BMC
Bioinformatics

Estimation of the methylation pattern distribution from deep sequencing data

Peijie Lin[1], Sylvain Forêt[2], Susan R Wilson[1,3] and Conrad J Burden[1*]

Abstract

Background: Bisulphite sequencing enables the detection of cytosine methylation. The sequence of the methylation states of cytosines on any given read forms a methylation pattern that carries substantially more information than merely studying the average methylation level at individual positions. In order to understand better the complexity of DNA methylation landscapes in biological samples, it is important to study the diversity of these methylation patterns. However, the accurate quantification of methylation patterns is subject to sequencing errors and spurious signals due to incomplete bisulphite conversion of cytosines.

Results: A statistical model is developed which accounts for the distribution of DNA methylation patterns at any given locus. The model incorporates the effects of sequencing errors and spurious reads, and enables estimation of the true underlying distribution of methylation patterns.

Conclusions: Calculation of the estimated distribution over methylation patterns is implemented in the R Bioconductor package `MPFE`. Source code and documentation of the package are also available for download at http://bioconductor.org/packages/3.0/bioc/html/MPFE.html.

Keywords: DNA methylation, Bisulfite sequencing, DNA methylation patterns, Epiallele

Background

Epigenetic regulations are involved in a broad range of biological processes, including development, tissue homeostasis, learning and memory, as well as various diseases such as obesity and cancer [1-3].

DNA methylation is one of the best studied epigenetic molecular mechanisms. It consists of the addition of a methyl group to the cytosine residues (C) of a DNA molecule. In animals, DNA methylation usually takes place in the CpG context: cytosines followed by a guanine (G) residue.

DNA methylation modulates gene expression through a variety of mechanisms. In vertebrates, methylation in the promoter region usually has a repressive effect on transcription initiation. By contrast, methylation of gene bodies is generally associated with an active transcriptional state and has been shown to play an important role in the control of alternative splicing [4,5].

The diverse and subtle effects of DNA methylation enable a given genome to produce different phenotypic outputs as part of a developmental program or in response to environmental factors. This has fundamental implications at the organismal level, where DNA methylation plays an important role in phenotypic plasticity [6]. This is also important at the cellular level to create diverse cell types, tissues and organs all based on the same genome. DNA methylation patterns can thus change from one cell type to another or within a cell under different conditions [7].

The diversity of methylation patterns in a sample can be studied with a single base pair resolution using the bisulphite sequencing technique [8]. When DNA is treated with bisulphite, the unmethylated cytosines are converted to uracils with a high (albeit not complete) efficiency, whereas the methylated cytosines remain as cytosines. A library is prepared from the bisulphite treated DNA by fragmenting to lengths of approximately 200 bp and PCR amplified. During this amplification process, uracils are replicated as thymines (T). The DNA library is then sequenced and the resulting reads are mapped to a reference. Within each read, CpG dinucleotides which have

*Correspondence: Conrad.Burden@anu.edu.au
[1]Mathematical Sciences Institute, Australian National University, Canberra ACT 2601, Australia
Full list of author information is available at the end of the article

Lin *et al. BMC Bioinformatics* (2015) 16:145

Page 2 of 11

been converted to TpG are recognised as unmethylated, and unconverted CpG dinucleotides are recognised as methylated.

A common type of analysis carried out with this type of data is to estimate, for each cytosine, the global methylation level of a sample, namely, the average methylation level across all DNA strands in all the cells represented in that sample. This is usually done by correcting for the fact that bisulphite conversion is not a complete reaction (for instance, see [4,9]).

However, this type of analysis only considers the methylation level at individual positions and is oblivious to the fact that cytosines present on a given read represent a broader snapshot of the methylation landscape on a particular strand of DNA. One can therefore gain a significantly deeper insight into the complexity of DNA methylation landscapes by reconstituting the methylation patterns that are physically present on the same sequence, and thus come from the same cell. Each read can be seen as containing a small number of binary labelled CpG sites: 1 for methylated, 0 for unmethylated and represents the methylation pattern of a given strand of DNA in one particular cell. This approach is of particular interest when studying complex biological samples that contain a mixture of cell types, for instance a tumour or a whole insect brain. It gives a much more detailed picture about the diversity of DNA methylation in a sample than simply looking at the methylation level at each position.

The study of the diversity and dynamics of DNA methylation patterns has recently generated a great deal of interest as it allows assessment of the methylation status in subsets of cells within a sample. Such studies have provided novel insights into the role of DNA methylation in cell differentiation and reprogramming, and the evolution of tumours (for instance [10-12]). It has also been proposed that the DNA methylation patterns present in heterogeneous samples could be effective tumour-specific biomarkers [13].

The approach of looking at methylation patterns instead of individual CpG sites can be hampered by the lack of sequencing coverage depth. For reasons of cost, most whole genome bisulphite sequencing studies have a mean genomic coverage in the 10-100 range. If the underlying sample contains hundreds of methylation patterns, they cannot be sampled representatively. Consequently, studies looking at DNA methylation diversity have focussed on short patterns (around 4 CpGs) [12].

As an alternative, in order to reach a high coverage, researchers have often focussed on specific loci, either by sequencing PCR amplicons of bisulphite converted DNA, or by reduced representation bisulphite sequencing, or by using a capture assay (for review see [14]). In this paper, we focus on the analysis of amplicon data, but the statistical framework introduced here could be applied to data

extracted from other types of sequencing approaches such as reduced representation and capture-based sequencing.

The problem of determining methylation patterns in the original sample from the observed data is non trivial. The purpose of this paper is to infer formally the probability distribution over all possible methylation profiles defined by the population of methylomes at a given locus. We show that taking into account the effect of the non-systematic errors of incomplete bisulphite conversion, as well as sequencing error, is critical as this removes a large number of spurious patterns present in the raw reads.

Results and discussion
Synthetic data
The probability distribution over methylation patterns at a given locus from from bisulphite sequencing data is estimated via an algorithm described below in the Methods section. The algorithm is implemented as an R Bioconductor [15] package MPFE (for **M**ethylation **P**atterns **F**requency **E**stimation).

As the true distribution over methylation patterns is always unknown in the laboratory, we have constructed a number of synthetic data sets to test the effectiveness of the algorithm. For a locus with n CpG sites we prescribe a true distribution θ_k, $k = 1, \ldots, 2^n$, over the 2^n possible methylation patterns. We simulate from these patterns a set of N_{read} initial reads, each labelled with its 'true' methylation pattern, and then redistribute the reads among patterns according to the statistical model described in the Methods section to simulate errors due to incomplete bisulphite conversion and sequencing errors.

The rates of conversion and sequencing error can be estimated by spiking-in non-methylated DNA of known sequence, or by looking at positions known in advance not to be methylated [16].

