Comparison with Akman et al. (2014)

Akman et al. [1] have developed the Bioconductor package BEAT (BS-Seq Epimutation Analysis Toolkit) for quantitative analysis of DNA methylation patterns from bisulphite sequencing data. The basis of their software is a statistical model for the distribution of the number of reads which register as being methylated from a pooled set of bisulphite-sequencing reads from CpG sites in a given region of a genome. The distribution is derived in terms of the total number of pooled reads from the region, a false positive rate and a false negative rate representing incorrect conversions or non-conversions, and an underlying methylation rate which is assumed to be common to all CpG sites within the region. Their purpose is to estimate the common underlying methylation rate for the region. Their model does not deal with the distribution over methylation patterns.

We argue that because of the assumption that all CpG sites within the pooled region share common values for the distributional parameters, their model is mathematically equivalent to the \( n = 1 \) CpG site version of the model set out in Section 3.1 of the Methods section of the current paper. More importantly, the binomial mixture model arrived at by Akman et al. by repeated application of the law of total probability is nothing more than a single binomial distribution. Once this is recognised, it is a trivial matter to estimate the underlying methylation rate from maximum likelihood without the need for the Bayesian analysis employed by Akman et al.

Table 1 lists the correspondence between the notation of Akman et al. and that used in the current paper. If we represent the number of reads registering as being methylated by the random variable \( K \), then Eq. (3) of the supplementary material of Akman et al. states that

\[
K \sim \sum_{j=0}^{n} \sum_{m=0}^{k} \text{Bin}(m; j, 1 - p_{-}) \text{Bin}(k - m; n - j, p_{+}) \text{Bin}(j; n, r).
\]
| Number of CpG reads | Akman et al. | Lin et al. |
|---------------------|-------------|------------|
| Number of reads registering as methylated | \( n \) | \( N_{\text{read}} \) |
| Underlying methylation rate | \( k \) | \( y \) |
| False positive rate | \( p_+ \) | \( \epsilon + \eta - 2\epsilon \eta \) |
| False negative rate | \( p_- \) | \( \eta \) |

Table 1: Correspondence between the notation of Akman et al. and that used in the current paper. The rates \( p_+ \) and \( p_- \) are the off-diagonal elements of the matrix Eq. (4) in the main paper.

From Eqs. (6) and (5) of the current paper we claim that this is equivalent to a single binomial distribution

\[
K \sim \text{Bin}(k; n, \phi),
\]

where

\[
\phi = r + p_+ - r(p_+ + p_-).
\]

The equivalence is demonstrated numerically for a specific case by the following R code:

```r
> n <- 20
> r <- 0.3
> pPlus <- 0.1
> pMinus <- 0.2
> #
> probAkman <- array(dim=n+1)
> for(k in 0:n){
+   probA <- 0
+   for(j in 0:n){
+     for(m in 0:k){
+       probA <- probA +
+         dbinom(m, size=j, prob=1-pMinus)*
+         dbinom(k-m, size=n-j, prob=pPlus)*
+         dbinom(j, size=n, prob=r)
+     }
+   }
+   probAkman[k+1] <- probA
+ }
> #
> phi <- r + pPlus - r*(pPlus + pMinus)
> probLin <- dbinom(0:n, size=n, prob=phi)
> #
> print(data.frame(Akman=probAkman, Lin=probLin,
+  relDifference=2*(probAkman - probLin)/
+  (probAkman + probLin), row.names=0:n))
```
Uniqueness of the maximum likelihood estimate of $\theta_1, \ldots \theta_{2^n}$

