MLML: Consistent simultaneous estimates of DNA methylation and hydroxymethylation

Jianghan Qu, Meng Zhou, Qiang Song, Elizabeth E. Hong and Andrew D. Smith

Molecular and Computational Biology, University of Southern California, Los Angeles, USA

ABSTRACT

Motivation: The two major epigenetic modifications of cytosines, 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), coexist with each other in a range of mammalian cell populations. Increasing evidence points to important roles of 5hmC in demethylation of 5mC and epigenomic regulation in development. Recently developed experimental methods allow direct single-base profiling of either 5hmC or 5mC. Meaningful analyses seem to require combining these experiments with bisulfite sequencing, but doing so naively produces inconsistent estimates of 5mC or 5hmC levels.

Results: We present a method to jointly model read counts from BS-seq, oxBS-seq and TAB-seq, providing simultaneous estimates of 5mC and 5hmC levels that are consistent across experiment types.

Availability: http://smithlab.usc.edu/plone/software/mlml-source-code/at_download/file

Contact: andrewds@usc.edu

Supplementary information: Supplementary material is available at Bioinformatics online.

1 INTRODUCTION

DNA methylation is an important epigenetic mark in mammals. In addition to the extensively studied 5mC modification, its oxidation product, 5hmC, has been observed at substantial levels in both somatic and embryonic stem cells (Tahiliani et al., 2009; Kriaucionis and Heintz, 2009). Recent studies of 5hmC in mouse TET knock-out models (Ito et al., 2010), mouse zygotic development (Iqbal et al., 2011), and multiple cell types (Globisch et al., 2010; Ito et al., 2011; Kinney et al., 2011; Sun et al., 2013), suggest that 5hmC is involved in epigenetic regulation.

The current most comprehensive and accurate method for profiling cytosine methylation is bisulfite sequencing (BS-seq). Treatment with sodium bisulfite converts unmethylated cytosines to uracils, but does not distinguish between 5mC and 5hmC (Huang et al., 2010), and consequently the yield of methylation from BS-seq is the sum of 5mC and 5hmC levels. Two recently developed techniques, oxidative bisulfite sequencing (oxBS-seq) (Booth et al., 2012) and Tet-Assisted Bisulfite sequencing (TAB-seq) (Yu et al., 2012), provide high throughput single-base resolution measurement of 5mC and 5hmC, respectively. Any two of BS-seq, TAB-seq or oxBS-seq can be combined to profile both the 5mC and 5hmC methyomes of a cell population, and especially when studying 5hmC, proper interpretation of results depends on having some estimate of the 5mC level. However, naive manipulation of read count frequencies from independent sequencing experiments often produces two kinds of “overshoot” problems in estimating 5mC and 5hmC levels. When combining BS-seq with TAB-seq, the 5mC level at a given CpG site can be estimated by subtracting the 5hmC level (TAB-seq) from the combined 5mC+5hmC level (BS-seq). The result can be negative, because of random sampling (or systematic error) in each experiment. Similarly, combining TAB-seq and oxBS-seq could lead to estimates of 5mC and 5hmC levels exceeding 100%. These overshoot sites may constitute a substantial proportion. In one dataset based on oxBS-seq technology, 17% of CpG sites captured by oxRRBS and RRBS experiments exhibited overshoot (Booth et al., 2012). To fully leverage the information in these data requires some method for making consistent estimates of 5mC and 5hmC levels.

We present Maximum Likelihood Methylation Levels (MLML) for simultaneous estimation of 5mC and 5hmC, combining data from any two of BS-seq, TAB-seq and oxBS-seq, or all three when available. Our estimates are consistent in that 5mC and 5hmC levels are non-negative, and never sum over one. In an important subset of cases, our estimates are not only consistent but also show significantly greater accuracy at sites with lower coverage.

2 METHODS

Each of BS-seq, TAB-seq and oxBS-seq provides some amount of information about both the 5mC and 5hmC levels. Our approach is to combine information from any pair or all three of these experiments, and arrive at maximum likelihood (ML) estimates for the 5mC and 5hmC levels. A similar method has been developed in the context of haplotype frequency estimation from pooled sequencing(Kessner et al., 2013). To explain our method we assume the data is from TAB-seq and BS-seq experiments for the same biological sample. The more general formulation is provided in supplementary information.