Incomplete conversion is specified by a non-conversion rate parameter ϵ equal to the probability that an unmethylated cytosine will fail to convert. Its value is typically of order 10^{-2}, and will be assumed in our simulations to be estimated independently from the rate of conversion of non-CpG cytosines. For instance, it could be calculated based on sequences known a priori not to be methylated. The sequencing error rates are specified by a vector $\eta = (\eta_1, \ldots, \eta_n)$, where $\eta_s \sim O(10^{-2})$ is the probability that site s is methylated but registers as unmethylated and vice versa. These two sources of error can generate spurious methylation patterns, that is, patterns for which the observed read count y_k is non-zero when the true frequency of the pattern, θ_k, is zero. The aim is to assess the ability of the algorithm to recover the true distribution θ_k from a set of observed read counts y_k, and, more specifically, determine whether spurious patterns can be identified.

Lin *et al. BMC Bioinformatics* (2015) 16:145

Page 3 of 11

The algorithm can be run in two possible modes: a slow (exact) mode, which performs a complete calculation over all possible 2^n methylation patterns, and an (approximate, default) fast mode which assumes $\theta_k = 0$ for any pattern for which the observed read count y_k is zero. Because of the exponential growth in the number of patterns, we find in general that use of the function in the slow mode becomes computationally prohibitive for $n > 8$. However, there is very little difference in the computed results between the two modes (see below). Furthermore, the number of realised patterns tends to be relatively small and the performance of the function is adequate even for large n.

To illustrate the concept of the algorithms, we begin with a small synthetic dataset corresponding to a locus with $n = 6$ CpG sites, and a distribution of pattern abundances θ_i contrived to mimic a combination of high and low abundance patterns as observed in a real dataset. Table 1 and Figure 1 show the results of creating the dataset assuming a non-conversion rate $\epsilon = 0.005$, sequencing errors $\eta = (0.008, 0.006, 0.006, 0.006, 0.006, 0.008)$ to model a signal that degrades towards the ends of the reads, and a total number of reads $N_{\text{read}} = 2000$ to generate read counts y_i. These data are then analysed using the same ϵ and η values in both the slow and fast modes to generate estimated pattern abundances $\hat{\theta}_i$.

For this synthetic dataset we find that for any pattern for which the 'observed' read count y_i is zero, the 'true' abundance θ_i and the estimated abundance $\hat{\theta}_i$ from the slow-mode version of the algorithm are also zero. For this reason, results are only shown for patterns for which the read count y_k is non-zero. For completeness, the plot produced for all 64 patterns by running our package in the slow mode is given in Additional file 1.

We observe that for almost all patterns the estimated distribution $\hat{\theta}_k$ is closer to the true distribution θ_k than the naive read proportion y_k/N_{read}, the only exception being a slight shift in the wrong direction for the pattern $m_{64}(111111)$.

We observe also that the algorithm has the effect of skewing the distribution of reads away from low-frequency, more highly methylated patterns towards the totally unmethylated pattern, thus reducing the false discovery rate. This algorithmic effect will generally be the case for any realistic dataset (see Additional file 2). Furthermore, our implementation of the algorithm registers whether, for any given pattern k, the estimate $\hat{\theta}_k$ is identically zero, that is, it makes a call as to whether the pattern is present or absent. In this example, no real pattern is classified as spurious. The algorithm correctly identifies 14 out of the 18 spurious patterns in the slow mode, and 13 out of the 18 in the fast mode. For the one extra pattern which was missed by the fast mode, namely pattern

Table 1 Comparison between estimates $\hat{\theta}_i$ of the true methylation distribution θ_i calculated with the slow and fast implementations of our algorithm for the dataset of Figure 1

Patterns	θ_i	y_i	y_i/N_{read}	$\hat{\theta}_i$ (slow)	$\hat{\theta}_i$ (fast)
000000	0.50	907	0.4535	0.4813	0.4812
000001	0.00	15	0.0075	0.0013	0.0013
000010	0.00	8	0.0040	0	0
000100	0.00	7	0.0035	0	0
001000	0.04	99	0.0495	0.0466	0.0466
001001	0.00	4	0.0020	0.0016	0.0014
001010	0.00	1	0.0005	0	0
001100	0.00	1	0.0005	0	0
010000	0.00	6	0.0030	0	0
011000	0.00	2	0.0010	0.0007	0.0004
011011	0.00	1	0.0005	0.0005	0.0002
011101	0.00	2	0.0010	0	0
011111	0.03	63	0.0315	0.0308	0.0306
100000	0.03	77	0.0385	0.0328	0.0329
101101	0.00	1	0.0005	0	0
101111	0.00	1	0.0005	0	0
110000	0.00	1	0.0005	0	0.0000
110111	0.00	2	0.0010	0	0
111001	0.00	1	0.0005	0	0
111011	0.00	1	0.0005	0	0
111100	0.00	3	0.0015	0	0
111101	0.20	393	0.1965	0.2001	0.2013
111110	0.00	3	0.0015	0	0
111111	0.20	401	0.2005	0.2044	0.2040

y_i is the number of observed reads for pattern i and N_{read} is the total number of reads.

110000, the estimated proportion $\hat{\theta}_i(\text{fast}) < 10^{-4}$ is very low.

In general, when the number of CpG sites n in an amplicon is large, it is expected that only a small faction of the 2^n possible patterns will be present. Furthermore, it is rare for a true pattern with $\theta_k > 0$ to have zero counts as a result of incomplete bisulphite conversion or sequencing errors. For instance, in the the current example none of the patterns with a positive true frequency θ_k had zero reads, and for the remaining patterns there is no substantial difference between the two implementations of the algorithm. From now on, the discussion focusses on the fast implementation.

Figure 2 and Additional file 3 show the results of applying the fast implementation of the algorithm to synthetic data modelled on biological data from a PCR amplicon in the honey bee *Apis mellifera* genome (see Biological data

