Notos -a galaxy tool to analyze CpN observed expected ratios for inferring DNA methylation types

Ingo Bulla, Benoît Aliaga, Virginia Lacal, Jan Bulla, Christoph Grunau, Cristian Chaparro

Bulla *et al. BMC Bioinformatics* (2018) 19:105
https://doi.org/10.1186/s12859-018-2115-4

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Notos - a galaxy tool to analyze CpN observed expected ratios for inferring DNA methylation types

Ingo Bulla[1,2], Benoît Aliaga[3], Virginia Lacal[4], Jan Bulla[4*], Christoph Grunau[3] and Cristian Chaparro[3]

Abstract

Background: DNA methylation patterns store epigenetic information in the vast majority of eukaryotic species. The relatively high costs and technical challenges associated with the detection of DNA methylation however have created a bias in the number of methylation studies towards model organisms. Consequently, it remains challenging to infer kingdom-wide general rules about the functions and evolutionary conservation of DNA methylation. Methylated cytosine is often found in specific CpN dinucleotides, and the frequency distributions of, for instance, CpG observed/expected (CpG o/e) ratios have been used to infer DNA methylation types based on higher mutability of methylated CpG.

Results: Predominantly model-based approaches essentially founded on mixtures of Gaussian distributions are currently used to investigate questions related to the number and position of modes of CpG o/e ratios. These approaches require the selection of an appropriate criterion for determining the best model and will fail if empirical distributions are complex or even merely moderately skewed. We use a kernel density estimation (KDE) based technique for robust and precise characterization of complex CpN o/e distributions without a priori assumptions about the underlying distributions.

Conclusions: We show that KDE delivers robust descriptions of CpN o/e distributions. For straightforward processing, we have developed a Galaxy tool, called Notos and available at the ToolShed, that calculates these ratios of input FASTA files and fits a density to their empirical distribution. Based on the estimated density the number and shape of modes of the distribution is determined, providing a rational for the prediction of the number and the types of different methylation classes. Notos is written in R and Perl.

Keywords: Epigenetics, DNA methylation, Kernel density estimation, CpG o/e ratio, CpN o/e ratio

Background

DNA methylation is an important bearer of epigenetic information

In eukaryotes, methylation occurs in the 5' position of the pyrimidine ring of cytosine, leading to 5-methyl-cytosine (5mC), which can subsequently be converted into hydroxy-5-methyl-cytosine [1]. The presence of 5mC can have an impact on gene expression [2], alternative splicing [3] and other biological processes. Compared to other bearers of epigenetic information, such as post-translational histone modifications and non-coding RNA, 5mC appears to be relatively stable and epimutation rates at this base rarely exceed 10^{-4} per generation [4]. The modification is also chemically very stable and survives common conservation methods for biological material. DNA methylation is therefore very often the target of choice when it comes to studying the impact of epigenetic information on the phenotype and the heritability of epiallels.

DNA methylation and CpN o/e ratios

Several techniques are available to study 5mC distribution. Nevertheless, the relatively high costs of DNA

*Correspondence: Jan.Bulla@uib.no
[4]Department of Mathematics, University of Bergen, P.O. Box 7803, 5020 Bergen, Norway
Full list of author information is available at the end of the article

methylation analyses have led to a bias in the results towards model organisms and towards the biomedical field. For the moment, it is not feasible to obtain comprehensive DNA methylation results for a large range of phylogenetic branches. This (i) is an obstacle to the introduction of epigenetics in fields in which historically the domain is not entirely accepted (e.g. ecology and evolution), and (ii) more importantly might lead to misinterpretation of results obtained in phylogenetically dissimilar (non-model) organisms. In many species, 5mC occurs either predominantly or exclusively in CpG pairs. This and the tendency of 5mC to deaminate spontaneously into thymine leads in methylated genomes to an underrepresentation of CpG over evolutionary time scales [5]. In human for instance, it was estimated that despite the existence of a specific repair mechanism that restores G/C mismatch, the mutation rate from 5mC to T is 10 to 50-fold higher than other transitions [6]. It was estimated that within 20 years, 0.17% of all 5mC in the human body, including germ cell generating tissue, were converted into thymine [7]. In molds, methylation can also be concentrated in CpA pairs and CpA o/e was used as an indicator of a process called repeat-induced-point-mutations (RIP) in which 5mC serves as mutagen, converting rapidly 5mC into thymine. Consequently, the ratio of observed to expected CpG pairs (CpG o/e) (and CpA o/e in fungi) was used to estimate the level of DNA methylation early on: in the methylated compartments of the genome, 5mCpN will tend to be mutated into TpN and the CpN o/e ratio will decrease (where 'N' stands for an arbitrary nucleotide). In contrast, in unmethylated genomes, the ratio will be close to 1. It should be noted that only those C to T transitions that are passed through the germline will have effects on CpN o/e ratios, i.e. technically CpN o/e distortions reflect past DNA methylation. Nevertheless, for more than 30 species CpG o/e were clearly related to contemporary methylation levels (see, e.g., [8–36]). In principle, it is therefore conceivable to infer methylation in DNA on the basis of CpN o/e, and to do this for any species for which genome and/or transcriptome sequence data are available [37]. DNA methylation prediction could then provide a starting point for more detailed biochemical DNA methylation analyses. The interest of transcription data would be that for many species, the available mRNA data outnumber largely the available genome sequences.

Robust description of CpN o/e ratios is challenging

In the following study we will focus on mRNA even though the method we will describe can be used on any type of DNA/RNA sequence. For the sake of clarity, in this manuscript, we will also use primarily methylation in the CpG context, although our approach can be applied to any (multiple)nucleotide frequency distribution. Simple Gaussian distributions can be used in some cases to describe CpG o/e distributions. But in many species, methylation distribution is heterogeneous, leading to complex mixtures in CpG o/e distributions over all genes, and the Gaussian mixture approach will fail. Many invertebrates, for instance, possess a mosaic type of methylation with large highly methylated regions intermingled with regions without methylation [38]. To our knowledge, no method exists that allows for a straightforward data processing of CpG o/e for non-specialists that is usable for all types of CpG o/e data. Here, we describe such a tool that we called Notos. We tested Notos on all data available in dbEST [39] since this database is one of the most widely used and covers a wide range of species. Notos integrates into Galaxy but is also available as suite of stand-alone scripts, it requires little computational resources, and the analysis is done within minutes. It is thus suitable for the routine first-pass prediction of DNA methylation in many biological settings.

Methods

Notos is a kernel density estimation (KDE) based tool. Its implementation is computationally efficient and allows for processing even large data sets on an ordinary personal computer. The analysis carried out by the Notos suite is composed of two steps and corresponds to two separate programs (see Fig. 1 for the work flow): First, the preparatory procedure CpGoe.pl calculates the CpG o/e ratios of the sequences provided by a FASTA file. Any CpN o/e can be calculated if supplied as parameter. Secondly, the core procedure KDEanalysis.r, which consists of an R script [40] carrying out two principal parts: data preparation and analysis of the distribution of the CpG o/e ratios using KDE. It is also possible to skip the preparatory procedure and directly provide KDEanalysis.r with CpG o/e ratios - or other data of comparable structure. We describe the two steps in the following.

Preparatory procedure: data input

The data necessary as input for the core procedures of Notos are CpG o/e ratios in form of a vector. These ratios correspond, in principle, to the number of CpGs observed in a sequence divided by the number of CpGs one would expect to observe in a randomly generated sequence with the same number of cytosine and guanine nucleotides.

Literature formulas

Several formulae for calculating this ratio have been established in the past years, all deriving some form of normalized CpG content. The presumably most popular versions (see, e.g., [41] and [42], respectively) are

$$CpGo/e = \frac{\#CpG}{\#C \cdot \#G} \cdot \frac{l^2}{l-1}$$

and

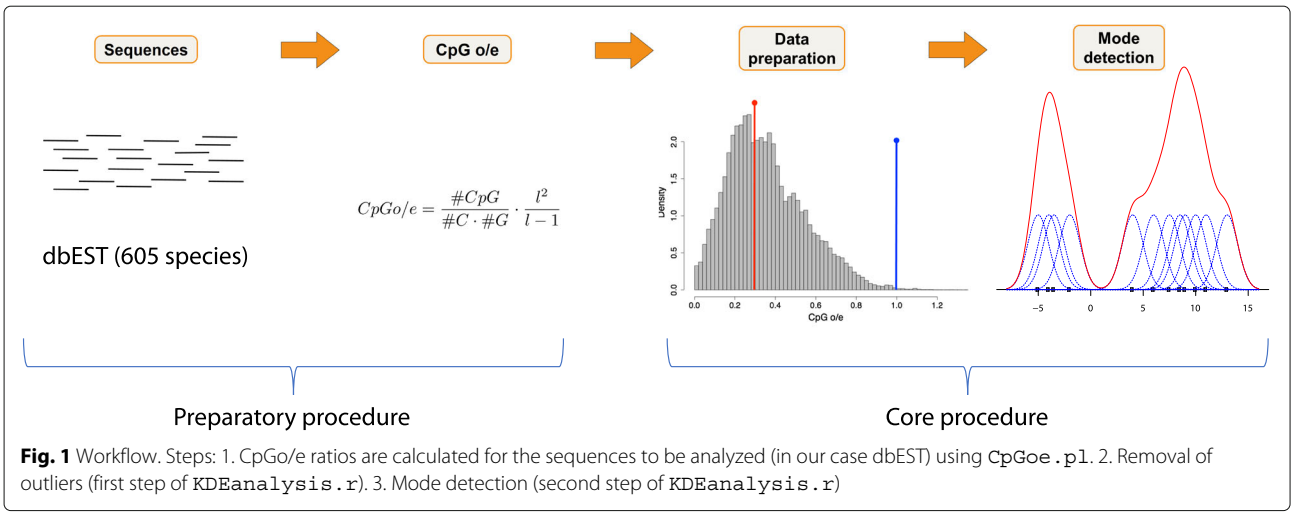

Fig. 1 Workflow. Steps: 1. CpGo/e ratios are calculated for the sequences to be analyzed (in our case dbEST) using `CpGoe.pl`. 2. Removal of outliers (first step of `KDEanalysis.r`). 3. Mode detection (second step of `KDEanalysis.r`)

