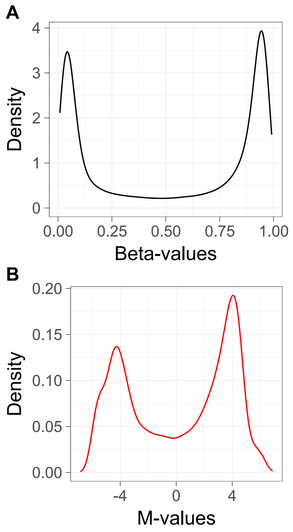
Statistical approaches to improve detection of DNA methylation

Benjamin K Johnson[1], Zachary B Madaj[1], Timothy Triche Jr.[2]

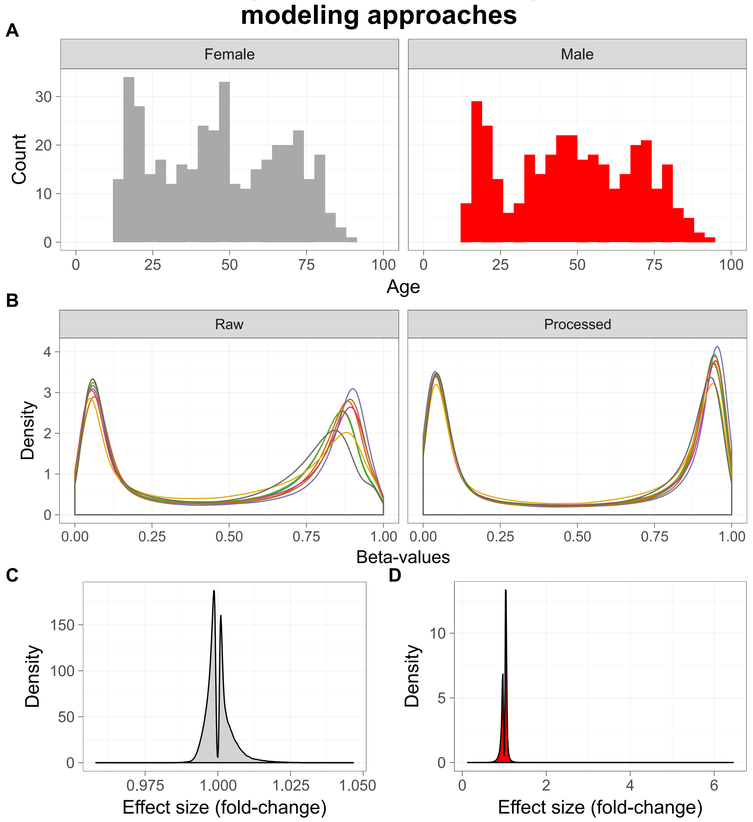
[1]Bioinformatics and Biostatistics Core, Van Andel Research Institute, Grand Rapids, MI, USA, [2]Center for Epigenetics, Van Andel Research Institute, Grand Rapids, MI, USA

DNA methylation arrays provide targeted epigenetic profiling and are typically modeled on Beta- or M-values

Epigenetic marks play critical roles in development, disease, and aging. For population-scale studies and exploratory analyses of DNA methylation differences/associations, short-read sequencing approaches can be cost prohibitive. Thus, hybridization-based bead arrays, in particular Illumina's HumanMethylation450 and HumanMethylationEPIC platforms, have become the dominant source of population-scale DNA methylation data. Data from over 120,000 samples are now publicly available, and many more have been assayed as part of pharmaceutical characterization efforts ranging from autoimmune to neurodegenerative diseases. Previous work has shown that beta regression outperforms competing methods in terms of sensitivity to detect differences in methylation[1]. **We extend prior work[1] to evaluate a range of modeling approaches on both A) beta-values (B-values: methylated/unmethylated + methylated loci) and B) M-values (logit2 transformed B-values).**