Lin *et al. BMC Bioinformatics* (2015) 16:145

Page 4 of 11
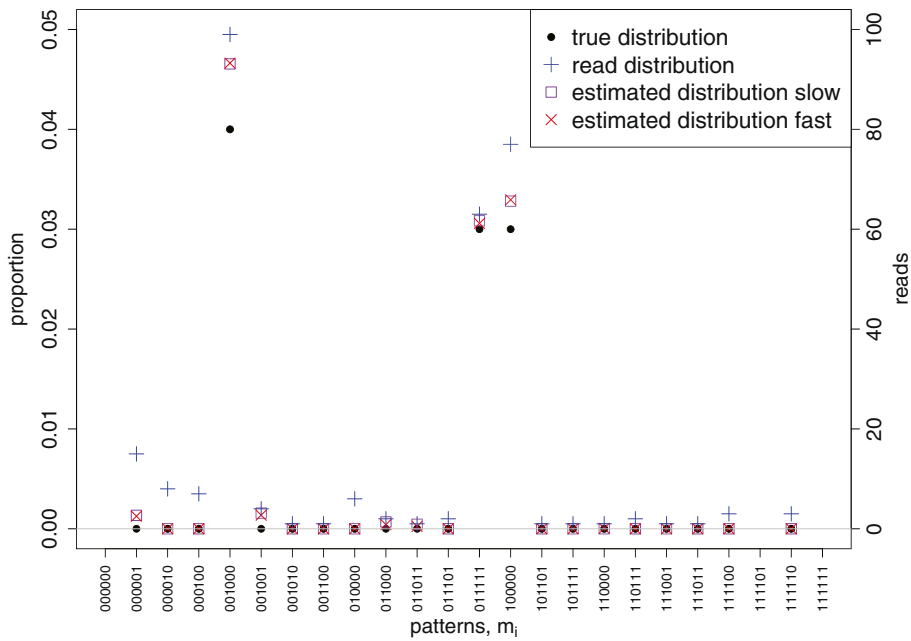
Figure 1 The 'true' distribution θ_k (black dots), observed read counts y_k (blue pluses) and estimated distribution $\hat{\theta}_k$ from the slow (purple squares) and fast (red crosses) implementations of our algorithm over methylation patterns for synthetic data with $n = 6$ CpG sites. The non-conversion rate is set to be $\epsilon = 0.005$, the sequencing errors are set to be $\eta = (0.008, 0.006, 0.006, 0.006, 0.006, 0.008)$, the total number of reads is $N_{\text{read}} = 2000$. Only patterns with non-zero reads are shown and methylation patterns are ordered lexicographically. Data for the three patterns (000000, 111101 and 111111) which are beyond the range of the plot can be found in Table 1.
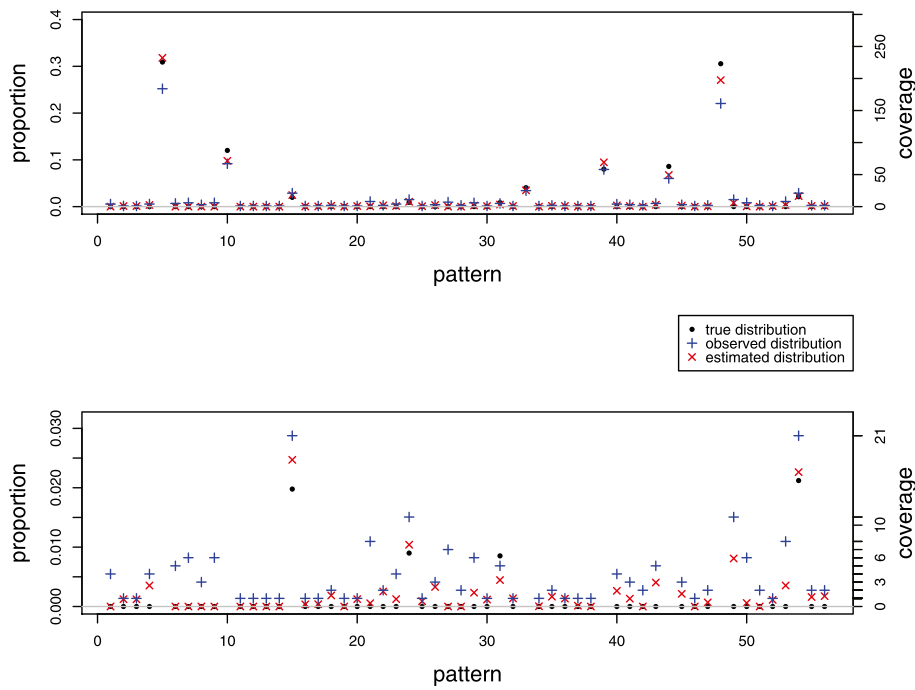
Figure 2 Comparison between the true distribution θ_i, the simulated read distribution y_i/N_{read}, and the estimated distribution $\hat{\theta}_i$ for a synthetic dataset based on amplicon data. Parameter values are $n = 9$ CpG sites, total number of reads $N_{\text{read}} = 730$, non-conversion parameter $\epsilon = 0.01$, and sequencing error rate $\eta = 0.02$. The lower plot is an expanded version of the upper plot showing only patterns with low frequencies.

Lin *et al. BMC Bioinformatics* (2015) 16:145

Page 5 of 11

section). To obtain the dataset, the function was applied once to a biological dataset of $N_{\text{read}} = 730$ reads from an amplicon corresponding to a locus with $n = 9$ CpG sites. To maintain a similar number of non-zero reads the 'true' distribution of the synthetic dataset was taken to be

$$\theta_i = \begin{cases} 0 & \text{if} \quad \hat{\theta}_i^{\text{init}} < 0.005, \\ \hat{\theta}_i^{\text{init}} & \text{otherwise,} \end{cases}$$

where $\hat{\theta}_i^{\text{init}}$ is the result of the initial application of the algorithm. Here the non-conversion parameter is $\epsilon = 0.01$, and the sequencing error rate $\eta = 0.02$ is taken to be uniform across all CpG sites.

Of the 46 spurious patterns in this dataset, 16 are correctly identified and no false identifications of spurious patterns are made. For the 30 failed identifications, the program estimates a lower estimate $\hat{\theta}_i$ than the read proportion y_i/N_{read}.

Given that the MPFE algorithm requires as input an estimate of the non-conversion rate ϵ and read error rate η, it is important to gauge the sensitivity of pattern frequency estimates to these parameters. Figure 3 shows the variation in estimated pattern abundances for the dataset of Figure 1 as the values of ϵ and η used by MPFE are perturbed from the 'true' values used to construct the synthetic data. We observe that the estimated abundances are relatively insensitive to ϵ for the actually occurring patterns, but drop rapidly with increasing ϵ for most of the spurious patterns. It is perhaps not surprising that $\hat{\theta}_i$ will be sensitive to the ϵ for those spurious patterns for which $\hat{\theta}_i > 0$: As explained in Additional file 2, increasing ϵ has the effect of skewing the distribution towards the less methylated patterns, which in general tend to be the more abundantly occurring patterns. A relevant point here is that the correction for non-conversion almost invariably moves the estimates $\hat{\theta}_i$ in the correct direction from the raw data for the spurious patterns, while leaving even the low-frequency non-spurious patterns almost unchnaged. Also, not surprisingly, the estimates $\hat{\theta}_i$ are sensitive to the read error rate η for both spurious and non-spurious patterns.

Detecting spurious patterns

The effectiveness of our algorithm in identifying spurious patterns can be further gauged by introducing a threshold on the estimated pattern abundance. In the following definitions we only consider methylation patterns with non-zero reads y_i.

We set a threshold $K \in [0, 1]$ as a binary classifier and, using the estimated distribution $\hat{\theta}_i$ as a test statistic, declare pattern i to be spurious when $\hat{\theta}_i \leq K$. Patterns are defined to be true or false positives or negatives according to the rules

TP if $\hat{\theta}_i \leq K$ and $\theta_i = 0$,   FP if $\hat{\theta}_i \leq K$ and $\theta_i > 0$,

TN if $\hat{\theta}_i > K$ and $\theta_i > 0$,   FN if $\hat{\theta}_i > K$ and $\theta_i = 0$.

True positive rates (TPR) and false positive rates (FPR) are defined in the usual way as

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}.$$

An analogous set of definitions applies using the raw data count proportions y_i/N_{read} as a test statistic. Note that this classification does not constitute classical hypothesis testing per se, as the threshold does not easily translate to a p-value under a well-defined null hypothesis: patterns are not independent, and specifying $\theta_i = 0$ for a given pattern does not determine the distribution of the estimator $\hat{\theta}_i$ without further assumptions about the remaining $2^n - 1$ patterns.

