Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics
& Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Analysis of DNA Methylation Epidemiological Data through a Generic Composite Statistical Framework

Ioannis Valavanis, Emmanouil Sifakis, Panagiotis Georgiadis, Sotirios Kyrtopoulos, Aristotelis A Chatziioannou*
Institute of Biology, Medicinal Chemistry & Biotechnology
National Hellenic Research Foundation,
Athens, Greece
*Corresponding author (email: achatzi@eie.gr), 48 Vassileos Constantinou Ave., Athens 11635, Greece

***Abstract*—DNA methylation events represent epigenetic heritable modifications that regulate gene expression by affecting chromatin remodeling. They are encountered more often in CpG rich promoter regions, while they do not alter the DNA sequence itself. High-volume DNA methylation profiling methods exploit microarray technologies and provide a wealth of data. This data solicits rigorous, generic, yet ad-hoc adjusted, analytical pipelines for the meaningful systems-level analysis and interpretation. In this work, the Illumina Infinium HumanMethylation450 BeadChip platform is utilized in an epidemiological cohort from Italy in an effort to correlate interesting methylation patterns with breast cancer predisposition. The composite computational framework proposed here builds upon well established, analytical techniques, employed in mRNA analysis. For analysis purposes, the log2(ratio) of the intensities of a Methylated probe (I_{Meth}) versus an UnMethylated probe ($I_{Un\text{-}Meth}$), quoted as M-value, is used. Intensity based correction of the M-signal distribution is systematically applied, based upon Intensity-related error measures from quality controls samples incorporated in each chip. Thus, batch effects are corrected, while probe-specific, intensity-related, error measures are considered too. Robust, (based on bootstrapping) statistical measures measuring biological variation at the probe level, are derived in order to propose candidate biomarkers. To this end, coefficient variation measurements of DNA methylation between controls and cases are utilized, alleviating simultaneously the impact of technical variation, and are juxtaposed to classical statistical differential analysis measures.**

Keywords-DNA methylation profiling; Epigenomics; Microarray; Intensity-based normalization; Statistical selection; Bootstrap correction

I. INTRODUCTION

DNA methylation of cytosine bases provides a layer of epigenetic modification in many eukaryotes with important implications for healthy and disease physiology [1]. In recent years, epigenomics and DNA methylation studies have thriven, to become a research field of critical importance for modern biological research [2]. The rapid progress in microarray technologies, enabling the interrogation of an ever larger numbers of DNA/RNA transcripts more efficiently and at a lower cost, has opened new avenues for epigenetic monitoring [3].

In general, two broad microarray-based assay categories have been developed to measure DNA methylations: the enrichment-based microarrays and the bisulfite sequencing microarrays [4]. Based on the latter, Illumina's Infinium HumanMethylation450 BeadChip is one of the newest microarray platforms and can detect CpG methylation changes in more than 480,000 cytosines distributed over the whole genome [5].

Despite the revolutionary character of the aforementioned microarray technology, inherent imperfections obscure the true biological signal by introducing measurement bias. Moreover, idiosyncratic particularities of the DNA methylation data render popular statistical tools and methodologies developed for transcriptomic analysis inapplicable in their current form, to these data [1],[4]. Therefore, preprocessing and analysis for targeted bisulfite sequencing microarrays remains a challenging, active area of ongoing research [4].

To date, several preprocessing and analysis approaches have been proposed in the literature. In particular, Teschendorff et al. managed to efficiently eliminate almost all unwanted variability by applying a multivariate regression model and adjusting for batch effects through the use of the Illumina BeadArray control probes [6]. Moreover, in the context of enrichment-based microarrays, Aryee et al. developed a generic normalization strategy, tailored to DNA methylation data, and an empirical Bayes percentage methylation estimator yielded accurate absolute methylation estimates [7]. Among others, they proposed within-sample and between-sample normalization approaches that, based on platform-specific, control features, can be used for loess regression fitting [8],[9] and subset quantile normalization [10], respectively. The recommended control probe loess procedure may be applied to other two-color tiling array DNA methylation protocols, while the subset quantile normalization is even more widely applicable as it is not tied to microarray data [7]. In a similar fashion, Sun et al. recommended empirical Bayes correction along with normalization for an effective batch effect removal [11]. Also, Sabbah et al. implemented 'SMETHILLIUM', a non-parametric, spatial normalization method for Illumina's HumanMethylation27 BeadChip [12]. Finally, the specially tailored to Illumina's Infinium HumanMethylation450 BeadChip computational package 'IMA' has been designed and developed in [13], in order to automate the exploratory analysis in epigenetic studies.