We demonstrate the uniqueness of the maximum likelihood estimator

$$\hat{\theta} = \arg\max_{\theta_1, \ldots, \theta_{2^n}} \sum_{k=1}^{2^n} y_k \log \left( \sum_{j=1}^{2^n} \theta_j M_{jk} \right),$$  \hspace{1cm} (1)$$

subject to the constraints

$$\sum_{k=1}^{2^n} \hat{\theta}_k = 1, \quad \hat{\theta}_k \geq 0,$$  \hspace{1cm} (2)$$

where $y_1, \ldots, y_{2^n}$ are the observed read counts and the matrix $M$ is defined by Equation (3) of the main paper. An equivalent formulation of this problem is to determine

$$\hat{\theta}_k = \sum_{j=1}^{2^n} \hat{\phi}_j (M^{-1})_{jk},$$  \hspace{1cm} (3)$$

where $\hat{\phi}$ is the argmax of the likelihood function

$$L(\phi) = \sum_{k=1}^{2^n} y_k \log \phi_k,$$  \hspace{1cm} (4)$$
subject to the constraints

$$\sum_{k=1}^{2^n} \hat{\phi}_k = 1, \quad \text{Eq. (5)}$$

which follows from Eq. (2) because $M$ has the properties of a Markov transition matrix, and

$$\sum_{j=1}^{2^n} \hat{\phi}_j (M^{-1})_{jk} \geq 0, \quad k = 1, \ldots, 2^n. \quad \text{Eq. (6)}$$

Note that $M^{-1}$ exists provided $\epsilon \neq 1$ and $\eta_s \neq \frac{1}{2}$ (see Eq. (4) of the main paper), which will hold in any realistic situation. An illustrated example of the constraint on $\phi$ for the case $n = 2$ is shown in Figure 8 of the main paper.

Suppose, without loss of generality, that the observed read counts are such that

$$y_1, \ldots, y_l > 0, \quad y_l+1, \ldots, y_{2^n} = 0, \quad \text{Eq. (7)}$$

for some $0 \leq l \leq 2^n$. Then $L(\phi)$ is independent of $\phi_{l+1}, \ldots, \phi_{2^n}$ and

$$\frac{\partial^2 L}{\partial \phi_j \partial \phi_k} = -\delta_{jk} \frac{y_j}{\phi_j^2} \quad \text{Eq. (8)}$$

is a diagonal matrix whose first $l$ diagonal elements are strictly negative, implying that $L(\phi)$ is a strictly concave function of $\phi_1, \ldots, \phi_l$. Furthermore the set

$$S = \{\phi_1, \ldots, \phi_l | \text{Eqs. (5) and (6) are satisfied} \} \quad \text{Eq. (9)}$$

is easily shown to be convex, i.e. if $u, v \in S$, then $\lambda u + (1 - \lambda)v \in S$ for all $0 \leq \lambda \leq 1$. It follows that

$$(\hat{\phi}_1, \ldots, \hat{\phi}_l) = \arg \max_{(\phi_1, \ldots, \phi_l) \in S} L(\phi) \quad \text{Eq. (10)}$$

is unique. Furthermore, $\hat{\phi} = (\hat{\phi}_1, \ldots, \hat{\phi}_{2^n})$ for any choice of $(\hat{\phi}_{l+1}, \ldots, \hat{\phi}_{2^n})$ will also achieve the maximum of $L(\phi)$, but will not in general satisfy the constraints of Eqs. (5) and (6).

Now note that the constraint Eq. (6) together with the fact that the $M_{ki} \geq 0$ implies that

$$\hat{\phi}_i = \sum_{j,k=1}^{2^n} \hat{\phi}_j (M^{-1})_{jk} M_{ki} \geq 0, \quad i = 1, \ldots, 2^n, \quad \text{Eq. (11)}$$

which, together with Eq. (5), implies that

$$\sum_{k=l+1}^{2^n} \hat{\phi}_k = c, \quad \sum_{k=1}^{l} \hat{\phi}_k = 1 - c, \quad \text{Eq. (12)}$$

