Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Ryu et al., J Biomet Biostat 2013, 4:3
DOI: 10.4172/2155-6180.1000164

Research Article Open Access

# Quantifying and Normalizing Methylation Levels in Illumina Arrays

**Duchwan Ryu[1]*, Hongyan Xu[1], Varghese George[1], Shaoyong Su[2], Xiaoling Wang[2] and Robert H Podolsky[1,3,4]**

[1]*Department of Biostatistics and Epidemiology, Georgia Regents University, Augusta, GA, USA*
[2]*Georgia Prevention Institute, Georgia Regents University, Augusta, GA, USA*
[3]*Center for Biotechnology and Genomic Medicine, Georgia Regents University, Augusta, GA, USA*
[4]*Department of Medicine, Georgia Regents University, Augusta, GA, USA*

## Abstract

The role of genome-wide patterns of methylation in human disease has drawn attention increasingly in recent years, because the methylome has the potential for large effects in disease etiology. Most analyses of methylation have utilized the percent signal that is methylated, known as $\beta$-value, or the logistic transformation of $\beta$, named M-value, as the summary measures. However, in general, these summary measures do not follow a Normal distribution and lead to statistical tests sensitive to outlying samples. In this paper, we proposed the N-value, a type of weighted logistic transformation of $\beta$ that accounts for signal variability among beads for analyses of differential methylation. Our analysis of 27K Illumina array data showed that the N-value follows a desirable shape of sample distribution, and its test is robust to outliers. Through a simulation study, we presented results that show the t-tests of the N-value is more consistent, and has greater power under the presence of heterogeneity of samples and in different sample sizes.

**Keywords:** Measure of methylation level; Methylated signal variability; Test for differential methylation; N-value

## Introduction

Interest in the role of genome-wide patterns of methylation in human disease has increased in recent years [1]. The epigenome, in general, and the methylome, more specifically, have the potential for large effects in disease etiology. DNA methylation has already been associated with many cancers [2], and different cell types are known to differ in their methylation patterns. Illumina has been developing genome-wide methylation arrays that enable epigenome-wide association studies of human disease [3]. These arrays are based on BeadChip technology, and the most recent ones contain probes for over 480K CpG sites. These sites cover 99% of RefSeq genes with multiple probes per gene, and 96% of CpG islands from the UCSC database.

This new array utilizes some of the site-specific probes from the previous generation chip that contains probes for 27K CpG sites [4]. Probes for the sites from the 27K array use the Infinium I assay, while the newer probes use the Infinium II assay. The Infinium I assay is based on separate beads for methylated and unmethylated DNA, and the Infinium II assay relies on a single bead for both methylated and umethylated DNA [3]. Both assays result in red- and green-channel intensities, which are normalized using a proprietary method in BeadStudio [5].

Regardless of the probe design, tests for differential methylation involve comparing the relative methylated/unmethylated signal among experimental conditions [5]. The standard output from BeadStudio provides estimates of the percent signal that is methylated, usually denoted by $\beta$. This value is a natural way to summarize the relative methylated/unmethylated signal. Although $\beta$ is a convenient way to summarize the extent of methylation for any given CpG site, statistical analyses aimed at detecting differential expression may be optimized using other measures. Recently, Du et al. [6] showed that using a logistic transformation of $\beta$, named M-value, may yield better results. However, neither $\beta$ nor M take into account, the variability in both the methylated and unmethylated signal. We investigated statistical approaches that utilize a weighted logistic transformation of the methylated and unmethylated signals, with the goal of improving the analyses aimed at detecting differential methylation. The method that we develop below takes into account the probe specific variances in summarizing methylation levels.

Another aspect to ensuring accurate results is the need for data normalization, prior to statistical testing for differential methylation. The Illumina software does normalize probe signals in calculating the methylated and unmethylated signals. While much work has focused on normalizing data obtained using BeadArrays for gene expression, less research has focused on normalization of the methylation arrays [5]. While not all studies normalize data beyond that supplied through BeadStudio, some studies have used various normalization strategies, including quantile [7,8], and mean normalization [9,10]. We observed that some arrays in our data had very different distributions of signals after normalization based on the Illumina software, and we were therefore, concerned about between array normalization. Sun et al. [11] has examined the use of various normalization methods for adjusting for batch effects. Teschendorff et al. [7] examined several approaches to normalization, examining several factors that could affect the analyses, including batch, DNA input and bisulfite conversion efficiency, as measured using control probes. The optimal normalization for their data was a linear regression that included batch, DNA input and bisulfite conversion efficiency. Bell et al. [12] normalized $\beta$ to follow the standard normal distribution, and then used this normalized $\beta$ in all analyses. We demonstrate below that our proposed method for summarizing methylation levels also has the advantage, in that it appropriately normalizes relative methylation levels.

## Materials and Methods

### Definitions of β-value, M-value and N-value

The Illumina software provides several measures for summarizing methylation levels: the average signal for both methylated and

***Corresponding author:** Duchwan Ryu, Department of Biostatistics and Epidemiology, Georgia Regents University, 1120 15th St., Augusta, GA 30912, USA, Tel: (706)721-6721; Fax: (706)721-6294; E-mail: dryu@gru.edu

**Received** January 30, 2013; **Accepted** April 08, 2013; **Published** April 13, 2013

**Citation:** Ryu D, Xu H, George V, Su S, Wang X, et al. (2013) Quantifying and Normalizing Methylation Levels in Illumina Arrays. J Biomet Biostat 4: 164. doi:10.4172/2155-6180.1000164

unmethylated "probes" (Infinium I utilizes separate probes, while Infinium II utilizes a single probe to generate both a methylated and unmethylated signal), the standard deviation of both methylated and unmethylated signals, and an estimate of the percent of chromosomes that are methylated, known as $\beta$-value. The $\beta$-value for the $i^{th}$ CpG site on the $j^{th}$ array, where $i$ =1,.. .,$I$ and $j$ =1,.. .,$J$, is defined as the ratio of the average signal of the methylated probe (methylated signal) divided by the total average signal of both the methylated and unmethylated probes (methylated and unmethylated signal),

$$\beta_{ij} = \frac{\max(\bar{y}_{ij,m}, 0)}{\max(\bar{y}_{ij,m}, 0) + \max(\bar{y}_{ij,u}, 0) + \alpha},$$

where $\bar{y}_{ij,m}$ is the methylated signal, $\bar{y}_{ij,u}$ is the unmethylated signal, and $\alpha$ is a constant offset ($\alpha$=100 by default). Du et al. [6] suggested the M-value,

$$M_{ij} = \log_2\left(\frac{\bar{y}_{ij}^{m} + \alpha}{\bar{y}_{ij}^{u} + \alpha}\right) = \log_2\left(\frac{\beta_{ij}}{1-\beta_{ij}}\right), \quad (1)$$

which is equivalent to a $\log_2$ logistic transformation of $\beta$ with a constant offset $\alpha$.

