sequencing DNA methylation data, we provide an improvement in performance over current testing procedures. Methylation analysis using genome information (MAGI) is applicable for both unreplicated and replicated data, and provides an effective analysis for studies with low sequencing depth. When compared with current tests, the annotation-informed tests provide an increase in statistical power and offer a significance-based interpretation of differential methylation.

Introduction

DNA methylation (herein, methylation) is an important, heritable, epigenetic modification that is known to influence gene expression, X-inactivation, and cellular differentiation in higher eukaryotes.[1-3] Next-generation sequencing (NGS) technologies, such as MethylC-seq,[4] make it feasible to investigate methylomes at the cytosine level and provide unparalleled insight into the role and function of methylation in a variety of organisms. Two recent studies[5,6] investigated the cost, genome coverage, maximum resolution, and quality measures for a variety of NGS approaches as applied to DNA methylation. Although the conclusions from these studies provided us guidance in choosing a technology for this application (i.e., MethylC-seq is typically considered the gold-standard for methylation analysis), there are a variety of NGS technologies for which the proposed approach is applicable.

It is well accepted that there can be vast differences in methylation patterns with respect to genomic regions (e.g., genes, promoters, intergenic regions).[7] Changes in methylation between conditions give rise to epigenomic discoveries that are uniquely related to the organization and control of the genome. Interestingly, current epigenomic investigations do not incorporate genome organization into the actual quantitative analysis for testing differences between methylation profiles. In fact, annotation is typically consulted after the quantitative results are obtained, and only for the purpose of gaining genomic context.

One of the most common approaches for testing differential methylation is a sliding window approach[8] that compares, at the cytosine level, the observed methylation levels to known/annotated regions of methylation.[9] This type of approach unfortunately leads to greater intraclass variability with less informative conclusions, simply because the windows are artifacts of the analysis and may in fact overlap multiple annotation regions simultaneously. As an improvement, we utilize existing annotation information to enhance the performance of testing for differences in methylation. The proposed approach is particularly useful for unreplicated data, and while we focus on the benefit of incorporating annotation using Fisher's Exact Test[10] (FET) for unreplicated data, the extension to replicated data are straightforward (see Methods).

Results and Discussion

We present methylation analysis for MethylC-seq data using two approaches, collectively referred to as Methylation Analysis using Genome Information (MAGI), both of which rely on an annotated genome. MethylC-seq involves a bisulfite treatment step, which converts unmethylated cytosines to uracils (and ultimately guanines), on each fragmented read prior to sequencing. As a result, a measurement of methylation level at each cytosine on the genome can be estimated by comparing the number of methylated and unmethylated cytosines on the sequenced reads. The first approach employs FET at the cytosine level using the number of (NGS) methylated reads, among a total number of reads, mapped to each cytosine in a known annotated region (**Fig. 1**). The false discovery rate[11] (FDR) is controlled by applying multiple testing corrections to the cytosine level tests within each region.[12] Results are summarized across each annotated region, and if the proportion of differentially methylated cytosines exceeds some ad hoc and arbitrary threshold (e.g., 10%), the region is declared differentially methylated. We refer to this standard approach as the $MAGI_C$ approach. $MAGI_C$ can be thought of as a special case of the sliding window approaches,[8,9] where the non-overlapping genomic "windows"

*Correspondence to: RW Doerge; Email: doerge@purdue.edu
Submitted: 06/27/2013; Revised: 01/29/2014; Accepted: 02/22/2014; Published Online: 03/31/2014
http://dx.doi.org/10.4161/epi.28322