Figure 4A shows the TPR curves for the data of Figure 1. It shows that using $\hat{\theta}_i$ results in a clear improvement in detecting which methylation patterns are likely to be a spurious artefact of incomplete conversion and reading error. The FPR curves for both test statistics are constantly zero for the same threshold range in the TPR graph.

Figure 4B shows the TPR curves for the synthetic data based on biological data from an amplicon analysed in Figure 2. Again we observe a clear improvement in detecting which methylation patterns are likely to be a spurious artefact of incomplete methylation and reading error.
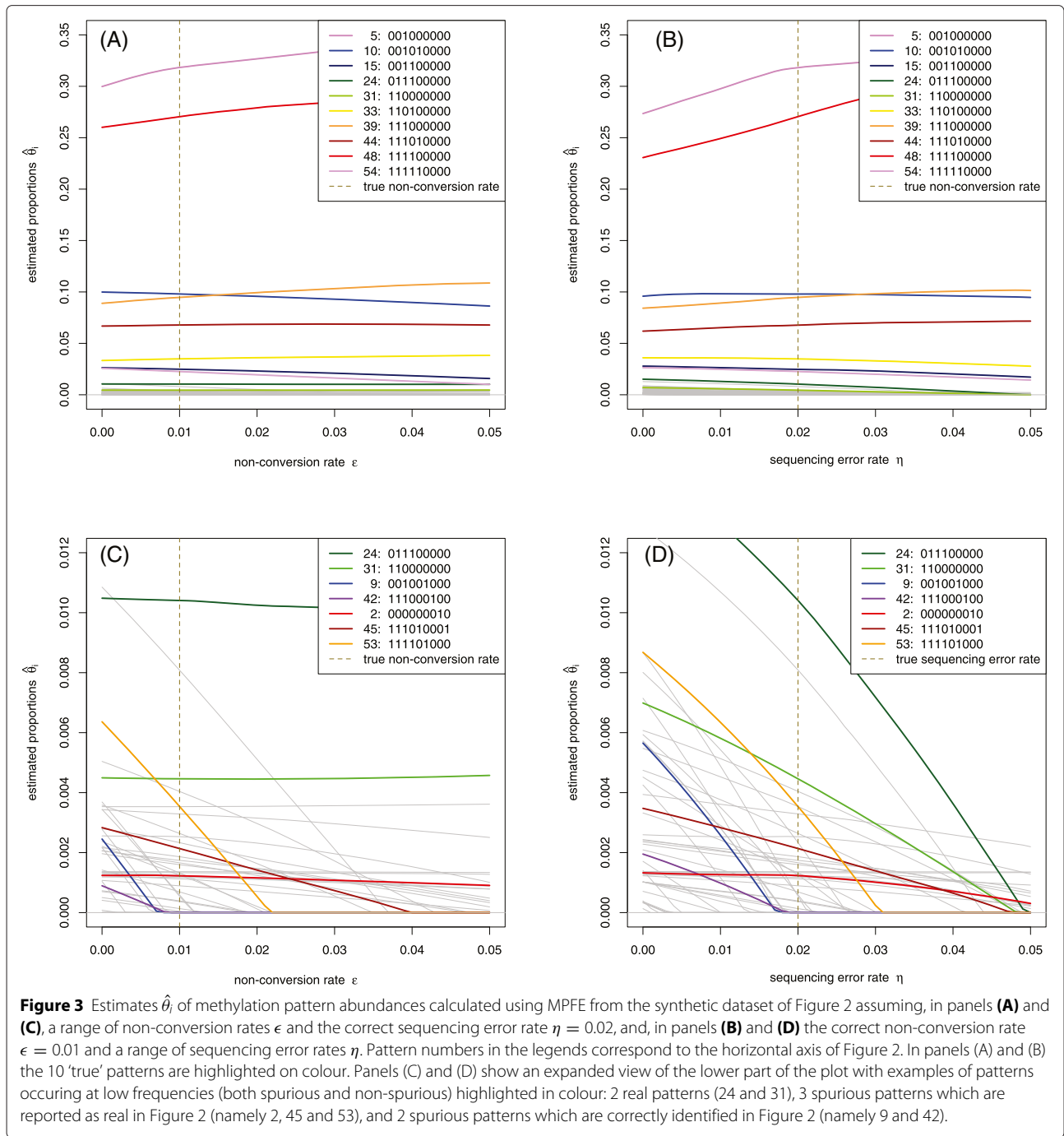
Biological data

Amplicons were obtained as described in [4]. Briefly, genomic DNA was extracted from brains of adult honeybee workers and treated with sodium bisulphite. A region of gene GB17113 (gene ID: 724724, a 6-phosphofructokinase) was then amplified by PCR and sequenced using the 454 technology. No ethical approval was required for this work.

Below we apply the fast algorithm to two examples from this dataset assuming a non-conversion rate of $\epsilon = 0.01$ and a global sequencing error rate of $\eta = 0.02$. The first example is shown in Figure 5 and Table 2. The parameter values are $n = 8$ CpG sites, 36 patterns with non-zero reads, and a total number of reads $N_{\text{read}} = 1793$.

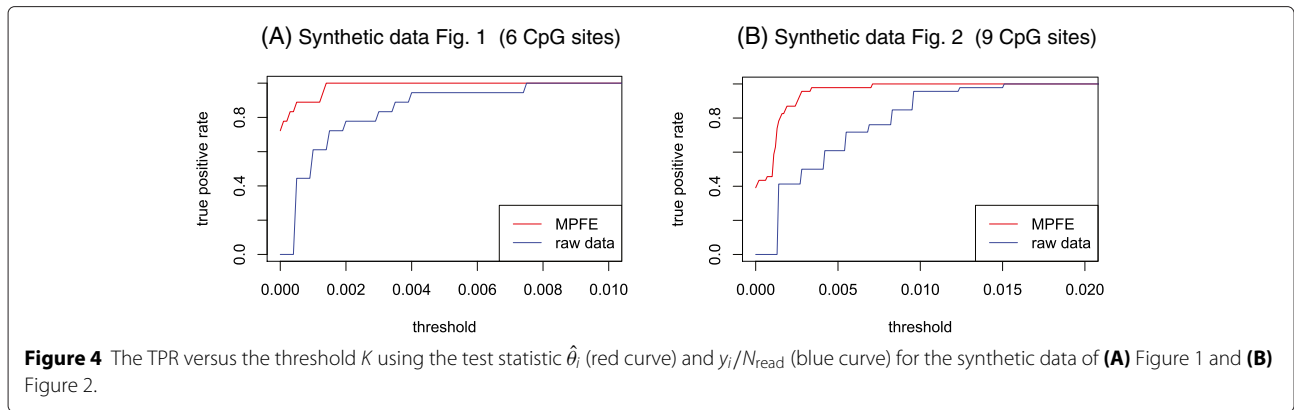There are several observations:

(i) 18 patterns (50%) are identified as spurious;
(ii) there are 11 patterns with only 1 read - our algorithm calls 9 of them as spurious, while predicts the other 2 patterns (m_{30} and m_{31}) to exist;
(iii) patterns m_3, m_4, m_9 and m_{11} with > 30 reads each has a read proportion $y_i/N_{\text{read}} \approx 2\%$, but are called as spurious.

