Conference on Applied Statistics in Agriculture | 2015 - 27th Annual Conference Proceedings

DIFFERENTIAL METHYLATION METHODS IN MULTI-CONTEXT ORGANISMS

Douglas Baumann
University of Wisconsin - La Crosse, dbaumann@uwlax.edu

Yuqing Su
Missouri University of Science and Technology

Iranga Mendis
Missouri University of Science and Technology

Gayla R. Olbricht
Missouri University of Science and Technology, olbrichtg@mst.edu

DIFFERENTIAL METHYLATION METHODS IN MULTI-CONTEXT ORGANISMS

Douglas D. Baumann[1], Yuqing Su[2], Iranga Mendis[2], Gayla R. Olbricht[2*]
[1]University of Wisconsin - La Crosse; [2]Missouri University of Science and Technology;
[*]*(olbrichtg@mst.edu) to whom correspondence should be addressed*

Abstract. DNA methylation is an epigenetic modification that has the ability to alter gene expression without any change in the DNA sequence. DNA methylation occurs when a methyl chemical group attaches to cytosine bases on the DNA sequence. In mammals, DNA methylation primarily occurs at CG sites, when a cytosine is followed by a guanine in the DNA sequence. In plants, DNA methylation can also occur in other cytosine sequences, such as when a cytosine is not followed directly by a guanine. Many of the statistical methods that have been developed to estimate methylation levels and test differential methylation in whole-genome bisulfite sequencing studies incorporate the observed correlation between methylation levels of neighboring cytosine sites. However, most of these methods have been applied to human studies, where only CG sites are investigated. In this study, we focus on plant studies and show that the correlation between methylation levels at neighboring sites depends on the DNA sequence immediately following the cytosine. We investigate the importance of accounting for these differences in the correlation structure by comparing the performance of three existing methods (MethylSig, MAGI, and M^3D) in plants.

1 Introduction

DNA methylation is an epigenetic modification in which methyl groups selectively bind to cytosines throughout the genome (Choy *et al.*, 2010). Changes in DNA methylation can lead to phenotypic differences between genetically identical subjects, different tissues in the same subject, different cells in the same tissue in the same subject, or even same cells in the same tissues in the same subject over time (Melka *et al.*, 2015; Roadmap Epigenetics Consortium *et al.*, 2015). Statistical methods have recently been developed to test for and model differential methylation in mammalian species, but methods for plant species have not yet been developed.

1.1 Methylation in Multi-Context Organisms

In mammals, methylation mainly occurs when a cytosine is followed by a guanine on the DNA sequence (CG sites) either through maintenance or through *de novo* methylation (Cao *et al.*, 2003). On the other hand, methylation can occur in plants at CG, CHG, or CHH sites, where H is any base other than guanine that follows a cytosine on the DNA sequence (Chan *et al.*, 2005). These different sequences where methylation can occur are referred to as

sequence "contexts". Methylation levels at different contexts are created and maintained by different mechanisms (Qian *et al.*, 2012; Zhong *et al.*, 2012). Many of the current statistical methods for analyzing DNA methylation data account for correlation in methylation levels between neighboring CG sites (Park *et al.*, 2014; Baumann and Doerge, 2014; Mayo *et al.*, 2014). However, the correlation in methylation levels appears to be context dependent in plants. Figure 1 shows the average decay in correlation between a reference cytosine's methylation level and that of downstream cytosines for the unreplicated *Arabidopsis thaliana* data presented in Lister *et al.* (2008) for three separate minimum sequencing depths. In this study, methylation levels in wild-type *columbia-0* and *met1* mutants, which are deficient in CG methylation maintenance, were compared. The CG correlation levels shown in Figure 1 differ greatly between the wild-type and mutant plants, while only minor differences are evident in CHG or CHH methylation correlations. Average starting correlations also differ between the contexts, though similar decay patterns are evident in each of the contexts and treatments.

1.2 Methods Under Comparison

Next-generation sequencing (NGS) technologies combined with bisulfite sequencing methods are able to produce full genome methylation profiles at single-base resolution. After sequencing and alignment, methylation data are typically summarized at each cytosine using the number of methylated reads observed, the number of total reads (sequencing depth) observed, methylation context, chromosome, chromosome position, condition (e.g., disease or healthy) and subject. Three recently developed methods for detecting differential methylation between conditions will be applied to NGS plant data and compared. Though each method accounts for spatial correlation between cytosines in some way, cytosine context is typically ignored or the method is applied to the CG context only. Each method additionally treats each non-overlapping gene region independently.

