Discovery of methylation loci and analyses of differential methylation from replicated high-throughput sequencing data

Thomas J. Hardcastle [1,*]

[1]Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, United Kingdom

ABSTRACT

Motivation: Cytosine methylation is widespread in most eukaryotic genomes and is known to play a substantial role in various regulatory pathways. Unmethylated cytosines may be converted to uracil through the addition of sodium bisulphite, allowing genome-wide quantification of cytosine methylation via high-throughput sequencing. The data thus acquired allows the discovery of methylation 'loci', contiguous regions of methylation consistently methylated across biological replicates. The mapping of these loci allows for associations with other genomic factors to be identified, and for analyses of differential methylation to take place.

Results: The `segmentSeq` **R** package has been extended to identify methylation loci from high-throughput sequencing data from multiple conditions. A statistical model is then developed that accounts for biological replication and variable rates of non-conversion of cytosines in each sample to compute posterior likelihoods of methylation at each locus within an empirical Bayesian framework. The same model is used as a basis for analysis of differential methylation between multiple experimental conditions with the `baySeq` **R** package. These analyses are demonstrated on a set of data derived from Dicer-like mutants in *Arabidopsis* to reveal complex interactions between the different Dicer-like mutants and their methylation pathways.

Availability: The `segmentSeq` and `baySeq` packages are available on the Bioconductor (`http://www.bioconductor.org`).

Contact: tjh48@cam.ac.uk

1 INTRODUCTION

Cytosine methylation, found in most eukaryotes and playing a key role in gene regulation and epigenetic effects, can be investigated at a genome wide level through high-throughput sequencing of bisulphite treated DNA. Treatment of denatured DNA with sodium bisulphite deaminates unmethylated cytosines into uracil; sequencing these data thus allows, in principle, not only the identification of every methylated cytosine but an assessment of the proportion of cells in which the cytosine is methylated. Moreover, by comparing these quantitative methylation data across experimental conditions, genomic regions displaying differential methylation can be detected.

The data available for methylation locus finding are generated from a set of sequencing *libraries*. Each library consists of a set of sequenced reads which can be aligned and summarised to report at each cytosine the number of sequenced reads in which the cytosine is methylated, and the number in which the cytosine is unmethylated (Hardcastle, 2013). Having aligned the methylated and unmethylated reads, we then require methods both for calling a methylation 'locus', a region on the genome where multiple cytosines show evidence for methylation, and finding differential methylation between samples, both for individual cytosines and for the methylation loci. However, in order to find biologically meaningful results, we must consider the natural variation in methylation between biological replicates. We adapt our previously described methods for defining siRNA loci from high-throughput sequencing of sRNAs (Hardcastle *et al.*, 2012), and for discovering differential expression in high-throughput sequencing of RNA from paired samples (Hardcastle and Kelly, 2013) to an analysis of differential behaviour in the methylome.

We demonstrate these methods in an analysis of methylation in all contexts in mutants of the Dicer-like proteins in *Arabidopsis* (Stroud *et al.*, 2013). The Dicer-like proteins are involved in the production of variously characterised small RNAs, known to play a role in initiation and maintenance of methylation. We demonstrate that the application of these methods reveal complex patterns of behaviour between the different Dicer-like mutants. These various patterns of behaviour associate differently with coding sequences, gene promoter regions and transposable elements, as well as showing divergent patterns of genome localisation which suggest functionally significant differences are being exposed by such analyses.

2 METHODS

Candidate loci and nulls

We begin our analysis of these data by defining a set of *candidate loci* which may plausibly represent some methylation loci. A candidate locus begins and ends at some cytosine with a minimal proportion p_{min} of methylation in at least one sequencing library. Considering all such

[*]to whom correspondence should be addressed

loci is computationally infeasible and so filters are required to exclude implausible candidates and reduce the computational effort required. If two cytosines with a proportion of methylation above p_{min} are within some minimal distance d_{min} they are assumed to lie within the same locus. We further restrict the set by removing from consideration any candidate locus containing a region greater than λ_{max} that contains no cytosine with a proportion of methylation above p_{min}. Candidate loci may be defined with respect to a single strand (by default) or include data from both strands.

We define the set of *candidate nulls*, regions which may represent a region without significant methylation, by considering the gaps between candidate loci. We refer to the regions separating each candidate locus from its nearest neighbour (in either direction) as 'empty'. Candidate nulls consist of the union of the set of 'empty' regions, the set of candidate loci extended into the empty region to their left, the set of candidate loci extended into the empty region to their right, and the set of candidate loci extended into the empty regions to both the left and right.