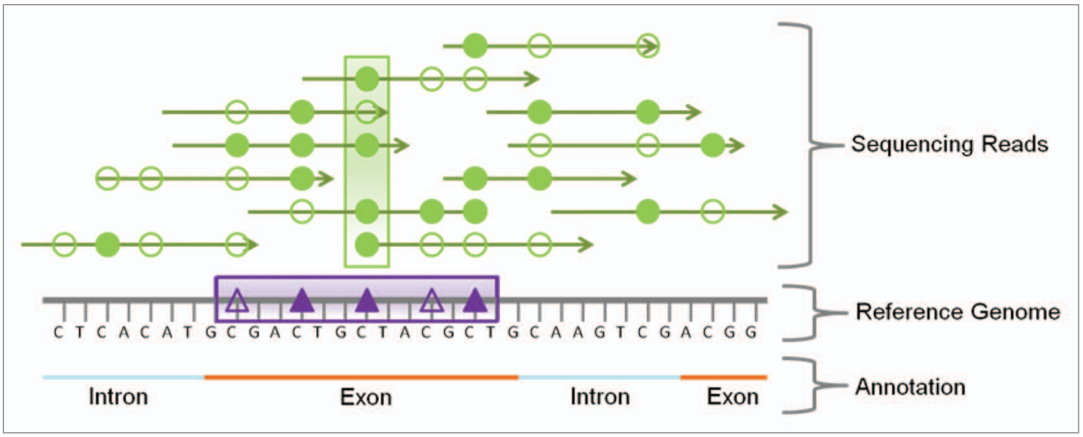

Figure 1. Representation of the data structure and testing framework for NGS differential methylation studies (forward strand shown). For each cytosine, the number of methylated reads (filled circles) and unmethylated reads (unfilled circles) are recorded. These values are also recorded in binary representation for each cytosine, where a filled triangle indicates that the proportion of methylated reads for the given cytosine has exceeded a predetermined threshold (e.g., 40%), and an unfilled triangle indicates that this proportion was not exceeded. Tests for differential methylation are performed for each individual cytosine using the read information with subsequent summarization over the region ($MAGI_C$), or with a single region-level test ($MAGI_G$) using the summarized read information.

represent regions of homogeneous methylation profiles; in this sense, $MAGI_C$ represents the most powerful, best-case sliding window scenario. Since each cytosine is tested individually, the region-level summary is a comparison of methylation patterns between treatment groups. While this exploration is centered on regions that define a gene, it is natural to extend the approach to other annotations (e.g., promoters, exons, intergenic regions, etc.). Since $MAGI_C$ compares the patterns of methylation between treatments, it is suitable for regions where a classification of "methylated" or "unmethylated" is not of interest (e.g., intergenic regions). In these cases, an appropriate partitioning of the region could be developed to further investigate long genomic regions if intraclass variability is of concern; if not, $MAGI_C$ could be easily adapted to incorporate the sliding window approach over these regions.

The second approach summarizes the methylation status for each cytosine based on the observed proportions of methylation (see Methods), which allows for a marginalization over each annotated region that is then tested with a single FET (**Fig. 1**). FDR is controlled across the genome with multiple testing corrections applied to annotated regions. We call this data-adaptive approach $MAGI_G$. By contrast to $MAGI_C$, the $MAGI_G$ approach minimizes the number of statistical tests that are conducted and benefits from a significance-based interpretation of differential methylation. In addition, the region-level summary from $MAGI_G$ represents a comparison of overall methylation levels between treatment groups.

Data were simulated for both unreplicated and replicated scenarios. For each simulation the FDR is controlled at 5%. The statistical power is estimated by assessing the average true positive rate for 1000 simulated data sets for each scenario (see Methods).

For unreplicated experiments, power gains for $MAGI_G$ are most evident when either correlation in consecutive cytosine-to-cytosine methylation status decreases (**Fig. 2**; "Medium" or "Low"), or the differences in methylation level at each cytosine decrease (**Fig. 2**; "Small" or "Medium"). In these cases, the power of the $MAGI_G$ method can be upward of 40% greater than that of the $MAGI_C$ approach. These simulations also illustrate modest power increases when the average sequencing depth increases from 7 to 15 (representing the range of depths typically employed[6,8]); when differences in cytosine methylation levels are large (**Fig. 2**; "Large") there is little gain in statistical power relative to additional sequencing depth.