MethylSig. The MethylSig method (Park *et al.*, 2014) tests for differential methylation at individual cytosines using a beta-binomial model for the number of methylated reads at a specific site. The model allows the sequencing depth and probability of methylation to vary between individuals and employs a likelihood ratio test to find differences in methylation levels between conditions. Local information for estimation of the beta-binomial parameters is incorporated through the use of triangular Kernel weights to obtain local maximum likelihood estimators.

MAGI. The Methylation Analysis using Genome Information (MAGI) approach (Baumann and Doerge, 2014) defines testing regions based on *a priori* genome annotation and implicitly assumes methylation homogeneity within these regions, but otherwise does not adjust for spatial correlations between cytosines. The tests performed include two variants of Fisher's Exact Test (FET) in unreplicated experiments or two variants of logistic regression in experiments with biological replication. The first variant of each type ($MAGI_c$) involves first

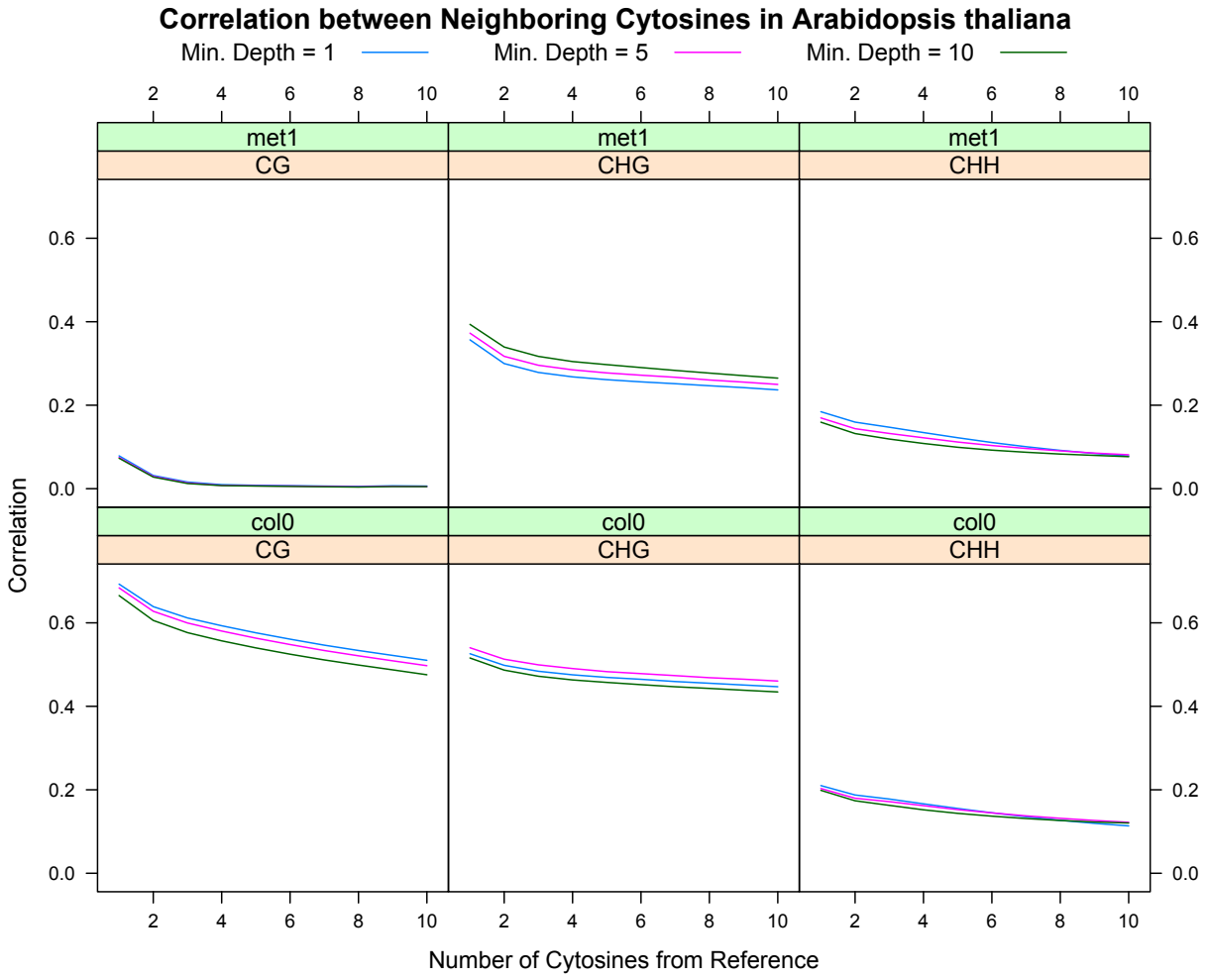

Figure 1: Average empirical Pearson correlation in methylation levels between a reference cytosine and downstream cytosines for two strains of *Arabidopsis thaliana* (col0 and met1) under different contexts and minimum sequencing depths. Differences in average correlation levels between treatments and between contexts indicate a need to model correlation as well as context in differential methylation studies in multi-context organisms.

testing differences between methylation levels at individual cytosines then summarizing these tests within each annotated region, similar to the sliding window approach proposed by Lister *et al.* (2009). The second variant ($MAGI_g$) first classifies each cytosine (within individual) as either methylated or not methylated based on an *a priori* threshold, then performs a single FET or logistic regression on the resulting data within each region, assuming the resulting data are binomially distributed. $MAGI_g$ is presented as the more powerful alternative and will be used for method comparisons.

M3D. The Maximum Mean Methylation Discrepancy (M^3D) method (Mayo *et al.*, 2014)

quantifies changes in the shapes of methylation profiles within local testing regions by applying a machine learning technique (MMD, (Gretton *et al.*, 2007, 2012)) to test homogeneity in underlying methylation generating distributions. A radial basis function (RBF) kernel function is employed to construct the MMD between data sets in each testing region which is then adjusted based on changes in coverage profiles. The resulting M^3D statistics are then compared to a null distribution of observed M^3D statistics between replicate pairs.