Classification of candidates by posterior likelihood

The data pertaining to the candidates defined above are the number of methylated and un-methylated cytosines sequenced and aligning to these loci for each sample. We then identify those candidates which represent at least part of a true locus of methylation given the observed data for each replicate group. Each sample will belong to a replicate group of samples from biological replicates, and so the samples may be thought of as the set $\{A_1, \cdots, A_m\}$ with a replicate structure defined by $\mathcal{R} = \{R_1, \cdots R_n\}$ where $j \in R_q$ if and only if sample A_j is a member of replicate group q.

For a replicate group R_q and segment i we consider the total number of methylated and unmethylated cytosines $u_{iq} = \sum_{j \in R_q} u_{ij}$ and $u'_{iq} = \sum_{j \in R_q} u'_{ij}$ respectively. This approach neglects the effect of non-conversion rates on the observed values for u_{qj} and u'_{qj}. We can find no closed form expression for the posterior solution if the effects of non-conversion rates on the distribution of the data are accounted for. However, we can normalise the observed data by the expected non-conversion rates by setting $u_{ij} = C_{ij} - \frac{Q_j}{1-Q_j} T_{ij}$ and $u'_{ij} = T_{ij} + \frac{Q_j}{1-Q_j} T_{ij}$.

We assume that these data are described by a binomial distribution with parameter p_{iq} which has a beta prior distribution with parameters (α, β); we use an uninformative Jeffreys prior of $\alpha = \beta = \frac{1}{2}$. The posterior distribution of the parameter p_{iq} is then a beta distribution with parameters $(\alpha + u_{iq}, \beta + u'_{qi})$. A segment is identified as a methylation 'locus' if the posterior likelihood that $p_{qj} > q$ exceeds some critical value. Similarly, we can classify candidate nulls as true representatives of a null region by identifying those candidates with a posterior likelihood that $p_{qj} < q$ exceeding some critical value. By default, we use use $q = 0.2$ and a critical value of 99%; that is, a methylation locus should have a 99% chance of exceeding a 20% proportion of methylation.

Consensus loci

Given a classification on the set of candidate loci and nulls, we identify a set of consensus loci given the classifications on sets of overlapping candidates in a similar manner to that described for siRNA loci (Hardcastle and Kelly, 2013). We begin by assuming that a true locus of methylation should not contain a null region within a replicate group in which the locus is methylated. Thus, if some candidate locus l_r is classified as a locus in replicate groups Ψ_r, and there exists some candidate null n_s that lies completely within l_r and is classified as a null in one or more of the replicate groups Ψ_r, we discard the locus l_r. Of the remaining candidate loci, we then rank those that remain by the number of replicate groups in which they are classified as a locus, settling ties by considering the longer candidate locus. The consensus loci are then formed by choosing all those candidate loci that do not overlap with some higher ranked candidate locus, giving a non-overlapping set of loci on each strand.

Likelihood of data given non-conversion rates

We can compute posterior likelihoods of methylation and differential methylation on the identified loci through application of empirical Bayesian methods in which we estimate by sampling a distribution on the parameters of a distribution assumed to apply to the data (Hardcastle, 2015). This approach allows the variability of data between replicates to be accounted for.

Ignoring issues of non-conversion, we would assume that the data in equivalently methylated samples are beta-binomially distributed as in a straightforward analysis of paired data (Hardcastle and Kelly, 2013). Accounting for non-conversion requires that the data within each sample j are assumed to be the sum of a binomial distribution with success parameter Q_j (the rate of non-conversion) and a beta-binomial distribution with parameters p (the expected proportion of methylated cytosines) and dispersion parameter ϕ. Then the likelihood of the observed data D_{jk} at a single locus i for a sample j is given by

$$\mathbb{P}(D_{ij}|Q_j, p, \phi) = \sum_{m=0}^{C_{ij}} \binom{T_{ij} + m}{m} Q_j^m (1 - Q_j)^{T_{ij}} \binom{C_{ij} + T_{ij}}{C_{ij} - m} \frac{B(\alpha + C_{ij} - m, \beta + T_{ij} + m)}{B(\alpha, \beta)} \tag{1}$$

where m is the number of unconverted unmethylated cytosines, C_{ij} and T_{ij} the number of observed methylated and unmethylated cytosines respectively. p_q represents the proportion of methylation and ϕ the dispersion of the beta-binomial, with $\alpha = p\frac{1-\phi}{\phi}$ and $\beta = (1-p)\frac{1-\phi}{\phi}$.

Posterior Likelihoods of Methylation

We can estimate posterior likelihoods of methylation for each replicate group and locus using the methods described in Hardcastle (2015). For a sampled locus, we estimate by maximum likelihood methods for each replicate group q the parameters $\{p, \phi\}$, in which the dispersion parameter ϕ is assumed to be preserved across replicate groups and p is not. By repeating (without replacement) the sampling of loci, we build an empirical joint distribution on the parameters for the methylation of loci within each replicate group. We similarly derive an empirical joint distribution on for null regions. Given these distributions, we are able to calculate posterior likelihoods of methylation for each locus and replicate group. Regions exhibiting various patterns of differential methylation can be similarly identified using the density function defined in Eqn 1 in the `baySeq` **R** package (Hardcastle, 2015).

