Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq

Johannes, Frank; Wardenaar, Rene; Colomé Tatché, Maria; Mousson, Florence; Mauritz, Petra; Mokry, Michal; Guryev, Victor; Timmers, H. Th. Marc; Cuppen, Edwin; Jansen, Ritsert

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/btq087

Publication date:
2010

Supplemental text

1) ChIP-seq data preparation

a. siRNA-mediated knockdown and chromatin immunoprecipitation

For the human basal transcription factor data, Hela tk- cells were transfected with a mixture of siRNAs (Dharmacon) targeting either BTAF1 or GAPDH mRNA. Two days later, subconfluent cultures of HeLa tk- cells were cross-linked by addition of 1% formaldehyde in PBS for 10 minutes at 37°C. Cells were lysed in buffer (50 mM Tris-HCl pH 7.9, 1% SDS, 10 mM EDTA, 1 mM DTT, and protease inhibitors). The lysate was sonicated 8 times for 30 seconds in a Bioruptor (Diagenode, Belgium) resulting in DNA fragments of 200 to 600 bp. Soluble material was supplemented with 0.1% Triton X-100 and 0.1% Na-deoxycholate and incubated for 6 hours with Dynabeads coupled to antibody against TBP (SL30). Samples were processed as previously described (Vermeulen et al., 2007). DNA concentration was measured using PicoGreen dsDNA reagent kit. In parallel knock-downs were controlled on both protein and RNA levels.

b. Library preparation

Chromatin was additionally sheared for 6 minutes using Covaris sonicator (6 x 16 mm AFA fiber Tube, duty cycle: 20%, intensity: 5, cycles/burst: 200, frequency sweeping) to obtain suitable shorter fragments (75-125 bp). After fragmentation, fragments were blunt-ended and phosophorylated at the 5'-prime end using the End-it Kit (Epicentre) according to the manufacturer's instructions. Ligation of double stranded adapters compatible with SOLiD sequencing was performed using Quick ligation kit (New England Biolabs) with 750 mM P1 ds and P2 ds adaptor (Applied Biosystems), 11.7 µl of 2x Quick ligation buffer, 1 µl Quick Ligase (NEB) in total volume of 23.4 µl. Samples were purified using Ampure beads (Agencourt) and run on a native 6% polyacrylamide gel. Fragments ranging from 140 to 180 bp were excised; the piece of gel containing DNA fragments was shredded and dispersed into 400 µl of Platinum PCR Supermix with 750 mM of each P1 and P2 PCR primer, 2,5 U of Pfu (Stratagene) and 5 U Taq (Bioline). Prior to ligation-mediated PCR the sample was incubated at 72^{O} C for 20 minutes in PCR mix to let the DNA diffuse out of the gel and to perform nick translation on non ligated 3'-ends of DNA fragments. After 17 cycles of amplification the library was purified using Ampure beads and was quality checked on 2100 Bioanalyzer (Agilent) for the absence of possible adapter dimers and heterodimers.

c. Templated beads preparation

To achieve clonal amplification of library fragments on the surface of sequencing beads, emulsion PCR (ePCR) was performed according to the manufacturer's instructions (Applied Biosystems). 600 pg of double stranded library DNA was added to 5.6 ml of PCR mix containing 1x PCR Gold Buffer (Applied Biosystems), 3000 U AmpliTaq Gold, 20nM ePCR primer 1, 3 µM of ePCR primer 2, 3.5 mM of each deoxynucleotide, 25mM $MgCl_2$ and 1.6 billion SOLiD sequencing beads (Applied Biosystems). PCR mix was added to SOLiD ePCR Tube containing 9 ml of oil phase and emulsified using ULTRA-TURRAX Tube Drive (IKA). Emulsion was dispensed into 96-well plate and cycled for 60 cycles. After amplification emulsion was broken with butanol, beads were enriched for template positive beads, 3'-end extended and covalently attached onto one quadrant of

sequencing slide and sequenced using SOLiD system version 2 to produce 35 bases long reads.

