Supplementary Material

February 12, 2014

1 A statistical model for BS-Seq data

Methylation/non-methylation counts for a single CpG position will always be prone to unavoidable errors from bisulfite conversion, DNA amplification and sequencing. Thus, statements about methylation states at individual CpG positions are inherently uncertain. It is therefore necessary to pool evidence about methylation across multiple successive CpG positions (regions). The BEAT package uses Bayesian, Beta-Binomial mixture modeling of the methylation rate in a given region. Its input are the region-based counts of methylated respectively unmethylated cytosines. It outputs a posterior probability distribution of the methylation rate in each region. On this basis, we define an objective criterion for the detection of epimutation events when comparing two samples.

1.1 The likelihood function (a mixture of Binomials)

For multi-cell samples, we assume that all counts at a single CpG position were obtained from pairwise different bisulfite converted DNA template strands and represent independent observations. This certainly holds in good approximation, because the number of available DNA template strands typically supersedes the read coverage at this position by far. For single cell samples, we encounter the opposite situation: There are at most two template DNA strands available, and for many CpG positions this number is reduced further through DNA degradation. Multiple reads covering one CpG position are therefore highly dependent. We combine multiple counts at one position to one single (non-)methylation call. For different CpG position, these calls are then independent observations. First, fix one region, i.e. some set of CpG positions. The number of counts at a given position is the number of reads mapping to that position. Let n denote the total number of counts at all CpG positions in the given region, and let k (respectively $n-k$) of them indicate methylation (respectively non-methylation). Let r be the (unknown) methylation rate at the given position. Then, assuming independence of the single counts as mentioned above, the actual number j of counts originating from methylated CpGs in this region follows a binomial distribution,

$$P(j \mid n, r) = Bin(j; n, r) \tag{1}$$

Let the false positive rate p_+ be the global rate of false methylation counts, which is identical to the non-conversion rate of non-methylated cytosines. Conversely, let the false negative rate p_- be the global rate of false non-methylation counts, which is identical to the inappropriate conversion rate of methylated cytosines. One can find an upper bound for p_+ by considering all methylation counts at non-CpG positions as false positives (resulting from non-conversion of presumably unmethylated cytosines). In the literature, false negative rates were not described, an estimate of $p_- = 0.01$ is reported in[1]. We chose a conservative value of $p_- = 0.2$, which takes into account potential errors originating from mapping artifacts or sequencing errors.

Due to failed or inapproriate conversion, the number k of counts indicating methylation differs from the actual number j of counts originating from methylated CpGs. Given the true number of methylation counts j, the the observed methylation counts k are the sum of the number m of correctly identified methylations and the number $k-m$ if incorrectly identified methylations (false positives). Hence, the probability distribution of k is a convolution of two binomial distributions,

$$\begin{aligned} P(k \mid j, n; p_+, p_-) &= \sum_{m=0}^{k} P(m \mid j, 1-p_-) \cdot P(k-m \mid n-j, p_+) \\ &= \sum_{m=0}^{k} \underbrace{Bin(m; j, 1-p_-)}_{=:C^1_{m,j}} \cdot \underbrace{Bin(k-m; n-j, p_+)}_{=:C^2_{n-j,k-m}} \end{aligned} \quad (2)$$

In (2), we use the convention that $Bin(m; j, p) = 0$ whenever $m > j$. Thus, given n reads, k methylation counts, the likelihood function for r is a mixture of Bionomial distributions,

$$\begin{aligned} P(k \mid n, r; p_+, p_-) &= \sum_{j=0}^{n} P(k, j \mid n, r, p_+, p_-) \\ &= \sum_{j=0}^{n} P(k \mid j, n, r, p_+, p_-) \cdot P(j \mid n, r, p_+, p_-) \\ &= \sum_{j=0}^{n} P(k \mid j, n, p_+, p_-) \cdot P(j \mid n, r) \\ &\overset{(1,2)}{=} \sum_{j=0}^{n} \sum_{m=0}^{k} C^1_{m,j} C^2_{n-j,k-m} \cdot Bin(j; n, r) \end{aligned} \quad (3)$$

1.2 The prior (a Beta mixture distribution)

In our Bayesian approach, we furthermore need to specify a prior for r to calculate the posterior distribution of r. Recall the beta distribution(s), which is a 2-parameter family of continuous probability distributions defined the unit

interval $[0,1]$,

$$Beta(r;\alpha,\beta) \propto r^{\alpha-1}(1-r)^{\beta-1} \text{ , for } \alpha,\beta>0,\, r\in(0,1) \text{ ,}$$