Lin *et al. BMC Bioinformatics* (2015) 16:145

Page 6 of 11

Figure 3 Estimates $\hat{\theta}_i$ of methylation pattern abundances calculated using MPFE from the synthetic dataset of Figure 2 assuming, in panels **(A)** and **(C)**, a range of non-conversion rates ϵ and the correct sequencing error rate $\eta = 0.02$, and, in panels **(B)** and **(D)** the correct non-conversion rate $\epsilon = 0.01$ and a range of sequencing error rates η. Pattern numbers in the legends correspond to the horizontal axis of Figure 2. In panels (A) and (B) the 10 'true' patterns are highlighted on colour. Panels (C) and (D) show an expanded view of the lower part of the plot with examples of patterns occuring at low frequencies (both spurious and non-spurious) highlighted in colour: 2 real patterns (24 and 31), 3 spurious patterns which are reported as real in Figure 2 (namely 2, 45 and 53), and 2 spurious patterns which are correctly identified in Figure 2 (namely 9 and 42).

Observations (i) and (ii) can be explained by the fact that the edit distance between the two patterns covered by a single read and any pattern observed to be highly abundant renders it unlikely that these patterns have arisen through sequencing errors or incomplete conversion. Observation (iii) arises because the spurious patterns with > 30 reads are just one sequencing error or one incomplete conversion away from the most abundant pattern, $m_0(00000000)$.

Figure 6 and Additional file 4 show the second example, with $n = 14$ CpG sites and 160 patterns with non-zero reads. The total number of reads is $N_{\text{read}} = 2347$. In this case 47 patterns are called as spurious. Note that these data indicate that the methylation statuses of neighouring CpG loci are correlated. This correlation was also observed by Lyco et al. [4] and has motivated our approach that has the advantage of being able to accommodate neighbouring spatial correlations
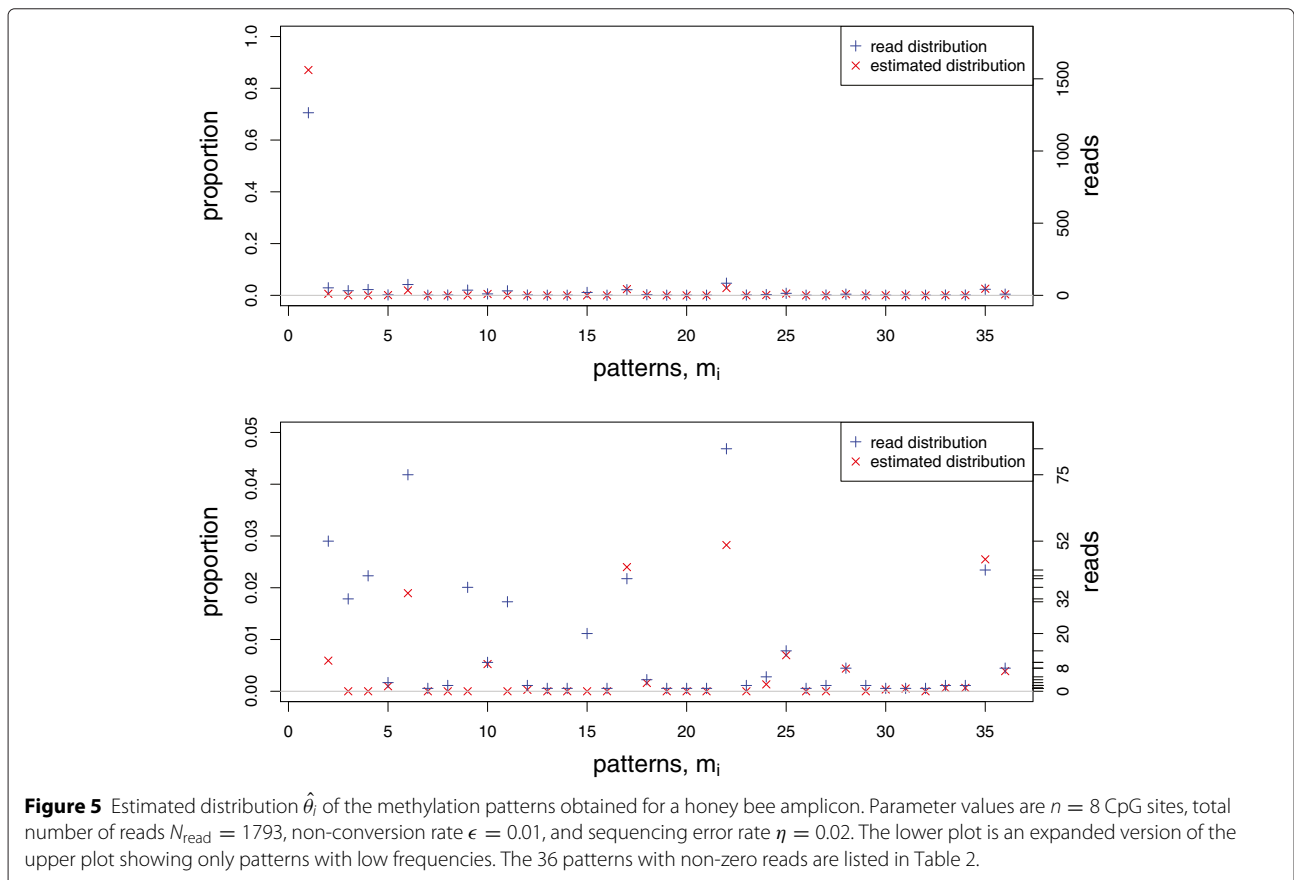
Lin *et al. BMC Bioinformatics* (2015) 16:145

Page 7 of 11

Figure 4 The TPR versus the threshold K using the test statistic $\hat{\theta}_i$ (red curve) and y_i/N_{read} (blue curve) for the synthetic data of **(A)** Figure 1 and **(B)** Figure 2.

between CpG positions without imposing any a priori structure.

Conclusions
We have developed an algorithm for estimating the true distribution of methylation patterns in at a genomic locus containing n CpG sites. The algorithm, based on a constrained multinomial model, accounts for statistical variation due to incomplete bisulphite conversion and sequencing errors. This method can be readily applied to

methylation patterns extracted from amplicon sequencing data analysed with with software such as BiQ-Analyser [17]. Further work would be required to extract methylation patterns counts from whole-genome sequencing data.

The analysis differs from previous treatments in that the estimated distribution is a joint probability distribution over patterns which preserves maximal information pertaining to interaction between different CpG sites, as opposed to a pointwise measure of methylation at each

Figure 5 Estimated distribution $\hat{\theta}_i$ of the methylation patterns obtained for a honey bee amplicon. Parameter values are $n = 8$ CpG sites, total number of reads $N_{\text{read}} = 1793$, non-conversion rate $\epsilon = 0.01$, and sequencing error rate $\eta = 0.02$. The lower plot is an expanded version of the upper plot showing only patterns with low frequencies. The 36 patterns with non-zero reads are listed in Table 2.

