Letter

Cell PRESS

'Leveling' the playing field for analyses of single-base resolution DNA methylomes

Matthew D. Schultz[1,2*], Robert J. Schmitz[2*], and Joseph R. Ecker[1,3]

[1] Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[2] Bioinformatics Program, University of California at San Diego, La Jolla, CA 92093, USA
[3] Howard Hughes Medical Institute, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA

Over the past few years the cost of DNA sequencing has plummeted while the numbers and lengths of sequencing reads have increased. This sequencing revolution has led to widespread adoption of methods to investigate genome-wide patterns of DNA methylation, collectively referred to as whole-genome bisulfite sequencing (WGBS). Single-base resolution DNA methylomes are now routinely being decoded by combining high-throughput sequencing with sodium bisulfite conversion, the gold-standard method for the detection of cytosine DNA methylation [1,2]. With increasing acquisition and analysis of DNA methylomes there is a growing need to reach a consensus on the definition(s) of the amount of methylation at a specific cytosine or region. 'Methylation level' is often poorly defined, and can vary significantly depending on the experimenter and the queries being addressed. Therefore, we propose a set of guidelines to be considered when analyzing 'methylation levels' from WGBS data.

Single-site methylation level

WGBS allows the interrogation of the methylation status at a single cytosine. This process uses sodium bisulfite to convert unmethylated cytosine to uracil and ultimately thymine via PCR [3]. These can then be detected by sequencing the converted product and mapping the data to a reference genome. Reads that contain a thymine where the reference genome contains a cytosine indicate that the reference cytosine is unmethylated, whereas reads that still retain a cytosine indicate that the reference cytosine is methylated. Although in a single cell a cytosine is either methylated or unmethylated, these experiments are forced to survey a population of cells because of the minimum quantity of input DNA required for sequencing and/or to survey cells that contain heterozygous regions. Consequently, these experiments yield a heterogeneous collection of sequencing reads, with some indicating that a specific cytosine is methylated and others indicating the same site is unmethylated. In the case of CG sites, it is generally safe to combine the read counts from both strands (i.e., count the symmetric sites as one unit) because the methylation between strands in this context is highly correlated. Often, a binomial test is used to determine if the observed methylation frequency is above the background expected from inefficiencies in the bisulfite conversion reaction and sequencing errors. When using this test, the amount of methylation at a given site is typically expressed as the ratio of reads with methylation (i.e., reads with a C at this position) out of the total number of reads covering the position (i.e., reads with a C or a T at this position) (Figure 1). We refer to this site-specific metric as the methylation level of the site.

Although this single-site quantity may be biologically relevant, researchers are often interested in the methylation levels over a region rather than at particular sites. There are a growing number of computer programs that identify differentially methylated regions (a source of variation in and of itself but which is not considered here due to space constraints), but once these regions are detected there are multiple methods to calculate methylation levels, often resulting in very different values even using the same WGBS data.