3 RESULTS

We test the analysis methods in a reanalysis of the Dicer-like mutants from the Stroud *et al.* (2013) dataset. We identify methylation loci in the *dcl2*, *dcl3*, *dcl4*, *dcl2/4* and *dcl2/3/4* mutants, together with wild-type samples . Non-conversion rates are estimated for each sample from the proportion of cytosine reads reported as methylated in reads aligning to mitochrondrial and chloroplast genomes.

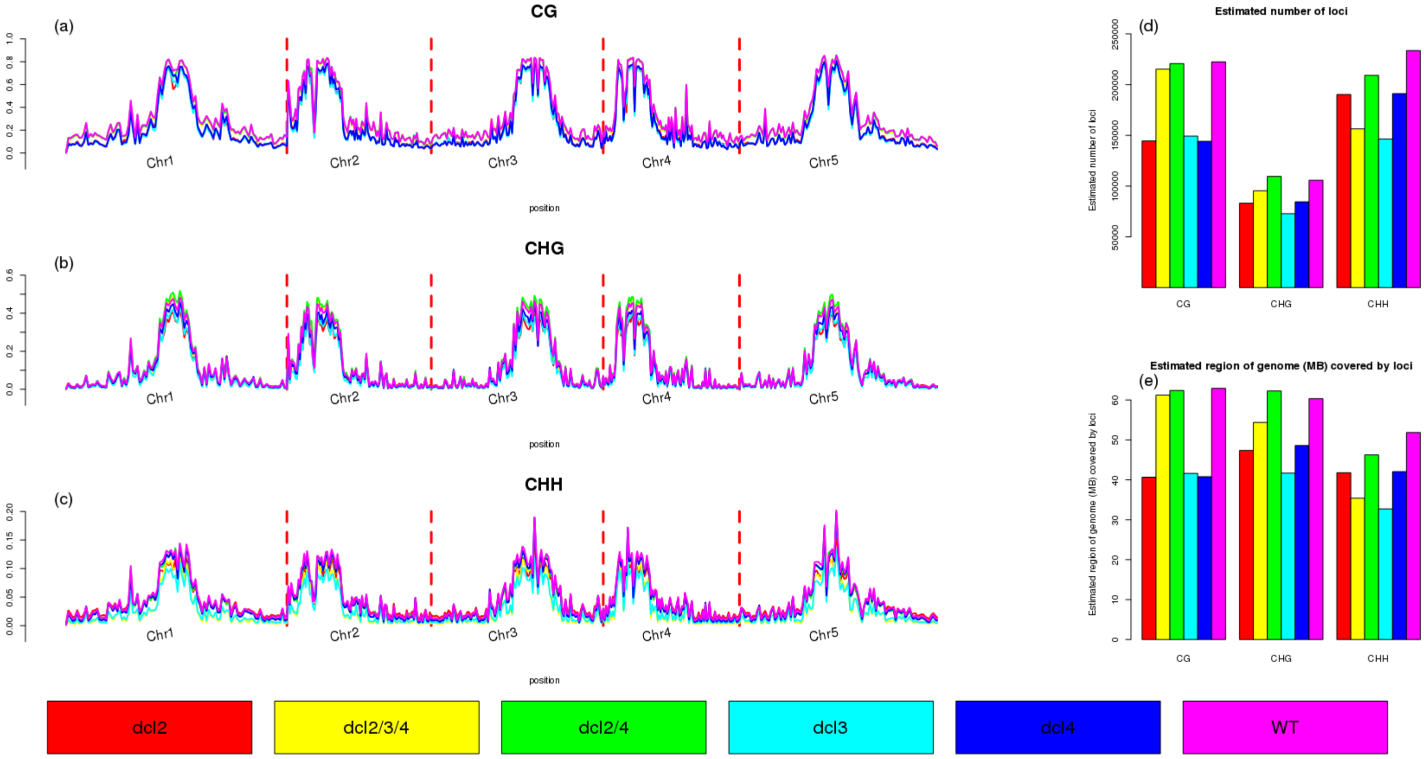

Fig. 1. Genome wide profiles of methylation for the various Dicer-like mutants, and wild-type, in CG (a) CHG (b) and CHH (c) contexts, adjusted for non-conversion rates. The estimated number of loci identified in each condition are shown in (d), while the estimated length of the genome covered by loci are shown in (e).