Lin *et al. BMC Bioinformatics* (2015) 16:145

Page 8 of 11

Table 2 Estimated methylation distribution $\hat{\theta}_i$ for the data of Figure 5

i	m_i	y_i	y_i/N_{read}	$\hat{\theta}_i$
1	00000000	1265	0.7055	0.8706
2	00000001	52	0.0290	0.0059
3	00000010	32	0.0178	0
4	00000100	40	0.0223	0
5	00000110	3	0.0017	0.0010
6	00001000	75	0.0418	0.0189
7	00001001	1	0.0006	0
8	00001010	2	0.0011	0
9	00010000	36	0.0201	0
10	00010100	10	0.0056	0.0053
11	00100000	31	0.0173	0
12	00100001	2	0.0011	0.0003
13	00101000	1	0.0006	0
14	00110000	1	0.0006	0
15	01000000	20	0.0112	0
16	01000001	1	0.0006	0
17	01100000	39	0.0218	0.0240
18	01100001	4	0.0022	0.0016
19	01101000	1	0.0006	0
20	01110000	1	0.0006	0
21	01110100	1	0.0006	0
22	10000000	84	0.0468	0.0282
23	10000001	2	0.0011	0
24	10000010	5	0.0028	0.0013
25	10000100	14	0.0078	0.0070
26	10001000	1	0.0006	0
27	10010000	2	0.0011	0
28	10010100	8	0.0045	0.0044
29	10100000	2	0.0011	0
30	11000100	1	0.0006	0.0003
31	11010000	1	0.0006	0.0005
32	11100000	1	0.0006	0
33	11100100	2	0.0011	0.0006
34	11110000	2	0.0011	0.0007
35	11110100	42	0.0234	0.0255
36	11110110	8	0.0045	0.0039

The 36 patterns with non-zero reads are labelled m_i, $i = 1, \ldots, 36$ and y_i are the observed read counts.

site. A pointwise methylation estimate can, of course, be recovered from our estimated distribution as a marginal distribution. The algorithm is implemented as the R Bioconductor package `MPFE`.

Numerical experiments with realistic synthetic data indicate that the algorithm is able to identify the majority of the spurious observed methylation patterns, that is, patterns which are not present in the original library but are observed in the reads because of incomplete bisulphite conversion or sequencing errors. In general, our estimates are closer to the true distribution than the naive estimates given by the relative proportion of observed read counts for almost all patterns in each simulation (see Figures 1, 2, and Table 1).

Application of the algorithm to biological data consisting of bisulphite treated amplicon reads for a honeybee genomic sequence predicts that a correspondingly high proportion of observed methylation patterns in real data may indeed be spurious. However, our results also reveal an important number of real methylation patterns in this biological sample. This complexity of the methylation landscape is virtually undetectable when one only considers position-wise methylation level, but becomes apparent through our method.

In the future, the statistical method presented here could be extended and refined to deal with systematic sources of errors (for instance non-homogeneous bisulfite conversion rates, differential amplification of different epialleles) and to extract information from incomplete data (reads where one or more positions of the pattern are missing, which are currently ignored).

Methods

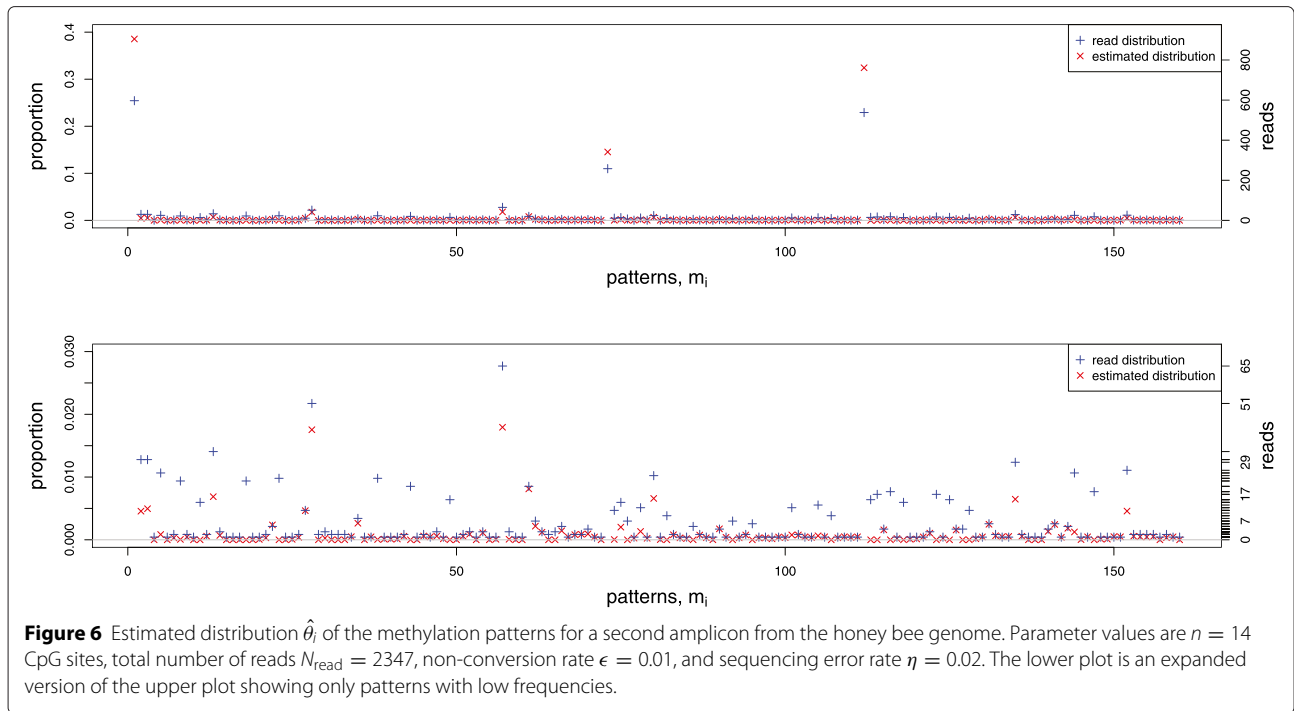### Statistical model of bisulphite sequencing

We take as the starting point of our statistical model a population of methylomes restricted to a locus containing n CpG sites. Each member of the population within a given class is represented by a vector of non-independent binary valued random variables $\mathbf{K} = (K_1, \ldots, K_n)$, where each $K_s \in \{0, 1\}$ labels the methylation state (0 for unmethylated, 1 for methylated) at the s-th CpG site at this locus. The population defines a methylation profile represented by the probability distribution of realising the pattern $\mathbf{k} = (k_1, \ldots, k_n)$ in a read randomly chosen from the population:

$$\theta_{\mathbf{k}} = \text{Prob}\,(\mathbf{K} = \mathbf{k}), \quad \mathbf{k} \in \{0, 1\}^n. \tag{1}$$

For convenience, from here on we will label the possible methylation patterns by the integers $k = 1, \ldots, 2^n$, and set $\theta_{\mathbf{k}} = \theta_k$, where $k - 1$ is the integer whose binary representation is the methylation pattern $\mathbf{k}$.

