Beta regression improves the detection of differential DNA methylation for epigenetic epidemiology

Timothy J. Triche, Jr.[1], Peter W. Laird[3,4,5,6], Kimberly D. Siegmund[2*§]

[1] Jane Anne Nohl Division of Hematology, USC/Norris Comprehensive Cancer Center,

[2] Department of Preventive Medicine, Keck School of Medicine of USC,

[3] USC Epigenome Center, USC/Norris Comprehensive Cancer Center,

[4] Department of Biochemistry and Molecular Biology, USC/Norris Comprehensive Cancer Center,

[5] Department of Surgery, USC/Norris Comprehensive Cancer Center, Los Angeles, CA 90089-9176, USA.

[6] Present address: Center for Epigenetics, Van Andel Research Institute, Grand Rapids, MI 49503 USA.

[§]Corresponding author

Email addresses:

TJT: ttriche@usc.edu

PWL: plaird@usc.edu

KDS: kims@usc.edu

Abstract

Background

DNA methylation is the most readily assayed epigenetic mark, possessing confirmed relationships with gene expression, imprinting, and chromatin accessibility. Given the increasingly widespread use of DNA methylation microarrays in population-scale epidemiological applications, we sought to determine which methods provided the greatest statistical power to reproducibly detect differences in DNA methylation across various conditions, using publicly available data sets on tissue type and aging.

Results

Beta regression, as proposed originally by Ferrari and Cribari-Neto, yielded more validated hits in each of our comparisons than any other method under consideration, both in a regression setting and in comparisons to two-group tests such as the Wilcoxon-Mann-Whitney, Student's t, and Welch's t tests. In large cohorts of whole blood samples, we corrected for compositional differences and batch effects, and found that marginal likelihood ratio tests from beta regression models uniformly dominate popular alternatives based on linear models. The superior sensitivity and specificity exhibited by beta regression in epidemiologically relevant cohort sizes corresponded to approximately a 2% increase in sensitivity at the same specificity when compared to linear models fitted on raw beta values (proportion of signal intensity due to the methylated allele), M-values, or rank quantile normalized values.

Conclusions

Investigators should consider beta regression to maximize statistical power in studies of DNA methylation using microarrays. At epidemiologically relevant sample sizes, with typical quality control procedures (compositional and batch effect correction), cross-cohort agreement uniformly favors beta regression over popular alternatives.

Keywords: DNA methylation, epigenetics, epidemiology, methodology, detection

Introduction

Epigenetic information – modifications to chromatin and its constituent DNA, as opposed to genetic information encoded in the DNA sequence itself – is now widely recognized as playing an important role in development, degeneration, and disease. In vertebrate genomes the most readily accessible epigenetic mark is cytosine methylation at CG dinucleotides, associated with transcriptional repression when found in promoter regions. In population-scale studies of multifactorial outcomes, improved sensitivity to detect true associations between DNA methylation changes and exposures or outcomes is of great interest. We therefore sought to explore the trade-offs involved in faithfully modelling the proportional measurements via maximum likelihood beta regression [1], vs. least-squares modelling after logit transformation [2], or a N(0,1)-quantile normalization [3], or no transformation.

Typical approaches to modelling DNA methylation at individual cytosines include ignoring the heteroskedasticity of proportional (0-1) measurements; N(0,1)-quantile normalizing the measurements [3]; logit-transforming the measurements and working on a fold-change scale [2], or discretizing proportions into "states" (for example, regions of unmethylated, variably methylated, and highly methylated cytosines) which are presented as surrogates for biologically relevant changes [4]. The latter discretization typically assumes extensive coverage (as with whole-genome bisulfite sequencing), making strong assumptions regarding local correlation on the genome. By contrast, we focus on regression approaches useful for sensitive modelling of associations at individual targeted cytosines, as a function of exposures, which could be continuous (e.g. age) and typically require sample sizes that would be cost-prohibitive for whole-genome bisulfite sequencing. We evaluated Beta regression and common approaches to linear models, with or without transformation, for various sample sizes and designs.