In a real data application we reanalyzed the unreplicated *Arabidopsis* methylcytosine data from Lister et al. (2008)[6] that compared wild-type (*Col-0*) lines to methylation-deficient mutants (*met1–3*). We considered all cytosines (with at least one read) in both samples, including those that demonstrated no evidence of methylation in either sample. Gene start and stop locations define the annotated genomic regions and were based on the Columbia reference genome.[13] We applied $MAGI_C$ and $MAGI_G$ to the MethylC-seq data for each gene, and assessed the differential methylation detection rate after FDR corrections for each method, as defined above. As $MAGI_C$ represents the optimal sliding window scenario due to methylation profiles varying substantially between genomic regions, and since only gene regions were investigated to facilitate interpretation of results, comparisons with the sliding window approach from Lister et al. (2008)[8] have been omitted. Due to fewer tests and more information per test, the $MAGI_G$ approach provides many more statistically significant results than the $MAGI_C$ approach (Methods, **Table 4**). These results reinforce that *met1–3* mutants have defective methylation maintenance when compared with the wild-type (*Col-0*) controls,[14] and are consistent with the average rate of genic methylation in the wild-type and mutant lines[15] and simulation power estimates (**Fig. 2**).

Single-cytosine analyses of methylation using NGS technologies encounter two primary challenges when summarizing to region-level results, namely dependence in methylation status between cytosines and the discrete nature of test statistics (and associated

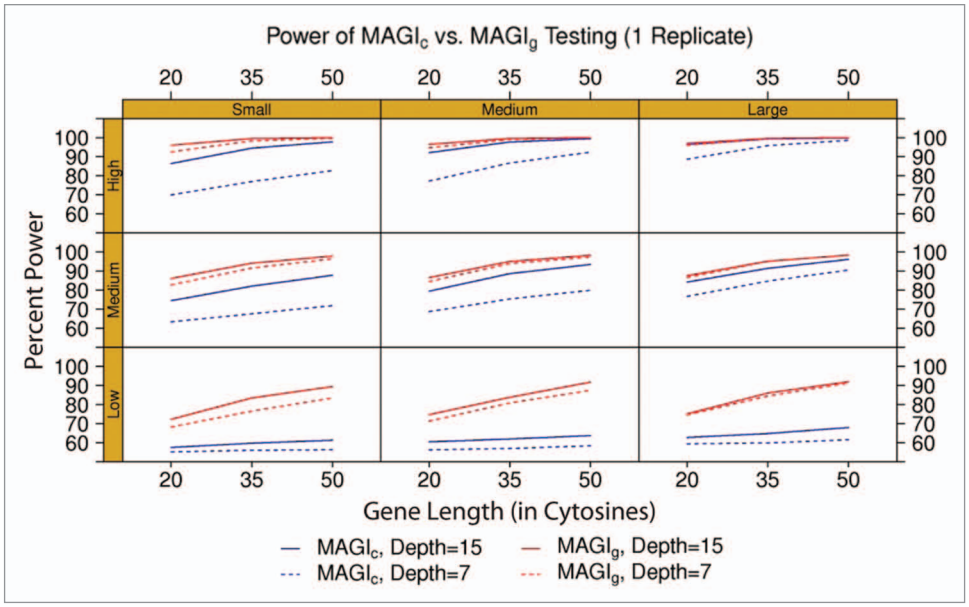

Figure 2. Simulation results in unreplicated settings. Panel columns represent the separation of binomial probabilities for read methylation (Methods). Panel rows represent the transition matrices used in the Hidden Markov Model (HMM) process used to generate cytosine methylation status (Methods). Increases in the statistical power of the $MAGI_G$ over $MAGI_C$ are evident across the simulation settings. Observed false discovery rates (FDRs) are lower in $MAGI_G$ (2–19%) when compared with $MAGI_C$ (4–43%). However, both FDRs increase with greater separation of binomial probabilities and decrease with greater correlation between cytosines. Modest statistical power increases are observed when the average sequencing depth is increased from 7 to 15 with similar observed FDRs.