The primary goal of this paper is to investigate the performance of current differential methylation analysis techniques when used in multi-context organisms such as plants. The three methods under comparison (MethylSig, MAGI, and M3D) were applied to an *Arabidopsis thaliana* seedlings dataset under a variety of simulated conditions. The data and simulation procedures are described in Section 2. Simulation results are presented in Section 3, followed by a discussion of the use of current methods on plant data in Section 4.

2 Data and Simulation Study

2.1 *Arabidopsis thaliana* Data Source

Methylation data for three biological replicates of bisulfite-sequenced columbia-0 seedlings *Arabidopsis thaliana* (Qian *et al.*, 2012; Zhong *et al.*, 2012; Law *et al.*, 2013) were accessed from NGSmethDB, an online, single-base resolution methylome browser and repository (Geisen *et al.*, 2014). NGSmethDB provides several levels of depth-filtered data, and a minimum depth of 10 reads per cytosine was chosen to allow for more sensitivity in testing for differential methylation (Baumann and Doerge, 2014).

2.2 Simulated Data

Data were simulated following the approach in Mayo *et al.* (2014) with modifications to leverage natural correlation patterns and methylation levels. For simplicity, 1000 gene regions (Lamesch *et al.*, 2011) were randomly selected from chromosome 1 of *Arabidopsis thaliana*. Of these, 200 were randomly selected to apply differential methylation changes. Using three biological replicates as a control group, a treatment group was simulated by first adding (or subtracting) random Poisson($\lambda = 1$) noise to the number of reads at each cytosine within each replicate. Uniform (from -0.1 to 0.1) random noise was added to cytosine methylation levels L_i, defined as the ratio of methylation reads to total reads mapped to a particular cytosine. Methylation levels of cytosines were adjusted within the 200 selected genes per Mayo *et al.* (2014). New treatment group methylation levels L_i^{trt} are simulated using control group methylation levels through $L_i^{trt} = (1-\alpha) * L_i^{control} + \alpha$ or $L_i^{trt} = (1-\alpha) * L_i^{control}$, for hyper- and hypo-methylation respectively, where $\alpha \in [0, 1]$ is used to control the degree of differential methylation between the control and treatment groups.