$$CpGo/e = \frac{\#CpG}{\#C \cdot \#G} \cdot l,$$

where l is the length of the sequence, and $\#C$, $\#G$, and $\#CpG$ denote the number of C's, G's, and CpG's, respectively observed in the sequence. Alternative formulations were, among other, given by [43] who proposed

$$CpGo/e = \frac{\#CpG/l}{(\#G + \#C \text{ content})^2}$$

and by [44] with

$$CpGo/e = \frac{\#CpG}{(GC \text{ content } / 2)^2}$$

In their version, the $\#G + \#C$ content is defined as the total number of C's and G's divided by the total number of nucleotides, and GCcontent is defined as the total number of C's and G's.

Notos

The script CpGoe.pl allows the calculation of CpG o/e ratios from a multi-FASTA sequence and uses the formulation of [41] (i.e. the first formula above) by default, the others are optional. Moreover, sequences having less than 200 unambiguous nucleotides are eliminated from the calculation in the default setting, since our test runs indicated that too short sequences led to large amount of zeros or other extreme values.

Core procedure: data cleaning and analysis via KDE

The core procedure KDEanalysis.r carries out two steps: first, data preparation, which is mainly necessary to remove data artifacts, and secondly mode detection via KDE. Both steps return the user results in form of CSV files and figures. In addition, they allow overriding the default settings, if this is required by the user. Note, however, that such changes should be carried out with care, since all settings have been calibrated through intensive testing procedures on several hundred species from the dbEST database. In the following paragraphs, we describe these two steps in detail.

Data preparation

The first step, data preparation, starts by removing all values equal to zero from the input data since these observations correspond to artifacts resulting from too short sequences or sequences that do not present any CpG dinucleotide. Then, extreme and outlying observations are removed, i.e. all values outside the interval $[\,Q25 - kIQR, Q75 + kIQR]$, where $Q25$, $Q75$, and IQR denote the 25% quantile, the 75% quantile, and the interquartile range, respectively. In order not to exclude too many observations, the threshold parameter $k > 1$ takes the smallest integer value ensuring that not more than 1% of the data are removed, whereby k cannot exceed the value five. We determined the value of 1% through testing on a large number of species, and found it to be a good compromise between the need to exclude as many outliers as possible and not changing the distributional properties of a sample in a substantial way.

The output of this step consists of a table with various summary statistics in CSV format, and a figure displaying the data before and after this step. Figure 2 corresponds to the output resulting from an arbitrarily selected species, the locust *Locusta migratoria*. The content of the resulting table is described in detail in the documentation of Notos, which can be found in the readme file or the help section of the galaxy interface. Additional files 1 and 2 contain results from this step for 603 species from dbEST.

Mode detection

KDE In the second step, we determine the number of modes by means of a KDE based procedure. The underlying statistical theory is well-established, and therefore described only briefly, for details see Additional file 3. In principle, it is assumed that the independent and

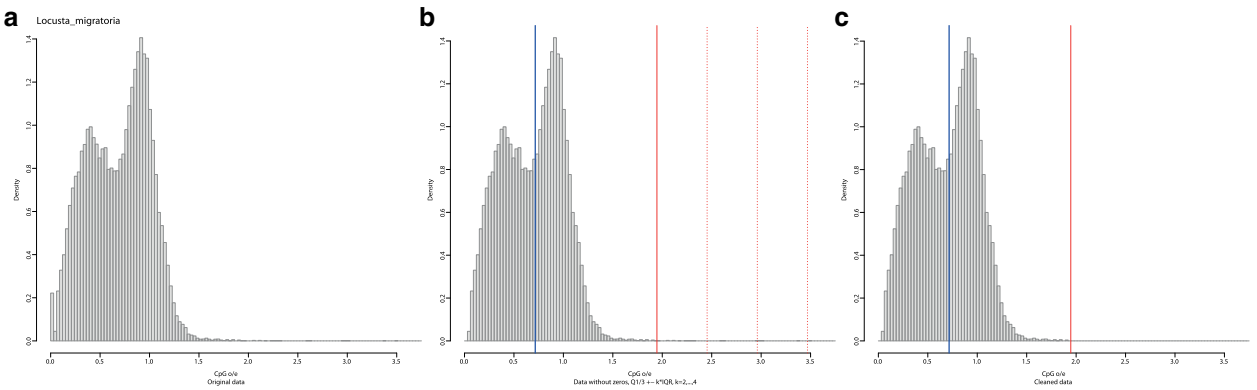

Fig. 2 Step 1: data cleaning of a sample of CpG o/e ratios from the locust *Locusta migratoria*. The left panel **a** shows the original data. The middle panel **b** displays the data after removal of all values equal to zero. The blue vertical line corresponds to the sample median. Red vertical lines indicate the possible thresholds for excluding outliers and extreme observations. The selected threshold ($k = 2$) is solid, alternative thresholds are dotted. The right panel **c** shows the cleaned data with the sample median and the selected threshold

identically distributed observations $x_1, x_n, \ldots, x_n$ constitute a sample with unknown density f. Then, the kernel density estimator $\hat{f}_h$ of f is given by

$$\hat{f}_h = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where $K(.)$ is the so-called kernel function. The kernel function is non-negative, has a mean value equal to zero, and the area under the function equals one, i.e., $K(.)$ satisfies the condition $\int_{-\infty}^{\infty} K(y)dy = 1$. Several families of kernel functions are available, and we considered the most common ones (Gaussian and Epanechnikov) for the implementation of our algorithms. Finally, we selected the probably most common Gaussian kernel function with $K(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$ due to the satisfactory results obtained in practice. In order to determine the value for the smoothing parameter, which is commonly termed bandwidth as well, we investigated different possible approaches, such as cross-validation, Silverman's rule [45], and Scott's variation of Silverman's rule [46]. Extensive testing on a large variety of species from different data sources suggested that the well-established bandwidth proposed by Scott provides the best results in terms of interpretability. In particular, it showed a satisfactory stability for species with either a very high or a very low number of observations.

Number of modes Subsequently, the number of modes is then determined by counting the number of local maxima of the estimated density, and a probability mass is assigned to each mode. The calculation of this probability mass is straightforward by integrating the density over the interval determined by the next-nearest local minima to the left and right, respectively, of the mode. If no local minimum is present to the left (right), the integration limits are set to minus (plus) infinity. The resulting probability masses for all modes sum up to one, and provide a single value which serves, roughly speaking, for determining the importance of a mode. Last, the obtained results are post-processed by a) merging modes that are closer than 0.2 (default value) to each other and b) removing modes that accumulate less than 1% (default value) of the probability mass of the estimated density. Multiple peaks suggest multiple sequence populations with different methylation types. The rational behind step a) is that very close modes reflect very similar types of methylation and hence probably have no biological significance. The value of 0.2 as minimum CpG o/e distance was empirically determined based on organisms with known mosaic-type methylation and double CpG o/e modes. We believe that relying entirely on confidence intervals is not a valid option for species with very high numbers of observations and as a consequence narrow confidence intervals. The choice of the probability mass threshold of 1% for step b) resulted again from extensive testing on a large number of species. A mode with 1% or less of probability mass lying outside of the core part of the density would most likely result from contamination. An optional feature of the KDE analysis is the estimation of confidence intervals for the position of the modes as well as confidence estimates for the number of modes. This is implemented through case resampling (non-parametric) bootstrap with 1,500 repetitions. Since this part is slightly computationally demanding, the bootstrap is optional and is accelerated by parallel execution via the doParallel package.