Figure 1 summarises the input data and identified loci. The genome-wide trends in methylation remain constant in all mutants relative to wild-type with substantial increases in methylation at the centromeric regions and at specific sites on the genome. There appears to be a minor global decrease in CG methylation (Fig. 1a) in *dcl4*, *dcl3* and *dcl2* mutants relative to the *dcl2/4* and *dcl2/3/4* mutants. In CHG methylation, a similar pattern is seen only at the centromeric regions. In CHH methylation the situation is more complex; *dcl3* and *dcl2/3/4* shows reduced methylation relative to wild-type in all regions of the genome, with *dcl3* paricularly reduced in the centromeric regions, while *dcl2* shows reduced methylation only in the centromeric regions. The total number of methylation loci in each condition may be estimated by summing the posterior liklelihoods of loci (Fig. 1d). Relative to wild-type, expected numbers of loci do not alter substantially for *dcl2/4* loci in any condition, or for CG methylation in *dcl2/3/4*, while all the single mutants show lower numbers of methylation in all contexts. The numbers of methylation loci discovered in the CHG context are substantially lower than for other contexts; however, the loci discovered are generally longer, as shown by the estimated portion of the genome covered by loci in each context (Fig. 1e), which shows roughly equivalent coverage for CG and CHG with a minor reduction in CHH context.

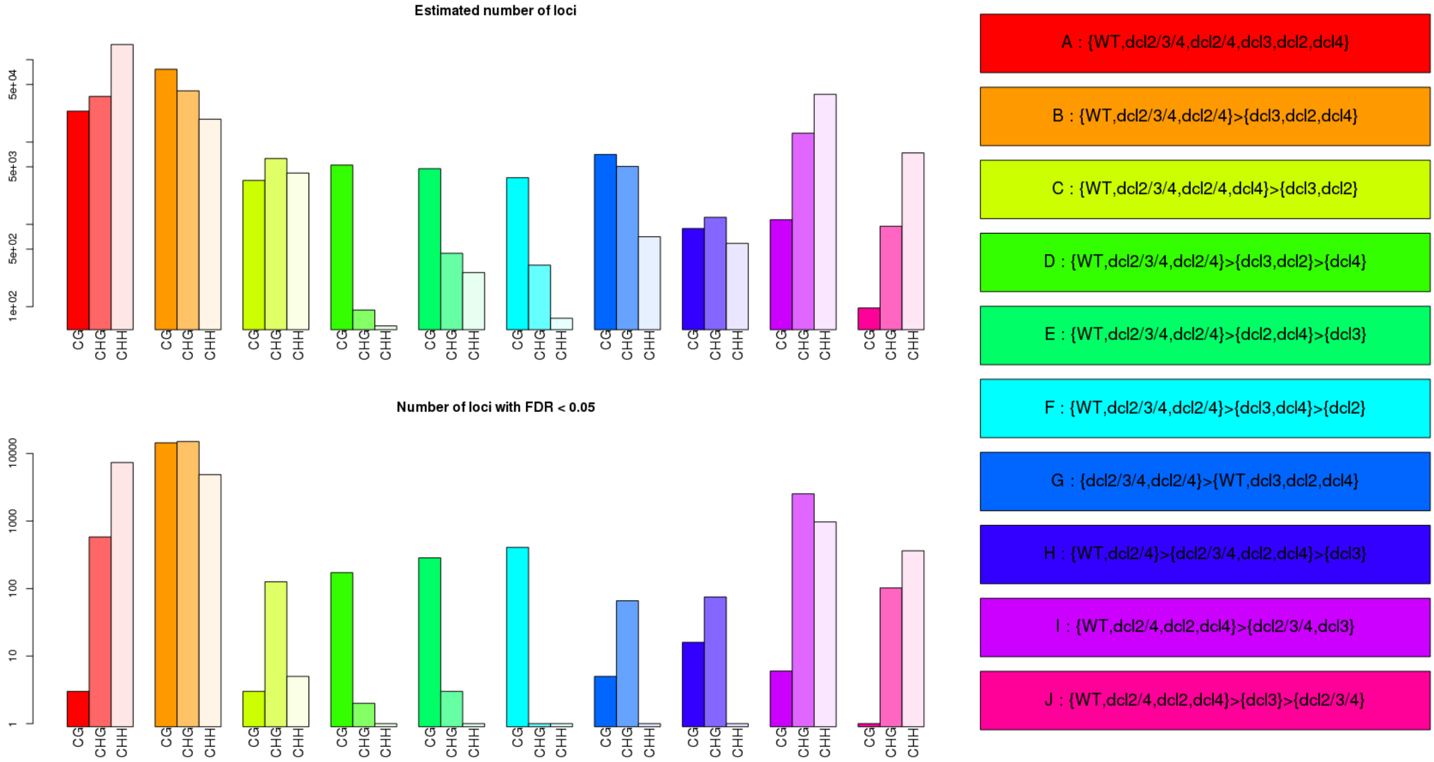

Fig. 2. Expected numbers (log-scale) of methylation loci and the number of loci (log-scale) which can be identified controlling FDR ¡ 0.05 for each of ten patterns of differential methylation, in each of CG, CHG and CHH contexts.

