Supplementary material

April 6, 2016

Supplementary Table S1

Table 1: DMRs detected by IMA. The number of regions and the corresponding number of sites has been given. Note that apart from other methods, IMA reports overlapping regions. Therefore, an indicative number of "unique" results has been shown. The latter were obtained by manual merging of overlapping regions, and the interpretation of the merged regions is questionable.

	IMA results				modified IMA (unique results)			
	TP		FP		TP		FP	
0.0	0	(0)	0	(0)	0	(0)	0	(0)
0.025	1747	(6288)	45	(223)	1096	(4251)	38	(207)
0.050	5847	(23959)	161	(949)	2981	(14199)	131	(770)
0.075	8380	(36279)	215	(1285)	3896	(19643)	172	(1086)
0.10	9882	(42627)	276	(1463)	4388	(22440)	228	(1280)
0.15	11315	(49299)	310	(1794)	4638	(24943)	248	(1486)
0.20	11822	(50032)	321	(1798)	4816	(25271)	255	(1543)

Supplementary Table S2

Table 2: The average computation time for each method, obtained from the simulation study. As bumphunter and seqlm can be parallelized, also the time on 10 parallel cores has been shown.

	Computation time in seconds	
	single core	10 parallel cores
bumphunter	60400	7400
Aclust	8200	–
seqlm	2020	310

Supplementary for MDL expressions

Here, we derive the expression $L(M)+L(D|M)$ for the segmentwise model introduced in the article. Let D be the methylation data matrix with p sites and n samples.

Let S be a fixed segmentation which consists of k segments $\{[s_1, e_1], \dots, [s_k, e_k]\}$. Let l_i denote the length of i-th region, i.e. $l_i := e_i - s_i + 1$. For a fixed segmentation, we have

$$L(M) = k \cdot \log p + \sum_{i=1}^{k} (l_i + 2) \cdot \gamma$$

where

- $k \cdot \log p$ represents the number of bits needed to code the segmentation. That is because it is sufficient to know the starting points of all segments and there are a total number of k starting points. Coding integers from the set $\{1, \dots, p\}$ takes $\log p$ bits each.
- $(l_i + 2) \cdot \gamma$ represents the number of bits needed to code the parameters for the linear model on segment i. Here γ represents the number of bits needed to code the real number with sufficient precision. This linear model has a baseline value for each site (there are l_i of those), the parameter β and the variance of residual errors σ^2. Altogether, there are $l_i + 1 + 1$ parameters.
- γ represents the number of bits needed to code the real number. Theoretical results [**?**] indicate that precision $1/\sqrt{m}$ is sufficient for discretizing the parameter which was estimated on m observations. The corresponding code length is $-\log \frac{1}{\sqrt{m}}$ or $0.5 \log m$. In our case, $m = n \cdot l_i$. As we want to code the parameters for all segments with equal precision, we will define γ as the maximum of $0.5 \log(n l_i)$ over $i = 1, \dots, k$.

We note that $L(M)$ can be expressed as the sum over description lengths of the models for each segment separately, i.e.

$$L(M) = \sum_{i=1}^{k} (\log p + (l_i + 2) \cdot \gamma)$$

So the description length function is additive.

Now, for $L(D|M)$ it was already mentioned that it can be expressed as the negative log-likelihood $L(D|M) = -\log \mathcal{L}(D|M)$. It can be expressed as following

$$-\log \mathcal{L}(D|M) = -\log \prod_{i=1}^{k} \mathcal{L}(D(s_i, e_i)|M_i) = -\sum_{i=1}^{k} \log \mathcal{L}(D(s_i, e_i)|M_i)$$

where $D(s_i, e_i)$ denotes the columns $s_i, \dots, e_i$ of the data matrix D. So also the $L(D|M)$ can be calculated for all segments separately and then summed together.

The expression for $\mathcal{L}(D(s_i, e_i)|M_i)$ is given by the gaussian density

$$\mathcal{L}(D(s_i, e_i)|M_i) = \prod_{k=1}^{n} \prod_{j=s_i}^{e_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_{kj} - \mu_j - \beta x_k)^2\right)$$

so

$$\log \mathcal{L}(D(s_i, e_i)|M_i) = \sum_{k=1}^{n} \sum_{j=s_i}^{e_i} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2 \ln 2}(y_{kj} - \mu_j - \beta x_k)^2\right)$$

Finally, the description length for a segment $[s_i, e_i]$ can be calculated as following

$$\begin{aligned} L(M_i) + L(D(s_i, e_i)|M_i) &= L(M_i) - \log \mathcal{L}(D(s_i, e_i)|M_i) \\ &= (\log p + (l_i + 2) \cdot \gamma) + \sum_{k=1}^{n} \sum_{j=s_i}^{e_i} \left(\frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2 \ln 2}(y_{kj} - \mu_j - \beta x_k)^2\right) \end{aligned}$$

Supplementary for Dynamic Programming

Here, we present the dynamic programming algorithm for identifying the optimal segmentation. As an input, an upper triangular matrix A with description lengths is needed.

Let D be the methylation data matrix with p sites and n samples. Suppose we have calculated the description lengths for all segments $[i, j]$ separately, according to the formulae given in the Supplementary for MDL expressions. Then, we fill the obtained values in the upper triangular matrix $A = (a_{ij})$ such that element a_{ij} denotes the description length of the model for segment $[i, j]$. The following algorithm identifies the segmentation with the smallest total description length.

Algorithm 1 Dynamic programming for identifying the optimal segmentation

```
Input: Matrix $A = (a_{ij})$ with description lengths
$L_0 \leftarrow 0$
for all $j \in \{1, ..., p\}$ do ▷ $p$ is the total number of sites
    $L_j \leftarrow +\infty$
    for all $i \in \{1, ..., j\}$ do
        if $L_j > L_{i-1} + a_{ij}$ then
            $I_j \leftarrow i - 1$
            $L_j \leftarrow L_{i-1} + a_{ij}$
        end if
    end for
end for
▷ Restoring the best segmentation $\mathcal{S}$
$k \leftarrow p$
$\mathcal{S} \leftarrow \emptyset$
while $k > 0$ do
    $\mathcal{S} \leftarrow \{[I_k + 1, k], \mathcal{S}\}$
    $k \leftarrow I_k$
end while
return $\mathcal{S}$
```