Focusing on an individual CpG site, let \( p_m \) denote the methylation level (a probability), \( p_h \) the hydroxymethylation, and \( p_u (= 1 - p_m - p_h) \) the level of unmethylated C. In the TAB-seq experiment, let \( h \) denote the number of C reads mapping over the CpG site, and let \( g \) denote the T reads mapping over the same CpG. The total reads covering the CpG site in the TAB-seq experiment is then \( h + g \). Similarly, let \( t \) denote the number of C reads mapping over the CpG site in the BS-seq experiment.
mapping over the site in the BS-seq experiment, while \( u \) denotes the number of \( T \) reads and the total reads covering the CpG in the BS-seq experiment is \( t + u \). If values for \( p_m \) and \( p_h \) are known, \( h \) and \( u \) are binomial random variables, i.e. \( h \sim \text{Bin}(h + t, p_h) \), and \( u \sim \text{Bin}(u + t, p_u) \):

\[
\begin{align*}
 f(h|p_h) & = \binom{h}{h} p_h^h (1 - p_h)^{t - h} \\
 f(u|p_m, p_h) & = \binom{u}{u} (1 - p_m - p_h)^u (p_m + p_h)^t.
\end{align*}
\]

Given observations of \( \{h, g, u, t\} \), when no overshoot would result we use the frequencies to estimate \( p = (p_m, p_h, p_u) \). In this case the frequencies directly give MLEs. At overshoot sites, we introduce latent variables and use expectation maximization (EM) to approximate the MLE for \( p \). Let \( t' (g') \) be the number of \( C (T) \) reads in BS-seq (TAB-seq) that correspond to 5mCs. Then \( t' - t' - (g - g') \) is the number of \( C (T) \) reads corresponding to 5hmC (unmethylated C). The complete data likelihood is then

\[
 L(p_m, p_h) = f(t', t - t', u + t + u, p) \times f(g', h, g - g'; h + g, p),
\]

where \( f(x, y, z; p) \) is a multinomial p.m.f. Estimates for \( p_u \) and \( p_m \) are then computed by EM algorithm to account for the latent \( t' \) and \( g' \) (see supplementary information). The ML estimates can be compared with binomial confidence intervals and show the corresponding frequency estimates if direct readouts (e.g. for 5hmC in the case of TAB-seq) are available. When estimates fail outside the specified confidence interval, sites are flagged as "strongly" inconsistent. An overabundance of such sites might suggest systematic error.

### 3 RESULTS

To understand the properties of our estimators and the frequency method, we used simulations with fixed coverage and precisely set levels for 5mC and 5hmC, assuming the experiments were BS-seq and TAB-seq. The case of BS-seq and \( \text{oxBS}-\text{seq} \) is symmetric with the estimates for \( p_h \) and \( p_m \) exchanged. For each valid combination of 5mC and 5hmC levels from \( \{0.1, 0.3, 0.5, 0.7\} \), we simulated from binomial distributions for both BS-seq and TAB-seq. Estimates for \( p_h \) and \( p_m \) were made using the ML method and the frequency method, which estimates \( p_h \) using \( h/(h + g) \) and \( p_m \) using \( \text{max}(0, t/(u + t) - h/(h + g)) \). The relative error \((|\hat{p} - p|/p) \) for both estimation methods was computed and then averaged over 100000 simulations for each parameter combination. The average estimation errors are presented in Table S1. Estimates of \( p_h \) are more accurate using MLML, especially at lower values of \( p_m \) and low coverage. For example, when the true values are \( p_m = 0.1, p_h = 0.1 \), the MLML reduces the average relative error by more than 23% at overshoot sites compared with frequency estimates when the coverage is 10X, and this reduction in error increases to 57% for such sites covered only 5X. The trend for errors of \( p_h \) estimates is shown in Fig. 1(a), indicating the accuracy advantage for MLML as a function of coverage. The simulation also revealed substantial amounts of overshoot sites under different 5mC and 5hmC level combinations, Fig. 1(b), supplementary tables.

### ACKNOWLEDGEMENT

We thank Fang Fang, Philip Uren, Benjamin Decato, and Jun Zhou for suggestions and testing of this software. We also thank an anonymous referee for suggesting the functionality of flagging sites that are so inconsistent as to suggest non-random error.

### Funding

This work was supported by US National Institutes of Health National Human Genome Research Institute grant R01 HG005238

### REFERENCES


![Fig. 1. Accuracy is improved at lower coverage using MLML. (BS-seq + TAB-seq).](image)