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

In the present work, we introduce a new, generic framework for the preprocessing and analysis of high-volume, DNA methylation microarray data originating from the novel, high-density Illumina's Infinium HumanMethylation450 BeadChip [5]. The challenges related with the efficient processing of such voluminous datasets are lingering, and no available 'gold-standard' methods have been proposed to tackle the issue of technical bias. The proposed framework is applied to a pool of 96 samples (controls and cases) in order to examine retrospectively the manifestation of breast cancer. Here, DNA methylation is measured using the log2 ratio of the two channels intensities per probe (Methylated and UnMethylated), referred to as M-value and used widely in microarray-based transcriptomic analysis. Correction of the M-signal distribution is introduced, taking into account Intensity-related error measures. These error measures are based on the variation of the DNA methylation measurements of quality controls (QCs) that correspond to the same biological sample. The experimental design of the 450K methylation array is such that accommodates the fine screening of the methylation in CpG rich regions. These are probed in terms of their differential methylation between controls and cases. In addition, bootstrap based statistical estimates are derived in the distributions of the scaled coefficient variation of methylation measurements, defined as the ratio of the coefficients of variation of the cases to their respective controls. In this way, a robust measure of the true variation, compared to the technical one, is derived, exploiting the measured methylation in the QCs. Additionally, classical statistical differential analysis methods, including unpaired or paired tests, are employed between the categories of donors who remained healthy or manifested the disease.

II. Dataset

Methylation analysis based on Illumina's Infinium technology was first introduced with the Infinium HumanMethylation27 BeadChip [14]. The dataset studied here contains methylation data extracted using the new Illumina's Infinium HumanMethylation450 BeadChip, that includes 485,577 probes (482,421 CpG sites, 3091 non-CpG sites and 65 random SNPs). The available Italian breast cancer dataset encompassed 114 samples, organized in 12-sample chips. Ninety-six (96) samples correspond to breast cancer cases and controls, matched with cases in terms of age, body mass index, pre-post menopause. The remaining 18 samples are quality control samples (QCs), corresponding to the same sample measured in different chips which can therefore be used for reliable estimation of the technical variation observed in this dataset. At probe level, 2 channels, referring to the degree of unmethylation and methylation, are used to measure average methylation of the corresponding CpG site, and correspond to two channel intensities available for each probe: I_{Meth} and $I_{Un\text{-}meth}$. Detection p-value measurements have been used to ensure that statistically non-significant detected signals in the chips are excluded from further analysis.

III. Methodology

A. Measure of methylation

To date, two methods have been proposed to measure DNA methylation: i) Beta-value, ranging from 0 to 1, which is used to measure the percentage of methylation (Beta= $I_{Meth}/(I_{Meth} + I_{Un\text{-}meth})$), and ii) M-value, where $M=\log_2 (I_{Meth}/I_{Un\text{-}Meth})$, also widely used in gene expression microarray analysis [15]. Although Beta-value has a direct biological interpretation, corresponding roughly to the percentage of a site that is methylated, M-value is statistically more valid as it is approximately homoscedastic, and is thus adopted here [15].

B. Intensity – based normalization

The normalization of the M-signal distribution is systematically applied, taking into account the average intensity level of both channels $I=(I_{Meth}*I_{Un\text{-}Meth})^{1/2}$ and the QCs incorporated in each chip. Our scope here is to alleviate the impact of technical bias in the signal estimates. The normalization takes place in two successive steps: i) within-chip and ii) across all probes.