To define the N-value incorporating the standard deviations of signals, we assume that the logarithm of the average signal is proportional to the logarithm of the standard deviation of signal. Let $S_{ij}^{m}$ denote the standard deviation of the methylated signal, and let $S_{ij}^{u}$ denote the standard deviation of the unmethylated signal for the $i^{th}$ site on the $j^{th}$ array. The linear relationship between average and standard deviation of signal in log-scale is then described using two regression models,

$$\log(\bar{y}_{ij}^{m}) = \Upsilon_0^{m} + \Upsilon_1^{m} \log(S_{ij}^{m}) + \epsilon_{ij}^{m},$$

$$\log(\bar{y}_{ij}^{u}) = \Upsilon_0^{u} + \Upsilon_1^{u} \log(S_{ij}^{u}) + \epsilon_{ij}^{u},$$

where $\Upsilon_0^{m}$, $\Upsilon_1^{m}$, $\Upsilon_0^{u}$ and $\Upsilon_0^{u}$ are regression coefficients, and $\epsilon_{ij}^{m}$ and $\epsilon_{ij}^{u}$ are normal random errors with zero means and constant variances, respectively for the methylated and unmethylated signals.

We define the N-value based on a scaled $\beta$. First, we obtain scaled estimates of average methylated signal, $\bar{y}_{ij}^{ms}$ and unmethylated signal, $\bar{y}_{ij}^{us}$ by utilizing estimated parameters $\hat{\Upsilon}_0^{m}$, $\hat{\Upsilon}_1^{m}$, $\hat{\Upsilon}_0^{u}$, and $\hat{\Upsilon}_1^{u}$ in the above regression models. In fact, the exponential residuals from the two regression models, $\exp(\hat{\epsilon}_{ij}^{m})$ and $\exp(\hat{\epsilon}_{ij}^{u})$, are scaled average signals, with respect to standard deviations of signals:

$$\bar{y}_{ij}^{ms} = \exp(\hat{\epsilon}_{ij}^{m}) = \frac{\bar{y}_{ij}^{m}}{\exp(\hat{\Upsilon}_0^{m})(S_{ij}^{m})^{\hat{\Upsilon}_1^{m}}},$$

$$\bar{y}_{ij}^{us} = \exp(\hat{\epsilon}_{ij}^{u}) = \frac{\bar{y}_{ij}^{u}}{\exp(\hat{\Upsilon}_0^{u})(S_{ij}^{u})^{\hat{\Upsilon}_1^{u}}}.$$

Next, the scaled average signals, $\bar{y}_{ij}^{ms}$ and $\bar{y}_{ij}^{us}$, are used to define a scaled $\beta$,

$$\beta_{ij}^{*} = \frac{\bar{y}_{ij}^{ms}}{\bar{y}_{ij}^{us} + \bar{y}_{ij}^{ms}} = \frac{1}{1 + \exp(\hat{\epsilon}_{ij}^{u} - \hat{\epsilon}_{ij}^{m})},$$

which, in turn, is used to define the N-value,

$$N_{ij} = \log_2\left(\frac{\beta_{ij}^{*}}{1-\beta_{ij}^{*}}\right). \quad (2)$$

The N-value is based on adjusting both the methylated and unmethylated signals relative to the expected signals, given the observed standard deviations for signals. While these normalized signals have little meaning in themselves, the induced $\beta$ measures the relative strength of the methylated and unmethylated signals in relation to the expected signals, given the standard deviations. The important measure in this case is the normalized $\beta$. Comparing equations (1) and (2) shows that the N-value can be viewed as a version of the M-value rescaled by the standard deviation. We will further show that this rescaling results in a normalization of the signals, adjusting the distribution of intensities that vary among samples with different standard deviations. The analysis of differential methylation using all three quantification methods ($\beta$, M, and N), is next considered using data from a pilot experiment.

## Obesity dataset

We utilize data collected from 7 obese samples (case samples) and 7 age-matched lean control samples, using the 27K Illumina array (27,578 CpG sites [13]; NCBI's Gene Expression Omnibus accession number GSE25301). These 14 subjects were identified from the participants (n=534) in the Lifestyle, Adiposity, and Cardiovascular Health in Youth (LACHY) study, using the following inclusion criteria: (1) African American ancestry; (2) male; (3) having leukocyte DNA available; (4) obese cases having a body mass index (BMI) ≥ 99$^{th}$ percentile for age and sex, and lean controls having BMI ≤ 10$^{th}$ percentile for age and sex. The LACHY study consisted of roughly equal numbers of African American and European American adolescents, aged 14 to 18 years, of both sexes recruited from high schools in the Augusta, Georgia area [14].

## Simulation settings

We conducted a simulation study to compare the performance of testing for differential methylation, using a t-test and each of the three measures of methylation level. In this simulation study, the percent of CpG sites designated as being differentially methylated was 5%, 10%, 15%, or 20%, with the specific sites being chosen randomly to be differentially methylated. Data for both cases and controls were simulated using the same distributions for all sites designated as not differentially methylated. Data for cases and controls were simulated from different distributions for all sites designated as being differentially methylated. Some differentially methylated sites were simulated with the case samples having a different distribution from the null distribution, and some sites were simulated with the control samples having a different distribution from the null distribution.