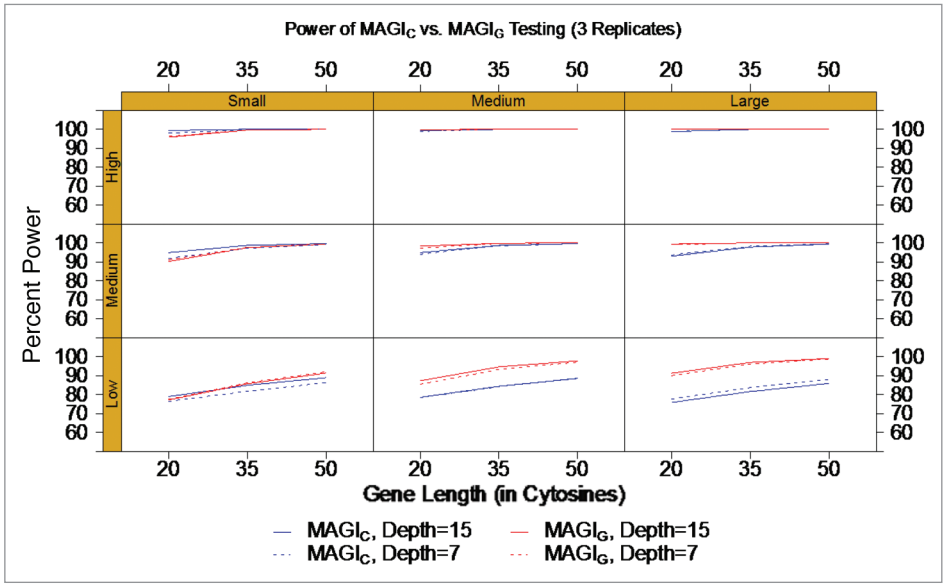

Figure 3. Simulation results in replicated settings. Panel columns represent the separation of binomial probabilities for read methylation (Methods). Panel rows represent the transition matrices used in the Hidden Markov Model (HMM) process used to generate cytosine methylation status (Methods). Increases in power of $MAGI_G$ were evident as the binomial probabilities of methylation increase in separation (i.e., from "Small" to "Large"). Very small power increases can be observed when increasing the average sequencing depth from 7 to 15. In general, replication may be sufficient to overcome the differences in sequencing depth, as well as the differences between $MAGI_C$ and $MAGI_G$.

P values). Several methods have been introduced to combine *P* values over regions under dependence,[16-18] including weighted approaches that could account for differences in sequencing depths,[19,20] but each method assumes that the *P* values were generated from multivariate t (or Z) distributions. On the other hand, FETs produce discrete, non-uniform *P* values under the null distribution, and as such, the distributional assumptions for combining *P* values are violated. The approaches recently proposed by Hebestreit et al. (BiSeq)[18] and Pedersen et al. (comb-p)[20] both employ variations on Stouffer's method[17] to

Table 1. Representation of the data structure and testing framework MAGI differential methylation studies

			Cytosine Index			
	Rep.	1	*c*	C_g		Summary Information
	1	$(m_{111\,g}, D_{111\,g})$	(m_{11cg}, D_{11cg})	(m_{11Cg}, D_{11Cg})	→	$\left(M_{11g} = \sum_{c=1}^{C} I(\frac{m_{11cg}}{D_{11cg}} \geq T_s), C_g\right)$
Trt 1	*j*	(m_{1j1g}, D_{1j1g})	(m_{1jcg}, D_{1jcg})	(m_{1jCg}, D_{1jCg})	→	$\left(M_{1jg} = \sum_{c=1}^{C} I(\frac{m_{1jcg}}{D_{1jcg}} \geq T_s), C_g\right)$
	J	$(m_{1J1g}, D_{121\,g})$	(m_{1Jcg}, D_{1Jcg})	(m_{1JCg}, D_{1JCg})	→	$\left(M_{1Jg} = \sum_{c=1}^{C} I(\frac{m_{1Jcg}}{D_{1Jcg}} \geq T_s), C_g\right)$
	1	$(m_{211\,g}, D_{211\,g})$	(m_{21cg}, D_{21cg})	(m_{21Cg}, D_{21Cg})	→	$\left(M_{21g} = \sum_{c=1}^{C} I(\frac{m_{21cg}}{D_{21cg}} \geq T_s), C_g\right)$
Trt 2	*j*	(m_{2j1g}, D_{2j1g})	(m_{2jcg}, D_{2jcg})	(m_{2jCg}, D_{2jCg})	→	$\left(M_{2jg} = \sum_{c=1}^{C} I(\frac{m_{1jcg}}{D_{1jcg}} \geq T_s), C_g\right)$
	J	(m_{2J1g}, D_{2J1g})	(m_{2Jcg}, D_{2Jcg})	(m_{2JCg}, D_{2JCg})	→	$\left(M_{2Jg} = \sum_{c=1}^{C} I(\frac{m_{1Jcg}}{D_{1Jcg}} \geq T_s), C_g\right)$