We next consider patterns of differential methylation in the loci. Using the `baySeq v2` (Hardcastle, 2015) with consensus priors we consider all possible models of equivalence and differential methylation between the mutants and wild-type. For each region of the genome, posterior likelihoods of difference are identified, and adjusted by the likelihood that the region is a methylation locus in at least one condition. From these posterior likelihoods, we can estimate the expected number of loci belonging to each pattern of equivalence and difference between the conditions. We can also select specific loci by controlling the false discovery rate (FDR) estimated from the posterior likelihoods.

Ten patterns (Figure 2) are identified with an estimated number of loci greater than one thousand and a number of loci with an FDR < 0.05 greater than fifty in at least one methylation context. Models I and J, represent the canonical changes in sRNA-linked methylation, in which there is loss of methylation in *dcl3* and *dcl2/3/4* relative to wild-type and all other mutants are found to a greater extent in CHG and CHH than in the CG context, conforming to the expectation that DCL3 is particularly relevant to the CHG and CHH methylation pathways.

However, substantial numbers of loci not conforming to the canonical models may also be found. Models D, E and F appear of particular interest; these all involve a loss of methylation in *dcl2*, *dcl3* and *dcl4* relative to wild-type, *dcl2/3/4* and *dcl2/4* mutants, with one of the *dcl2*, *dcl3*, *dcl4* showing a greater loss of methylation than the remaining two. Notably, these patterns of methylation are prevalent only in CG context, in contrast to the very similar B model, in which there is an equal loss of methylation in *dcl2*, *dcl3* and *dcl4*, and which is found in high numbers in all three contexts.

Given the sets of loci identified in each context for each pattern with an FDR < 0.05, we use the block-bootstrap method of Bickel *et al.* (2010) to identify overlap with annotation features (Figure 3); for robustness we limit these analyses to those cases where 20 or more loci can be identified with an FDR < 0.05. In the CG-context, models D, E and F overlap significantly with coding sequence regions; however, the related model B shows fewer overlaps in coding sequences than expected and an enrichment in transposable element overlaps; other models do not have sufficient identifiable loci for analysis. Almost all patterns of differential methylation in CHG and CHH contexts for which there are sufficient identifiable loci show significant depletion in overlap with coding sequences. Loci identified from models H and I show significant enrichment in overlaps with promoters which also overlap transposable elements in the CHG context, with loci identified as belonging to models I and J showing such enrichment in the CHH context.

Divergence can also be observed between the loci associated with the models when considering overlap with transposable element superfamilies (Figure 4). Loci associated with model B show significant enrichment of overlap with nearly all transposable elements and contexts with the exception of the RAth, DNA/Mariner and RC/Helitron (in CHG and CHH) contexts. In CHG and CHH contexts, the loci associated with models I and J show signficant enrichment of overlap with most superfamilies, particularly with the RAth superfamilies,