Linear models are quite flexible and can be solved with relative ease even for highly underdetermined data matrices (e.g. via penalized least-squares methods). Similarly, any generalized linear model can be iteratively solved to find constrained maximum likelihood estimates of predictor effects on a response, with certain distributional assumptions. Beta regression, as proposed by [5, 6], provides a regression framework that will be familiar to practitioners comfortable with generalized linear models. However, the model allows for the coupled mean and variance terms to both depend on covariates. It is not feasible, with a single uniformly applied transformation across responses, to simultaneously normalize the proportional measurements and completely decouple its variance from its mean (although transformations which render the data approximately homoskedastic, e.g. [2], have proven useful). Thus, directly modelling both terms as a function of the data is an attractive approach.

In addition, changes in the variability of DNA methylation are also recognized as predictors of interest for many diseases [7] and more recently, as nearly inevitable side effects of aging [8]. A dispersion (variability) parameter is integral to the beta regression model, permitting either or both of the mean and dispersion to depend upon covariates. Moreover, one of the most popular biospecimens for population-scale epigenomics work – whole blood – is in fact a complicated mixture of myeloid, lymphoid, and dendritic cells [9], each with characteristic DNA methylation marks [4]. While cell sorting is a useful method to reduce this ambiguity, it is not always feasible to separate fractions (e.g. [10]). Greater sensitivity is thus useful for detecting phenotypic associations of interest, which may only be present in a fraction of cell types. At the same time, regression analysis allows us to model the influence of other factors, such as fraction of cell type (e.g. [11, 12]) , which may influence the estimated associations. Therefore, we employed several publicly available data sets to explore various approaches to modelling variation in DNA methylation.

First, we evaluated Beta regression in the context of dichotomous exposure. We used the myeloid/lymphoid classification of sorted blood cell fractions to evaluate common two-group tests (Student's t test, the Welch t test, Mann-Whitney-Wilcoxon) on variously transformed or untransformed data, and to compare the sensitivity of these to that of likelihood ratio tests based on the Beta regression model, with either a shared estimate of variability (the "Beta-T" test) or a fitted per-group estimate of variability (the "Beta-Welch" test).

Second, we analyzed two large studies associating aging, as a continuous exposure, with DNA methylation levels in whole blood [13][14]. These large studies allowed us to use one (HuGeF) as a discovery set and the other (Hannum) as a test set when gauging sensitivity of different methods. As we know blood to be a complex mixture of different cell types varying with age, we used the compositional estimation procedure from Houseman [15] in addition to factors for batch, plate, gender, and ethnicity (where the latter was available) to reasonably approximate the steps which would be expected in an epidemiologically relevant cohort study. Sensitivity and specificity was estimated as the proportion of "true positives" (replicated hits) at a given specificity (putative error level in the discovery cohort) and plotted in Figure 1.

Methods

Models for differential DNA methylation

The 'betareg' R package (version 3.0.4), as described in [1], was employed for the Beta regression fits. Instead of the usual formulation, Beta(a,b), with a, b > 0, mean μ =a/(a+b) and variance σ^2 = μ (1- μ)/(a+b+1) as a function of the mean, the reparameterized distribution Beta(μ,φ) is used, where φ = a+b is a precision parameter, independent of the mean. For small φ, the variance is large, and for large φ, the variance is small. Briefly, given link functions $g_1(\mu)$ and $g_2(\varphi)$, we model the mean μ and precision φ as a function of predictor matrices X and Z, respectively, to find maximum likelihood estimates of the coefficients β and γ. Assume the Beta value for locus *j* from individual *i* is y_{ij} ~ Beta(μ_j,φ_j). For simplicity, we drop the index *j*. Then,

$$g_1(\mu) = \eta_1 = X^T \beta$$
$$g_2(\varphi) = \eta_2 = Z^T \gamma$$

with the links for $g_1(\mu)$ and $g_2(\varphi)$ being logit(μ) and log(φ), respectively. The log-likelihood is then expressed for responses of N subjects as