Output Similarly to the first step, the script KDEanalysis.r returns a figure to the user. Figure 3 shows this graphical output for the four species *Locusta migratoria*, *Alligator mississippiensis*, *Antheraea mylitta*, and *Citrus*

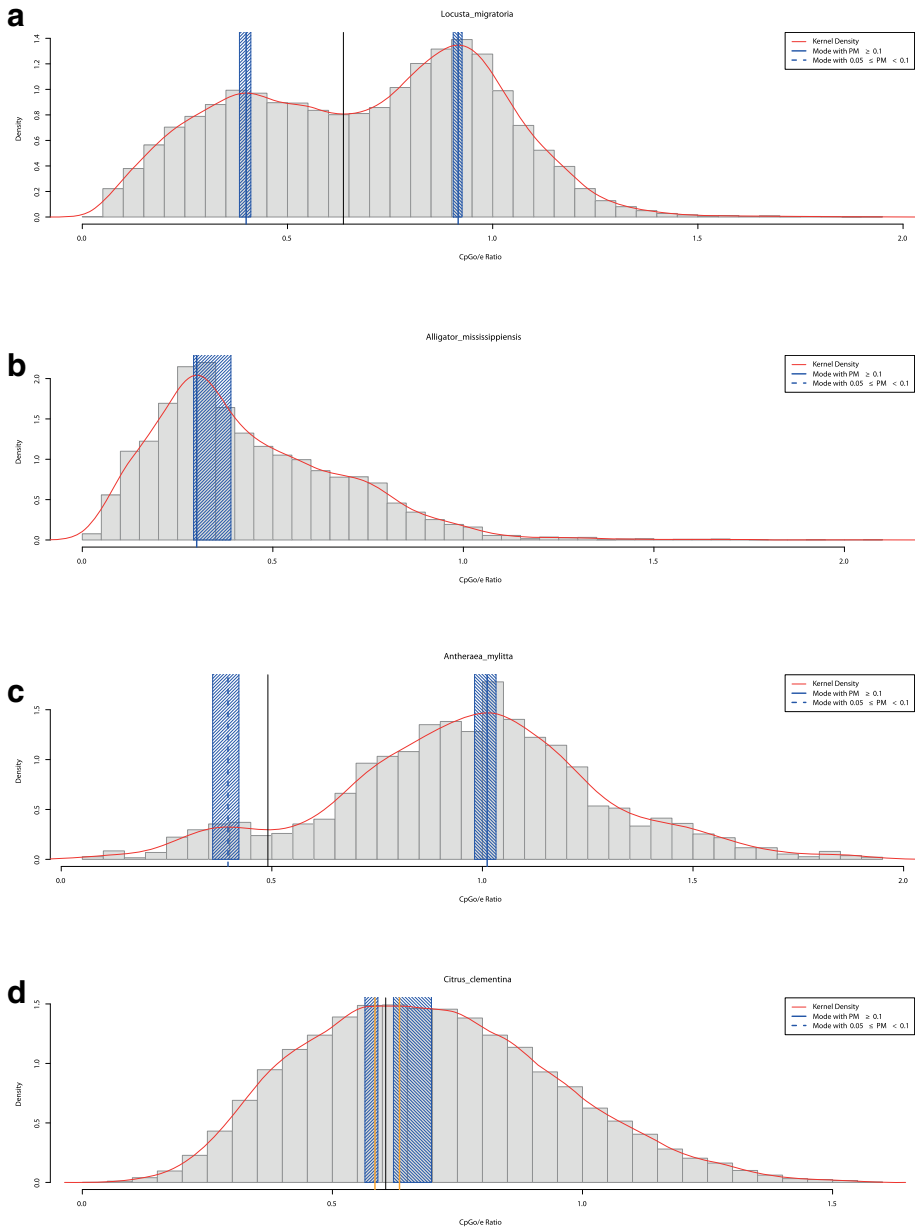

Fig. 3 Step 2: kernel density estimation for samples of CpG o/e ratios from four species. The red line corresponds to the density estimated via KDE. Full vertical blue lines indicate modes with PM $\geq$ 0.1. Shaded blue areas around the modes correspond to bootstrap confidence intervals with a default level of 95%. From top to bottom, the panels show results for *Locusta migratoria* (**a**), *Alligator mississippiensis* (**b**), *Antheraea mylitta* (**c**), and *Citrus clementina* (**d**)

clementina. The top panel a with *L. migratoria* shows two clearly distinct modes (blue vertical lines), their corresponding confidence intervals (shaded blue), and the fitted density (red). Moreover, a thin black vertical line indicates a local minimum, which serves for separating the probability masses attributed to each mode. In the case of *A. mississippiensis* (panel b), only one mode is present. Note that the confidence interval is strongly skewed, which results from the skewed empirical distribution used for the parametric bootstrap. For *A. mylitta*, one can observe that one of the two modes is assigned less than ten percent of probability mass, indicated by the dashed vertical line for the left mode in panel c. Last, *C. clementina* (panel d) possesses two modes relatively close to each other, i.e., the distance lies below the above mentioned threshold of 0.2. For this reason, the two modes may be interpreted as being too close for indicating biologically relevant differences in methylation types, which is underlined by their orange color. For results concerning other species from dbEST, see Additional file 4.

Moreover, the user obtains one table with various statistics related to the modes and their probability masses (see Additional file 5 for the results for 605 species from dbEST). Optionally, a second table linked to the results obtained from the bootstrap procedure is generated (cf. Additional file 6). The content of these two tables is also described in detail in the readme section of the Galaxy interface. The output from the bootstrap procedure deserves two additional remarks. Firstly, from a practical perspective, the number of modes identified in the bootstrap samples allows insight into the stability (and potential instability) of the number of identified modes. For example, at least one of the modes detected in the original sample should be considered weakly developed if a high proportion of bootstrap samples possesses a lower number of modes than the original sample. Alternatively, a frequently occurring higher number of modes in the bootstrap samples than in the original sample indicates that additional modes could develop with an increasing sample size - however, an increasing sample size may also have the opposite effect. Secondly, from a technical perspective, it may be non-trivial to assign modes identified in a bootstrap sample to the corresponding modes from the original sample, e.g., if several weakly developed modes are present in the original sample. In order to obtain reliable confidence intervals, two safeguards are implemented. On the one hand, bootstrap samples having a different number of modes than the original sample are excluded. On the other hand, samples with modes subject to strong changes (default value: 20%) in the probability mass compared to the original sample are excluded as well.

Implementation

A Galaxy package has been created that allows the automated installation of the Notos suite in a Galaxy server. The suite installs an interface for CpGoe.pl which provides the calculation of the CpG o/e ratio as well as an interface for KDEanalysis.r which calculates the distribution of CpG o/e ratios using KDE. Empirical testing showed that at least about 500 sequences are necessary to obtain a reliable parametrization of the KDE for CpG o/e frequency distributions.

Results

The test of Silverman [45] constitutes a classical, popular way to investigate multimodality. In the context of DNA methylation patterns, model-based approaches essentially founded on mixtures of Gaussian distributions have become a very popular approach to investigate questions related to the number of modes or underlying subpopulations [47–50]. This popularity may result, inter alia, from the easy accessibility of statistical software allowing the treatment of mixture models, such as `flexmix`, `mclust`, or `mixtools` [51–53]. While the test of Silverman provides a rather simple criterion in form of a *p*-value rejecting (or not) the null hypothesis of a certain number of modes, model-based approaches require the selection of an appropriate criterion for determining the best model. The most prominent among established criteria are, e.g., the Akaike Information criterion (AIC) and its extensions, the Bayesian information criterion (BIC), and the Integrated Completed Likelihood (ICL) (see, e.g., [54, 55], and the references therein).

Comparison

We investigated the performance of the Silverman test, the different criteria, and Notos on our data base with 603 species from dbEST. Table 1 shows the results from 17 arbitrarily chosen species, which display patterns that are representative of the full sample. The principal results are the following:

(i) The test of Silverman selects a low number of modes in most cases, with a few exceptions where the number of modes reaches high values. Overall, the number of detected modes is often difficult to explain or confirm by visual inspection of the sample, and the biological interpretation is (very) limited. Furthermore,

Table 1 This table shows the number of modes selected by different approaches and methods for 17 selected species: the test of Silverman (2nd column), model-based approaches, based on the criteria AIC, BIC, and ICL (3rd to 5th column) and Notos (last column). The maximum number of modes is limited to ten, all mixture models were estimated by the R-package `mclust`

Species	Silv.	AIC	BIC	ICL	Notos
Acropora palmata	1	10	5	1	2
Actinidia chinensis	1	8	8	1	1
Aegilops speltoides	1	7	2	1	1
Aiptasia pallida	2	6	3	1	2
Alligator mississippiensis	1	7	4	1	1
Antheraea mylitta	4	6	3	1	1-2
Aspergillus oryzae	1	5	1	1	1
Bombus terrestris	2	5	3	1	2
Citrus clementina	1	8	4	1	1-2
Citrus limon	1	5	3	1	1
Danio rerio	1	8	8	1	1
Daphnia pulex	1	9	5	1	1
Drosophila melanogaster	2	8	3	1	1
Locusta migratoria	2	9	9	1	2
Nematostella vectensis	2	9	6	1	2
Pinctada maxima	1	10	4	1	2
Rattus norvegicus	1	10	8	1	1

(ii) The model selection criteria AIC and BIC generally produce non-interpretable results: both criteria allow for models with too many parameters, which regularly results in the selection of models with a far too high number of modes and no biological interpretability. This effect is illustrated in panel a of Fig. 4 which shows the fitted density and the location of the component-specific means for *L. migratoria*, determined by the AIC solution. The discrepancy between the relatively clearly visible bimodal shape and the selected model with nine components is rather large. This high number of modes results from the very good fit to the empirical density for this sample containing a high number of observations. Panel b of Fig. 4 illustrates the non-satisfactory performance of the BIC by means of *A. mississippiensis*. This species shows a single, clearly pronounced mode at approximately 0.3, and is strongly skewed to the right. This strong skewness leads to the additional identification of two components at about 0.6 and 1.0. Moreover, an additional component is identified at ~ 0.15 for compensating for another small deviation from normality.