Hyper- and hypo-methylation of cytosines was determined by first calculating mean cytosine methylation levels within each gene in the control group and hypo-methylating the cytosines within the corresponding gene in the treatment group if the gene mean exceeded 0.5 (otherwise cytosines were hyper-methylated). Differential methylation was simulated for α values between 0.2 and 0.8, and under several different context settings (CG-only, CHG-only, CHH-only, and context-blind methylation). For simulations investigating context-specific methylation, mean cytosine methylation levels were calculated within each gene only for the specified context, and methylation changes were only applied to cytosines of those contexts. This simulation strategy allows the current methods to be tested both when differential methylation in a region occurs at sites of all contexts and also when it changes for different contexts.

2.3 Method Evaluation

Two different analysis approaches were implemented on each of the 1000 independent gene regions for each of the three statistical methods under comparison to determine the importance of accounting for sequence context in detecting differential methylation. Context-independent analyses included all cytosines regardless of context when implementing the three methods. The context-specific analyses involved analyzing each context separately for the three methods by only including the cytosines of one specific context within a given analysis. The false discovery rate was controlled at $\alpha = 0.05$ for all analyses (Benjamini and Hochberg, 1995).

Performance of the methods was evaluated by comparing the true positive rates. For $MAGI_g$ and M^3D, the true positive rate is defined as the proportion of the 200 truly differentially methylated regions that were identified as differentially methylated by the method. For MethylSig, the calculation is done on a site level rather than a region level. Comparing the performance of the methods for the context-independent analysis will give insight into which of the methods perform best using the methods as is without considering sequence context. Comparing the performance of the context-independent and context-specific analysis will determine whether applying the methods in a context-specific way improves the ability to detect sites or regions that are truly differentially methylated.

3 Results

3.1 Context-Independent Analyses

Figure 2 shows the comparison of true positive rates (TPR) between the three methods for varying degrees of differential methylation (α values) when the context-independent analysis is employed. The four plots within Figure 2 show the results corresponding to where

the true underlying differential methylation occurs (CG only, CHG only, CHH only, or all contexts). When differential methylation occurred at all contexts within a region, $MAGI_g$ and M^3D performed similarly well with a TPR over 0.90 regardless of the size of differences in methylation between the groups. MethylSig had a consistently lower TPR. When only sites of one context were differentially methylated within a region, the three methods varied in their performance. MethylSig performed best when only CG sites were differentially methylated, M^3D performed best when only CHG sites were differentially methylated, and $MAGI_g$ performed best when only CHH sites were differentially methylated. In general, the TPR were lower for the context-independent analysis when differential methylation only occurred at specific contexts. This indicates that a direct application of the three methods without considering context can lead to lower TPR if differential methylation is truly context-dependent.