$$l(\mu, \varphi \vee Y) = \sum_{i=1}^{N} l\left(\mu, \varphi \vee y_{ij}\right)$$

such that each individual observation contributes to the loglikelihood the quantity

$$l\left(\mu, \varphi \vee y_{ij}\right) = log\Gamma(\varphi) - log\Gamma(\mu\varphi) - log\Gamma\left((1-\mu)\varphi\right) + (\mu\varphi - 1)log\left(y_{ij}\right) + \left(\left((1-\mu)\varphi\right) - 1\right) log\left(1 - y_{ij}\right),$$

where Γ is Euler's Gamma function (Γ(n) = (n-1)!).

We implemented two changes that we found necessary to improve speed and stability. First, we employed the "squeezing" technique suggested by [6] to eliminate 0 and 1 values without altering the rank order of observations. To squeeze the data we shifted the values around zero and multiplied by 0.99 (= (Beta – 0.5)*0.99 + 0.5). Second, we used an approach similar to that of 'limma' [18], with predefined design matrices for the null and alternative hypotheses, for model fitting. Fitting a 2-3 variable model to each CpG site on the HumanMethylation450 array took under 6 hours on a machine with 6 processors and 48GB of RAM, when a likelihood ratio test was conducted at each locus.

For linear regression, the 'limma' package was employed, using the marginal t-test statistics for each coefficient of interest. Models were fit using untransformed and transformed Beta values. Untransformed Beta values, the ratio of intensities M/(M+U), where M measures methylation and U lack of methylation intensities, were labelled '(B)' or 'betas'. We also considered M-values, log2(M/U), labelled '(M)', owing to their purportedly superior sensitivity, as described in [2]. Lastly, we evaluated the normalizing transformation described in [3], whereby the observed quantiles at a given locus are replaced with their corresponding values for a standard Normal distribution (i.e., N(0,1)) labelled '(NQ)'.

Error rate control

We used independent data sets to validate loci associated with age, allowing us to compare sensitivity of the different analysis approaches. For the blood lineage comparisons, a validation set was not available. For this data set, we permuted the samples, blocked by subject, showing control of the type I error through permutation.

Data sets

All analyses in this paper were applied to publicly available data. When available (in fractionated normal blood cells and the HuGeF-EPIC Italy cohort), raw IDAT files were downloaded and processed using the 'methylumi ' package to correct background fluorescence and equalize within–array dye-bias [19]. Where raw data was not available, internal control probes were used by GenomeStudio to equalize dye bias and correct background fluorescence. Probes were mapped to genomic coordinates using manufacturer annotations verified in the 'FDb.InfiniumMethylation.hg19' package, masking common SNPs, repeat regions, and sex chromosome probes. This mask is also used for the Cancer Genome Atlas project HumanMethylation450 data, and excludes probes with common SNPs at, or up to 15 bases 3' to, the interrogated locus, as well as probes with at least 15 bases proximal to the interrogated locus falling into an annotated repeat region of the genome. Regressions were fit separately per probe (excluding masked loci), and (as noted) p-values were adjusted using Benjamini & Hochberg's procedure [20] to control the FDR at 0.05.

A data set of 48 samples of sorted cell fractions (24 myeloid/ 24 lymphoid samples) from 6 adult male subjects was used to compare methods for testing differential DNA methylation in two-group designs (GEO series GSE35069) [12]. Reinius et al. (2012)[12] showed that cell type-specific differences dominated individual-specific differences in these samples, so we modelled DNA methylation as a function solely of cell type (myeloid or lymphoid). IDAT files were downloaded from (http://publications.scilifelab.se/kere_j/methylation) and processed as described above.