Our aim is to estimate the distribution θ_k representing the relative abundance of methylation pattern k from high throughput sequencing data consisting of a set of integer valued read counts Y_k. In a typical experiment, the number n of CpG sites in an amplicon may be up $O(10^2)$, and

Lin *et al. BMC Bioinformatics* (2015) 16:145

Page 9 of 11

Figure 6 Estimated distribution $\hat{\theta}_i$ of the methylation patterns for a second amplicon from the honey bee genome. Parameter values are $n = 14$ CpG sites, total number of reads $N_{\text{read}} = 2347$, non-conversion rate $\epsilon = 0.01$, and sequencing error rate $\eta = 0.02$. The lower plot is an expanded version of the upper plot showing only patterns with low frequencies.

the total number of read counts $N_{\text{read}} = \sum_{k=1}^{2^n} Y_k$ may be up to $O(10^6)$.

The model takes into account two sources of error. First, the bisulphite conversion of unmethylated cytosine to uracil is not 100% efficient. There is a probability ϵ that an unmethylated CpG site will register as being methylated, where $\epsilon \sim O(10^{-2})$ can be estimated from the cytosines known not to be methylated (mitochondrial genome, chloroplastic genome, spike-ins, etc). The second type of error is caused by sequencing. For many applications this may be assumed for practical purposes to be site independent. However, to allow for effects such as degradation of the read quality towards the ends of the reads, we will assume there is a site-dependent probability $\eta_s \sim O(10^{-2})$ that if site s is unmethylated it will register as methylated and vice versa. It follows that if the true methylation pattern of any given read is **K**, but the read registers as being pattern **L**, then

$$\text{Prob}(L = \ell | K = k) = M_{k\ell}, \quad k, \ell = 1, \dots 2^n, \qquad (2)$$

where, as above, we adopt the conventon that $k - 1$ is the integer whose binary representation is the methylation pattern $\mathbf{k} = (k_1, \dots, k_n)$ and $\ell - 1$ is the integer whose binary representation is the methylation pattern $\mathbf{l} = (\ell_1, \dots, \ell_n)$, and

$$M_{k\ell} = (E_1)_{k_1\ell_1}(E_2)_{k_2\ell_2} \dots (E_n)_{k_n\ell_n}, \qquad (3)$$

where, for each $s = 1, \dots n$,

$$E_s = \begin{pmatrix} 1-\epsilon & \epsilon \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1-\eta_s & \eta_s \\ \eta_s & 1-\eta_s \end{pmatrix}$$
$$= \begin{pmatrix} 1-\epsilon-\eta_s+2\epsilon\eta_s & \epsilon+\eta_s-2\epsilon\eta_s \\ \eta_s & 1-\eta_s \end{pmatrix}. \qquad (4)$$

is a 2×2 matrix whose rows and columns are labelled 0 (for unmethylated) and 1 (for methylated). In fact the matrix M whose elements are $M_{k\ell}$ is the Kronecker product $E_1 \otimes \dots \otimes E_n$.

Eqs. (1) and (2) imply that the probability that a random read will be the pattern ℓ is

$$\text{Prob}(L = \ell) = \sum_{k=1}^{2^n} \theta_k M_{k\ell} = \phi_\ell, \qquad (5)$$

say. Assuming each read to be independent, for an experiment with a given total number of reads N_{read} the observed set of read counts represented by the random variable Y_ℓ has a multinomial distribution:

$$\text{Prob}(Y_\ell = y_\ell | \phi) = \frac{N_{\text{read}}!}{y_1! \, y_2! \dots y_{2^n}!} \phi_1^{y_1} \dots \phi_{2^n}^{y_{2^n}}. \qquad (6)$$

In a recent applications note, [9] develop a statistical model for the distribution of the number of reads which register as being methylated in a pooled set of bisulphite-sequencing reads from CpG sites in a given region of a genome. Their model is mathematically equivalent to the $n = 1$ version of the above model, and as such can be simplified to a single binomial distribution (see Additional file 2).

Lin *et al. BMC Bioinformatics* (2015) 16:145

Page 10 of 11

Parameter estimation

The parameters of the distribution over methylation profiles, θ_k, are estimated by maximising the log likelihood:

$$L(\theta|y) = \sum_{k=1}^{2^n} y_k \log \left(\sum_{j=1}^{2^n} \theta_j M_{jk} \right), \qquad (7)$$

subject to the constraint that $(\theta_1, \ldots, \theta_{2^n})$ lies in the $(2^n - 1)$-dimensional simplex

$$S = \left\{ \theta : \sum_{k=1}^{2^n} \theta_k = 1, \theta_k \geq 0 \right\}. \qquad (8)$$

One may be tempted to use the usual formula, $\hat{\phi}_\ell = y_\ell/N_{\text{read}}$, for the maximum likelihood estimate of multinomial parameters, and simply invert the matrix M to recover $\hat{\theta}_k$. However this will not work for any realistic data because the matrix M shrinks the simplex to a smaller volume. In practice many of the y_ℓ are zero, which leads to a naive estimate $\hat{\phi}_\ell$ on the boundary of the unshrunken simplex in ϕ-space, and this boundary is not included in the shrunken simplex (see Figure 7 for the $n = 2$ case, in which the simplex is tetrahedron).

Instead we maximise the log likelihood over the allowed domain numerically by using the R function `constrOptim()`. Unfortunately the performance of this function becomes prohibitively slow for $n \gtrsim 8$ as the dimensionality of the parameter space grows exponentially. However, we have noticed in numerical simulations that if the observed counts y_k are zero for a subset of the possible patterns, the corresponding estimates $\hat{\theta}_k$ are zero (or in rare cases, very close to zero) for the same subset. Thus we have implemented an algorithm which only searches over that part of the boundary of S constrained by $\theta_k = 0$ for those k such that $y_k = 0$. The algorithm remains reasonably efficient on a standard desktop computer provided the number of observed methylation patterns does not exceed about 200. This allows analysis of most realistic datasets while still addressing the biologically relevant question of identifying spurious methylation patterns which are the result of incomplete methylation.

In Additional file 2 we prove that both the exact maximum likelihood estimate, and the aproximate maximum likelihood estimate over reduced part of the boundary described above are unique. However, in both cases the function `constrOptim()` is of finite accuracy in locating the maximum of the log-likelihood. Thus, if the located maximum $\hat{\theta}$ is close to the boundary of the simplex, our algorithm also calculates the value of the log-likelihood at several nearby points on the boundary. If this results in a log-likelihood bigger than or equal to the maximum reported by `constrOptim()`, the appropriate point on the boundary is taken as the maximum likelihood estimate, and those patterns m_k for which $\hat{\theta}_k \equiv 0$ are reported as being spurious reads.