(iii) This drawback cannot be overcome by selecting the number of modes based on the ICL. This criterion almost always determines a single mode, which is sensible from a clustering perspective, but not desirable for mode identification, as panel c of Fig. 4 shows.

Interpretation

In conclusion, while conventional methods can perform well in many cases, they will also often fail to produce biologically interpretable results. For the 603 species from dbEST, the information criteria mentioned above as well as the test of Silverman fall short for approximately 60% of the data in this regard. In contrast, Notos performed well with all tested data sets. After having firmly established that Notos provides robust descriptions of mode locations and mode numbers, we attempted to establish a link between these parameters. As outlined above, a CpG o/e ratio around 1 is assumed to occur in non-methylated sequences and a ratio below 1 in methylated sequences.

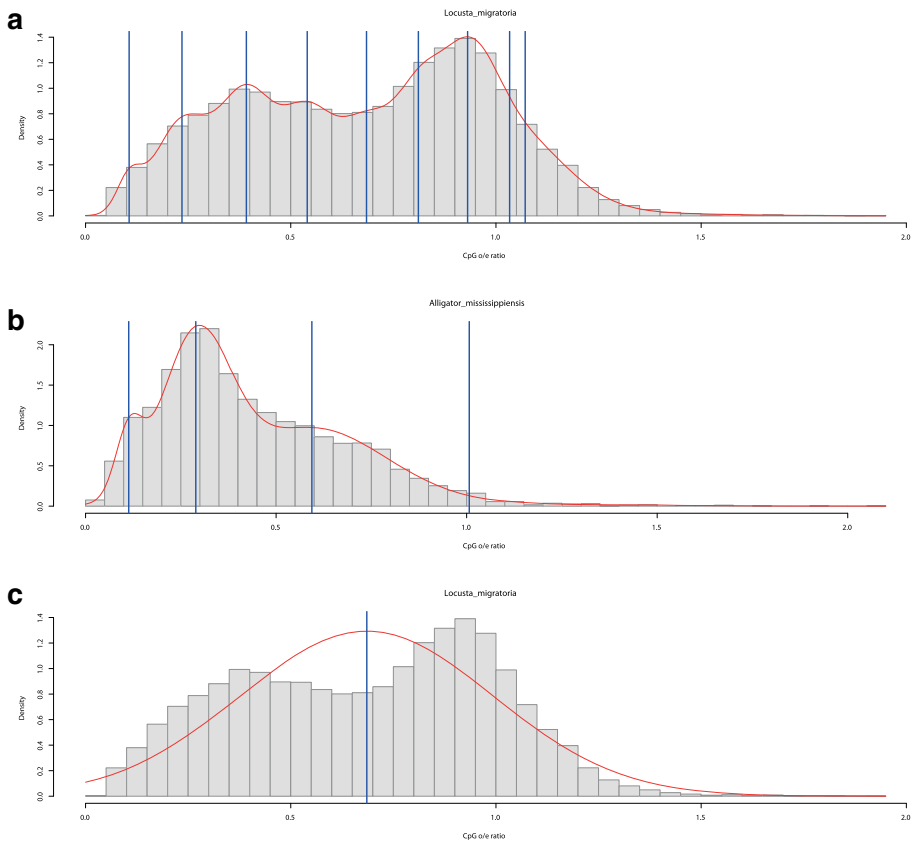

Fig. 4 Examples for model-based clustering and model selection with Gaussian mixtures of CpG o/e ratios. The red line corresponds to the estimated density via KDE. Full vertical blue lines indicate the location of means belonging to each component of the mixture distribution (estimated by the R-package `mclust`). The top panel **a** shows the model selected by the AIC for *Locusta migratoria*, while the lowest panel **c** displays the corresponding ICL solution. The middle panel **b** displays the model selected by the BIC for *Alligator mississippiensis*

Consequently, if both situations are detected, both types of sequences co-exist in the studies sequence population. Based on comparison of Notos results with available literature data on DNA methylation, we tentatively assigned a threshold value of 0.75 to differentiate presumably methylated (<0.75) from presumably non-methylated (≥0.75) sequences. This is slightly higher than the 0.6, conventionally used e.g. for the detection of generally unmethylated CpG islands [56]. Based on DNA methylation data from the literature, our prediction on gene body methylation has a positive predictive value of 91% (for details, see [57]).

Case studies

To illustrate the use of Notos in two CpN contexts, we will present in the following results for the classical model species *Neurospora crassa*. *N. crassa* is a mold that belongs to the ascomycota. DNA methylation in this species is well described: only repetitive sequences such as relicts of transposons but not protein coding genes are methylated [58]. Methylation in these regions is associated with a genome defence system called repeat-induced point mutations (RIP) (reviewed in [59]). This system targets specifically CpA dinucleotides [60] where C is converted into T. CpA depletion is considered as a sign of RIP in other fungal species as well [61]. We therefore anticipated that CpG o/e and CpA o/e ratios in coding sequences would be around or above 1 (no methylation), while CpA o/e ratios, but not CpG o/e ratios, would be clearly below 1 in repeats indicating methylation in this context. We used the Neurospora_crassa.ASM18292v1.31.dna_sm.genome.fa genome assembly and the corresponding Neurospora_crassa.ASM18292v1.31.gff3 annotation file from http://fungi.ensembl.org/Neurospora_crassa/Info/Index to extract 40,826 sequences for repeats and 10,432 sequences of spliced exons. A minimum length of 1 kb was used. As expected, a distribution with a single mode at a maximum at 0.9-1.1 was observed for CpG and CpA o/e ratios in spliced exons (panels a and b, respectively of Fig. 5). In contrast, the mono-modal CpA o/e ratio distribution in repeats peaked at 0.47, while for CpG o/e the single mode was shifted towards 1.5 (panels c and d of Fig. 5). The results of this straightforward and rapid analysis correspond therefore entirely to what is known about DNA methylation in *N. crassa*.

Discussion

DNA methylation is a conserved feature of many genomes. Since it remains neutral in its protein coding potential its use for adding additional epigenetic information to the DNA has been evolutionary stable. Nevertheless, the type of encoded information and consequently the type of DNA methylation can vary considerably, and many species have no or very little DNA methylation. It is thus of great practical value to be able to propose a well-founded hypotheses or at least educated guess about the type of methylation in a biological model before choosing an experimental strategy to study it in more detail. Notos was generated to produce such testable hypothesis.

Technical alternatives

It could be argued that other wet-bench based methods deliver comparable results about the presence and the type of methylation. It is for instance straightforward to digest DNA with methylation sensitive restriction enzymes [62] and to separate the resulting fragments by electrophoration. A digestion smear would indicate absence of methylation. But this requires producing sufficient amounts of high-quality DNA, which is not always possible (e.g. protected or rare species, degraded DNA, samples that are difficult to obtain). Digestion is also difficult to quantify. Extensions of the digestion method are methylation sensitive amplified length polymorphism (MS-AFLP) [63], reduced-restriction bisulfite sequencing (RRBS) [64] or reference-free reduced representation bisulfite sequencing (epiGBS) [65]. These methods are very powerful and can be used with or without a reference genome (that is not necessarily available for non-model species). A caveat of RRBS is however that it was designed for the methylation type of vertebrates that typically possess methylation free CpG islands. It might not work well with other methylation types. Similarly to the simple digestion method, all these methods need physical access to high quality DNA and require already considerable investment (currently from several hundreds to thousands of euros). The same applies for more exhaustive and more expensive affinity based methods (such as MeDIP) [66] or whole genome bisulfite sequencing (WGBS) [67]. In many cases, a biochemical analysis of DNA methylation will hence be difficult and would require time and labor-intensive acquisition of DNA as well as investment in optimization of the analysis. Especially researchers with little biomolecular knowledge will hesitate to engage in investigations on DNA methylation even though they possess a perfect expertise about their species of interest and epigenetic insights would present advancements to them. These technical difficulties have led to a distortion in the available methylation information. A review of the available data in databases and in the literature showed that at least 300 methylomes are available for Human, mouse and the model plant *Arabidopsis thaliana* but only 63 for a total of 16 other species [68–86].