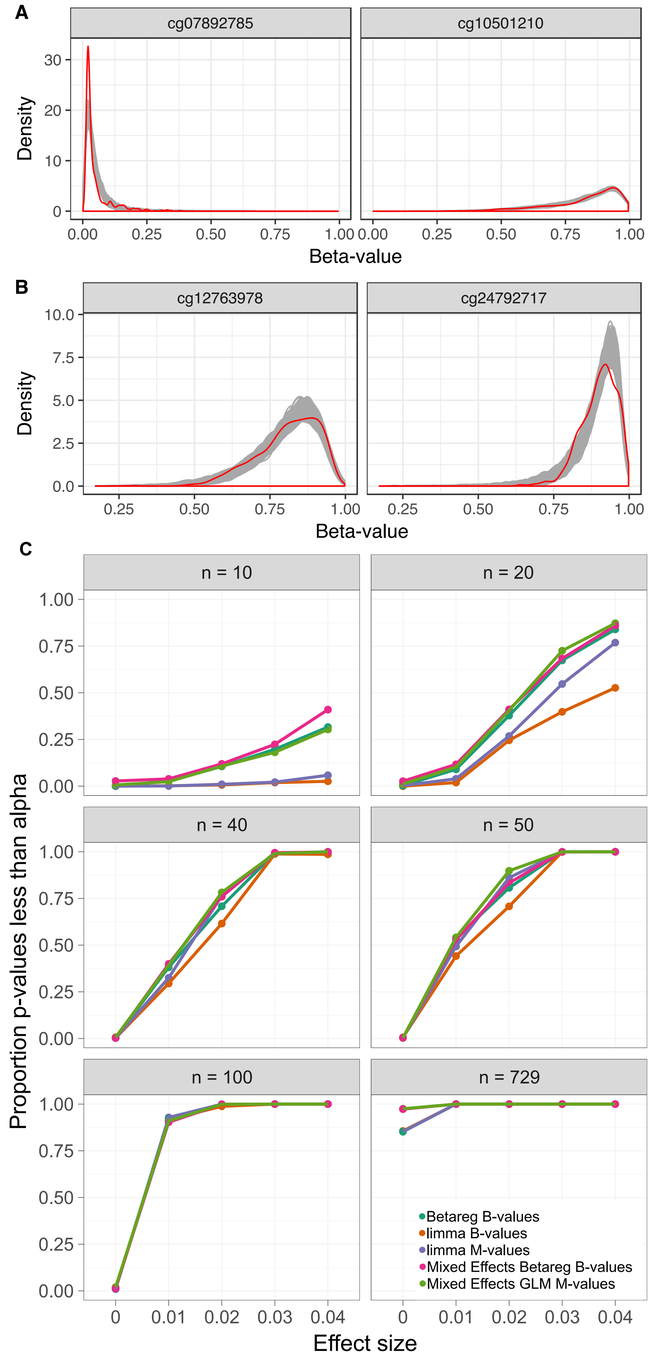


Population-scale differential methylation analysis can detect a range of effect sizes using standard modeling approaches



A population-scale study focused on differences in methylation during aging was downloaded from GEO (GSE87571; N = 732). **A)** The distribution of females and males were similar across age groups. **B)** Data were pre-processed using SeSAMe[2] "best-practices" to correct for various known sources of bias[3]. Standard modeling approaches (limma fitted M-values) were used to establish effect sizes for **C)** age and **D)** gender to be used in subsequent simulations.

Simulation of CpG loci across a range of effect and sample sizes shows mixed-effects and beta regression provide increased sensitivity to detect differential methylation



Candidate CpG loci were selected for simulation based on standard modeling approaches (limma fitted M-values) and a nominal p-value < 0.05 for both age and gender (data not shown). Individual loci were randomly selected from the candidates based on effect size. Simulated beta-values were derived using fitted model objects and each locus was simulated 1000 times. **A)** Examples of age and **B)** gender loci used for simulation. Observed beta-values are in red with simulated beta-values in gray. **C)** The simulated beta-values from age-derived effect sizes were fitted using limma, beta regression (betareg), and mixed modeling (ME) approaches on age and gender (slide was fitted as a random effect for mixed models). Significance testing was performed using likelihood ratio tests or using moderated t-statistics. Alpha = 0.05 / 1000.

Models fit to the full dataset shows mixed-effects and beta regression provide increased sensitivity to detect differential methylation

Modeling Approach	Age significant probes (q < 0.05)	Gender significant probes (q < 0.05)	Percent of total significant probes (Age)	Percent of total significant probes (Gender)
Limma (M-values)	224,624	76,453	54.2%	18.4%
Limma (B-values)	216,884	75,868	52.3%	18.3%
ME GLM (M-values)	**232,989**	**79,938**	**56.3%**	**19.3%**
ME Betareg (B-values)	**228,953**	**80,789**	**55.3%**	**19.5%**
Betareg (B-values)	**228,507**	**87,943**	**55.1%**	**21.2%**

Models were fitted to 414,154 probes with age, gender, and cellular composition ($CD4^+$ T-cells, $CD8^+$ T-cells, NK, monocytes, granulocytes, B-cells) as fixed effects. Slide was fitted either as a fixed effect in limma and betareg or as a random effect in the generalized linear mixed models (ME GLM and ME Betareg). Significance testing was performed using likelihood ratio tests for mixed models and betareg, while moderated t-statistics were used for limma. Modeling results were filtered for loci that did not converge prior to false discovery rate correction. All modeling approaches were subjected to false-discovery rate correction (Benjamini-Hochberg) and significance was determined as q < 0.05.

Conclusions

Of the methods explored, generalized linear mixed modeling approaches and beta regression showed the greatest sensitivity to detect methylation differences in both simulations and when fitted to the full data set. Simulation studies across a range of effect and sample sizes demonstrate that below 20 samples, overall power to detect small methylation differences (e.g. age effects) is diminished. Following the simulations, the models fitted to the entire data set (n=729; 414,154 probes) resulted in similar overall numbers of significant probes with a modest 1-2% increase in sensitivity.

Acknowledgements

We would like to thank the authors and individuals that contributed to the publicly available dataset used for this analysis (GEO: GSE87571). The Van Andel Research Institute High-Performance Computing/Scientific Computing group for providing support and resources for the Bioinformatics and Biostatistics Core computing resources. Megan Bowman, Ph.D. for outstanding discussions and critical reading of this work.

Citations and Software Versions

1. Triche TJ, Laird PW, Siegmund KD, 2016, Beta regression improves the detection of differential methylation for epigenetic epidemiology. *BioRxiv.* doi: https://doi.org/10.1101/054643.
2. SeSAMe: Sensible Step-wise Analysis of Methylation Data - https://github.com/zwdzwd/sesame
3. Zhou W, Laird PW, Shen H, 2017, Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Research.* doi: https://doi.org/10.1093/nar/gkw967
4. Data analysis packages used: R (v 3.5), plyr (v 1.8.4), dplyr (v 0.7.5), ggplot2 (v 2.2.1), RColorBrewer (v 1.1-2), limma (v 3.37.2), betareg (v 3.1-0), glmmTMB (v 0.2.1.0), minfi (v 1.26.2), lmtest (v 0.9-36)