for some $0 \leq c \leq 1$. Suppose $c > 0$. Then since $L(\phi)$ is independent of $\phi_{l+1}, \ldots, \phi_{2^n}$ and $\log \phi_k$ is an increasing function of $\phi_k$, Eq. (4) can be increased
by simultaneously incrementing $\hat{\phi}_j \rightarrow \hat{\phi}_j + \epsilon$ for some $j \leq l$ and $\hat{\phi}_k \rightarrow \hat{\phi}_k - \epsilon$ for some $k > l$ such that $\hat{\phi}_k > 0$, which contradicts Eq (10). Thus we must have $c = 0$, and hence

$$\hat{\phi} = (\hat{\phi}_1, \ldots, \hat{\phi}_l, 0 \ldots, 0)$$

as the unique maximum of Eq. (4) satisfying Eqs. (5) and (6), from which the unique $\hat{\theta}$ is recovered using Eq. (3).

The above argument applies to the exact “slow algorithm” referred to in the main paper. Most of the results reported are obtained with an approximate “fast algorithm”, which uses a truncated $l \times l$ submatrix whose elements $\tilde{M}_{jk}$ are equal to $M_{jk}$ for $j, k = 1, \ldots, l$. An equivalent formulation of the fast algorithm is to determine

$$\hat{\theta}_k = \sum_{j=1}^{l} \hat{\phi}_j (\tilde{M}^{-1})_{jk},$$

where $\hat{\phi}$ is the argmax of the likelihood function

$$\tilde{L}(\phi) = \sum_{k=1}^{l} y_k \log \phi_k,$$

subject to the constraints

$$\sum_{j,k=1}^{l} \hat{\phi}_j (\tilde{M}^{-1})_{jk} = 1,$$

and

$$\sum_{j=1}^{l} \hat{\phi}_j (\tilde{M}^{-1})_{jk} \geq 0, \quad k = 1, \ldots, l.$$  

The computational efficiency arises because the number of constraints reduces considerably for large $n$ and physically realistic values of $l$.

It is easy to check that $\partial^2 \tilde{L}(\phi)/\partial \phi_j \partial \phi_k$ is a diagonal matrix with strictly negative diagonal elements, implying that $\tilde{L}(\phi)$ is a concave function, and that the constraints Eq. (16) and (17) define a convex set. Thus the fast algorithm is also guaranteed to give a unique result.

**Reduction in the number of patterns present**

One may be tempted to ask whether the maximum likelihood estimate described above is guaranteed to lead to a reduction in the number of methylation patterns reported relative to the number of patterns naïvely observed. We argue here that if read errors are ignored (i.e. we set $\eta = 0$), the estimated distribution over patterns $\hat{\theta}_i$ is shifted, relative to the naïve distribution $y_i/n_{\text{read}}$, towards the totally unmethylated pattern $(0, \ldots, 0)$. As a consequence the set of patterns
reported to be present is guaranteed to be a subset of or equal to the set of observed patterns.

The result is straightforward for the case of \( n = 1 \) CpG site. In this case we have two patterns: unmethylated (0), and methylated (1). The constraints Eq. (5) and Eq. (6) are

\[
\hat{\phi}_0 + \hat{\phi}_1 = 1 \quad (\hat{\phi}_0, \hat{\phi}_1) \left( \begin{array}{c} 1 \\ 0 \\ 1 - \epsilon \end{array} \right) \geq 0, \tag{18}
\]

and hence the maximum likelihood estimate is, from Eq. (4),

\[
\hat{\phi}_0 = \underset{0 \leq \phi_0 \leq 1 - \epsilon}{\text{argmax}} \{ y_0 \log \phi_0 + y_1 \log(1 - \phi_0) \}, \quad \hat{\phi}_1 = 1 - \hat{\phi}_0. \tag{19}
\]

The region over which \((\hat{\phi}_0, \hat{\phi}_1)\) is constrained and its transformation back to the reported abundances \((\hat{\theta}_0, \hat{\theta}_1)\) are illustrated in Figure 1A.