2) Comparison with other clustering algorithms

a. Example data results

Clustering or classification methods are common in high-dimensional data problems and have been extensively employed in array-based gene expression analysis (Quackenbush, 2001), and more recently in DNA methylation profiling (Marjoram et al., 2006; Houseman et al., 2008; Siegmund et al., 2004). Their primary purpose is to group signals based on similar intensity patterns. The unique feature of the classification approach outlined here is that the clustering procedure itself is highly structured and consistent with alternative biological models that could have generated the data. The best fitting model therefore provides potentially valuable insights into the global behavior of chromatin modifications across conditions (e.g. tissues). This dimension of our approach has no counterpart in existing methods, and it is difficult to find meaningful grounds for comparison. Nonetheless, to provide some type of reference, we compare the performance of one of our models (model 2) with three commonly employed clustering algorithms in biological analysis, hierarchical clustering, K-means partitioning, and the model-based multivariate normal clustering as implemented through the *Mclust* R package (Fraley and Raftery, 2007). Details concerning the specification of these alternative methods are provided in the methods section.

First we considered the datasets discussed above and applied each of the alternative methods to these data**.** We found that the classification results vary quite substantially between these methods (**Figure S3**), and do not appear to agree well with our biological expectations of the distribution of the data. This may be in part attributable to the relatively small number of **RDE**, which are difficult to detect without sensible constraints in the classification procedure, especially when sub-sampling of the data is involved (see below). However, based on these clustering results alone we cannot draw any firm conclusions about the relative merits of each of these methods as the 'state of nature' is unknown. We therefore extended the comparison to several simulated datasets.

b. Simulation results

We generated two types of datasets (A and B). Dataset A was simulated directly from model 2, and hence fully met the distributional assumptions of this model. In contrast, dataset B deliberately violated the assumption of bivariate normality of the **RDE** components (see below). We applied each of the alternative classification methods to these data and compared their performance directly to model 2. For each data example we report a component-specific false positive and false negative rate (+/- standard errors) based on 50 simulation replicates (see below). The results from this comparison are shown in **Figure S4**. As expected model 2 performs well in the case of dataset A, but also appears to maintain a reasonable false positive and false negative rate for dataset B. The alternative methods considered here consistently reported exaggerated false positive or false negative rates. This trend is particularly apparent for components 3 and 4 (the two **RDE** components), which we predicted to be difficult to detect due to their small size. We also noticed substantial variation in the classification obtained with hierarchical clustering as well as multivariate clustering of the *Mclust* package. While this may be

directly attributable to the sub-sampling scheme that is necessary for the efficient implementation of these two methods, in the case of *Mclust* it may also stem from the built-in model-selection procedure (see below), whose best-fitting model may not consistently approximate the true model.

c. Implementation details for alternative algorithms

Mclust: We utilized the multivariate modeling R package *mclust* as an alternative model-based classification approach (Fraley and Raftery). In this package, we made use of *mclust*'s extensive model selection procedure, which compares several models with different covariance constraints. The best model is automatically selected based on its BIC value and used to obtain a final classification of the data. Application of this package to very large datasets proves to be quite inefficient and parameter estimates were therefore obtained from a random subset of the data (N = 2000), as recommended by the authors (Fraley and Raftery, 2007).

Hierarchical clustering: Hierarchical clustering was perform using the standard R (R Development Core Team) function, *hclust* (default settings), followed by the function *cutree* to extract four separate clusters. Since the algorithm requires the specification of a distance matrix, large sample sizes become prohibitive. The same random subset of the data as above (N = 2000) was therefore used to obtain an initial clustering of the data. However, in the case of hierarchical clustering there is no 'direct' way to move from the classification of the subsample to the full sample. We solved this problem by estimating bivariate normal parameters for each of the clusters detected in the subsample and used these estimates along with empirical mixing weights corresponding to the relative size of the clusters to the whole dataset.