1) **Within-chip**: Available QCs (1-2 per chip) are used in order to calculate an error estimator in M-signals, across all intensity levels. Error for a certain probe type is estimated, considering the average M of all probe values in the QCs. All probes measured at a certain intensity level (the intensity space I is partitioned in percentiles) are utilized for the calculation of the error at this intensity level. Probe estimates for all arrays (cases, controls and QCs) are then updated, i.e. M-value of a probe is recalculated by subtracting the respective error calculated for the corresponding intensity level. Algorithmic steps at this stage are presented below:

Stage 1: For each chip containing at least a QC do:

a. *Identify Intensity Segments (I-S) (percentiles) at QC sample (or average QCs probe values if >1 QC samples in the chip).*

b. For each I-S

Find probes members of the I-S in the given QC and calculate the error in this I-S:

$Error_{I\text{-}S}$ = Average ($Error_k$), k: number of probes in the I-S

$Error_k = M_k - M_{0(k)}$ ($M_{0(k)}$: average M of k probe in all QCs)

c. *For each I-S, identify all j probes of QCs and other samples in each chip, which intensity lies in the same I-S percentile and update M_j values*:

$M_{j\_corrected\_1} = M_j - Error_{I\text{-}S}$

Whenever a chip contains no QC, then the estimation of the error is performed, by relating it to another chip, which contains at least one QC, based in the similarity of their intensity distributions. To this end, *k*-means clustering of all samples across chips is employed, using the average intensity measurements per probe. Clusters are formed corresponding to groups of samples with similar I distributions. A chip without QCs is thus linked with another chip with QCs and its signal values are then updated using the I-S error estimates of this chip. Since a chip without QCs contains 12 samples, a majority voting scheme is applied to resolve the selection of the appropriate chip, based on the similarity of the incorporated samples after completion of *k*-means clustering.

2) **Across all probes**: A second normalization, per probe this time, is applied exploiting the standard deviation of the M-values across all *t* QCs for any probe. M-values of any probe across all samples are then updated, by subtracting the probe based error estimate.

Stage 2: Using M-values, as updated in Stage 1, do:

a. *For each probe p, calculate an estimator of $Error_p$ from all available t QCs:*

$$\mathrm{Error}_p = \mathrm{std}(M_p(1),M_p(2)..,M_p(t)),$$

where *t* is the number of QCs for probe *p*

b. *Across all samples, update probe p:*

$$M_{p\_final} = M_{p\_corrected\_2} = M_{p\_corrected\_1} - \mathrm{Error}_p$$

C. Scaled coefficient variation measurements

In order to identify probes that are reliable candidates for differential DNA methylation among physiological categories (cases vs. controls), the notion of scaled coefficient variation (CV_{scaled_p}) (Eq. 1, 2) for each probe is introduced. This represents a robust measure of the real inter-class variability observed in a probe in the whole population (controls and cases), when compared to that observed among QCs, which measures solely the technical variation. The greater this coefficient for a probe p is, the greater the real (beyond technical variation) differential expression is. Thus:

$$CV_{scaled\_p} = \mathrm{abs}\,[\,CV_{controlsUcases\_p} / CV_{QCs\_p}\,] \qquad (1)$$

$$CV_{samples(1,..k)\_p} = \mathrm{std}(M_p(1,..k)) / \mathrm{mean}(M_p(1,..k)) \qquad (2)$$

Based on the distribution of CV_{scaled} values across all probes, a z-test is applied in the CV_{scaled} distribution, to assign a p-value to each CV_{scaled} value. A higher CV_{scaled} is related to a lower p-value, implying more intense inter-class variability, for this probe.

D. Statistical Selection

In order to derive statistically significant differential DNA methylation patterns between controls and cases at the probe level, apart from the typical unpaired t-test (control vs. cases), a paired t-test is also applied. The paired t-test is based on the total 47 cases-control matched pairs, defined through the use of additional variables (see Dataset Section). Therefore, two sets of p-values are computed for each probe and then are comparatively assessed (see Results & Discussion Section).