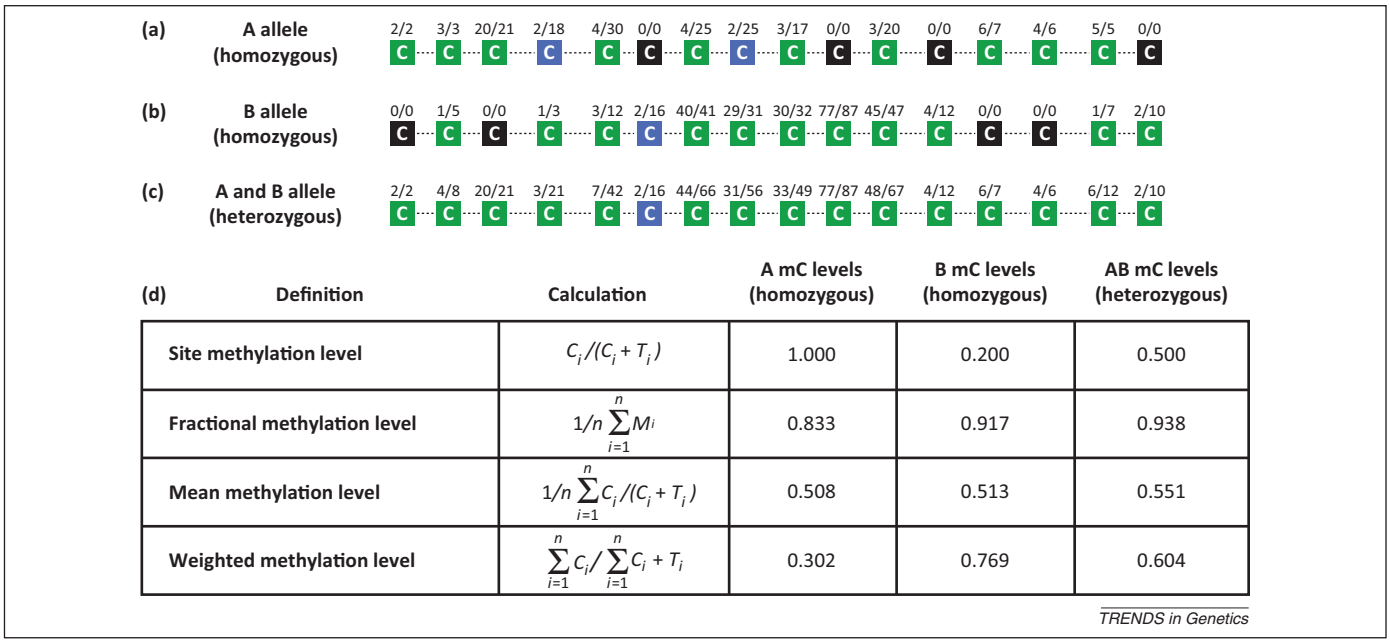
Fraction of methylated cytosines

The simplest way to combine this site-specific methylation information in a region is to calculate the fraction of cytosines that show a statistically significant amount of methylation (as determined above) (i.e., the fraction of methylated cytosines). This metric is useful if one is interested in the potential for methylation at sites in various regions because a significant binomial test indicates that at least one cell/allele in the population has a methylated cytosine in that region. As previously mentioned, the sequencing data at each site represent a survey of the methylation states across a population, and the fraction of methylated cytosines metric does not capture information about differences in the methylation level at each site. This may be important because a shift in the proportion of cells with methylation at particular sites could indicate a fundamental phenotypic change (e.g., in cancer [4] or development [5]). Furthermore, applying this method globally to heterozygous regions can be problematic because greater numbers of detected methylated cytosines will pass a binomial test if one of the alleles is methylated (Figure 1). This pitfall can be overcome by filtering sequencing reads for those that contain heterozygous genetic variants that can link methylation information to a particular allele.

Mean methylation level

One way to include additional information in the methylation level metric of a region is to take the arithmetic mean

Corresponding author: Ecker, J.R. (ecker@salk.edu)
Keywords: DNA methylation; bisulfite sequencing; methylation level; epigenomics.
* These authors contributed equally to this work.



(d)

Definition	Calculation	A mC levels (homozygous)	B mC levels (homozygous)	AB mC levels (heterozygous)
Site methylation level	$C_i/(C_i + T_i)$	1.000	0.200	0.500
Fractional methylation level	$1/n \sum_{i=1}^{n} M_i$	0.833	0.917	0.938
Mean methylation level	$1/n \sum_{i=1}^{n} C_i/(C_i + T_i)$	0.508	0.513	0.551
Weighted methylation level	$\sum_{i=1}^{n} C_i / \sum_{i=1}^{n} C_i + T_i$	0.302	0.769	0.604

TRENDS in Genetics

Figure 1. An example scenario of a methylated region and multiple methods for calculating 'methylation levels.' Examples of regions containing different profiles of DNA methylation for homozygous regions of the **(a)** A allele and **(b)** the B allele and **(d)** the heterozygote. Green 'C' squares indicate methylated cytosines, and blue 'C' squares indicate unmethylated cytosines as determined by a binomial test. Black 'C' squares indicate that no sequencing reads cover the particular cytosine. The fractions above each cytosine represents the number of reads with a cytosine divided by the total number cytosines and thymines for that position. **(d)** Using the regions in a–c, 'methylation levels' were calculated using each discussed method. C = read supporting methylated cytosine, T = read supporting unmethylated cytosine, i = position of cytosine, n = total number of cytosine positions, M = an indicator variable that is one when the position was identified as methylated by the binomial test. The results from the 2nd position starting from the left were used to calculate the 'site methylation level'. In addition, for any position containing a blue 'C', square the reads containing a cytosine were not included in the calculation of methylation levels as described in the text. Abbreviation: mC, methylcytosine.