Gaussian mixtures

When analyzing CpG o/e ratios related to DNA methylation, the model selection criteria AIC and BIC are regularly used for determining whether a model with two Gaussian components should be preferred to a simple

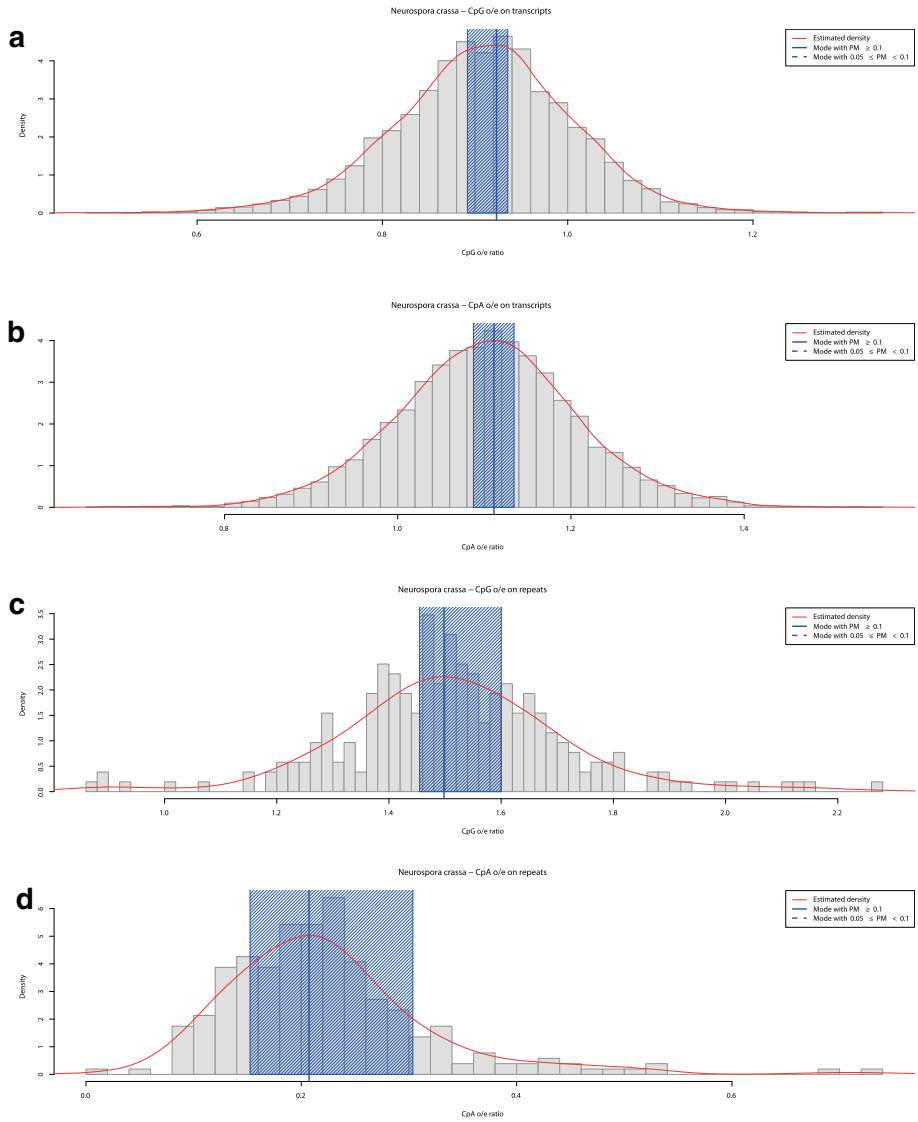

Fig. 5 CpN o/e analyzed by Notos for *Neurospora crassa*. The red line corresponds to the estimated density via KDE. Full vertical blue lines indicate modes with PM ≥ 0.1. Shaded blue areas around the modes correspond to bootstrap confidence intervals with a default level of 95%. The panels show kernels of transcripts for CpG o/e (**a**) and CpA o/e (**b**), and for repeats (**c** and **d**), respectively. In this case CpG and CpA o/e ratios were calculated for spliced exons and repeat regions of the *N. crassa* genome. Both o/e frequency distributions are clearly unimodal, but for the CpA o/e in repeats there is a shift towards 0.5 which is concordant with DNA methylation only in this context (repeats and CpA) in this species

normal distribution. This approach is at least questionable for two reasons. Firstly, model selection should be carried out taking a large number of possible models into account, and not just two (conveniently) selected alternatives. In our setting, it seems natural to consider models with more than two components as well, since the restriction to one or two components seems hard to justify from a biological perspective. This leads, however, to solutions that are (very) difficult to interpret. Secondly, models with two components may describe entirely different phenomena: on the one hand, the second component may result from a well-developed second mode. On the other hand, the second component may just result from minor deviations from normality, such as skewness or excess kurtosis. The latter behavior of both criteria results from the tendency to provide a good fit of the estimated density to the empirical data and put less emphasis on the clustering aspect, a fact investigated in more detail, e.g., by Baudry et al. [87].

Other approaches investigated

Investigating confidence intervals and their properties (width, overlap) may provide additional insight, but requires a case-by case investigation which may then

lead to subjective conclusions. We also tried to find a better balance between mode (or component) identification and non-normality by fitting mixtures of non-Gaussian distributions, e.g., via a GAMLSS-based approach [88]. This turned out to be an approach most likely suitable for in-depth analysis of a limited number of data sets. However, automatized treatment of a high number of data sets is problematic, mainly due to computational difficulties requiring manual intervention.

Conclusion

Notos allows for robust description of CpN o/e distributions and mode detection. In the future, it seems advisable to also take other aspects into account, for example skewness and kurtosis, but also simple location measures such as the location of or distance between several modes. On the long run, DNA methylation patterns should also be investigated on sequence-level, since the reduction to a CpN o/e ratio comes along with a loss of information, such as location of the (non-)methylated regions. Such an approach would, nevertheless, require the development of suitable models, and their estimation would be by far more computationally intensive than the procedures carried out by Notos. We anticipate that already the availability of Notos will make it possible to calibrate the CpN o/e distributions with existing experimental data so that precise estimations of DNA methylation can be obtained based on Notos data.

Additional files

Additional file 1: CpG o/e ratios from dbEST analyzed by Notos: data preparation output - graphics. This file shows the figure produced by the data cleaning step. (PDF 1850 kb)

Additional file 2: CpG o/e ratios from dbEST analyzed by Notos: data preparation output - table. The data preparation step of Notos carried out for 603 species from dbEST provides the tab-separated file 'outliers_cutoff.csv'. In the following we provide brief explanation on the content of the columns of this file. Future improvements of Notos may lead to changes, hence consult the the readme section of the galaxy interface.

- Name: name of the file analyzed
- prop.zero: proportion of observations equal to zero excluded (relative to original sample)
- prop.out.2iqr: proportion of values equal excluded if 2·IQR was used, relative to sample after exclusion of zeros (0 - 100)
- prop.out.3iqr: proportion of values equal excluded if 3·IQR was used, relative to sample after exclusion of zeros (0 - 100)
- prop.out.4iqr: proportion of values equal excluded if 4·IQR was used, relative to sample after exclusion of zeros (0 - 100)
- prop.out.5iqr: proportion of values equal excluded if 5·IQR was used, relative to sample after exclusion of zeros (0 - 100)
- used: IQR used for exclusion of outliers / extreme values
- no.obs.raw: number of observations in the original sample
- no.obs.nozero: number of observations in sample after excluding values equal to zero
- no.obs.clean: number of observations in sample after excluding outliers / extreme values (CSV 75.8 kb)

Additional file 3: Details on kernel density estimation. This file contains additional details on the underlying theory of kernel density estimation. (PDF 273 kb)

Additional file 4: CpG o/e ratios from dbEST analyzed by Notos: mode detection output - graphics. This file shows the graphical output from the density estimation step with activated option for the bootstrap procedure. (PDF 29500 kb)

Additional file 5: CpG o/e ratios from dbEST analyzed by Notos: mode detection output - basic statistics. The density estimation step of Notos carried out for 603 species from dbEST provides the tab-separated file 'modes_basic_stats.csv'. In the following we provide brief explanation on the content of the columns of this file. We are hereby using the following notation: σ – standard deviation, μ – mean, ν – median, Mo – mode, Q_i – the i-th quartile, q_s – the s % quantile. Future improvements of Notos may lead to changes, therefore consult the the readme section of the galaxy interface.

- Name: name of the file analyzed
- Number of modes: number of modes without applying any exclusion criterion
- Number of modes (5% excluded): number of modes after exclusion of those with less then 5% probability mass
- Number of modes (10% excluded): number of modes after exclusion of those with less then 10% probability mass
- Skewness: Pearson's moment coefficient of skewness $E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$
- Mode skewness: Pearson's first skewness coefficient $\frac{\mu-Mo}{\sigma}$
- Nonparametric skew: $\frac{\mu-\nu}{\sigma}$
- Q50 skewness: Bowley's measure of skewness / Yule's coefficient $\frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$
- Absolute Q50 mode skewness: $(Q_3+Q_1)/2-Mo$
- Absolute Q80 mode skewness: $(q_{90}+q_{10})/2-Mo$
- Peak i, $i=1,...,10$: location of peak i
- Probability Mass i, $i=1,...,10$: probability mass assigned to peak i
- Warning close modes: flag indicating that modes lie too close. The default threshold is 0.2
- Number close modes: number of modes lying too close, given the threshold
- Modes (close modes excluded): number of modes after exclusion of modes that are too close
- SD: sample standard deviation σ
- IQR 80: 80% distance between the 90% and 10% quantile
- IQR 90: 90% distance between the 95% and 5% quantile
- Total number of sequences: total number of sequences / CpG o/e ratios used for this analysis step (CSV 186 kb)

Additional file 6: CpG o/e ratios from dbEST analyzed by Notos: mode detection output - bootstrap statistics. The optional bootstrap procedure of the density estimation step of Notos carried out for 603 species from dbEST provides the tab-separated file 'modes_bootstrap.csv'. In the following we provide brief explanation on the content of the columns of this file. Future improvements of Notos may lead to changes, thus consult the the readme section of the galaxy interface.