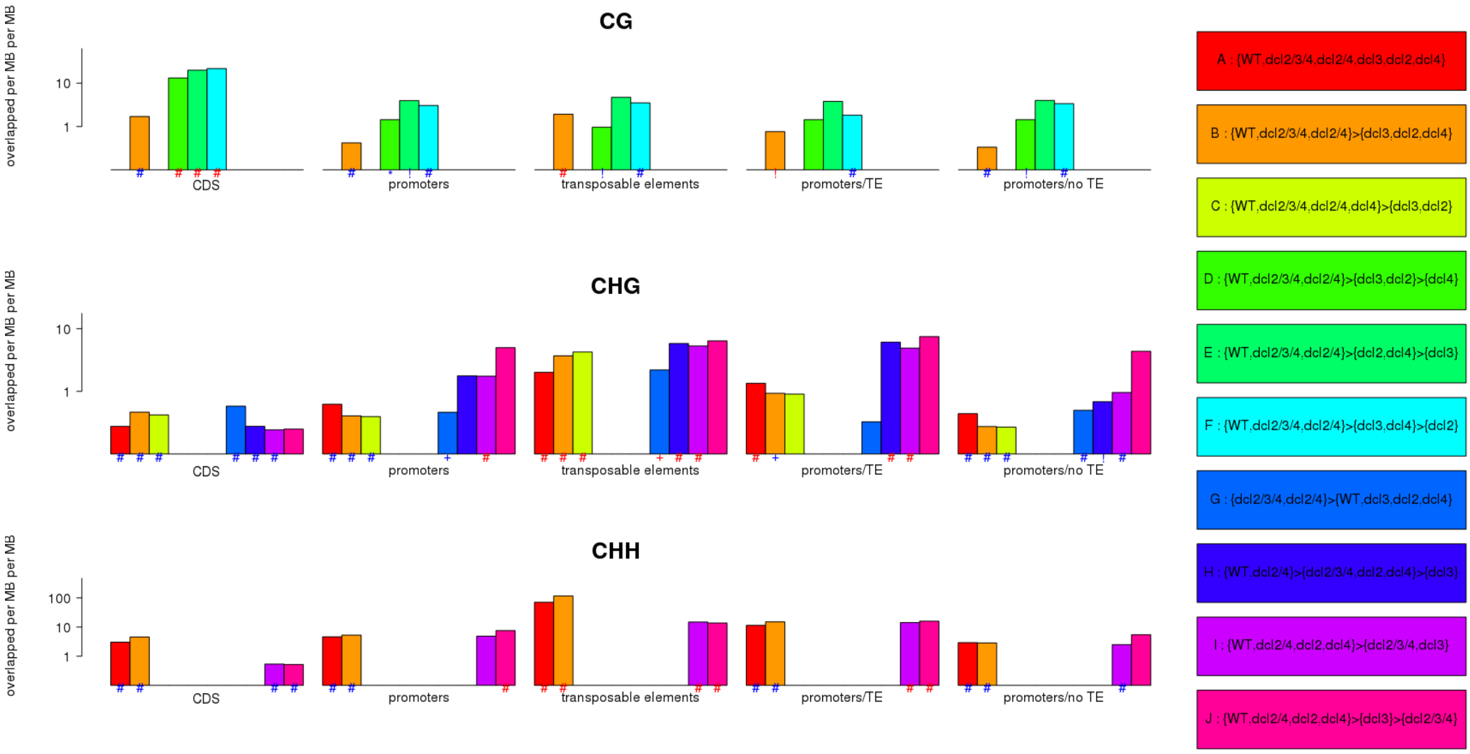

Fig. 3. Overlaps of annotation feature with ten patterns of differential methylation in CG, CHG and CHH contexts. Significance scores are calculated using a block-bootstrap method (Bickel *et al.*, 2010) which accounts for clustering of loci on the genome; the number of overlaps per megabase of identifable loci per megabase of annotation feature is the nearest surrogate for the significance score which has a straightforward interpretation. Significance scores are shown below the bars; Significance levels are shown in blue for under-representation and red for over-representation, the level of significance is indicated by the symbol such that # = $0 < p < 10^{-5}$, + = $10^{-4} < p < 10^{-5}$, * = $10^{-3} < p < 10^{-4}$, ! = $10^{-2} < p < 10^{-3}$.

but depleted overlaps for LTR/Gypsy and LTR/Copia elements. In CHG methylation, several other transposable element superfamilies show significant enrichment of overlap with models C (LTR/Gypsy, LTR/Copia, LINE/L1, DNA/En-Spm, DNA/Tc1, DNA/Harbinger), G (LTR/Copia) and H (RC/Helitron), suggesting that these models may indeed represent functionally divergent pathways of methylation.

We finally examine the genome localisation of the loci associated with the ten models of differential methylation (Figure 5). Loci with high likelihood of representing the model are most predominant in centromeric regions for the majority of models and contexts, however, models D, F, G and I in the CG context are depleted in the centromeric regions; model J is also depleted in the centromeric regions of CHG and CHH loci.