For each treatment *i*, replicate *j*, cytosine *c*, and gene *g*, the number of methylated reads (m_{ijcg}) and the sequencing depth (total number of reads mapped to the cytosine, D_{ijcg}) are recorded. $MAGI_C$ tests for differential methylation at each cytosine using a Fisher's Exact Test (no replicates) or a logistic regression (replicates); if the proportion of positive base-pair decisions exceeds a predefined threshold, the subset is declared differentially methylated. $MAGI_G$ first summarizes the read information for each cytosine within each treatment and replicate, and then performs tests on this summarized information. For treatment *i* and replicate *j* for gene *g*, M_{ijg} represents summary information on the number of cytosines for which m_{ijcg}/D_{ijcg} exceeds a predetermined threshold T_S. Given M_{ijg} and the subset length C_g, tests similar to those used for the base-pair level framework ($MAGI_C$) can be employed.

combine *P* values from bisulfite methylation data, but it is unclear whether the distributional assumptions are reasonable in either case. Although BiSeq improves upon the assumptions from Stouffer's method by first smoothing the methylation data over a genomic range, the models employed require larger data sets than are required by either MAGI approach. Further, both MAGI approaches can be applied to studies without replication. In order to more closely satisfy distributional assumptions, researchers often rely on increased average sequencing depth. Increasing average sequencing depth in unreplicated studies has distinct advantages when testing individual cytosines, but the benefits all but disappear when differential methylation is considered over specific annotated regions (**Fig. 2**). This is due in part to the use of nominal significance thresholds (e.g., $\alpha = 0.05$). When dealing with low coverage at individual cytosines, the discreteness of the *P* value space covered by FET typically leads to a reduction in statistical power when compared with other unconditional exact tests (e.g., Barnard's Test[21]). Unfortunately, this discreteness translates to a loss in power when the significance threshold is fixed. Fortunately, these issues dissipate when the column marginals are large (as is the case for $MAGI_G$ testing); this is due to the discreteness of FET being less pronounced. Combining dependent, discrete *P* values over genomic regions is an approach that we are currently investigating since it has the potential to further improve statistical inference in differential methylation studies beyond the gains observed in $MAGI_G$ testing.

To explore the effects of low coverage of individual bases on detection of methylation difference, we filtered cytosines that had observed read depths of less than a specific low-count threshold (i.e., 5, 7, or 10) for either sample (*Col-0* and *met1–3*). Overall, appreciable changes in methylation detection for each method were found, indicating that moderate low-count filtering (filtering level 5) is a reasonable approach to increase detection rate for the $MAGI_C$ approach and to distill the results from the $MAGI_G$ approach (Methods, **Table 4**). Excessive filtering (i.e., filtering levels 8 and 10) yields little benefit to $MAGI_C$, however, and may be too extreme for $MAGI_G$. The dramatic differences between the $MAGI_C$ and $MAGI_G$ results highlights the inferential distinctions between the two methods. Specifically, $MAGI_C$ may be better suited to exploring differences in methylation patterns, while $MAGI_G$ is more appropriate when testing for differences in methylation prevalence. In both cases, genomic context provides useful boundaries for region-level summaries.