Two data sets, one from the HuGeF-EPIC Italy consortium (unpublished, GEO GSE51032, N=845) and one from Hannum (2013, GEO series GSE40279, N=656), were used to test for association between CpG methylation and age, a quantitative variable. We fitted models including age, gender, plate, batch, ethnicity (in Hannum only, not available for HuGeF) and estimated cell counts (monocytes, granulocytes, B-cells, CD4+ T-cells, CD8+ T-cells, CD56+ NK cells) to each dataset, using the HuGeF-EPIC Italy dataset for discovery and Hannum for validation.

Results

Table 1 presents the number of loci discovered to have differential DNA methylation in myeloid and lymphoid cell fractions, in a data set with 48 samples of sorted blood cells from 6 healthy males. The "Beta-T" and "Beta-Welch" tests were most sensitive among the methods compared, discovering 194,767 (51%) and 204,189 (54%) out of 381,552 loci. However, in these data, the marginal gain over the number of loci discovered by any other approach was modest. A large number of loci were identified by all approaches (177,446 loci, 91% of the loci discovered by "Beta-T" and 87% of those found using "Beta-Welch"). Permutations confirmed that the error rate was adequately controlled by all methods. Although there was variation in empirical error rate between the two permutations, this is likely a function of the small sample size (6 subjects). The rank order of the methods was not altered when we computed either Benjamini-Hochberg FDR-adjusted p-values or q-values for multiple comparisons correction (data not shown).

Table 1 - Autosomal loci associated with lineage in blood

	GSE35069 p_{BH}< .05 (%)	Permutations, p<0.05 (true α Set 1	Set 2
Student t: B	194,565 (50.9%)	7,059 (1.9%)	17,009 (4.5%
Student t: NQ	186,940 (49.0%)	7,226 (1.9%)	20,314 (5.3%
Student t: M	194,001 (50.8%)	6,692 (1.8%)	19,057 (5.0%
Welch t: B	193,189 (50.6%)	6,833 (1.8%)	16,874 (4.4%
Welch t: NQ	186,444 (48.9%)	7,193 (1.9%)	20,231 (5.3%
Welch t: M	192,870 (50.5%)	6,589 (1.7%)	18,902 (5.0%
Mann-Whitney	187,653 (49.2%)	6,582 (1.7%)	21,767 (5.7%

Beta-T test	194,767 (51.0%)	6,957 (1.9%)	20,844 (5.4%)
Beta-Welch	**204,189 (53.5%)**	14,521 (3.8%)	20,928 (5.4%)

Notes:
B=Beta-value
M=M-value
NQ=N(0,1)-quantile normalized

Set 1 and Set 2 are two different permutation data sets, obtained by swapping lymphoid and myeloid samples for three of six subjects, maintaining the original balance of ages (verified by personal correspondence with Dr. Kere) between the donors.

Figure 1 presents the relative sensitivity and specificity (judged by cross-cohort replication at a given p-value cutoff in the discovery cohort) for the various methods under consideration, after correcting for cellular composition (the same figure constructed without correcting for cellular composition is presented as supplementary figure 1). We note that beta regression uniformly outperforms all competing methods herein.

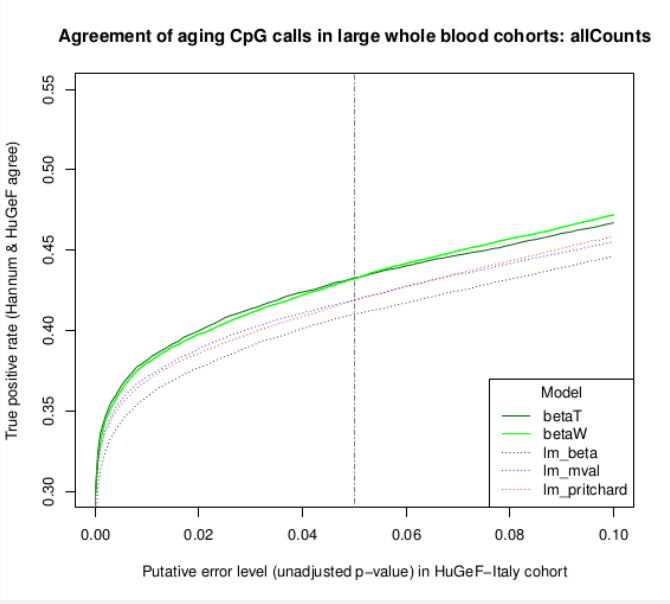

Figure 1: replication ("true positive") versus putative error rate ("false positive") for methods compared in Results.