K-means clustering: The K-means algorithm was implemented via the R (R Development Core Team) function *kmeans* with a specification for a four cluster solution. Since the solution is very dependent on the initial bivariate cluster centers, a set of 1000 random starting values for the cluster centers was chosen.

d. Implementation details for simulation analysis

We generated two types of datasets (A and B), each of dimension (n x d = 30000 x 2). Dataset A was simulated from model 2, and hence fully met the distributional assumptions of this model. In contrast, dataset B included violations to the assumptions of bivariate normality of the **RDE** components by letting the observations $\mathbf{w}_j$ be drawn from the bivariate skewed normal distribution (Wuertz), with the skew in the direction away from the diagonal. This situation may be encountered when the differences between the two conditions are very strong for several regions within the **RDE** components. The precise parameter specifications for each of these simulation models are provided below. We generated 50 replicates of datasets A and B and applied the different classification methods at each simulation run. For the l^{th} simulation run $(l = 1,\ldots,50)$ and the i^{th} mixture component $(i = 1,\ldots,4)$, we determined the False Positive rate (FP) and the False Negative rate (FN) as $FP_l = 1 - \#(T_{il} \cap D_{il})/\#D_{il}$ and $FN_l = 1 - \#(T_{il} \cap D_{il})/\#T_{il}$, where T and D denote the sets of true and detected component memberships, respectively. Estimates $\hat{FP}$ and $\hat{FN}$ as well as their standard errors $\hat{SE}_{FP}$ and $\hat{SE}_{FN}$, were obtained

empirically by considering the expected values and the standard deviations of the simulation distributions. As the component indexing may vary from run to run (label-switching problem), care was taken to convert the component labeling in accordance with the expected spatial location of the cluster means in the scatterplot (see **Figure 1, Figure S1**). For each clustering method we made the following specifications:

Parameter settings for simulation study: Dataset A was directly generated from model 2. The variance-covariance matrices $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$, and $\mathbf{\Sigma}$ were equivalent to the estimated variance-covariance matrices from the *Arabidopsis* methlation data. The mean vector, $\boldsymbol{\mu}$, and the mixing weights λ_1, λ_2, λ_3 and λ_4 were defined arbitrarily, and were fixed at $\boldsymbol{\mu} = (\mu_1, \mu_2) = (-0.5,1)$, $\lambda_1 = 0.7$, $\lambda_2 = 0.28$, $\lambda_3 = 0.01$, $\lambda_4 = 0.01$. The variance-covariance matrices $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$, and $\mathbf{\Sigma} = \mathbf{\Sigma}_3 = \mathbf{\Sigma}_4$ were set at:

$$\Sigma_1 = \begin{bmatrix} 0.59 & 0.54 \\ 0.54 & 0.59 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 0.33 & 0.28 \\ 0.28 & 0.33 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 0.56 & 0.42 \\ 0.42 & 0.56 \end{bmatrix}$$

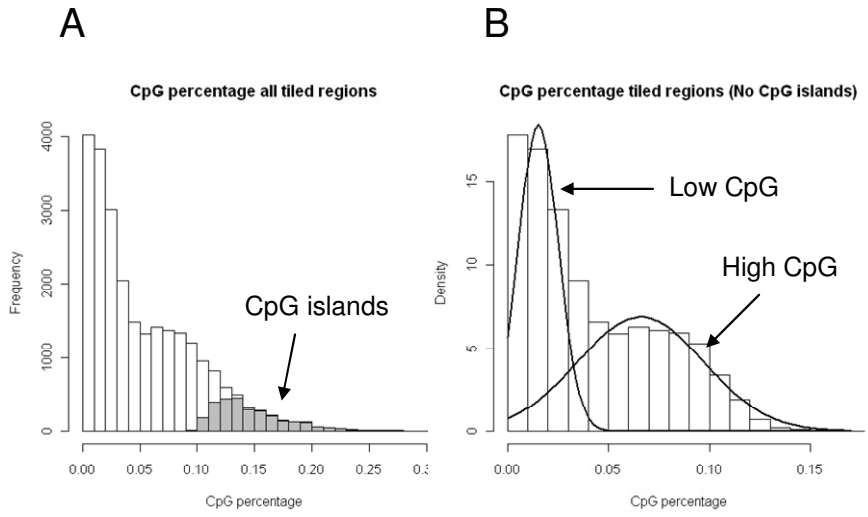
The parameter specifications for dataset B were equivalent to those for dataset A, except that the random values $\mathbf{w}_j$ were not drawn from a mixture of bivariate normal distributions but from a mixture of bivariate skewed normal distribution, as implemented through the *fMultivar* package in R. To define this distribution, the skewness parameter $\mathbf{a} = (\alpha_x, \alpha_y)$ for components 3 and 4 were taken to be, $\mathbf{a} = (-3,3)$ and $\mathbf{a} = (3,-3)$, respectively. This had the effect of producing skewness away from the diagonal of the bivariate plot, which is consistent with the idea that many regions in **RDE** components 3 and 4 show large differences, while the bulk of the data in those components shows a less divergent pattern.