- Name: name of the file analyzed
- Number of modes (NM): number of modes detected for the original sample
- % of samples with same NM: proportion of bootstrap samples with the same number of modes (0 - 100)
- % of samples with more NM: proportion of bootstrap samples a higher number of modes (0 - 100)
- % of samples with less NM: proportion of bootstrap samples a lower number of modes (0 - 100)
- no. of samples with same NM: number of bootstrap samples with the same number of modes
- % BS samples excluded by prob. mass crit.: proportion of bootstrap samples excluded due to strong deviations from the probability masses determined for the original sample (0 - 100) (CSV 29.8 kb)

Abbreviations
AIC: Akaike information criterion; BIC: Bayesian information criterion; CpG o/e: Observed to expected ratio of di-nucleotides composed of cytosine, followed by guanine in 5'-3' direction; CpN o/e: Observed to expected ratio of di-nucleotides composed of cytosine, followed by any nucleotide in 5'-3' direction; dbEST: Database of expressed sequence tags; ICL: Integrated completed likelihood; KDE: Kernel density estimation

Funding
This work has been supported by Campus France and the Norges forskningsrad (program AURORA, nr. 34040YK) to C. Grunau and J. Bulla, the grant Felleslegat til fordel for biologisk forskning ved Universitetet i Bergen to J. Bulla, the ANR grant ANR-10-BLAN-1720 (EpiGEvol) to C. Grunau, a PhD grant for disabled students by the French Ministry of Education and Research to B. Aliaga, and a DFG return grant to I. Bulla (BU 2685/4-1).

Availability of data and materials
All data generated or analyzed during this study are included in this published article and its supplementary information files. Notos is available in the Galaxy ToolShed and can also be downloaded from http://ihpe.univ-perp.fr/acces-aux-donnees/.

Authors' contributions
IB, BA, VL, and JB performed the experiments and developed and tested the mathematical algorithms, CG designed the experiments. IB and CC implemented the software. All authors contributed to writing the manuscript. JB coordinated the work. All authors read and approved the final manuscript.

Authors' information
The authors are grateful to Rémi Emans for technical support, and David Duval and Céline Cosseau for helpful discussions. The work on this tool was initiated during a meeting that had received funding of the French-Norwegian travel program AURORA. Therefore, Notos (the son of Aurora, the goddess of dawn in Roman mythology) was chosen as name. In addition, the name refers to the simplicity (no-to-does) of the tool.

Ethics approval and consent to participate
No ethics approval was required for this study.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Institut für Mathematik und Informatik, Universität Greifswald, Walther-Rathenau-Str. 47, 17487 Greifswald, Germany. [2]Theoretical Biology and Biophysics, Group T-6, Los Alamos National Laboratory, Los Alamos, New Mexico, USA. [3]Univ. Perpignan Via Domitia, IHPE UMR 5244, CNRS, IFREMER, Univ. Montpellier, 58 Avenue Paul Alduy, 66860 Perpignan, France. [4]Department of Mathematics, University of Bergen, P.O. Box 7803, 5020 Bergen, Norway.

Received: 1 July 2017 Accepted: 13 March 2018
Published online: 27 March 2018

References
1. Bhutani N, Burns DM, Blau HM. DNA demethylation dynamics. Cell. 2011;146(6):866–72. NIHMS150003.
2. Wan J, Oliver VF, Wang G, Zhu H, Zack DJ, Merbs SL, Qian J. Characterization of tissue-specific differential DNA methylation suggests distinct modes of positive and negative gene expression regulation. BMC Genomics. 2015;16(1):49.
3. Lev Maor G, Yearim A, Ast G. The alternative role of DNA methylation in splicing regulation. Trends Genet. 2015;31(5):274–80.
4. van der Graaf A, Wardenaar R, Neumann DA, Taudt A, Shaw RG, Jansen RC, Schmitz RJ, Colomé-Tatché M, Johannes F. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. Proc Natl Acad Sci USA. 2015;112(21):6676–81. arXiv:1410.5723.
5. Jabbari K, Bernardi G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. Gene. 2004;333(SUPPL.):143–9.
6. Fryxell KJ, Moon WJ. Cpg mutation rates in the human genome are highly dependent on local gc content. Mol Biol Evol. 2005;22(3):650–8. https://doi.org/10.1093/molbev/msi043.
7. Cooper DN, Krawczak M. Cytosine methylation and the fate of cpg dinucleotides in vertebrate genomes. Hum Genet. 1989;83:181–8.
8. Nicholson S, Nickerson M, Dean M, Song Y, Hoyt P, Rhee H, Kim C, Puterka G. The genome of diuraphis noxia, a global aphid pest of small grains. BMC Genomics. 2015;16:429.
9. Cunningham C, Ji L, Wiberg R, Shelton J, McKinney E, Parker D, Meagher R, Benowitz K, Roy-Zokan E, Ritchie M, Brown S, Schmitz R, Moore A. The genome and methylome of a beetle with complex social behavior, nicrophorus vespilloides (coleoptera: Silphidae). Genome Biol Evol. 2015;7(12):3383–96.
10. Wang X, Fang X, Yang P, Jiang X, Jiang F, Zhao D, Li B, Cui F, Wei J, Ma C, Wang Y, He J, Luo Y, Wang Z, Guo X, Guo W, Wang X, Zhang Y, Yang M, Hao S, Chen B, Ma Z, Yu D, Xiong Z, Zhu Y, Fan D, Han L, Wang B, Chen Y, Wang J, Yang L, Zhao W, Feng Y, Chen G, Lian J, Li Q, Huang Z, Yao X, Lv N, Zhang G, Li Y, Wang J, Wang J, Zhu B, Kang L. The locust genome provides insight into swarm formation and long-distance flight. Nat Commun. 2014;5:2957.
11. Oxley P, Ji L, Fetter-Pruneda I, McKenzie S, Li C, Hu H, Zhang G, Kronauer D. The genome of the clonal raider ant cerapachys biroi. Curr Biol. 2014;24(4):451–8.
12. Chipman A, Ferrier D, Brena C, Qu J, Hughes D, Schröder R, Torres-Oliva M, Znassi N, Jiang H, Almeida F, Alonso C, Apostolou Z, Aqrawi P, Arthur W, Barna J, Blankenburg K, Brites D, Capella-Gutiérrez S, Coyle M, Dearden P, Du Pasquier L, Duncan E, Ebert D, Eibner C, Erikson G, Evans P, Extavour C, Francisco L, Gabaldón T, Gillis W, Goodwin-Horn E, Green J, Griffiths-Jones S, Grimmelikhuijzen C, Gubbala S, Guigó R, Han Y, Hauser F, Havlak P, Hayden L, Helbing S, Holder M, Hui J, Hunn J, Hunnekuhl V, Jackson L, Javaid M, Jhangiani S, Jiggins F, Jones T, Kaiser T, Kalra D, Kenny N, Korchina V, Kovar C, Kraus F, Lapraz F, Lee S, Lv J, Mandapat C, Manning G, Mariotti M, Mata R, Mathew T, Neumann T, Newsham I, Ngo D, Ninova M, Okwuonu G, Ongeri F, Palmer W, Patil S, Patraquim P, Pham C, Pu L, Putman N, Rabouille C, Ramos O, Rhodes A, Robertson H, Robertson H, Ronshaugen M, Rozas J, Saada N, Sánchez-Gracia A, Scherer S, Schurko A, Siggens K, Simmons D, Stief A, Stolle E, Telford M, Tessmar-Raible K, Thornton R, van der Zee M, von Haeseler A, Williams J, Willis J, Wu Y, Zou X, Lawson D, Muzny D, Worley K, Gibbs R, Akam M, Richards S. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede strigamia maritima. PLoS Biol. 2014;12(11):1002005.
13. Chen F, Chuang T, Lin H, Hsu M. The evolution of the coding exome of the arabidopsis species–the influences of dna methylation, relative exon position, and exon length. BMC Evol Biol. 2014;14:145.
14. Terrapon N, Li C, Robertson HM, Ji L, Meng X, Booth W, Chen Z, Childers CP, Glastad KM, Gokhale K, Gowin J, Gronenberg W, Hermansen RA, Hu H, Hunt BG, Huylmans AK, Khalil SMS, Mitchell RD, Munoz-Torres MC, Mustard JA, Pan H, Reese JT, Scharf ME, Sun F, Vogel H, Xiao J, Yang W, Yang Z, Yang Z, Zhou J, Zhu J, Brent CS, Elsik CG, Goodisman MAD, Liberles DA, Roe RM, Vargo EL, Vilcinskas A, Wang J, Bornberg-Bauer E, Korb J, Zhang G, Liebig J. Molecular traces of alternative social organization in a termite genome. Nat Commun. 2014;5:3636.
15. Dixon G, Bay L, Matz M. Bimodal signatures of germline methylation are linked with gene expression plasticity in the coral acropora millepora. BMC Genomics. 2014;15:1109.
16. Kocher S, Li C, Yang W, Tan H, Yi S, Yang X, Hoekstra H, Zhang G, Pierce N, Yu D. The draft genome of a socially polymorphic halictid bee, lasioglossum albipes. Genome Biol. 2013;14(12):142.
17. Simola D, Wissler L, Donahue G, Waterhouse R, Helmkampf M, Roux J, Nygaard S, Glastad K, Hagen D, Viljakainen L, Reese J, Hunt B, Graur D, Elhaik E, Kriventseva E, Wen J, Parker B, Cash E, Privman E, Childers C, Muñoz-Torres M, Boomsma J, Bornberg-Bauer E, Currie C, Elsik C, Suen G,