We assume that a fraction of λ_m postions are essentially methylated, i.e., and that their rate r follows a $Beta(r;\alpha=r_m\cdot w_m,\beta_m=(1-r_m)\cdot w_m)$ distribution, having an expectation value for r of $\frac{\alpha_m}{\alpha_m+\beta_m}=r_m$. Here, we set $r_m=0.7$. The additional parameter w_m weights the strength of the prior relative to the strength of the likelihood. Since (the confidence into/ the knowledge about) our prior distribution of methylation rates is rather weak, we want our procedure to be strongly data-driven, therefore we choose a low w_m, $w_m=0.5$. A fraction of $\lambda_u=1-\lambda_m$ is essentially unmethylated, and their rate is assumed to follow a $Beta(r;\alpha_u=r_u\cdot w_u,\beta_u=(1-r_u)\cdot w_u)$ distribution, having an expectation value for r of $\frac{\alpha_u}{\alpha_u+\beta_u}=r_u$, where we set $r_u=0.2$ and $w_u=0.5$. Thus, the prior distribution $\pi(r)$ is a 2-Beta mixture distribution,

$$\pi(r;\alpha_m,\beta_m,\alpha_u,\beta_u,\lambda_m) \quad = \quad \sum_{s\in\{m,u\}} \lambda_s Beta(r;\alpha_s,\beta_s) \tag{4}$$

The pragmatic reason for choosing a Beta mixture as a prior distribution is the fact that the Beta distribution is the conjugate prior of the Binomial distribution[2], such that for some normalizing constant $D^{\alpha,\beta}_{j,n}$,

$$Bin(j;n,r)\cdot Beta(r;\alpha,\beta) \quad = \quad D^{\alpha,\beta}_{j,n}\cdot Beta(r;;j+\alpha,n-j+\beta) \tag{5}$$

1.3 The posterior distribution (a Beta mixture distribution)

By virtue of Equation (5), we can write down the posterior distribution of r analytically (Equation 6). This has the advantage that we can answer all questions on the posterior distribution of r efficiently and up to an arbitrary precision. Efficiency is an issue, because we need to calculate posterior distributions for all regions, which can easily amount to millions.

$$\begin{aligned}
&P(r \mid k,n;\, p_+,p_-;\alpha_m,\beta_m,\alpha_u,\beta_u,\lambda_m) \qquad (6)\\
&\quad = \; N^{-1}\cdot P(k \mid n,r;\, p_+,p_-)\cdot \pi(r;\alpha_m,\beta_m,\alpha_u,\beta_u)\\
&\quad \overset{(3,4)}{=} \; N^{-1}\cdot \sum_{j=0}^{n}\sum_{m=0}^{k} C^1_{m,j}C^2_{n-j,k-m}\cdot Bin(j;n,r)\cdot \sum_{s\in\{m,u\}} \lambda_s Beta(r;\alpha_s,\beta_s)\\
&\quad \overset{(5)}{=} \; N^{-1}\cdot \sum_{j=0}^{n}\sum_{m=0}^{k} C^1_{m,j}C^2_{n-j,k-m}\cdot \left(\sum_{s\in\{m,u\}} \lambda_s D^{\alpha_s,\beta_s}_{j,n} Beta(r;j+\alpha_s,n-j+\beta_s)\right)
\end{aligned}$$

In the above equation, N is a normalization constant,

$$N = \sum_{j=0}^{n} \sum_{m=0}^{k} C^1_{m,j} C^2_{n-j,k-m} \cdot \sum_s \lambda_s D^{\alpha_s,\beta_s}_{j,n} \tag{7}$$

The ingredients for the construction of the posterior distribution are visualized in Figure (1).

For each region under consideration, we obtain an individual posterior distribution $P(r \mid k, n, p_+, p_-)$. With this posterior at hand, a point estimate of the methylation rate is given by the expectation value $\hat{r}$ of r,

$$\hat{r} = \int_0^1 r \cdot P(r \mid k, n, p_+, p_-)\, dr \tag{8}$$

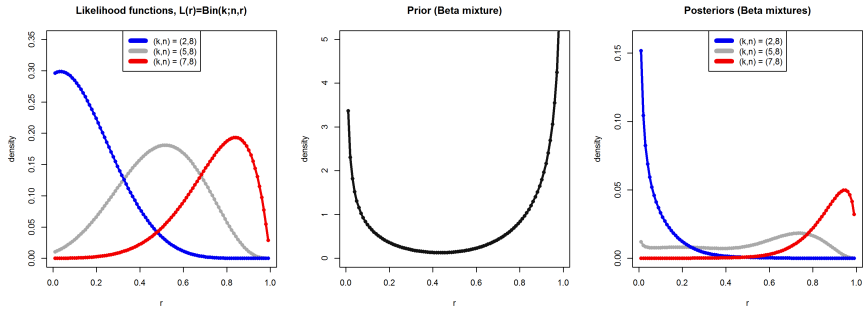

Figure 1: Plot of the likelihood functions for three different observations (k, n) (left), the Beta mixture prior distribution (middle) and the corresponding three posteriors (right). The number n of counts is fixed to 8, of which $k = 2$ (blue), $k = 5$ (grey) and $k = 7$ (red) are methylation counts. In this example we set $p_+ = 0.4$ and $p_- = 0.2$. The prior is defined in Equation (4).