We simulated the average signals and standard deviations of signal, which were then used to calculate the $\beta$-, M, and N-value as above. The methylated and unmethylated signals were simulated separately, since these two signals had slightly different distributions in the obesity data. We simulated standard deviations of methylated signal, $S_{ij}^{m*}$ and unmethylated signal, $S_{ij}^{u*}$, for the CpG site $i$, and the sample $j$, from lognormal distributions such that

$$\log S_{ij}^{m*} \sim N(\mu_s^{m}, \sigma_s^{m2}) \text{ And } \log S_{ij}^{u*} \sim N(\mu_s^{u}, \sigma_s^{u2}).$$

The means and standard deviations used in the simulation were estimated using the obesity data. Intensities were simulated next, by using simple linear regression models of log intensity on

log standard deviation for the intensities of the obesity data such that $\bar{y}_{ij}^m = \Upsilon_0^m + \Upsilon_1^m S_{ij}^m + \epsilon_{ij}^m$ and $\bar{y}_{ij}^u = \Upsilon_0^u + \Upsilon_1^u S_{ij}^u + \epsilon_{ij}^u$, where $\epsilon_{ij}^k \sim N(0, \sigma_\epsilon^{k2})$ and $(\Upsilon_0^m, \Upsilon_1^m)$ and $(\hat{\Upsilon}_0^u, \hat{\Upsilon}_0^u)$ are regression coefficients.

The least square estimators of these regression coefficients, $(\hat{\Upsilon}_0^m, \hat{\Upsilon}_1^m)$ and $(\hat{\Upsilon}_0^u, \hat{\Upsilon}_1^u)$, follow normal distributions such that $(\hat{\Upsilon}_0^m, \hat{\Upsilon}_1^m) \sim N(\mu_\Upsilon^m, \Sigma_\Upsilon^m)$ and $(\hat{\Upsilon}_0^u, \hat{\Upsilon}_1^u) \sim N(\mu_\Upsilon^u, \Sigma_\Upsilon^u)$, where $\mu_\Upsilon^m, \mu_\Upsilon^u, \Sigma_\Upsilon^m$ and $\Sigma_\Upsilon^u$ are means and variances of the estimated regression coefficients. We estimated the variances of regression errors through the best quadratic unbiased estimators denoted by $\hat{\sigma}_\epsilon^{m2}$ and $\hat{\sigma}_\epsilon^{u2}$. Similarly, we have estimated means and variances of regression coefficients, $\hat{\mu}_\Upsilon^m$ and $\hat{\Sigma}_\Upsilon^m$ for methylated signals, and $\hat{\mu}_\Upsilon^u$ and $\hat{\Sigma}_\Upsilon^u$ for unmethylated signals based on the regressions fit separately for each sample.

Random samples of regression coefficients were then generated by $(\Upsilon_0^{m*}, \Upsilon_1^{m*}) \sim N(\hat{\mu}_\Upsilon^m, \hat{\Sigma}_\Upsilon^m)$ for methylated probes and $(\Upsilon_0^{u*}, \Upsilon_1^{u*}) \sim N(\hat{\mu}_\Upsilon^u, \hat{\Sigma}_\Upsilon^u)$ for unmethylated probes. These simulated regression coefficients were then used to simulate the intensities using the regression equations

$$\log \hat{y}_{ij}^{m*} = \Upsilon_0^{m*} + \Upsilon_1^{m*} \log S_{ij}^{m*} + \epsilon_{ij}^{m*},$$

$$\log \hat{y}_{ij}^{u*} = \Upsilon_0^{u*} + \Upsilon_1^{u*} \log S_{ij}^{u*} + \epsilon_{ij}^{u*}$$

where $\epsilon_{ij}^{m*} \sim N(0, \hat{\sigma}_\epsilon^{m2})$ and $\epsilon_{ij}^{u*} \sim N(0, \hat{\sigma}_\epsilon^{u2})$.

The distribution of signal intensity is heterogenous across samples in the obesity data. The simulation procedure described above does not result in the heterogeneity observed in the data. We, therefore, also simulated data with increased heterogeneity, in which the data from one subject, subject *o*, came from modified distributions. For the methylated probes, the standard deviations were generated by $\log S_{io}^{m*} \sim N[\mu_s^m, (0.3\sigma_s^m)^2]$. To maintain the linear relationship between the standard deviations and the intensities, regression coefficients were sampled from $(\Upsilon_{0,o}^{m*}, \Upsilon_{1,o}^{m*}) \sim N(\hat{\mu}_\Upsilon^m + 0.7, \hat{\Sigma}_\Upsilon^m)$. These regression coefficients were then used to generate the sample intensities as above.

The unmethylated probes were simulated using larger standard deviations than the methylated probes, such that $\log S_{io}^{u*} \sim N[\mu_s^u, (0.5\sigma_s^u)^2]$. The linear relationship between standard deviation and intensity was generated using $(\Upsilon_{0,o}^{u*}, \Upsilon_{1,o}^{u*}) \sim N(\hat{\mu}_\Upsilon^u + 1, \hat{\Sigma}_\Upsilon^u)$.

Differences between cases and controls were generated by shifting the standard deviations in the case samples. In one set of simulations, these standard deviations were shifted by 1, and in another set of simulations these standard deviations were shifted by -1.

We examined 1,000 simulations for all scenarios. Two sample sizes for cases and controls were considered: 10 and 20, under two well known significance levels $\alpha$ (Type I error): 0.05 and 0.01.

## Results

### Signal intensity mean and standard deviation

Two measurements are made for each CpG site, one methylated and the other unmethylated. Any measure of relative methylation levels will ultimately depend on these signal intensities. As such, we first examined the distribution of the mean signal intensity, as well as the standard deviation of the signal intensity for each CpG site. Comparing the distribution of the mean signal intensity among samples showed that the samples are not homogenous, with the signal standard deviation showing a similar pattern (Figure 3). Of note was case sample #5, which has a wider range in mean and standard deviation of the signal. Further, the signal distributions showed greater variability among cases than among controls. A linear relationship between the mean and the standard deviation of the signals appeared to underlie these differences (Figure 4). We utilized this linear relationship to obtain our proposed N-value (Equation 2).