Goodisman M, Keller L, Liebig J, Rawls A, Reinberg D, Smith C, Smith C, Tsutsui N, Wurm Y, Zdobnov E, Berger S, Gadau J. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. Genome Res. 2013;23(8):1235–47.
18. Glastad K, Hunt B, Goodisman M. Evidence of a conserved functional role for dna methylation in termites. Insect Mol Biol. 2013;22(2):143–54.
19. Fneich S, Dheilly N, Adema C, Rognon A, Reichelt M, Bulla J, Grunau C, Cosseau C. 5-methyl-cytosine and 5-hydroxy-methyl-cytosine in the genome of biomphalaria glabrata, a snail intermediate host of schistosoma mansoni. Parasit Vectors. 2013;6(1):167.
20. Sarda S, Zeng J, Hunt B, Yi S. The evolution of invertebrate gene body methylation. Mol Biol Evol. 2012;29(8):1907–16.
21. Albalat R, Martí-Solans J, Cañestro C. Dna methylation in amphioxus: from ancestral functions to new roles in vertebrates. Brief Funct Genomics. 2012;11(2):142–55.
22. Zhan S, Merlin C, Boore J, Reppert S. The monarch butterfly genome yields insights into long-distance migration. Cell. 2011;147(5):1171–85.
23. Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera E, Cash E, Cavanaugh A, Denas O, Elhaik E, Favé M, Gadau J, Gibson J, Graur D, Grubbs K, Hagen D, Harkins T, Helmkampf M, Hu H, Johnson B, Kim J, Marsh S, Moeller J, Muñoz-Torres M, Murphy M, Naughton M, Nigam S, Overson R, Rajakumar R, Reese J, Scott J, Smith C, Tao S, Tsutsui N, Viljakainen L, Wissler L, Yandell M, Zimmer F, Taylor J, Slater S, Clifton S, Warren W, Elsik C, Smith C, Weinstock G, Gerardo N, Currie C. The genome sequence of the leaf-cutter ant atta cephalotes reveals insights into its obligate symbiotic lifestyle. PLoS Genet. 2011;7(2):1002007.
24. Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt B, Ingram K, Falquet L, Nipitwattanaphon M, Gotzek D, Dijkstra M, Oettler J, Comtesse F, Shih C, Wu W, Yang C, Thomas J, Beaudoing E, Pradervand S, Flegel V, Cook E, Fabbretti R, Stockinger H, Long L, Farmerie W, Oakey J, Boomsma J, Pamilo P, Yi S, Heinze J, Goodisman M, Farinelli L, Harshman K, Hulo N, Cerutti L, Xenarios I, Shoemaker D, Keller L. The genome of the fire ant solenopsis invicta. Proc Natl Acad Sci USA. 2011;108(14):5679–84.
25. Smith C, Smith C, Robertson H, Helmkampf M, Zimin A, Yandell M, Holt C, Hu H, Abouheif E, Benton R, Cash E, Croset V, Currie C, Elhaik E, Elsik C, Favé M, Fernandes V, Gibson J, Graur D, Gronenberg W, Grubbs K, Hagen D, Viniegra A, Johnson B, Johnson R, Khila A, Kim J, Mathis K, Munoz-Torres M, Murphy M, Mustard J, Nakamura R, Niehuis O, Nigam S, Overson R, Placek J, Rajakumar R, Reese J, Suen G, Tao S, Torres C, Tsutsui N, Viljakainen L, Wolschin F, Gadau J. Draft genome of the red harvester ant pogonomyrmex barbatus. Proc Natl Acad Sci USA. 2011;108(14):5667–72.
26. Smith C, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie C, Elhaik E, Elsik C, Fave M, Fernandes V, Gadau J, Gibson J, Graur D, Grubbs K, Hagen D, Helmkampf M, Holley J, Hu H, Viniegra A, Johnson B, Johnson R, Khila A, Kim J, Laird J, Mathis K, Moeller J, Muñoz-Torres M, Murphy M, Nakamura R, Nigam S, Overson R, Placek J, Rajakumar R, Reese J, Robertson H, Smith C, Suarez A, Suen G, Suhr E, Tao S, Torres C, van Wilgenburg E, Viljakainen L, Walden K, Wild A, Yandell M, Yorke J, Tsutsui N. Draft genome of the globally widespread and invasive argentine ant (linepithema humile). Proc Natl Acad Sci USA. 2011;108(14):5673–8.
27. Park J, Peng Z, Zeng J, Elango N, Park T, Wheeler D, Werren J, Yi S. Comparative analyses of dna methylation and sequence evolution using nasonia genomes. Mol Biol Evol. 2011;28(12):3345–54.
28. Xiang H, Zhu J, Chen Q, Dai F, Li X, Li M, Zhang H, Zhang G, Li D, Dong Y, Zhao L, Lin Y, Cheng D, Yu J, Sun J, Zhou X, Ma K, He Y, Zhao Y, Guo S, Ye M, Guo G, Li Y, Li R, Zhang X, Ma L, Kristiansen K, Guo Q, Jiang J, Beck S, Xia Q, Wang W, Wang J. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat Biotechnol. 2010;28(5):516–20.
29. Gavery M, Roberts S. Dna methylation patterns provide insight into epigenetic regulation in the pacific oyster (crassostrea gigas). BMC Genomics. 2010;11:483.
30. Zeng J, Yi S. Dna methylation and genome evolution in honeybee: gene length, expression, functional enrichment covary with the evolutionary signature of dna methylation. Genome Biol Evol. 2010;2:770–80.
31. Walsh T, Brisson J, Robertson H, Gordon K, Jaubert-Possamai S, Tagu D, Edwards O. A functional dna methylation system in the pea aphid, acyrthosiphon pisum. Insect Mol Biol. 2010;19 Suppl 2:215–28.
32. Elango N, Hunt B, Goodisman M, Yi S. Dna methylation is widespread and associated with differential gene expression in castes of the honeybee, apis mellifera. Proc Natl Acad Sci USA. 2009;106(27):11206–11.
33. Yi S, Goodisman M. Computational approaches for understanding the evolution of dna methylation in animals. Epigenetics. 2009;4(8):551–6.
34. Suzuki M, Kerr A, De Sousa D, Bird A. Cpg methylation is targeted to transcription units in an invertebrate genome. Genome Res. 2007;17(5):625–31.
35. Liaud M, Valentin C, Brandt U, Bouget F, Kloareg B, Cerff R. The gapdh gene system of the red alga chondrus crispus: promoter structures, intron/exon organization, genomic complexity and differential expression of genes. Plant Mol Biol. 1993;23(5):981–94.
36. Ellis J, Griffin H, Morrison D, Johnson A. Analysis of dinucleotide frequency and codon usage in the phylum apicomplexa. Gene. 1993;126(2):163–70.
37. Yi SV, Goodisman MaD. Computational approaches for understanding the evolution of DNA methylation in animals. Epigenetics Off J DNA Methylation Soc. 2009;4(8):551–6.
38. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet. 2008;9(6):465–76.
39. Boguski MS, Tolstoshev TMJLCM. dbEST-database for "expressed sequence tags". Nat Genet. 1993;4:332–3.
40. R Core Team. R: A language and environment for statistical computing. Vienna: R foundation for statistical computing; 2017. https://www.R-project.org/.
41. Matsuo K, Clay O, Takahashi T, Silke J, Schaffner W. Evidence for erosion of mouse CpG islands during mammalian evolution. Somat Cell Mol Genet. 1993;19(6):543–55.
42. Gardiner-Garden M, Frommer M. CpG Islands in vertebrate genomes. J Mol Biol. 1987;196(2):261–82.
43. Zeng J, Yi SV. DNA Methylation and Genome Evolution in Honeybee: Gene Length, Expression, Functional Enrichment Covary with the Evolutionary Signature of DNA Methylation. Genome Biol Evol. 2010;2:770–80. gbe/evq060.
44. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci USA. 2006;103(5):1412–7.
45. Silverman BW. Density estimation for statistics and data analysis. Monogr Stat Appl Probab. 1986;37(1):1–22.
46. Scott DW. Multivariate Density Estimation. Theory, Practice, and Visualization. New York: Wiley; 1992, p. 511.
47. Gavery MR, Roberts SB. DNA methylation patterns provide insight into epigenetic regulation in the Pacific oyster (Crassostrea gigas). BMC Genomics. 2010;11(1):483.
48. Fneich S, Dheilly N, Adema C, Rognon A, Reichelt M, Bulla J, Grunau C, Cosseau C. 5-methyl-cytosine and 5-hydroxy-methyl-cytosine in the genome of Biomphalaria glabrata, a snail intermediate host of Schistosoma mansoni. Parasites Vectors. 2013;6:167.
49. Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, Schröder R, Torres-Oliva M, Znassi N, Jiang H, Almeida FC, Alonso CR, Apostolou Z, Aqrawi P, Arthur W, Barna JCJ, Blankenburg KP, Brites D, Capella-Gutiérrez S, Coyle M, Dearden PK, Du Pasquier L, Duncan EJ, Ebert D, Eibner C, Erikson G, Evans PD, Extavour CG, Francisco L, Gabaldón T, Gillis WJ, Goodwin-Horn EA, Green JE, Griffiths-Jones S, Grimmelikhuijzen CJP, Gubbala S, Guigó R, Han Y, Hauser F, Havlak P, Hayden L, Helbing S, Holder M, Hui JHL, Hunn JP, Hunnekuhl VS, Jackson LR, Javaid M, Jhangiani SN, Jiggins FM, Jones TE, Kaiser TS, Kalra D, Kenny NJ, Korchina V, Kovar CL, Kraus FB, Lapraz F, Lee SL, Lv J, Mandapat C, Manning G, Mariotti M, Mata R, Mathew T, Neumann T, Newsham I, Ngo DN, Ninova M, Okwuonu G, Ongeri F, Palmer WJ, Patil S, Patraquim P, Pham C, Pu LL, Putman NH, Rabouille C, Ramos OM, Rhodes AC, Robertson HE, Robertson HM, Ronshaugen M, Rozas J, Saada N, Sánchez-Gracia A, Scherer SE, Schurko AM, Siggens KW, Simmons DN, Stief A, Stolle E, Telford MJ, Tessmar-Raible K, Thornton R, van der Zee M, von Haeseler A, Williams JM, Willis JH, Wu Y, Zou X, Lawson D, Muzny DM, Worley KC, Gibbs RA, Akam M, Richards S. The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede Strigamia maritima. PLoS Biol. 2014;12(11):1–24.