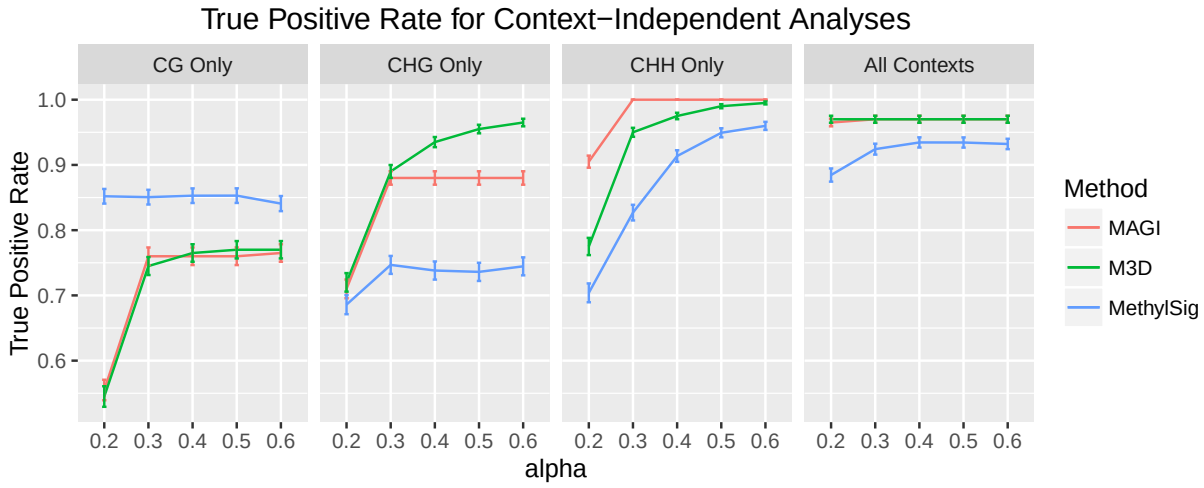

Figure 2: True positive rates verses the α level for controlling the degree of differential methylation for each of the three different methods ($MAGI_g$ - red, M^3D - green, and MethylSig - blue) using the context-independent analysis. Standard error bars for the proportions are included. The four plots show results for differential methylation at only specific contexts within a gene (CG only, CHG only, and CHH only) as well as when all cytosines within a gene are differentially methylated (All Contexts).

3.2 Context-Specific vs. Context-Independent Analyses

Figure 3 shows the comparison of the true positive rates (TPR) between context-specific and context-independent analyses for varying degrees of differential methylation (α values) in each of the three methods (shown in separate rows). The four plots (in columns) for each method show the results corresponding to where the true underlying differential methylation occurs (CG only, CHG only, CHH only, or all contexts). When differential methylation