E. Bootstrapping-based correction of p-values

Bootstrapping-based correction is here applied, so as to immunize statistical findings against the detrimental effect of multiple hypothesis testing. Goal is to examine whether the p-values obtained either from a statistical test or extracted based on CV measurements are indeed that extreme or they could represent random false selections. Thus, the p-value in the original p-value distribution is compared to those computed from a series of distributions of p-values, ranked in ascending order, derived by bootstrapping. The procedure is repeated and the number of times, where the original p-value is found less extreme compared to the one derived by the bootstrap distribution, represents the corrected p-value with respect to the original one. In order to incorporate into the corrected value the significance of the initial statistical test, we designate a Bayesian-type product term, that multiplies both the p-value estimate derived by the bootstrapping, and that of the original statistical test, The value extracted is next and normalized by the average estimate of this product for all probes. The algorithmic steps for this correction are presented in the following:

Input: A series of m *p*-values ($p_1,..p_m$) and a number of bootstrap repetitions (*nboo*t, ought to be a big number: 1,000, 100,000 etc.)[1]

Steps

a. *Construct nboot distributions (m elements each) by bootstrapping (selection and replacement) the original p-value distribution*[2]*.*

b. *For each p_i*

   i. *Find its position in the original distribution (e.g. percentile):* $p_{i\_original\_position}$

   ii. *For each iteration k, use the k^{th} p-values bootstrap distribution vector:*

      Derive the p-value $p_{i(k)}$ of the $p_{i\_original\_position}$ percentile of the k^{th} bootstrap distribution

      If $p_{i\_original\_position} > p_{i(k)}$ increase the *counter*

   iii. *Correct each p_i* :

$$p_i\_corrected = \frac{p_i * (counter / nboot)}{\dfrac{\sum_i p_i * (counter / nboot)}{n}}$$

IV. RESULTS AND DISCUSSION

Regarding the intensity-based correction of the M-values, the estimation of the error across all intensity levels proved to be greater at lower intensity levels. This implies lower statistical power, which would require higher replicate numbers for reliable signal estimation at these levels. This was a consistent finding, observed in all 12-sampled chips (an example is illustrated in Fig. 1). Intensity-based correction of the M-values was followed by the calculation of the scaled CV_{scaled} measurements and then statistical selection testing (paired and unpaired tests).

[1] When *p*-values are derived, based on the CV_{scaled} distribution, CV_{scaled} measurements are also input to the algorithm.

[2] When *p*-values are corrected, based on the CV_{scaled}, the original CV_{scaled} distribution is bootstrapped and comparisons in b.i and b.ii steps are performed based on CV_{scaled} measurements (original or bootstrapped).

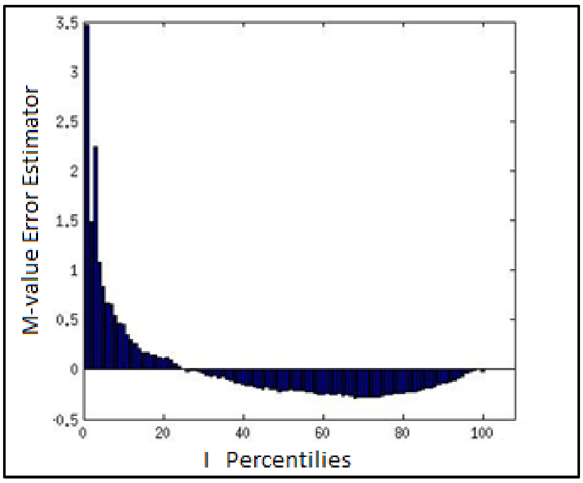

Figure 1. M-value error estimator acrosss average intensity levels for one of the available 12-sample chips.

Two sets of p-values were extracted either from the conventional statistical tests or the CV_{scaled} measurements and were further corrected exploiting bootstrap resampling (nboot=100000). Regarding statistical thresholds, top 1% and top 5% of probes were adopted, using their corrected p-values. In the case of the CV_{scaled} measurements, the top 10% performing probes were selected, utilizing the corresponding p_corrected values. As it can be seen in Figure 2, where the histogram of the CV_{scaled} measurements is illustrated, the impact of the various normalization steps of our composite framework is such that median of the CV_{scaled} distribution is 0.9964 (very close to 1). This supports the plausibility of the processing steps applied, as a CV_{scaled} measurement of 1, implies that the variation observed between cases and controls is the same with that observed in QCs (technical variation). Extrapolating, one can conclude that the top 50% probes in this distribution, demonstrate higher variation than the technical one, something that is strengthening when narrowing the selection threshold. Furthermore, if coupled with another statistical selection method (e.g. statistical selection using paired or un-paired t-test), it can filter out unreliable probes in terms of signal quality, retaining sound candidates as regards their true differential DNA methylation expression. In this study, this strategy was adopted, resulting in selected probe subsets corresponding to CpG sites, that were qualified by both approaches (Table I).