1.4 Epimutation calling

It is customary to provide a Bayesian measure of uncertainty of this estimate $\hat{r}$, a so-called credible interval. A credible interval is an interval which contains the estimate $(\hat{r})$ and in which a prescribed probability mass of the posterior is located. One can construct a 90% credible interval $[m, M]$ as the shortest interval containing $\hat{r}$ such that $P(r \in [n, M] \mid k, n, p_+, p_-) = 0.9$. Moreover, we call a region *highly methylated* if

$$P(r > 0.7 \mid k, n, p_+, p_-) > c \tag{9}$$

for some stringency level c which we set to 0.75 here. The false negative methylation calling rates were set to $p_- = 0.1$ for all samples, and the false positive

calling rates were determined by $p_+ = 1 - \text{CH}$ methylation rate for each sample separately. A region is said to show *increased methylation* if

$$P(r > 0.5 \mid k, n, p_+, p_-) > c \tag{10}$$

Analogously, a region is called *sparsely methylated* if

$$P(r < 0.3 \mid k, n, p_+, p_-) > c \tag{11}$$

and a region with *decreased methylation* satisfies

$$P(r < 0.5 \mid k, n, p_+, p_-) > c \tag{12}$$

By definition, any highly methylated region has increased methylation, and every sparsely methylated region shows decreased methylation. For $c > 0.5$, high and sparse methylation calls are mutually exclusive. Regions that are neither highly nor sparsely methylated are called *ambiguous*. For each pair (k, n), Figure (2) shows the estimated methylation rate, and the corresponding methylation calls. Note that according to our model, (strict) high methylation calls can only be made if $n \geq 5$. The size of the regions has to be chosen appropriately, to contain enough reads that guarantee a sufficient detection power for epimutation events. The region size largely depends on the coverage in the samples, and typically lies in the range of $d = 100$ to $d = 10,000$.

We define a demethylating epimutation event (i.e., a *epidemethylation)* in a given region, by requiring the reference sample to be highly methylated, and the other sample to display decreased methylation. Vice versa, an *epimethylation* event is called if the given region is sparsely methylated in the reference and displays increased methylation in the second sample.

References

[1] Meissner, Alexander and Gnirke, Andreas and Bell, George W and Ramsahoye, Bernard and Lander, Eric S and Jaenisch, Rudolf (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis Nucleic acids research 33(18): p 5868–5877 PMID: 16224102.

[2] Gelman, Andrew and Carlin, John B. and Stern, Hal S. and Rubin, Donald B. (2003) Bayesian Data Analysis, Second Edition CRC Press.

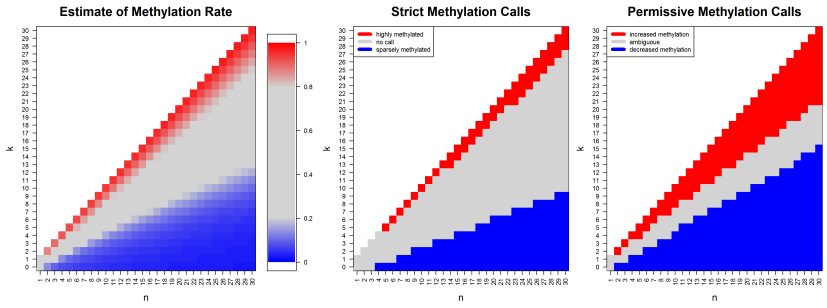

Figure 2: Illustration of the the results of our statistical modeling applied to regions of size $d = 1000$ in the Liver sample. In each plot, n (on the x-axis) denotes the total number of counts mapping to that region, of which k (on the y-axis) are counts indicating methylation. Left: Using the Liver-specific estimates of the false positive rate $p_+ = 0.2$ and the false negative rate $p_- = 0.1$ and the methylation prior in Equation (4), we obtain for each admissible pair (k, n) a methylation rate estimates $\hat{r}$. Colors correspond to methylation rate, ranging from deep blue (zero methylation) to deep red (full methylation). Middle: The red respectively blue area defines the pairs (k, n) which satisfy our criteria for high respectively sparse methylation. Right: The red respectively blue area defines the pairs (k, n) which satisfy our criteria for increased respectively decreased methylation. Note that strict methylation calls are only made when at least $n = 5$ counts were observed.