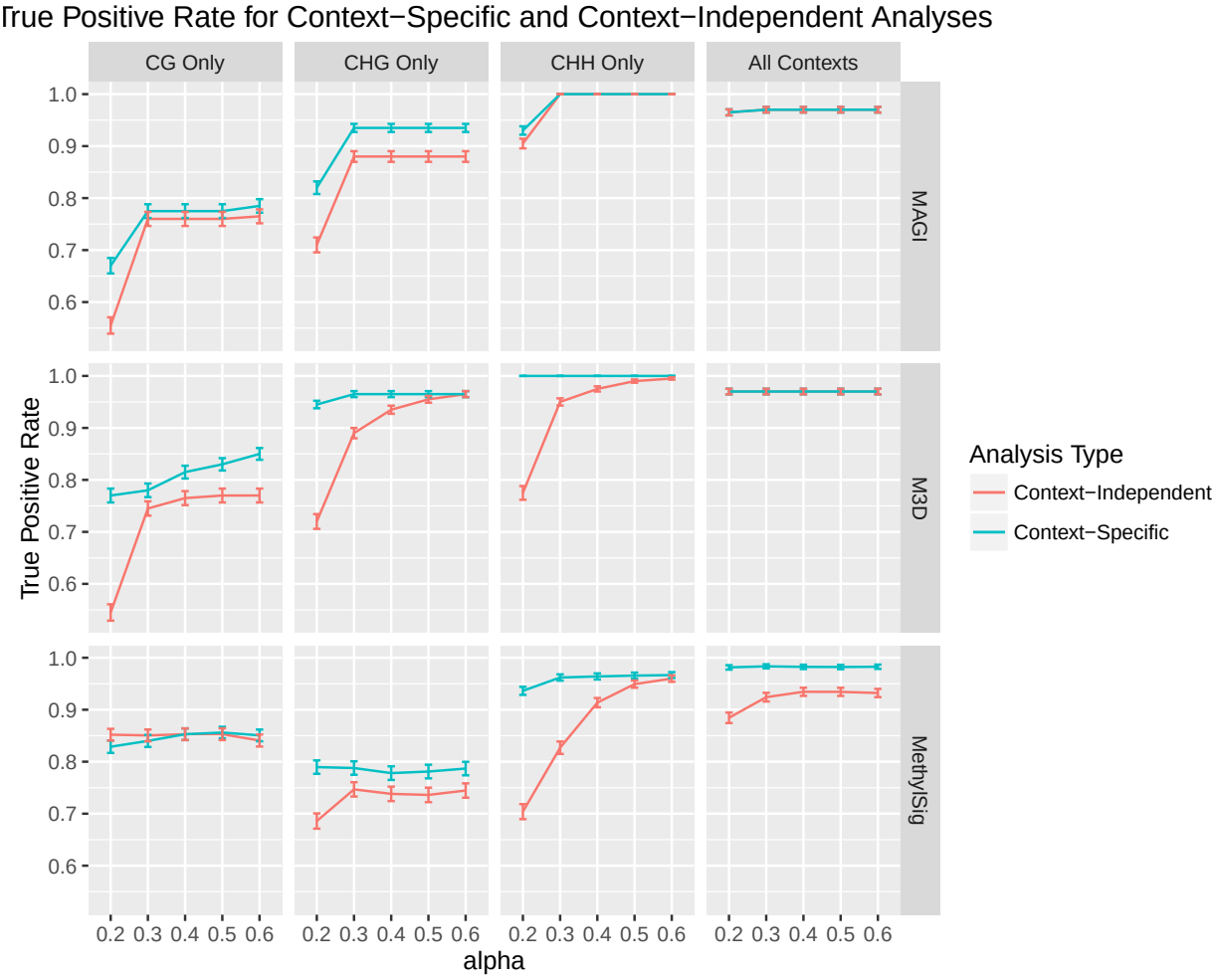

Figure 3: True positive rates verses the α level for controlling the degree of differential methylation for the context-specific (blue) and context-independent (red) analyses. Standard error bars for the proportions are included. Results for the three methods (MethylSig, M3D, and MAGI) are given in different rows. The four plots for each method show results for differential methylation at only specific contexts within a gene (CG only, CHG only, and CHH only) as well as when all cytosines within a gene are differentially methylated (All Contexts).

occurred at all contexts within a region, the context-specific and context-independent analyses performed identically for $MAGI_g$ and M^3D. The context-specific analysis showed an improvement over the context-independent analysis in MethylSig in this scenario with TPR over 0.90 similar to $MAGI_g$ and M^3D. Since MethylSig tests each site individually rather than at the gene level, this improvement may be due to inherent differences in correlation between methylation levels of different contexts that affect the local estimates. These may affect testing at individual sites more than the region level tests used in $MAGI_g$ and M^3D.

When differential methylation occurred only at a specific context within a region (CG only, CHG only, or CHH only), the context-specific analyses outperformed the context-independent analysis across almost all settings for all three methods. The largest improvements are seen when the average methylation differences are small, but notable differences are observed even at large average methylation differences in some settings (e.g., $MAGI_g$ results for CHG only). In the few cases in which the context-specific analyses did not clearly outperform the context-independent analysis (MethylSig and $MAGI_g$ - CG only, $MAGI_g$ - CHH only), the TPR were nearly identical.

4 Discussion

The focus of this work was to evaluate the performance of existing methods for differential methylation detection in multi-context organisms such as plants. Since the correlation between methylation levels of neighboring cytosines depends on the sequence context, it is important to consider the context in differential methylation analyses; however current statistical methods (Baumann and Doerge, 2014; Mayo *et al.*, 2014; Park *et al.*, 2014) fail to do so. Through simulation studies, we investigated the performance of three existing methods on plant data using two approaches: combining all contexts (context-independent) into one analysis and running separate (context-specific) analyses for each context. The true underlying DNA methylation differential status was simulated under conditions of both context dependence and independence to thoroughly test the methods.

Results of the context-independent analyses indicate that as long as differential methylation is not context dependent, the M^3D and $MAGI_g$ methods work well and can be applied without modification. The underperformance of MethylSig may be indicative that testing at a gene/region level is more powerful than at individual sites, as noted in the MAGI manuscript (Baumann and Doerge, 2014). If differential methylation occurs in a context-specific manner (CG only, CHG only, or CHH only), there is room for improvement among the three current methods tested and there is not a single method that works best across all three sequence contexts.

Results comparing the context-specific and context-independent analyses indicate that running a separate analysis for each context is important when working with organisms like plants where methylation can occur in multiple sequence contexts. The TPR is generally much higher and at worst has a similar TPR as the analysis when all contexts are combined into one analysis (context-independent). These differences are likely due to the correlation differences between sites of different contexts. Since it may not be known *a priori* whether all contexts or only a specific context is differentially methylated within a gene/region, the separate (context-specific) analyses for each context are recommended. By testing separately, the analyses can also provide information about which of the contexts are differentially methylated in a region, which the context-independent analysis can not provide.

Several model modifications are being considered to improve performance when testing for differential methylation in organisms where methylation occurs in multiple sequence contexts. In plants, DRM and CMT3 methyltransferase classes each maintain CHG and CHH contexts (Chan *et al.*, 2005), so considering correlation between different methylation contexts could increase statistical power in these models. To this end, we are investigating the incorporation of an additional kernel element into $MAGI_g$, MethylSig, or M^3D to combine context profiles under different correlation structures. In addition, the effects of sequencing depth are currently being considered. Although we expect a lower minimum depth to reduce statistical power in each of the three methods, the effects are expected to less detrimental to $MAGI_g$ due to $MAGI_g$'s thresholding protocol when compared to the methylation-level smoothing techniques of MethylSig or M^3D.