In Table I, the results of the un-paired and the paired t-test selection are juxtaposed, with the latter method intensifying the significance effect in the selection process, namely, lower p-value estimates for the same selection threshold (top1% or top 5%). The top 1% criterion in paired t-test was combined with the $CV_{scaled} > 1$ criterion, in order to select a robust subset of CpG sites that are both significantly differentially methylated and immunized against technical variation. Both criteria provided a subset of 2398 CpG sites (Table I, Column 1) corresponding, according to the Illumina's Infinium HumanMethylation450 BeadChip annotation, to 1717 unique UCSC gene ids.

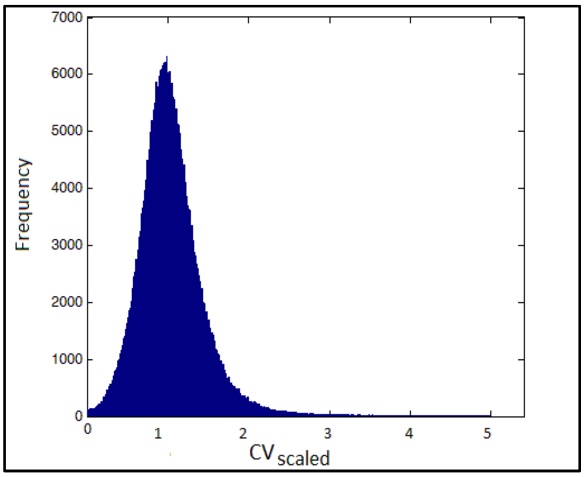

Figure 2. Histogram of all CV_{scaled} measurements: [mean median]=[1.2582 0.9964].

TABLE I. NUMBER OF SELECTED PROBES PER STATISTICAL METHODS ALTERANTIVE AND THEIR INTERSECTIONS

Statistical Method 1 and thresholds	**Statistical Method 2 and thresholds**	**Common probes by Methods 1 &2**
Un-paired t-test, top 5% (24279 in total, p_corrected ≤ 0.0515)	Paired t-test, top 5% (24279 in total, p_corrected ≤ 0.0401)	16828
Un-paired t-test, top 1% (4856 in total, p_corrected ≤ 0.0132)	Paired t-test, top 1% (4856 in total, p_corrected ≤ 0.0063)	3067
Un-paired t-test, top 1% (4856 in total, p_corrected ≤ 0.0132)	$CV_{scaled} > 1$ (240499 in tolal)	2305
Paired t-test, top 1% (4856 in total, p_corrected ≤ 0.0063)	$CV_{scaled} > 1$ (240499 in tolal)	2398
Un-paired t-test, top 1% (4856 in total, p_corrected ≤ 0.0132)	CV_{scaled}, top 10% (48558 in total, p_corrected ≤ 0.0815, most significant p_corrected corresponds to CV_{scaled}=1.5212)	483
Paired t-test, top 1% (4856 in total, p_corrected ≤ 0.0063)	CV_{scaled}, top 10% (48558 in total, p_corrected ≤ 0.0815, most significant p_corrected corresponds to CV_{scaled}=1.5212)	478

In order to better understand the molecular mechanisms implicated in the promotion of breast cancer, pathway analysis was performed. This was done through StRAnGER web application [16], which performs functional analysis in high-throughput genomic datasets, starting from a list of significant molecular targets (transcripts, genes, proteins, etc). The disease is allegedly linked with aberrant epigenomic, regulatory functions, as a result of chromatin remodeling, due to DNA methylation events. In StRAnGER, established statistical tests are coupled with bootstrapping, thus enabling the derivation of a final population of statistically significant ontological terms, that comprise a set of over-represented terms, compared to all other terms of the ontology utilized. The list of 1717 unique gene ids yielded 187 over-represented GO terms for the default settings (hypergeometrc test $p \leq 0.05$, 10000 bootstrap iterations), the most significant of which ($p \leq 10^{-5}$) are presented in Table II. The cellular processes which CpG sites present a pronounced differential methylation pattern, are related mainly with developmental and transcriptional regulatory actions. This is line with the fact that chromatin remodeling represents a critical regulatory function, as well as the fact that gradual deviation of a cell to cancerous phenotype, necessitates aberrant operation of cellular developmental and morphological programs and circuitries. The fact that among such programs, neural specific developmental processes are quoted, implies that upon the carcinogenic transformation, the cells lose their differentiation potential and they regain their transformative pluripotency.