Methods

MethylC-seq data can be represented at the cytosine level as the cumulative number of methylated and unmethylated sequencing

Table 2. Cytosine-specific methylation status transition matrices for methylated genes

(A) High			(B) Medium			(C) Low		
	UM	M		UM	M		UM	M
UM	0.35	0.65	UM	0.50	0.50	UM	0.35	0.65
M	0.15	0.85	M	0.15	0.85	M	0.35	0.65

"M" and "UM" represent methylated and unmethylated status, respectively. Unmethylated gene transition matrices are formed similarly, with elements on each diagonal interchanged. Transition matrix (**A**) forms chains with longer homogeneous strings of methylated cytosines, while matrices (**B**) and (**C**) allow more unmethylated cytosines to be generated when the gene is methylated.

Table 3. Binomial probabilities for assigning methylated status (MR) to a read for unmethylated and methylated cytosines (UC and MC, respectively)

Setting	Separation	P(MR|UC)	P(MR|MC)
1	Large	0.10	0.80
2	Medium	0.15	0.70
3	Small	0.15	0.60

Setting 1 indicates a large separation of read probabilities, and settings 2 and 3 decrease this level of separation.

Table 4. Exploration and impact of low-coverage filtering on significance results from $MAGI_C$ and $MAGI_G$ for 33,759 analyzed gene regions

Filtering Level	% Filtered	$MAGI_C$	$MAGI_G$	Intersection
No Filtering	0	216	3146	181
5	40	612	2926	310
7	51	669	2132	275
10	67	651	1528	246

Arabidopsis data (*Col-0* vs. *met1–3*) from Lister et al. (2008) were analyzed using both $MAGI_C$ and $MAGI_G$ with varying degrees of low-coverage filtering. Significance thresholds of 0.10 and 0.05 were employed as example thresholds for each method, respectively. The "Filtering Level" represents the threshold for which individual cytosines are removed from downstream analyses, while % filtered indicates the percentage removed as a result of the filtering. Specifically, if the coverage of a given cytosine is below this threshold in either sample, the cytosine information is not used. As the filtering becomes more strict (i.e., higher filtering level), the number of significant subsets decreases using $MAGI_G$, and increases using $MAGI_C$. A balance between increased detection for $MAGI_C$ and decreased detection for $MAGI_G$ occurs when the filtering level is set to 5.

bases covering a specific cytosine. For both unreplicated data and replicated data the analysis can be performed in two ways: focus on cytosine level tests and summarize to the genomic region, or summarize the cytosine level information and test once over the whole region (**Table 1**).

Cytosine level analysis ($MAGI_C$)

We employ Fisher's Exact Test (FET) when testing unreplicated data due to its lack of asymptotic assumptions and generally similar performance when compared with either Wald's Test or the methods proposed by Audic and Claverie.[22-24] Under the assumptions of fixed marginals, the FET for cytosine level differential methylation ($MAGI_C$) tests the hypotheses

$H_{0,cg}: \theta_{cg} = 1$ vs. $H_{1,cg}: \theta_{cg} \neq 1$,

where $\theta_{cg} = [\pi_{1cg}(1-\pi_{2cg})]/ [\pi_{2cg}(1-\pi_{1cg})]$ and π_{icg} is the true methylation level for the c^{th} cytosine in the g^{th} gene for treatment i (Trt_i). If the estimated odds ratio, where m_{ijcg} and D_{ijcg} are defined as in **Table 1**, differs significantly from 1, then cytosine c in gene g is differentially methylated. When the results from all cytosines in a given region are taken together, if the proportion of differentially methylated cytosines is above a pre-specified threshold (say, 0.10), the region is said to be differentially methylated. When biological replication is available, logistic regression can be applied, the hypotheses are similar to the unreplicated case, and the logistic model is

$\log(\pi_{cg}/[1-\pi_{icg}]) = \alpha + \beta_{cg}*Trt_i$.

The test for cytosine level differential methylation relies on the hypotheses

$H_{0,cg}: \beta_{cg} = 0$ vs. $H_{1,cg}: \beta_{cg} \neq 0$.