5 Summary

In this work, the importance of sequence context in detecting differential methylation in next generation bisulfite sequencing data was investigated. Sequence context is important for organisms like plants where methylation can occur at CG, CHG, and CHH contexts, yet current statistical literature does not address this issue. Three current statistical methods were applied to an *Arabidopsis thaliana* data set that was simulated based on real data. Two different analyses were employed for each method, a context-independent analysis where all cytosines were analyzed together regardless of context and context-specific analysis where cytosines of different contexts were analyzed separately. It was shown that it is more powerful to analyze each context separately, which can also provide additional information about the type of methylation differences within a region. Improvements to current methods to address context-specific issues are currently being investigated.

Acknowledgments

We would like to thank the University of Missouri and University of Wisconsin-La Crosse for funding and resources.

References

Baumann, D. and Doerge, R. (2014). MAGI: Methylation analysis using genome information. *Epigenetics*, **9**(5), 698–703.

Benjamini, Y. and Hochberg, Y. (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.

Cao, X., Aufsatz, W., Zilberman, D., Mette, M., Huang, M., Matzke, M., and Jacobsen, S. (2003). Role of the drm and cmt3 methyltransferases in RNA-directed DNA methylation. *Current Biology*, **13**(24), 2212–2217.

Chan, S., Hendersen, I., and Jacobsen, S. (2005). Gardening the genome: DNA methylation in *Arabidopsis thaliana*. *Nature Reviews Genetics*, **6**, 351–360.

Choy, M.K., Movassagh, M., Goh, H.G., Bennett, M., Down, T., Foo, R. (2003). Genome-wide conserved consensus transcription factor binding motifs are hypermethylated. *BMC Genomics*, **11**(1), 519.

Geisen, S., Barturen, G., Alganza, A., Hackenberg, M., and Oliver, J. (2014). NGSmethDB: an updated genome resource for high quality, single-cytosine resolution methylomes. *Nucleic Acids Research*, **42**(D1), D53–D59.

Gretton, A., Borgwardt, K. M., Rasch, M., Scholkopf, B., and Smola, A. J. (2007). A kernel method for the two-sample-problem. *Advances in neural information processing systems*, **19**, 513.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Scholkopf, B., , and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, **13**, 723–773.

Lamesch, P., Berardini, T., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D., Garcia-Hernandez, M., Karthikeyn, A., Lee, C., Nelson, W.D. ND Ploetz, L., Singh, S., Wensel, A., and Huala, E. (2011). The *Arabidopsis* information resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, **40**(D1), D1202–D1210.

Law, J., Du, J., Hale, C., Feng, S., Krajewski, K., Palanca, A., Strahl, B., Patel, D., and Jacobsen, S. (2013). Polymerase iv occupancy at RNA-directed DNA methylation sites requires shh1. *Nature*, **498**(7454), 385–389.

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

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

Mayo, T., Schweikert, G., and Sanguinetti, G. (2014). M3D: a kernel-based test for spatially correlated changes in methylation profiles. *Bioinformatics*.

Melka, M.G., Castellani, C.A., OReilly, R., Singh, S.M. (2015). Insights into the origin of DNA methylation differences between monozygotic twins discordant for schizophrenia. *J. Mol. Psychiatry*, **3**(7), 7.

Park, Y., Figueroa, M., Rozek, L., and Sartor, M. (2014). MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics*, **30**(17), 2414–2422.

Roadmap Epigenetics Consortium *et al.*(2015). Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317-330.

Qian, W., Miki, D., Tang, K., Liu, R., and Zhu, J. (2012). A histone acetyltransferase regulates active DNA demethylation in *Arabidopsis*. *Science*, **336**(6087), 1445–1448.

Zhong, X., Hale, C., Law, J., Johnson, L., Feng, S., Tu, A., and Jacobsen, S. (2012). Ddr complex facilitates global association of RNA polymerase v to promoters and evolutionarily yound transposons. *Nature Structural and Molecular Biology*, **19**(9), 870–875.