The modular methodological framework presented here, including the 2-stepped intensity-based normalization of the M values and the exploitation of the estimates of the scaled coefficients of variation of the probe values, is quite generic. Whenever the experimental design is such as to afford quality controls to enable derivation of reliable technical variation measures, the methodology proposed here is straightforward applicable even for the case of mRNA data analysis as well. As future work, already in the phase of implementation, the development of a web pipeline is envisaged to provide access to the algorithmic framework presented, through the setup of appropriate interfaces and the accommodation of various experimental platforms.

V. Conclusions

In this paper, a composite computational framework for the analysis of high volume (Illumina technological platform) methylation data was presented. An intensity-based normalization method of M-values was proposed, while a scaled coefficient variation ratio term was introduced so as to tackle the issue of the detrimental role of technical bias, when assessing real inter-class variability (controls vs. cases). The methods proposed here exploit the technical controls available in the specific dataset. They are, however, quite generic to accommodate other designs, even such that dispense with quality controls. As a third step, a bootstrap based p-value correction algorithm was applied either to the statistical selection results or the scaled coefficient variation measurements, to enhance the reliability of the statistical values derived. The framework was applied to an italian breast cancer dataset and the preliminary results are promising and convincing, as it can be surmised from their functional evaluation. The selected CpG sites were further subjected to a statistical enrichment pathway analysis, revealing cellular functions, congruent either with established aspects of the breast cancer physiology or those of epigenomic regulation. Moreover, they have already provided a long pile of candidates, which are gradually deepening our understanding about the complexity of the carcinogenic process. They also represent possible targets for further experimentation or systems level interpretation. The composite framework proposed here, is generic enough so as to be extended to or accommodate other tangible high-throughput analysis tasks as those of various microarray technologies.

Acknowledgment

The work, presented in this paper, has been generously co-funded by the EU/FP7/ EnviroGenomarkers project.

References

[1] P.W. Laird, “Principles and challenges of genome-wide DNA methylation analysis,” Nat.Rev.Genet., vol. 11, pp. 191–203, February 2010.

[2] A. Schumacher, P. Kapranov, Z. Kaminsky, J. Flanagan, A. Assadzadeh, P. Yau, C. Virtanen, N. Winegarden, J. Cheng, T. Gingeras and A. Petronis, “Microarray-based DNA methylation profiling: technology and applications,” Nucleic Acids Res., vol. 34, pp. 528–542, January 2006.

[3] B. van Steensel and S. Henikoff, “Epigenomic profiling using microarrays,” BioTechniques, vol. 35, pp. 346–50, 352–4, 356–7, August 2003.

[4] K.D. Siegmund, “Statistical approaches for the analysis of DNA methylation microarray data,” Hum.Genet., vol. 129, pp. 585–595, June 2011.

[5] J. Sandoval, H. Heyn, S. Moran, J. Serra-Musach, M.A. Pujana, M. Bibikova, and M. Esteller, “Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome,” Epigenetics, vol. 6, pp. 692–702, June 2011.

[6] A.E. Teschendorff, U. Menon, A. Gentry-Maharaj, S.J. Ramus, S.A. Gayther, S. Apostolidou, A. Jones, M. Lechner, S. Beck, I.J. Jacobs, and M. Widschwendter, “An epigenetic signature in peripheral blood predicts active ovarian cancer,” PLoS One, vol. 4, pp. e8274, December 2009.

[7] M.J. Aryee, Z. Wu, C. Ladd-Acosta, B. Herb, A.P. Feinberg, S. Yegnasubramanian, and R.A. Irizarry, “Accurate genome-scale percentage DNA methylation estimates from microarray data,” Biostatistics, vol. 12, pp. 197–210, April 2011.

[8] W.S. Cleveland, “Robust locally weighted regression and smoothing scatterplots,” J. Am. Stat. Assoc., vol. 74, pp. 829–836, December 1979.