### Distributions of measures of methylation level

Although the distributions of the original signals showed clear differences among samples, these differences need not translate into differences in the three summary statistics, $\beta$, M, and N. As noted above, $\beta$ ranged from 0 to 1, with 0 indicating that none of the DNA molecules in that sample were methylated, and 1 indicating that all of the DNA molecules in that sample were methylated. Although $\beta$ provided an easily interpretable measure of methylation level, its distribution for each sample was highly skewed and slightly bimodal (Figure 1). General concerns about analyzing proportions suggest that a logistic transformation might be appropriate (e.g. M-value, Equation 1). The distribution of the M-value for each sample was bimodal, with the two peaks being more prominent than that for $\beta$. While the distribution of $\beta$ and M for case sample #5 did not appear to be as drastically different from the other samples, the variability in the distribution of both $\beta$ and M was greater among the cases than among the controls. This heterogeneity in distributions suggested the need for normalization.

Given the linear relationship observed between mean signal intensity and the standard deviation of the signal, the sample specific signal for any given CpG site could be considered as the deviation from the expected relationship between mean and standard deviation (Figure 4). Using these adjusted signal intensities, we defined the N-value (Equation 2). The distribution of the N-value was symmetric, and more importantly, showed less variability among samples than that observed for $\beta$ and M (Figure 1).

Being a proportion, the variance in β among samples was likely to be associated with the mean of $\beta$, which was what we observed (Figure 2). This relationship for $\beta$ occurred mostly when $0.2 \le \beta \le 0.8$. The M-value also showed a relationship between the standard deviation and mean among the samples (Figure 2). Our proposed N-value was quite different, with no obvious relationship between the mean N-value and the standard deviation of the N-value (Figure 2). This property should result in the N-value being more appropriate for testing for differences in methylation levels based on a test of mean differences such as a t-test.

### Comparison of differential methylation analyses

We compared the results of our analyses to detect differential methylation, using the three summary measures, $\beta$, M, and N. We used t-tests to test for differential methylation on a per site basis, using all three summary measures. Using all samples, the analysis using $\beta$ resulted in 2,091 CpG sites at the 0.05 significance level (Table 1). Analysis using M identified 2,115 sites, and analyses of N identified 1,508 sites. The number of sites identified when case #5 is excluded from the analyses, changed with $\beta$ identifying 1,901 sites, M 2,137 sites, and N 1,549 sites. Although both $\beta$ and M identified a larger number

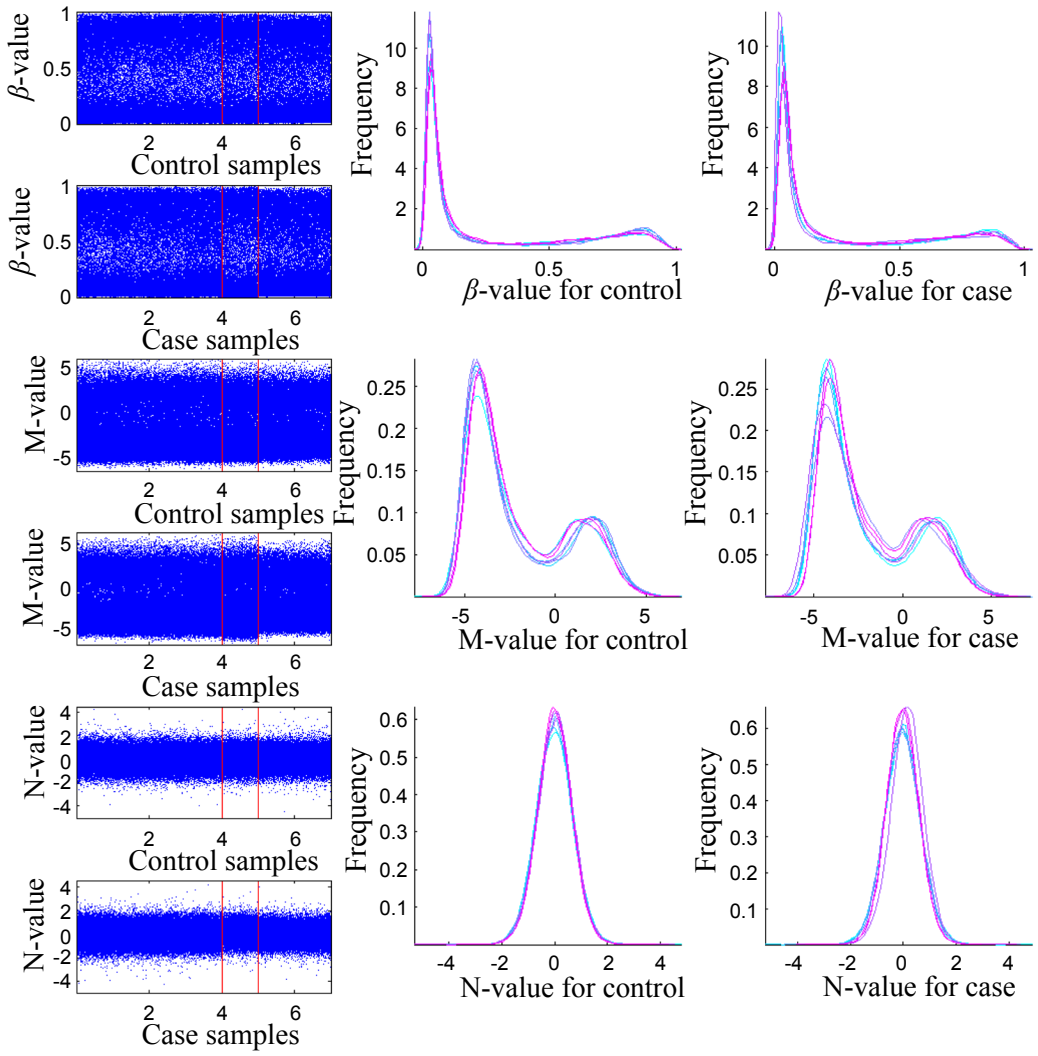


**Figure 1:** Distributions of $\beta$, M, and N for obesity data. The first column shows dot plots of the respective summary statistic of methylation levels for each sample, with controls and cases plotted separately. The middle and the right columns show kernel density estimates of each sample separately for cases and controls. Samples are color coded as follows: — sample #1, — sample #2, — sample #3, — sample #4, — sample #5, — sample #6, and — sample #7.