Additional file 2 also contains an argument that for realistic datasets and parameter values, the estimated distribution $\hat{\theta}_i$ is in general skewed towards less-methlyated states relative to the naïve estimate y_i/n_{read}, and that the set of patterns reported to be present is a subset of (or the same set as) the set of patterns naïvely observed. These observations hold rigorously for any dataset if the algorithm is run with the read-error rate η_s set to zero for all sites s.
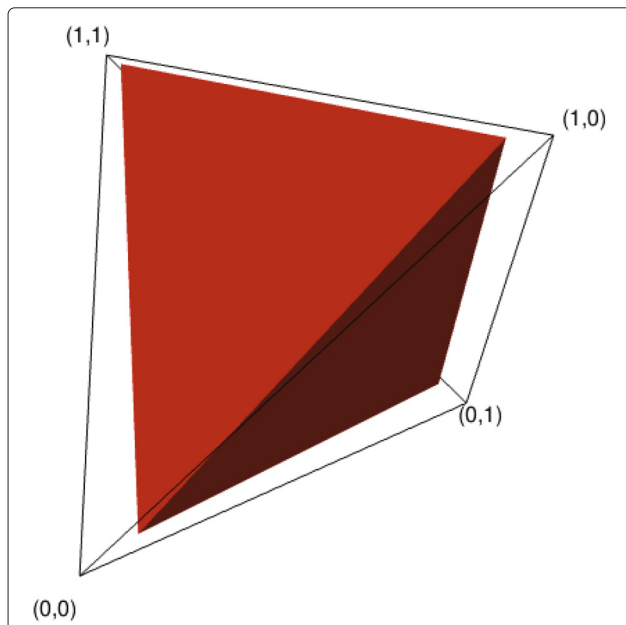
Figure 7 Boundary of the allowed simplex Eq. (8) for the parameters θ, (black wire frame) and corresponding shrunken simplex containing allowed values of ϕ (red tetrahedron) for the case $n = 2$ CpG sites, $\epsilon = 0.05$, $\eta_s = 0.02$. Numbers at the corners are the $2^n = 4$ possible methylation patterns).

Additional files

Additional file 1: Estimated pattern abundances using the exact algorithm. Estimates $\hat{\theta}_i$ calculated with the exact, slow implementations of our algorithm for the synthetic dataset of Figure 1. Methylation patterns are labelled lexicographically from $m_1 = 000000$ to $m_{64} = 111111$. Data for patterns $m_1, m_{62} = 111101$, and m_{64} are beyond the range of the plot, but are listed in Table 1.

Additional file 2: Technical details. Analysis of the relationship of the statistical model used by Akman et al. [9] to our statistical model, proof of uniqueness of maximum likelihood estimate and argument that the number of reported methylation patterns is reduced by the MPFE algorithm.

Additional file 3: Table of results for the second synthetic data example. True and estimated methylation distributions θ_i and $\hat{\theta}_i$ for the data of Figure 2. The 56 patterns with non-zero reads are labelled by their binary representation and y_i are the observed read counts.

Lin *et al. BMC Bioinformatics* (2015) 16:145

Page 11 of 11

Additional file 4: Table of results for the second biological data example. Estimated methylation distribution $\hat{\theta}_i$ for the data of Figure 6. The 160 patterns with non-zero reads are labelled by their binary reresentation and y_i are the observed read counts.

Abbreviations

CpG: Cytosine-phosphate-guanine dinucleotide; DNA: Deoxyribonucleic acid; FN: False negative; FP: False positive; FPR: False positive rate; PCR: Polymerase chain reaction; TN: True negative; TpG: Thymine-phosphate-guanine dinucleotide; TPR: True positive rate.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SF, CJB, SRW conceived and designed the project. SF prepared the datasets. PL, CJB developed the mathematical and statistical model, carried out the analysis and developed the software. PL, SF, CJB wrote the manuscript. All authors read and approved the final manuscript.

Author details

[1]Mathematical Sciences Institute, Australian National University, Canberra ACT 2601, Australia. [2]Research School of Biology, Australian National University, Canberra ACT 2601, Australia. [3]School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2052, Australia.

References

1. Cantone I, Fisher AG. Epigenetic programming and reprogramming during development. Nat Struct Mol Biol. 2013;20(3):282–9.
2. Day JJ, Sweatt JD. Dna methylation and memory formation. Nat Neurosci. 2010;13(11):1319–23.
3. Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. Nature. 2007;447(7143):433–40.
4. Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, Maleszka R. The honey bee epigenomes: differential methylation of brain DNA in queens and workers. PLoS Biol. 2010;8(11):1000506.
5. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, et al. Ctcf-promoted rna polymerase ii pausing links dna methylation to splicing. Nature. 2011;479(7371):74–9.
6. Kucharski R, Maleszka J, Foret S, Maleszka R. Nutritional control of reproductive status in honeybees via dna methylation. Science. 2008;319(5871):1827–30.
7. Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell stem cell. 2010;6(5):479–91.
8. Clark SJ, Harrison J, Paul CL, Frommer M. High sensitivity mapping of methylated cytosines. Nucleic Acids Res. 1994;22(15):2990–7.
9. Akman K, Haaf T, Gravina S, Vijg J, Tresch A. Genome-wide quantitative analysis of DNA methylation from bisulphite sequencing data. Bioinformatics. 2014;30(13):1933–4.
10. He J, Sun X, Shao X, Liang L, Xie H. Dmeas: Dna methylation entropy analysis software. Bioinformatics. 2013;29(16):2044–5. doi:10.1093/bioinformatics/btt332.
11. Li S, Garrett-Bakelman F, Perl AE, Luger SM, Zhang C, To BL, et al. Dynamic evolution of clonal epialleles revealed by methclone. Genome Biol. 2014;15(9):472. doi:10.1186/s13059-014-0472-5.
12. Shao X, Zhang C, Sun MA, Lu X, Xie H. Deciphering the heterogeneity in dna methylation patterns during stem cell differentiation and reprogramming. BMC Genomics. 2014;15:978. doi:10.1186/1471-2164-15-978.
13. Mikeska T, Candiloro ILM, Dobrovic A. The implications of heterogeneous dna methylation for the accurate quantification of methylation. Epigenomics. 2010;2(4):561–73. doi:10.2217/epi.10.32.
14. Lee EJ, Luo J, Wilson JM, Shi H. Analyzing the cancer methylome through targeted bisulfite sequencing. Cancer Lett. 2013;340(2):171–8. doi:10.1016/j.canlet.2012.10.040.
15. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. 2004;5:80.
16. Chen GG, Diallo AB, Poujol R, Nagy C, Staffa A, Vaillancourt K, et al. Bisqc: an operational pipeline for multiplexed bisulfite sequencing. BMC Genomics. 2014;15:290. doi:10.1186/1471-2164-15-290.
17. Lutsik P, Feuerbach L, Arand J, Lengauer T, Walter J, Bock C. Biq analyzer ht: locus-specific analysis of dna methylation by high-throughput bisulfite sequencing. Nucleic Acids Res. 2011;39(Web Server issue):551–6. doi:10.1093/nar/gkr312.