3) Annotation-based genome partitioning: An example

a. Implementation details

In the main text we discussed an annotation-based genome partitioning strategy (Section 2.1.6). This approach uses information about particular sequence contexts in the analysis framework. We illustrate this approach using the mouse promoter methylation data. Specifically, we explored the effect of distinguishing between promoters that are classified as CpG islands, high CpG content (but not CpG island), and low CpG content. This partitioning strategy may be useful insofar that it has been argued that ChIP-chip signals can influenced by the CpG content of particular probes or sequences (Roys et al., 2007). The resulting signal distributions may therefore require fitting distributions with means and variances that are specific to each probe or sequence set.

The three different promoter classes considered here were obtained as follows: We first selected promoter sequences that met the CpG island criterion according to Takai, and Jones (2002), see Figure A. When plotting a histogram of the CpG content of the remaining promoter sequences, we noticed a bi-model distribution (see Figure B). This suggested to us that there are probably two additional populations of promoter sequences, one corresponding to high CpG content (but not CpG island) and the other corresponding to low CG content. We used univariate normal mixture classification scheme to assigned these remaining promoter sequences to the high and low CpG content groups. All together this resulted in three annotation sets. Their respective sample sizes are shown in the Table below.



b. Model comparisons

For each annotation set we fitted our three mixture models (model 1, model 2 and model 3), and selected the best fitting model. The 'composite' log-likelihood (as well the AIC value) were compared to a full model in which no distinction was made between the three annotation sets. From this analysis we find overwhelming evidence against partitioning the data into different sequence contexts (AIC full data = 656143 versus AIC portioned

data = 1328802) . This means that in the case of the promoter mouse methylation data, a simultaneous treatment of all promoters provides more information.

Annotation sets	*n*	Best fitting model	Total parameters	log-likelihood	AIC
Full data	371694	model 3	14	-328057.8	656143
CpG islands	42164	model 3			
high CpG content	164289	model 3	42	-664358.9	1328802
low CpG content	165241	model 3			

4) Supplemental Tables

Below we provide a representative summary of available methods for ChIP-chip (Table S1) and ChIP-seq (Table S2) analysis. We classified these methods based on the following criteria (see columns).

1. Author: author of original publication.
2. Year: year of publication.
3. Name: name of the algorithm, if applicable.
4. ChIP sample comparisons: Does the method have capabilities to perform formal comparisons between two or more ChIP-samples.
5. Input-normalized data: Does the method handle Input-normalized data.
6. Method: Which data analytic approach is taken. Here we distinguish between global clustering, window-based, and hybrid (Hidden Markov Models).
7. FDR: Does the method allow for FDR control.
8. Application: Is the application of the method specific to a particular platform or chromatin modification, or can it be applied generally.