| Measure | | Site with p ≤ 0.05 | | Total |
|---|---|---|---|---|
| | (a) | (b) | (c) | |
| | with case #5 | regardless #5 | without case #5 | |
| β-value | 809 (29.85%) | 1,282 (47.31%) | 619 (22.84%) | 2,710 |
| M-value | 592 (21.69%) | 1,523 (55.81%) | 614 (22.50%) | 2,729 |
| N-value | 361 (18.90%) | 1,147 (60.05%) | 402 (21.05%) | 1,910 |

The differentially methylated CpG sites have been identified by two-sample t-test, with p ≤ 0.05 for 7 control samples and 7 case samples over 27,578 CpG sites. By including an outlying sample, case #5, the identified number of CpG sites is changed. Columns (a), (b) and (c) summarize the number of CpG sites identified to be differentially methylated under t-tests, with or without case #5: (a) t-tests identify corresponding sites only when case #5 is included; (b) t-test identify the sites regardless of inclusion of case #5; and (c) t-test identify the sites only when case #5 is excluded. The total column presents the number of identified CpG sites, with or without case #5 by three summary measures.

**Table 1**: Number of identified CpG sites as being differentially methylated in obesity data.

of candidate CpG sites under both analyses, these analyses showed the lowest consistency, as determined by the percent of the CpG sites that have p ≤ 0.05 in both analyses (Table 1). This improved consistency suggested that the N-value results are more stable, mainly due to the normalization that is inherent in N compared to both $\beta$ and M.

## Simulation study

Our simulations were based on simulating the mean signal and signal standard deviation based on our observations in the obesity study. We simulated data, with and without the sample heterogeneity

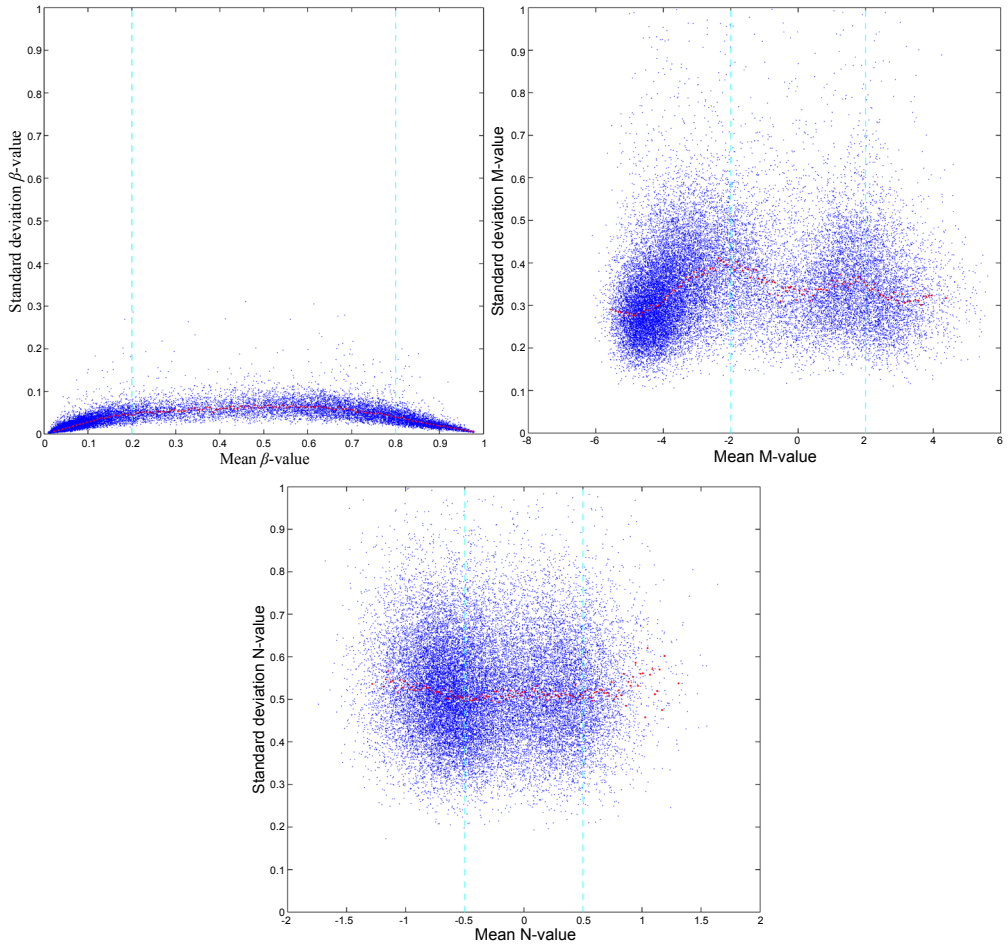


**Figure 2:** Relationship between means and standard deviations of summary measures of methylation levels among samples. β-value (top, left), M-value (top, right) and N-value (bottom). Light blue blocked lines distinguish upper, middle, and lower regions of values, which have been used in Du et al. [6]. The red line represents the Lowess fit.