of the methylation levels at sites within the region (mean methylation level) (Figure 1). Although an improvement over the fraction of methylated cytosines, this method does not take into account variable sequencing coverage across the sites in a region. Because each site contributes equally to the overall average, this method assumes that the information content at each site is also equal; however, more deeply sequenced sites will provide more accurate estimates of the mean methylation level in the region.

Weighted methylation level

Consequently, one may wish to weight the amount each site contributes to the level in a region by the sequencing depth at each site (weighted methylation level) (Figure 1). For example, imagine two sites in a region, one with 90 methylated reads from 100 total reads and another with one methylated read out of two total reads. The methylation level for this region will be very different depending on whether more weight is given to the first site or if equal weight is given to both sites (Figure 1).

Additional considerations and normalization methods

A key point in the calculation of both the mean and weighted methylation level is that although sites that were deemed unmethylated by the binomial test should still be included in these calculations, they should not contribute any methylated reads to the computation (Box 1). In other words, the number of methylated reads at a site that failed to pass the binomial test should be set to zero regardless of the number of methylated reads detected at that site. It is also important to note that all of the metrics described only consider cytosines within a specific region across many samples. One may be tempted to normalize the methylation levels of sites in a region by the size of that region (i.e., by base pairs), but this does not take into account the differences in base composition among different segments of the genome. Consequently, a significant difference in methylation may be caused simply by a difference in base composition (i.e., one region has fewer cytosines), which is generally uninformative.

Box 1. Additional considerations for calculating methylation levels

In some cases, when calculating mean or weighted methylation level, cytosine counts from positions that were not determined to be statistically significant by the binomial test are included. We do not include these counts because the purpose of the binomial test is to determine if there is an amount of methylation at a site that is above experimental noise. Consequently, if the test fails, cytosines found at that site should be assumed to be unmethylated, unconverted bases. If these sites are included, it will inflate the methylation level estimate. The magnitude of this change will depend on the size and coverage of the region as well as the number of unmethylated sites, but to be conservative we omit them. Admittedly, some sites that pass the binomial test will include these same unconverted cytosines, but once they have passed this test there is at least some methylation at the site. If there are concerns about unconverted cytosines, the estimated non-conversion rate can be subtracted from the mean or weighted methylation levels. Furthermore, because WGBS is strand-specific, we prefer to ignore reads supporting an A or G base call at cytosine positions because these represent sequencing errors and cannot accurately be used in the calculation. For these reasons, we feel that the C and T read counts at statistically significant sites represent our best estimate of the true methylation level.

Concluding remarks

We propose that the weighted methylation level should be the default metric for studying DNA methylomes because it is most broadly applicable. This is not to say that the alternative metrics described here are without value (e.g., as described above for fractional methylation level). Therefore, it is critical for scientists to define precisely the question they are attempting to answer with their DNA methylome analysis so that they can correctly compute a methylation level that captures the biology of interest and so that the results can be accurately evaluated.

Acknowledgments

R.J.S. was supported by a National Institutes of Health (NIH) postdoctoral fellowship. This work was supported by the Mary K. Chapman Foundation, the National Science Foundation (MCB-0929402 and MCB1122246), the NIH (U01-ES017166 and R01-MH094670), the Howard Hughes Medical Institute (HHMI) and the Gordon and Betty Moore foundation (GBMF) to J.R.E. J.R.E. is a HHMI–GBMF Investigator.

References

1 Cokus, S.J. *et al.* (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 452, 215–219
2 Lister, R. *et al.* (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133, 523–536
3 Clark, S.J. *et al.* (1994) High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.* 22, 2990–2997
4 Hansen, K.D. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* 43, 768–775
5 Lister, R. *et al.* (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* 471, 68–73

http://dx.doi.org/10.1016/j.tig.2012.10.012 Trends in Genetics, December 2012, Vol. 28, No. 12