Discussion

As a relatively young field, genome-scale investigation of DNA methylation (and indeed epigenetic marks in general) has spawned multiple approaches to assess the significance of focal differences in DNA methylation (1bp-1kbp), as well as broader concordant DNA methylation patterns, often associated with chromatin accessibility states. Both focal and broad changes may be associated with a phenotype of interest, and both are strongly associated with tissue-specific changes involved in development and disease.

Our results suggest that, for data from Illumina Infinium methylation chips, Beta regression offers the greatest sensitivity to detect these changes in DNA methylation, and for the analysis of age-related changes, this result carried through to validation in independent datasets. However, whole blood is known to be a mixture of cell types and the fractions of those cell types vary over age [16]. Therefore it is expected that a number of the validated loci may not have differences in DNA methylation with age within individual cell fractions, but rather it is the changing composition of cells types with age that allow cell-type specific marks to result in different average DNA methylation levels with aging. Thus caution is needed when interpreting results from complex cell mixtures.

If, as reported for whole blood, cell type-specific aging is faster in one or more cell fractions than the majority of cells [15], one might expect the cell-type specific DNA methylation differences with age to be attenuated in the mixed cell population. We therefore corrected for the changing estimated cell counts at the single-sample level, and found that this decreased both the overall number of hits and the proportion of replicated hits across blood cohorts, suggesting that its proposed critical importance is indeed a valid concern in the analysis of studies of epigenetic epidemiology.

The data sets we selected for analysis had large sample sizes, as will be typical in an epidemiologic genome-wide DNA methylation study. In previous work, we conducted a simulation study to evaluate control of the false-discovery rate in small samples for the methods evaluated in this paper (see Supplemental file 1). These results indicated that very small (<25 samples per group) or unbalanced comparisons (splits more uneven than approximately 80%/20%) designs, could lead to loss of control of the error rate for Beta regression. This might be due to instability of the unconstrained maximum likelihood estimates in small samples. Various approaches to bias correction and reduction have been

proposed [1], but these can increase computation times by at least an order of magnitude. In order to establish control of the error rate for the (unvalidated) two-group tests analyzed in this paper, having different genomic distribution and potentially different properties than studied in our simulations, we analyzed two permutations of the samples, randomly switching the group assignments for age-balanced sets of subjects. We recommend this type of verification for error-rate control when an independent data set is not available.

As always, the investigator must decide what the goal of the experiment or study should be, and what the most appropriate tools may be to accomplish the goal. Computational resources and time frames for completion can guide such choices, but ultimately the nature of a particular study should determine the method(s) of analysis. Exploratory analyses may be better served by the speed of linear methods, which allow for rapid iteration of models fitted to millions of data points.

By contrast, an additional day of computation may not be a significant consideration for a cohort study investigating changes in DNA methylation at specific loci in normal tissues as associated with exposures, where sensitivity might well be a central consideration, and all resulting findings subject to independent orthogonal validation. Such situations may favor sensitivity over speed of computation.

Similarly, when characterizing regional changes at a given significance threshold, either for a sequential area statistic or by modelling smoothed runs of multiple test statistics, additional sensitivity may help to unearth biologically interesting patterns that are difficult to discern otherwise. Future work will investigate an omnibus test statistic for both variable and fixed effects which appears to outperform competitors, and may yield more fruitful leads than individual tests of either type. Nonetheless, the sequential nature of regional differential methylation calls requires permutation to establish a robust FDR estimate, and this permutation step itself provides additional control of the putative error rate at the regional level. Therefore, we contend that greater sensitivity at the local (individual CpG) level is of interest even if broad, subtle effects are anticipated.