| n | R | Without Heterogeneity (α=0.05) | | | Without Heterogeneity (α=0.01) | | |
|---|---|---|---|---|---|---|---|
| | | β | M | N | β | M | N |
| 10 | 5 | 38.77 (1.07) | 42.65 (0.95) | 50.32 (0.69) | 63.77 (2.04) | 69.02 (1.77) | 80.26 (1.02) |
| | 10 | 55.44 (0.92) | 59.42 (0.81) | 66.97 (0.63) | 77.55 (1.37) | 81.39 (1.13) | 89.11 (0.61) |
| | 15 | 67.13 (0.76) | 70.58 (0.67) | 76.96 (0.52) | 84.94 (0.97) | 87.71 (0.81) | 93.04 (0.41) |
| | 20 | 73.40 (0.70) | 76.60 (0.59) | 82.07 (0.42) | 88.27 (0.80) | 90.57 (0.63) | 94.74 (0.33) |
| 20 | 5 | 48.81 (0.82) | 50.71 (0.71) | 52.43 (0.69) | 79.26 (1.23) | 81.52 (1.02) | 84.56 (0.78) |
| | 10 | 65.37 (0.71) | 67.09 (0.65) | 68.72 (0.58) | 88.22 (0.72) | 89.68 (0.60) | 91.57 (0.46) |
| | 15 | 75.67 (0.56) | 77.10 (0.51) | 78.43 (0.47) | 92.47 (0.47) | 93.44 (0.39) | 94.72 (0.32) |
| | 20 | 80.89 (0.47) | 82.23 (0.44) | 83.43 (0.40) | 94.26 (0.40) | 95.11 (0.32) | 96.14 (0.25) |
| n | R | With Heterogeneity (α=0.05) | | | With Heterogeneity (α=0.01) | | |
| | | β | M | N | β | M | N |
| 10 | 5 | 40.51 (1.21) | 32.88 (1.27) | 46.41 (0.83) | 65.63 (2.38) | 62.37 (2.99) | 73.59 (1.40) |
| | 10 | 57.64 (1.09) | 49.11 (1.26) | 63.18 (0.71) | 79.35 (1.59) | 76.47 (2.12) | 84.57 (0.87) |
| | 15 | 68.59 (0.91) | 61.25 (1.09) | 74.11 (0.59) | 86.00 (1.11) | 83.98 (1.59) | 90.26 (0.57) |
| | 20 | 74.89 (0.76) | 67.72 (0.98) | 79.42 (0.49) | 89.14 (0.90) | 87.22 (1.27) | 92.43 (0.48) |
| 20 | 5 | 51.73 (0.90) | 48.49 (1.05) | 52.31 (0.68) | 81.26 (1.34) | 80.26 (1.52) | 84.23 (0.80) |
| | 10 | 68.04 (0.74) | 65.23 (0.86) | 68.59 (0.59) | 89.59 (0.78) | 88.94 (0.85) | 91.38 (0.48) |
| | 15 | 77.50 (0.61) | 75.41 (0.65) | 78.33 (0.48) | 93.25 (0.50) | 92.89 (0.56) | 94.61 (0.33) |
| | 20 | 82.55 (0.49) | 80.54 (0.58) | 83.36 (0.41) | 94.92 (0.40) | 94.52 (0.47) | 96.03 (0.26) |

The average (standard deviation) of the true positive percent in 1,000 simulations is shown. The significance level (Type I error) of the test is α, the per group sample size is n, and the percent CpG sites that are differentially methylated is denoted by R. Results shown are for the scenario in which the case standard deviation was shifted by 1

**Table 2:** True positive rates for all CpG sites in simulation study without and with sample heterogeneity.

that was observed in our data. Our proposed N-value had the highest right decision makings, say true positive rates, based on using t-tests, where the significance level 0.1 brought a higher true positive rate than the significance level 0.05 (Table 2). All three summary measures, $\beta$, M and N, were affected by the introduction of heterogeneity into the data. Interestingly, the true positive rate for $\beta$ increased when sample heterogeneity was added to the simulations, whereas M and N both exhibited decreased true positive rates. Overall, N tended to have higher true positive rates than $\beta$ and M, although there was no significant difference between $\beta$ and N for sample sizes of 20, when sample heterogeneity was present under significance level 0.05 (Table 2).

In addition to having better true positive rates, we also examined the consistency between results, with and without sample heterogeneity. All three summary measures showed increasing robustness with increasing sample size (Table 3). Regardless of sample size, our proposed N-value always showed greater consistency than $\beta$ and M, while $\beta$ always showed greater consistency than M.

Interestingly, while the true positive rate was affected by the proportion of sites differentially methy-lated (Table 2), the consistency between results for simulations with and without sample heterogeneity was slightly affected by the proportion of sites differentially methylated (Table 3).