50. Bewick AJ, Vogel KJ, Moore AJ, Schmitz RJ. Evolution of DNA Methylation across Insects. Mol Biol Evol. 2016;34(3):264. https://doi.org/10.1093/molbev/msw264.
51. Benaglia T, Chauveau D, Hunter D, Young D. mixtools: An r package for analyzing mixture models. J Stat Softw Artic. 2009;32(6):1–29. https://doi.org/10.18637/jss.v032.i06.
52. Fraley C. MCLUST: Software for model-based cluster analysis. J Classif. 1999;16(2):297–306. https://doi.org/10.1007/s003579900058.
53. Leisch F. Flexmix: A general framework for finite mixture models and latent class regression in r. J Stat Softw Artic. 2004;11(8):1–18. https://doi.org/10.18637/jss.v011.i08.
54. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans Pattern Anal Mach Intell. 2000;22(7):719–25. https://doi.org/10.1109/34.865189.
55. Bozdogan H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika. 1987;52(3):345–70. https://doi.org/10.1007/BF02294361.
56. Illingworth R, Bird A. Cpg islands–'a rough guide'. FEBS Lett. 2009;583(11): 1713–20.
57. Aliaga B, Bulla I, Mouahid G, Duval D, Grunau C. The evolution of gene body dna methylation in eucaryotes. Sci Rep. 2018. Under review.
58. Selker E, Tountas N, Cross S, Margolin B, Murphy J, Bird A, Freitag M. The methylated component of the neurospora crassa genome. Nature. 2003;422(6934):893–7.
59. Selker E. Repeat-induced gene silencing in fungi. Adv Genet. 2002;46: 439–50.
60. Cambareri E, Jensen B, Schabtach E, Selker E. Repeat-induced g-c to a-t mutations in neurospora. Science. 1989;244(4912):1571–5.
61. Clutterbuck A. Genomic evidence of repeat-induced point mutation (rip) in filamentous ascomycetes. Fungal Genet Biol. 2011;48(3):306–26.
62. Mandel J, Chambon P. Dna methylation: organ specific variations in the methylation pattern within and around ovalbumin and other chicken genes. Nucleic Acids Res. 1979;7(8):2081–103.
63. Yamamoto F, Yamamoto M, Soto J, Kojima E, Wang E, Perucho M, Sekiya T, Yamanaka H. NotI-msell methylation-sensitive amplied fragment length polymorhism for dna methylation analysis of human cancers. Electrophoresis. 2001;22(10):1946–56.
64. Meissner A, Gnirke A, Bell G, Ramsahoye B, Lander E, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. Nucleic Acids Res. 2005;33(18):5868–77.
65. van Gurp T, Wagemaker N, Wouters B, Vergeer P, Ouborg J, Verhoeven K. epigbs: reference-free reduced representation bisulfite sequencing. Nat Methods. 2016;13(4):322–4.
66. Weber M, Davies J, Wittig D, Oakeley E, Haase M, Lam W, Schubeler D. Chromosome-wide and promoter-specific analyses identify sites of differential dna methylation in normal and transformed human cells. Nat Genet. 2005;37(8):853–62.
67. Lister R, Ecker J. Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res. 2009;19(6):959–66.
68. Beeler SM, Wong GT, Zheng JM, Bush EC, Remnant EJ, Oldroyd BP, Drewell Ra. Whole-genome DNA methylation profile of the jewel wasp (Nasonia vitripennis). G3 (Bethesda, Md.) 2014;4(3):383–8. https://doi.org/10.1534/g3.113.008953.
69. Drewell RA, Bush EC, Remnant EJ, Wong GT, Beeler SM, Stringham JL, Lim J, Oldroyd BP. The dynamic DNA methylation cycle from egg to sperm in the honey bee Apis mellifera. Development. 2014;141(13): 2702–11.
70. 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):1–12.
71. Zemach A, Mcdaniel I, Silva P, Zilberman D. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. Science (New York, NY). 2010;11928(May 2008):1186366–1.
72. Herb BR, Wolschin F, Hansen KD, Aryee MJ, Langmead B, Irizarry R, Amdam GV, Feinberg AP. Reversible switching between epigenetic states in honeybee behavioral subcastes. Nat Neurosci. 2012;15(10):1371–3.
73. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, Ukomadu C, Sadler KC, Pradhan S, Pellegrini M, Jacobsen SE. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci USA. 2010;107(19):8689–94. https://doi.org/10.1073/pnas.1002720107.
74. Bonasio R, Li Q, Lian J, Mutti NS, Jin L, Zhao H, Zhang P, Wen P, Xiang H, Ding Y, et al. Genome-wide and caste-specific dna methylomes of the ants camponotus floridanus and harpegnathos saltator. Curr Biol. 2012;22(19):1755–64.
75. Hunt BG, Glastad KM, Yi SV, Goodisman MA. Patterning and regulatory associations of dna methylation are mirrored by histone modifications in insects. Genome Biol Evol. 2013;5(3):591–8.
76. Glastad KM, Hunt BG, Soojin VY, Goodisman MA. Epigenetic inheritance and genome regulation: is dna methylation linked to ploidy in haplodiploid insects? Proc R Soc Lond B Biol Sci. 2014;281(1785):20140411.
77. Xiang H, Zhu J, Chen Q, Dai F, Li X, Li M, Zhang H, Zhang G, Li D, Dong Y, et al. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat Biotechnol. 2010;28(5):516–20.
78. Xiang H, Li X, Dai F, Xu X, Tan A, Chen L, Zhang G, Ding Y, Li Q, Lian J, et al. Comparative methylomics between domesticated and wild silkworms implies possible epigenetic influences on silkworm domestication. BMC Genomics. 2013;14(1):646.
79. Falckenhayn C, Boerjan B, Raddatz G, Frohme M, Schoofs L, Lyko F. Characterization of genome methylation patterns in the desert locust schistocerca gregaria. J Exp Biol. 2013;216(8):1423–9.
80. Wang X, Li Q, Lian J, Li L, Jin L, Cai H, Xu F, Qi H, Zhang L, Wu F, et al. Genome-wide and single-base resolution dna methylomes of the pacific oyster crassostrea gigas provide insight into the evolution of invertebrate cpg methylation. BMC Genomics. 2014;15(1):1119.
81. Rondon R, Grunau C, Fallet M, Charlemagne N, Sussarellu R, Chaparro C, Montagnani C, Mitta G, Bachère E, Akcha F, et al. Effects of a parental exposure to diuron on pacific oyster spat methylome. Environ Epigenetics. 2017;3(1):dvx004.
82. Adema CM, Hillier LW, Jones CS, Loker ES, Knight M, Minx P, Oliveira G, Raghavan N, Shedlock A, do Amaral LR, et al. Whole genome analysis of a schistosomiasis-transmitting freshwater snail. Nat Commun. 2017;8:1–11.
83. Suzuki MM, Yoshinari A, Obara M, Takuno S, Shigenobu S, Sasakura Y, Kerr AR, Webb S, Bird A, Nakayama A. Identical sets of methylated and nonmethylated genes in ciona intestinalis sperm and muscle cells. Epigenetics Chromatin. 2013;6(1):38.
84. Suzuki MM, Kerr AR, De Sousa D, Bird A. Cpg methylation is targeted to transcription units in an invertebrate genome. Genome Res. 2007;17(5): 625–31.
85. Raddatz G, Guzzardo PM, Olova N, Fantappié MR, Rampp M, Schaefer M, Reik W, Hannon GJ, Lyko F. Dnmt2-dependent methylomes lack defined dna methylation patterns. Proc Natl Acad Sci. 2013;110(21):8627–31.
86. Dabe EC, Sanford RS, Kohn AB, Bobkova Y, Moroz LL. Dna methylation in basal metazoans: Insights from ctenophores. Integr Comp Biol. 2015;55(6):1096–110.
87. Baudry JP, Raftery AE, Celeux G, Lo K, Gottardo R. Combining mixture components for clustering. J Comput Graph Stat. 2010;19(2):332–53. https://doi.org/10.1198/jcgs.2010.08111.
88. Stasinopoulos D, Rigby R. Generalized additive models for location scale and shape (gamlss) in r. J Stat Softw Artic. 2007;23(7):1–46. https://doi.org/10.18637/jss.v023.i07.