Conclusions

Beta regression offers greater sensitivity for the discovery phase of genome-wide DNA methylation association studies, particularly in large sample sizes with potentially confounding per-group effects on dispersion. The Beta regression framework appears to draw its enhanced power from the natural representation of such effects, which is more challenging in a linear model due to the decoupling of the variance from the mean. The ability of the Beta regression framework to simultaneously model changes in the mean and dispersion grants a significant advantage when modelling factors (such as aging) that appear to influence both quantities (and indeed confounding factors such as cellular composition). Investigators dissecting complex epigenetic associations with continuous phenotypes should include beta regression in their tools for discovery, balancing computational convenience and conservative error rates of linear modelling approaches against the need for sensitivity and flexibility in modelling dispersion.

Conflicts of Interest

The authors declare no competing financial interests.

Acknowledgements

Research reported in this publication was supported by the National Institutes of Health: National Human Genome Research Institute Award Number R01HG006705 (KS) and National Cancer Institute Award Numbers R01 CA097346 (KS) and R01 - CA170550 (PWL). The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health. The authors would like to acknowledge Juha Kere, for IDAT files of his data; and the USC High Performance Computing Center, for computational resources.

Authors' contributions

TJT designed and conducted the analyses, and wrote the manuscript. KDS conceived of the experiment and performed a pilot study in a separate dataset. PWL supervised the analyses and offered insights regarding the biological relevance of the results.

References

1. Grün B, Kosmidis I, Zeileis A: Extended Beta regression in R: shaken, stirred, mixed, and partitioned. *Journal of Statistical Software* 2012, 48(11):1-25.

2. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM: Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 2010, 11:587.

3. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK: DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology* 2011, 12:R10.

4. Hodges E, Molaro A, Dos Santos CO, Thekkat P, Song Q, Uren PJ, Park J, Butler J, Rafii S, McCombie WR *et al*: Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol Cell* 2011, 44(1):17-28.

5. Ferrari S, Cribari-Neto F: Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics* 2004, 31:799-815.

6. Smithson M, Verkuilen J: A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol Methods* 2006, 11(1):54-71.

7. Hansen KD, Langmead B, Irizarry RA: BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology* 2012, 13(10):R83.

8. Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, Diez J, Sanchez-Mut JV, Setien F, Carmona FJ *et al*: Distinct DNA methylomes of newborns and centenarians. *Proceedings of the National Academy of Sciences of the United States of America* 2012, 109(26):10522-10527.

9. Naik SH, Perie L, Swart E, Gerlach C, van Rooij N, de Boer RJ, Schumacher TN: Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* 2013, 496(7444):229-232.

10. Beyan H, Down TA, Ramagopalan SV, Uvebrant K, Nilsson A, Holland ML, Gemma C, Giovannoni G, Boehm BO, Ebers GC *et al*: Guthrie card methylomics identifies temporally stable epialleles that are present at birth in humans. *Genome Research* 2012, 22(11):2138-2145.

11. Guintivano J, Aryee MJ, Kaminsky ZA: A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics : official journal of the DNA Methylation Society* 2013, 8(3):290-302.

12. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlen SE, Greco D, Soderhall C, Scheynius A, Kere J: Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PloS one* 2012, 7(7):e41361.

13. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y *et al*: Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* 2013, 49(2):359-367.

14. Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MP, van Eijk K, van den Berg LH, Ophoff RA: Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biology* 2012, 13(10):R97.

15. Chu M, Siegmund KD, Hao Q-L, Crooks GM, Tavaré S, Shibata D: Inferring relative numbers of human leucocyte genome replications. *British Journal of Haematology* 2008, 141(6):862-871.

16. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT: DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 2012, 13:86.

17. Nocedal J, Wright SJ: Numerical Optimization, 2nd Edition edn. Berlin, New York: Springer-Verlag; 2006.

18. Smyth GK: Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology* 2004, 3.

19. Triche TJ, Jr., Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD: Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Research* 2013, 41(7):e90.

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