[9] W.S. Cleveland and S.J. Devlin, “Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting,” J. Am. Stat. Assoc., vol. 83, pp. 596–610, September 1988.

[10] Z. Wu and M.J. Aryee, “Subset quantile normalization using negative control features,” J.Comput.Biol., vol. 17, pp. 1385–1395, October 2010.

[11] Z. Sun, H.S. Chai, Y. Wu, W.M. White, K.V. Donkena, C.J. Klein, V.D. Garovic, T.M. Therneau, and J.P. Kocher, “Batch effect correction for genome–wide methylation data with Illumina Infinium platform,” BMC Med.Genomics, vol. 4, pp. 84, December 2011.

[12] C. Sabbah, G. Mazo, C. Paccard, F. Reyal, and P. Hupe, “SMETHILLIUM: spatial normalization method for Illumina infinium HumanMethylation BeadChip,” Bioinformatics, vol. 27, pp. 1693–1695, June 2011.

[13] D. Wang, L. Yan, Q. Hu, L.E. Sucheston, M.J. Higgins, C.B. Ambrosone, C.S. Johnson, D.J. Smiraglia, and S. Liu, “IMA: An R package for high–throughput analysis of Illumina’s 450K Infinium methylation data,” Bioinformatics, January 2012.

[14] M. Bibikova, J. Le, B. Barnes, S. Saedinia-Melnyk, L. Zhou, R. Shen, and K.L. Gunderson, “Genome-wide DNA methylation profiling using Infinium(R) assay,” Epigenomics, vol. 1, pp. 177–200, October 2009.

[15] P. Du, X. Zhang, C.C. Huang, N. Jafari, W.A. Kibbe, L. Hou, and S.M. Lin, “Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis,” BMC Bioinformatics, vol. 11, pp. 587, November 2010.

[16] A.A. Chatziioannou and P. Moulos, “Exploiting Statistical Methodologies and Controlled Vocabularies for Prioritized Functional Analysis of Genomic Experiments: the StRAnGER Web Application,” Front.Neurosci., vol. 5, pp. 8, January 2011.

TABLE II. OVER-REPRESENTED GO TERMS BASED ON GENES CORRESPONDING TO THE 2,398 SELECTED CPG SITES

GO Term	GO Annotation/Gene Description	GO Category	OT p-value	Enrichment
GO:0005515	protein binding	F	1.10E-09	667/7070
GO:0007507	heart development	P	3.00E-08	34/162
GO:0007156	homophilic cell adhesion	P	8.86E-08	28/125
GO:0045995	regulation of embryonic development	P	1.34E-07	6/8
GO:0030528	transcription regulator activity	F	2.24E-07	67/457
GO:0015020	glucuronosyltransferase activity	F	3.71E-07	10/23
GO:0007411	axon guidance	P	3.88E-07	23/98
GO:0042613	MHC class II protein complex	C	7.59E-07	9/20
GO:0043565	sequence-specific DNA binding	F	8.10E-07	82/616
GO:0045165	cell fate commitment	P	8.81E-07	14/45
GO:0045202	synapse	C	9.30E-07	44/268
GO:0005244	voltage-gated ion channel activity	F	1.39E-06	29/149
GO:0046872	metal ion binding	F	1.68E-06	281/2780
GO:0007399	nervous system development	P	1.85E-06	58/400
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	P	2.10E-06	55/374
GO:0007275	multicellular organismal development	P	2.14E-06	111/923
GO:0021513	spinal cord dorsal/ventral patterning	P	2.94E-06	4/5
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	P	3.17E-06	41/254
GO:0007155	cell adhesion	P	3.58E-06	73/551
GO:0045596	negative regulation of cell differentiation	P	4.08E-06	10/28
GO:0016020	membrane	C	6.16E-06	425/4513
GO:0048646	anatomical structure formation involved in morphogenesis	P	8.29E-06	8/20
GO:0030900	forebrain development	P	9.10E-06	16/66

OT p-values correspond to the p-value yielded by hypergeometric test. The enrichment score equals to the number of times a GO terms appears due to differentially methylated CpG sites and corresponding gene, divided by the number of times the GO terms appears due to all genes in the human background annotation available to the Stranger platform [16]. GO Category may correspond to molecular function (F), biological process (P) cellular component (C).