If the maximum of the function in chain brackets in Eq (19), namely \(y_0/n_{\text{read}}\), lies in the interval \([0, 1 - \epsilon]\), the transformation \(M^{-1}\) back to \(\theta\)-space clearly skews the estimated abundances towards the unmethylated state. If the unconstrained maximum lies in the interval \([\epsilon, 1]\), the constrained maximum returns the value \(\hat{\phi}_0 = 1 - \epsilon\), which transforms to the estimate \(\hat{\theta}_0 = 1\), and any methylated reads are interpreted as spurious. More specifically, the reported estimate is

\[
(\hat{\theta}_0, \hat{\theta}_1) = \begin{cases} \left( \frac{y_0}{n_{\text{read}}}, \frac{y_1}{n_{\text{read}}} - \epsilon \right) & \text{if } \frac{y_0}{n_{\text{read}}} \leq 1 - \epsilon \\ (1, 0) & \text{if } 1 - \epsilon < \frac{y_0}{n_{\text{read}}} \leq 1. \end{cases} \tag{20}
\]

To see how the result generalises to higher numbers of CpG sites, consider the \( n = 2 \) case illustrated in Figure 1B. The fully methylated pattern \((1, 1)\) is a fixed point of the transformation \(M\). Points in each triangular face of the black wire tetrahedron correspond to reported distributions in \(\theta\)-space for which the pattern \(i\) indicated at the opposite corner occurs with probability \(\hat{\theta}_i = 0\). Thus the transformation \(M^{-1}\) of the red tetrahedron to the black wire frame skews the distribution towards less-methylated patterns. Furthermore, whenever the maximum \(y_i/n_{\text{read}}\) of the likelihood function Eq (4) lies outside the red tetrahedron but inside the wire frame, the constrained maximum lies on the surface of the red tetrahedron, which maps to a point in one of the triangular faces of the wire tetrahedron, and possibly to an edge or corner. In this case at least one pattern is reported to be spurious\(^1\).

For an arbitrary number \( n \) of CpG sites, the only fixed point of the \(2^n \times 2^n\) matrix \(M\) is vector \((1, \ldots, 1)\) corresponding to the fully methylated pattern. Again the transformation \(M^{-1}\) skews the distribution towards less-methylated patterns, and any observed pattern with a small number of observed counts can potentially be reported as spurious.

\(^1\)Note that in the extreme, unphysical limit \(\epsilon \to 1\), the red tetrahedron shrinks to the corner \((1, 1)\) and all observed patterns except the fully unmethylated pattern \((0, 0)\) become “spurious.”
Figure 1: (A) Schematic of the region of φ-space constrained by Eq. (5) and Eq. (6) shown in red and its mapping back to θ-space for the case of an amplicon with \( n = 1 \) CpG site and read-error rate \( η = 0 \); (B) the constrained region of φ-space (red solid tetrahedron) and its mapping back to θ-space (black wire tetrahedron) for the case of an amplicon with \( n = 2 \) CpG site and read-error rate \( η = 0 \); and (C) the same as (A) with \( η > 0 \).
The situation when \( \eta > 0 \) is more complicated (see Figure 1C for the case of \( n = 1 \) CpG site). In principle, an observed distribution \( y_i/n_{\text{read}} \) can be skewed either towards the fully unmethylated pattern \((0,\ldots,0)\) or the fully methylated pattern \((1,\ldots,1)\) by the transformation \( M^{-1} \). In practice, however, and particularly in the case of non-vertebrates, observed distributions are typically weighted in favour of the fully unmethylated pattern and patterns within a small Hamming distance of the fully unmethylated pattern. Such distributions will be skewed towards the fully unmethylated pattern and the number of reported patterns will decrease, as observed in Figure 3 of the main paper.

References

[1] Akman, K., Haaf, T., Gravina, S., Vijg, J. and Tresch, A. (2014) Genome-wide quantitative analysis of DNA methylation from bisulphite sequencing data, Bioinformatics, Advance access publication. doi:10.1093/bioinformatics/btu142