Genome region level analysis ($MAGI_G$)

The $MAGI_G$ approach summarizes methylation status for a genomic region by first inferring a binary representation of methylation status for each cytosine in a given gene, within (treatment) groups. If the proportion of methylated reads is above a given threshold τ (e.g, 0.40), the cytosine is considered methylated. The threshold τ can be set a priori, or determined empirically. Here, τ is defined as the mean of the two cluster centroids as determined through k-means ($k = 2$) clustering on the observed methylation proportions for each chromosome and strand. The gene level information for both groups is then summarized into a 2 × 2 table, where the rows represent methylated and unmethylated cytosine status and the columns represent treatment groups. The FET for this scenario tests

$H_{0,g}: \theta_g = 1$ vs. $H_{1,g}: \theta_g \neq 1$,

where $\theta_g = [\pi_{1g}(1-\pi_{2g})]/[\pi_{2g}(1-\pi_{1g})]$ and π_{ig} is the true methylation level for the g^{th} gene for treatment i. If the estimated odds ratio, where M_{ijg} and C_g are defined as in **Table 1**, differs significantly from 1, gene g is differentially methylation. When biological replication is available, logistic regression can be applied, hypotheses are similar to the unreplicated case, and the logistic model is

$\log(\pi_{ig}/[1-\pi_{ig}]) = \alpha + \beta_g*Trt_i$.

The test for base-pair level differential methylation relies on the hypotheses

$H_{0,g}: \beta_g = 0$ vs. $H_{1,g}: \beta_g \neq 0$.

Simulations

Both the cytosine ($MAGI_C$) and region level ($MAGI_G$) approaches are assessed using a series of simulations. In each simulation, 1000 genomic subsets, with "region lengths" as measured by the number of cytosines, are generated using a random *Poisson* (λ) where λ = 20, 35, or 50. These lengths were chosen to represent a variety of region lengths, with an emphasis on shorter regions. The region lengths are identical for two treatments with three subjects simulated in each treatment (only one subject was used in the unreplicated settings). The number of treatments and subjects in each treatment was chosen because they represent the common choices in exploratory experiments, though the logistic regression approach can accommodate more treatments and unequal samples. For each cytosine in each subject and independent of treatment group,

sequencing depth is simulated via a random *Poisson* (λ) where λ = 7 or 15, representing the range of current sequencing depths attained in the literature. The Poisson distribution was chosen in order to facilitate potential estimation of the rate parameter in unreplicated studies, and could be easily adapted in MAGI to incorporate or simulate additional variability in sequencing depth (via a Negative Binomial or similar distribution).

Methylation status for each read was generated by first assigning, within each treatment group, a methylation status to each subset via a random *Binomial* (2, 0.5) process (akin to a coin-tossing process for each treatment group). Then, given a subset's methylation status, subject-level base-pair-specific cytosine methylation status was simulated under a Hidden Markov Model (HMM) framework.[25] The HMM approach was used to account for the correlated nature of methylation of cytosines within a subset. The transition probabilities are defined in **Table 2**. The transition matrices were chosen to span a variety of methylation patterns, ranging from relaxed to strict methylation status based on the subset's status. Finally, given a cytosine's methylation status, individual read status is simulated via a random *Binomial (n, p)* distribution, where *n* is the sequencing depth at the given cytosine, and *p* is the probability of a methylated read given the cytosine methylation status (see **Table 3**). This process was repeated 1000 times.

Results under biological replication

Replicated data (i.e., three samples) were simulated using settings similar to the unreplicated scenario and analyzed using a logistic regression. Statistical power was comparable across varying sequencing depths, indicating that increased depth in the presence of replication may give rise to diminishing returns on investment. $MAGI_G$ gained power when differences in cytosine methylation levels are large; interestingly, this effect was not seen in $MAGI_C$. In the cases with higher correlation in consecutive cytosine-to-cytosine methylation status, the $MAGI_C$ and $MAGI_G$ approaches are comparable (**Fig. 3**).

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Author Contributions

D.D.B. and R.W.D. conceived the concept for this work. D.D.B. performed the simulations, analyses, and drafted the paper. R.W.D. oversaw the work, and finalized the paper.