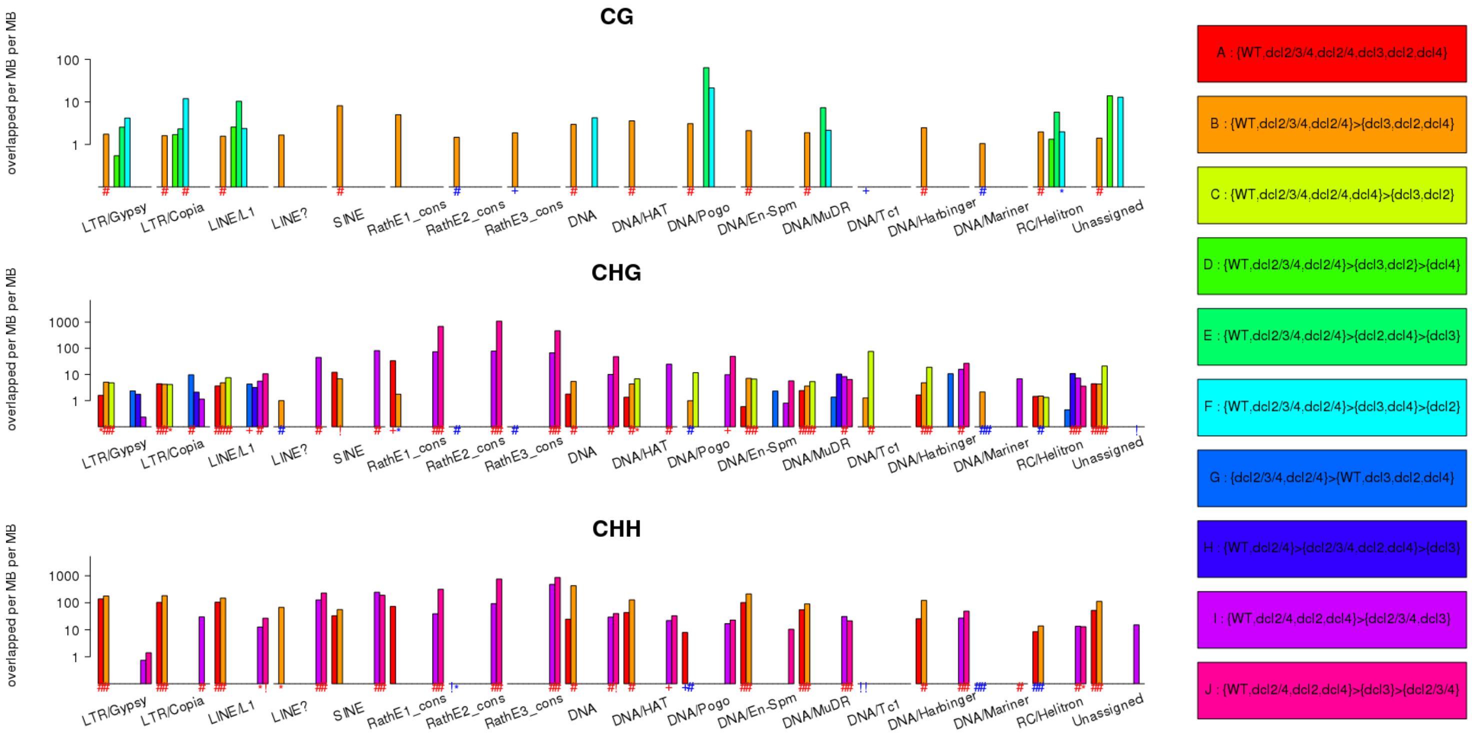

Fig. 4. Overlaps of transposable element superfamily with ten patterns of differential methylation in CG, CHG and CHH contexts. Significance scores are calculated using a block-bootstrap method (Bickel *et al.*, 2010) which accounts for clustering of loci on the genome; the number of overlaps per megabase of identifable loci per megabase of superfamily is the nearest surrogate for the significance score which has a straightforward interpretation. Significance scores are shown below the bars; Significance levels are shown in blue for under-representation and red for over-representation, the level of significance is indicated by the symbol such that # = $0 < p < 10^{-5}$, + = $10^{-4} < p < 10^{-5}$, * = $10^{-3} < p < 10^{-4}$, ! = $10^{-2} < p < 10^{-3}$.