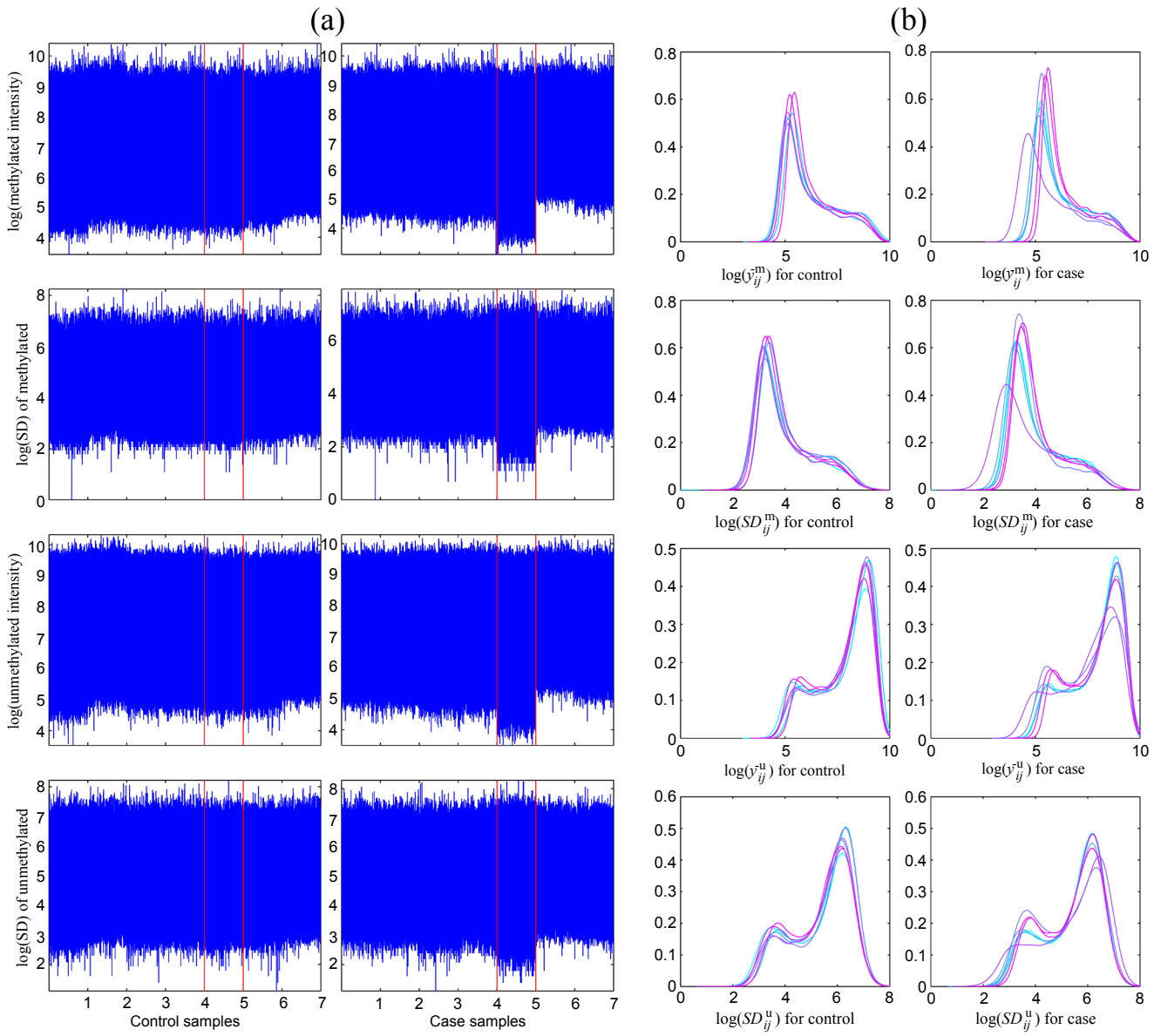


**Figure 3:** Distribution of signal intensities and standard deviations. (a) Each plot is a dot plot showing the means and standard deviation of the signal for each target site, with control samples on the left and case samples on the right. Red vertical lines indicate the data for either control sample #5 or case sample #5 (b) Kernel density estimates are shown that correspond to the dot plots in (a) Samples are color coded as follows:— sample #1, — sample #2, — sample #3, — sample #4, — sample #5, — sample #6, and — sample #7.

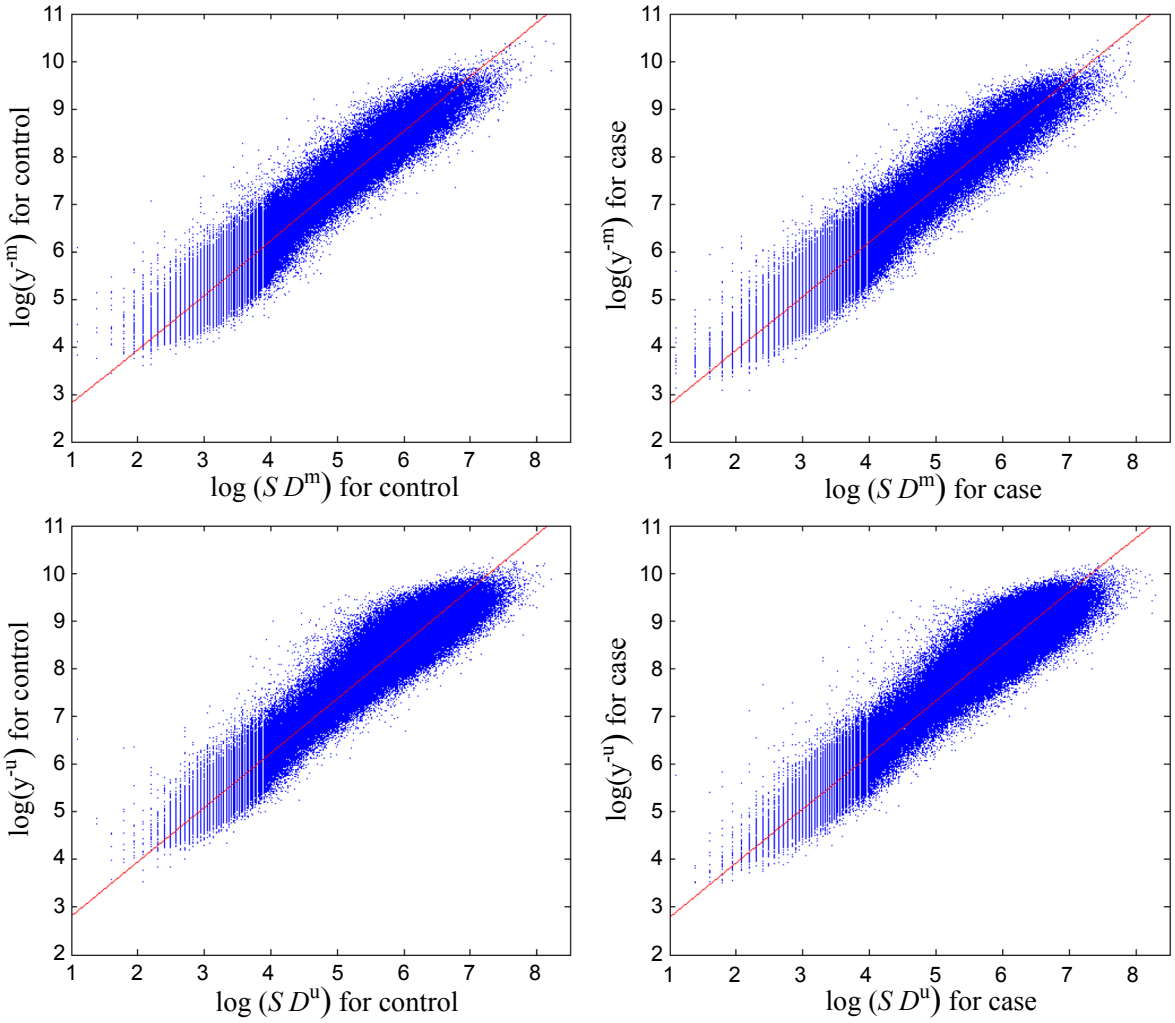


**Figure 4:** Regression for intensities on standard deviations. Each plot is a scatterplot of mean signal against the signal standard deviation, with the fit regression line included (p-value<0.01 and R2 ≥ 0.9).