References

1. Riggs AD. X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genet 1975; 14:9-25; PMID:1093816; http://dx.doi.org/10.1159/000130315
2. Holliday R, Pugh JE. DNA modification mechanisms and gene activity during development. Science 1975; 187:226-32; PMID:1111098; http://dx.doi.org/10.1126/science.1111098
3. Finnegan E. *Plant Developmental Biology – Biotechnological Perspectives.* DNA methylation: a dynamic regulator or genome organization and gene expression in plants. 2010; Springer Berlin Heidelberg.
4. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol 2010; 28:1097-105; PMID:20852635; http://dx.doi.org/10.1038/nbt.1682
5. Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, Jäger N, Gnirke A, Stunnenberg HG, Meissner A. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol 2010; 28:1106-14; PMID:20852634; http://dx.doi.org/10.1038/nbt.1681
6. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 2008; 133:523-36; PMID:18423832; http://dx.doi.org/10.1016/j.cell.2008.03.029
7. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell 2006; 126:1189-201; PMID:16949657; http://dx.doi.org/10.1016/j.cell.2006.08.003
8. Choufani S, Shapiro JS, Susiarjo M, Butcher DT, Grafodatskaya D, Lou Y, Ferreira JC, Pinto D, Scherer SW, Shaffer LG, et al. A novel approach identifies new differentially methylated regions (DMRs) associated with imprinted genes. Genome Res 2011; 21:465-76; PMID:21324877; http://dx.doi.org/10.1101/gr.111922.110
9. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009; 462:315-22; PMID:19829295; http://dx.doi.org/10.1038/nature08514
10. Fisher R. The logic of inductive inference. JR Stat Soc 1935; 98:39-82; http://dx.doi.org/10.2307/2342435
11. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc, B 1995; 57:289-300
12. Efron B. Simultaneous inference: when should hypothesis testing problems be combined? Annals of Applied Statistics 2008; 2:197-223; http://dx.doi.org/10.1214/07-AOAS141
13. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res 2008; 36:D1009-14; PMID:17986450; http://dx.doi.org/10.1093/nar/gkm965
14. Ogrocká A, Polanská P, Majerová E, Janeba Z, Fajkus J, Fojtová M. Compromised telomere maintenance in hypomethylated Arabidopsis thaliana plants. Nucleic Acids Res 2013; 2013:gkt1285; PMID:24334955
15. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet 2007; 39:61-9; PMID:17128275; http://dx.doi.org/10.1038/ng1929
16. Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet 2012; 13:705-19; PMID:22986265; http://dx.doi.org/10.1038/nrg3273
17. Brown M. A method for combining non-independent, one-sided tests of significance. Biometrics 1975; 31:987-92; http://dx.doi.org/10.2307/2529826
18. Hebestreit K, Dugas M, Klein HU. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics 2013; 29:1647-53; PMID:23658421; http://dx.doi.org/10.1093/bioinformatics/btt263
19. Stouffer S, Suchman E, DeVinney L, Star S, Williams R Jr. Adjustment during Army Life. The American Soldier 1949; (1) Princeton University Press, Princeton.
20. Pedersen BS, Schwartz DA, Yang IV, Kechris KJ. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated *P*-values. Bioinformatics 2012; 28:2986-8; PMID:22954632; http://dx.doi.org/10.1093/bioinformatics/bts545
21. Barnard G. A new test for 2x2 tables. Nature 1945; 156:783-4; http://dx.doi.org/10.1038/156783b0
22. Man MZ, Wang X, Wang Y. POWER_SAGE: comparing statistical tests for SAGE experiments. Bioinformatics 2000; 16:953-9; PMID:11159306; http://dx.doi.org/10.1093/bioinformatics/16.11.953
23. Ruijter JM, Van Kampen AH, Baas F. Statistical evaluation of SAGE libraries: consequences for experimental design. Physiol Genomics 2002; 11:37-44; PMID:12407185
24. Romualdi C, Bortoluzzi S, Danieli GA. Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests. Hum Mol Genet 2001; 10:2133-41; PMID:11590130; http://dx.doi.org/10.1093/hmg/10.19.2133
25. Rabiner L. A tutorial on hidden markov models and selected applications in speech recognition. Proc IEEE 1989; 77:257-86; http://dx.doi.org/10.1109/5.18626