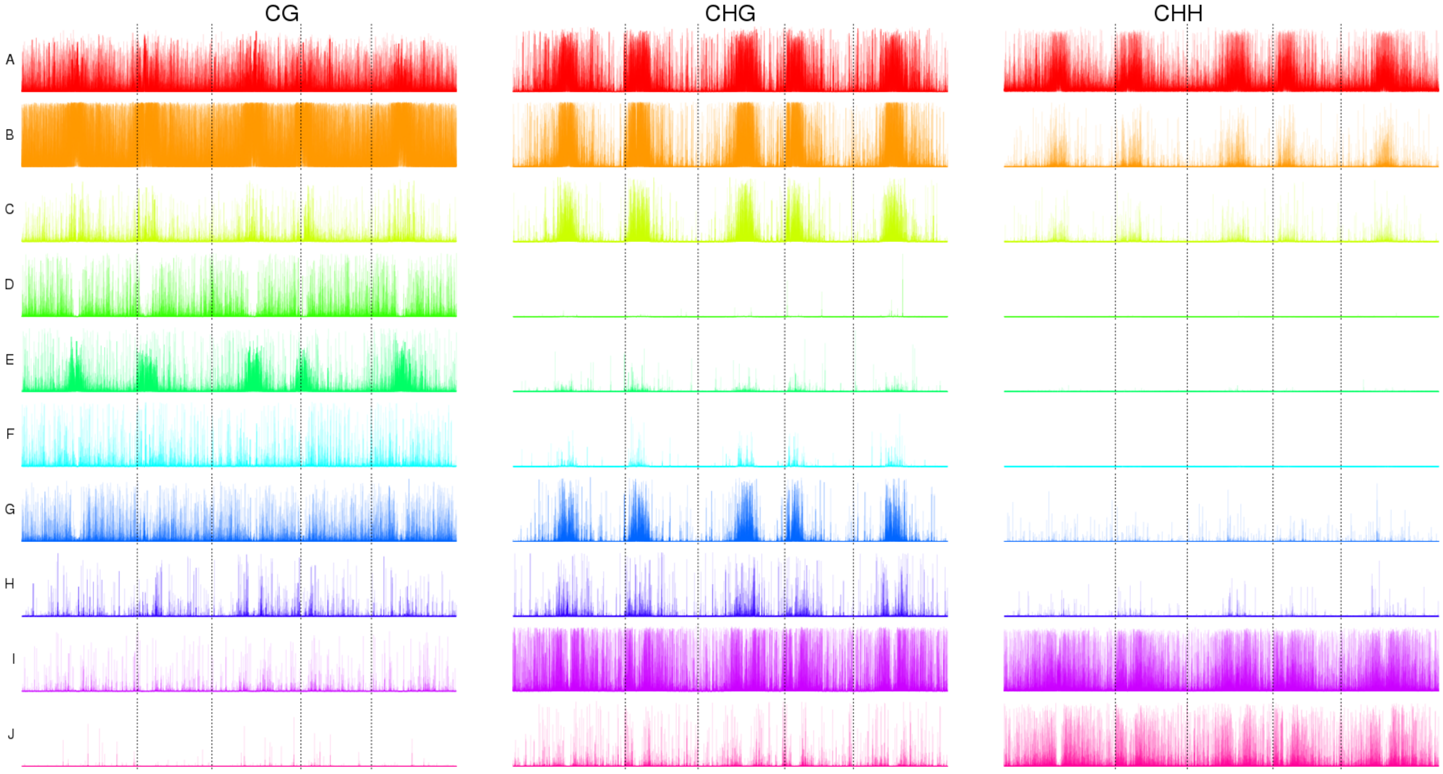

Fig. 5. Posterior likelihoods of belonging to models A-J in each methylation cotext across the genome. Mean likelihoods are calculated with traveling windows of 101, 1001, 10001, 100001 bases; the transparency of the plotted value diminishes with increasing window size.

4 CONCLUSIONS

The methods described here allow for the identification of methylation loci from multiple sequencing data, the estimation of likelihoods, for each replicate group, that a region is truly methylated above background levels, and ultimately the detection of differential methylated regions. The key advantage provided by these methods is the ability to account for biological replication within the analyses, and thus avoid spurious results due to spontaneous variation in methylation. The analyses also incorporate methods to account for observed rates of non-conversion of unmethylated cytosines to thymines, and thus to increase the specificity of identified regions of methylation. When applied in the empirical Bayesian framework of `baySeq v2` (Hardcastle, 2015), the model provided also allows for robust multivariate analyses of methylation data.

We demonstrate these methods on the Dicer-like mutants in *Arabidopsis* from the Stroud *et al.* (2013) dataset. Analyses of these data identify loci with high sensitivity, and allow the detection of complex patterns of differential methylation. These patterns show different associations with coding sequences, promoter regions and transposable elements, and show divergent localisations on the genome, suggesting that even subtle variations in the changes in methylation may have functional significance.

The implementation of the methods in the `segmentSeq` and `baySeq` **R** packages ensures compatibility with the analyses of sRNA-seq, mRNA-seq *et cetera* already developed in these packages. The results acquired by high-throughput sequencing of methylation can thus be readily incorporated with these other -omic data in a systems level analysis.

Funding: This work was supported by European Research Council Advanced Investigator Grant ERC-2013-AdG 340642.

REFERENCES

Bickel, P. J., Boley, N., Brown, J. B., Huang, H., and Zhang, N. R. (2010). Subsampling methods for genomic inference. *The Annals of Applied Statistics*, **4**(4), 1660–1697.

Hardcastle, T. J. (2013). High-throughput sequencing of cytosine methylation in plant DNA. *Plant Methods*, **9**(1), 16.

Hardcastle, T. J. (2015). Generalised empirical Bayesian methods for discovery of differential data in high-throughput biology. *bioRxiv*.

Hardcastle, T. J. and Kelly, K. A. (2013). Empirical Bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution. *BMC Bioinformatics*, **14**(1), 135.

Hardcastle, T. J., Kelly, K. A., and Baulcombe, D. C. (2012). Identifying small interfering RNA loci from high-throughput sequencing data. *Bioinformatics*, **28**(4), 457–63.

Stroud, H., Greenberg, M., Feng, S., Bernatavichute, Y., and Jacobsen, S. (2013). Comprehensive Analysis of Silencing Mutants Reveals Complex Regulation of the Arabidopsis Methylome. *Cell*, **152**(1), 352–364.