| n | R | (α=0.05) | | | (α=0.01) | | |
|---|---|---|---|---|---|---|---|
| | | β | M | N | β | M | N |
| 10 | 5 | 47.94 (1.06) | 34.95 (0.88) | 68.05 (0.92) | 45.12 (1.76) | 23.23 (1.36) | 62.06 (1.55) |
| | 10 | 53.40 (0.93) | 35.48 (0.78) | 72.50 (0.79) | 49.78 (1.43) | 23.59 (1.02) | 63.03 (1.19) |
| | 15 | 56.29 (0.80) | 35.81 (0.68) | 76.31 (0.66) | 50.66 (1.33) | 23.55 (0.88) | 65.68 (1.00) |
| | 20 | 58.54 (0.73) | 35.26 (0.62) | 76.91 (0.59) | 51.64 (1.11) | 22.96 (0.79) | 64.49 (0.92) |
| 20 | 5 | 64.06 (0.90) | 55.51 (1.10) | 82.00 (0.68) | 67.72 (1.43) | 49.97 (1.57) | 90.46 (0.71) |
| | 10 | 71.30 (0.74) | 61.32 (0.99) | 87.49 (0.48) | 71.77 (1.02) | 51.61 (1.24) | 93.51 (0.48) |
| | 15 | 75.14 (0.58) | 64.53 (0.81) | 90.98 (0.36) | 72.16 (0.82) | 51.85 (0.97) | 94.97 (0.35) |
| | 20 | 77.83 (0.57) | 65.32 (0.81) | 92.69 (0.31) | 73.10 (0.73) | 50.63 (0.89) | 95.12 (0.32) |

The average (standard deviation) of the percent of sites with p ≤ 0.05, in both a simulation with and a simulation without sample heterogeneity is shown, summarized for 1,000 simulations. Tests have been performed under the given significance level α. The per group sample size is n, and the percent CpG sites that are differentially methylated is denoted by R. Results shown are for the scenario in which the case standard deviation was shifted by 1

**Table 3:** Consistency of test results between analyses with and without heterogeneity.

## Discussion

We proposed the N-value as a weighted logistic transformation of $\beta$, with the advantage that N also normalizes data across arrays. Using our existing obesity data to compare N with both $\beta$ and M, we showed that the use of N yielded more stable results. Although results were more stable with N, it produced fewer discoveries. The effect sizes in an experiment comparing obese subjects with matched controls are expected to be small, and the number of sites showing differential

methylation is likely to be small. The p-value in this case is more likely to be close to uniformly distributed. As such, the smaller number of discoveries with N in the obesity data may be more reflective of true signals. The potential increased accuracy of N relative to $\beta$ and M was confirmed in our simulations. Further, the simulations demonstrated that N produces results that are more robust to the type of heterogeneity that we observed. Of note is that analyses of M result in increased accuracy, relative to $\beta$ only when no sample heterogeneity was present, while $\beta$ was more robust to sample heterogeneity. Our results suggest that N should be preferred to both $\beta$ and M, with $\beta$ being preferred to M.

Our results were entirely based on data obtained using the Infinium I assay. Most of the probes on the 450K array are based on the Infinium II assay, which may suggest that our results are not applicable here. The distribution of $\beta$ is known to differ between Infinium I and Infinium II assays, and the distribution of the signal values will also differ between Infinium I and Infinium II. However, our proposed N-value only depends on separate methylated and unmethylated signals, in which there is a relationship between the signal standard deviation and the signal mean. Importantly, our method accounts for the uncertainty in mean signal, and appropriately adjusts for the variability in signal. Therefore, the N-value should show improved performance with the Infinium II assay as well.

Although we used linear regression to define the N-value here, any sort of regression relationship could be used, since the normalized signals were defined as residuals from the regression of the mean signal on signal standard deviation. The assumption about linear relationship between mean signal and signal standard deviation is one that can easily be examined with each data set, and the regression function adjusted appropriately. We have examined this relationship in preliminary data obtained from the 450K chip, and the relationship between mean signal and signal standard deviation was linear. Such results do suggest that the N-value is a potentially important summary statistic to be used in testing for differential methylation in methylome-wide association studies. Ultimately, the efficacy of different methods will need to be evaluated based on confirmatory biological findings based on analyses using the different summary measures.

## Acknowledgments

This work was supported by intramural funds from Georgia Regents University. The authors appreciate many conversations with Dr. Huidong Shi and for his feedback about our work, and are thankful to the editor and reviewers for comments to enhance the manuscript.

## References

1. Rakyan VK, Down TA, Balding DJ, Beck S (2011) Epigenome-wide association studies for common human diseases. Nat Rev Genet 12: 529-541.
2. Kulis M, Esteller M (2010) DNA methylation and cancer. Adv Genet 70: 27-56.
3. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, et al. (2011) High density DNA methylation array with single CpG site resolution. Genomics 98: 288-295.
4. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, et al. (2009) Genome-wide DNA methylation profiling using infinium assay. Epigenomics 1: 177-200.
5. Siegmund KD (2011) Statistical approaches for the analysis of DNA methylation microarray data. Hum Genet 129: 585-595.
6. Du P, Zhang X, Huang CC, Jafari N, Kibbe W, et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11: 587.
7. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, et al. (2010) Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res 20: 440-446.
8. Kim YJ, Yoon HY, Kim SK, Kim YW, Kim EJ, et al. (2011) EFEMP1 as a novel DNA methylation marker for prostate cancer: Array-based DNA methylation and expression profiling. Clin Cancer Res 17: 4523-4530.
9. Choufani S, Shapiro JS, Susiarjo M, Butcher DT, Grafodatskaya D, et al. (2011) A novel approach identifies new differentially methylated regions (DMRs) associated with imprinted genes. Genome Res 21: 465-476.
10. Hinoue T, Weisenberger DJ, Lange CPE, Shen H, Byun HM, et al. (2012) Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res 22: 271-282.
11. Sun Z, Chai H, Wu Y, White W, Donkena KV, et al. (2011) Batch effect correction for genome-wide methylation data with Illumina Infinium platform. BMC Med Genomics 4: 84.
12. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, et al. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol 12: R10.
13. Wang X, Zhu H, Snieder H, Su S, Munn D, et al. (2010) Obesity related methylation changes in DNA of peripheral blood leukocytes. BMC Med 8: 87.
14. Gutin B, Johnson MH, Humphries MC, Hatfield-Laube JL, Kapuku GK, et al. (2007) Relationship of visceral adiposity to cardiovascular disease risk factors in black and white teens. Obesity 15: 1029-1035.