From each of the two tables we chose two methods that would be best suited for a comparison with our mixture modeling approach. As we have pointed out in the main text of the manuscript, most available methods are designed for the analysis of single ChIP experiments. This makes it difficult to formally compare them with our approach. To make the comparison as meaningful as possible, we focused on *general* methods that could be applied to *Input-normalized data*, and that could be compared on the basis on an *FDR* score. We highlight in yellow those that could be considered for this purpose. Among them, we choose ChIPmonk and ChIPmix for the ChIP-chip data, and CisGenome and SISSRs for the ChIP-seq data

Table S1: ChIP-chip methods

Author	Year	Name	ChIP sample comparisons	Input-normalized data	Method	FDR	Application
Siegmund et al.	2004	none*	yes	no	Global clustering	no	specific
Buck et al.	2005	ChIPOTle	no	yes	Window-based	yes	general
Ji et al.	2005	TileMap	yes	yes	Hybrid	yes	general
Li et al.	2005	none	no	yes	Hybrid	no	specific
Marjoram et al.	2006	none*	yes	no	Global clustering	no	specific
Andrews	2007	ChIPmonk	yes	yes	Window-based	yes	general
Keles	2007	TileHGMM	no	yes	Window-based	yes	specific
Down et al.	2008	Batman	no	yes	Window-based	no	specific
Houseman et al.	2008	none*	yes	no	Global clustering	no	specific
Martin-Magniette	2008	ChIPmix	no	yes	Global clustering	yes	general

* A clustering approach that applies to the very special Methyl-light technology. In this case, the signal distribution is quite different from that seen in on other platforms. It involves a two-part distribution with a single spike peak around zero and a Gaussian distribution for positive signal values. The resulting datasets are also considerably smaller.

Table S2: ChIP-seq methods

Author	Year	Name	ChIP sample comparisons	Input-normalized data	Method	FDR	Application
Johnson et al.	2007	PeakFinder	no	yes	Window-based	no	general
Albert et al.	2008	GeneTrack	no	no	Window-based	no	general
Fejes et al.	2008	FindPeaks	no	no	Window-based	yes	general
Jothi et al.	2008	SISSRs	no	yes	Window-based	yes	general
Valouev et al.	2008	QuEST	no	yes	Window-based	yes	general
Zhang et al.	2008	MACS	no	yes	Window-based	yes	general
Ji et al.	2008	CisGenome	no	yes	Window-based	yes	general
Rowzowsky et al.	2009	PeakSeq	no	yes	Window-based	yes	general
Zang et al.	2009	SICER	no	no	Window-based	no	specific
Xu et al.	2009	ChIPDiff	yes	no	Hybrid	no	specific

5) Calculation of False Discovery Rate (FDR) and False Positive Rate (FPR) for RDE mapping

When focus is on detecting **RDE**, our parametric classification approach can be conveniently used to calculate a conservative False Discovery Rate (FDR) or False Positive Rate (FPR). To see this, define $\mathbf{D} = d(\mathbf{w}) = \mathbf{x} - \mathbf{y}$, so that $\mathbf{D}$ represents a new variable measuring the differences between sample vectors $\mathbf{x}$ and $\mathbf{y}$. Since the estimated membership probabilities of each pair $\mathbf{w}_j$ are known, the membership probabilities of $\mathbf{D}_j$ are also known. We treat the conditional mixture distribution $\mathbf{D}^* = \mathbf{D} \mid \mathbf{D} \hat{\in} \{C_1 \cup C_2\}$ as the null distribution of no differentially enrichment, where C_1 and C_2 denote component 1 (**RSNE**) and component 2 (**RSE**), respectively, and $\hat{\in}$ is an estimated membership based on the highest posterior probability loading. Given a set of lower and upper quantiles, $\tilde{q}_l$ and $\tilde{q}_u$, we have a conservative estimate of the FDR and FPR as follows:

$$\text{FDR} \mid \tilde{\mathbf{q}} \approx \left(\#\mathbf{D}^* > \tilde{q}_u + \#\mathbf{D}^* < \tilde{q}_l\right) \Big/ \left(\#\mathbf{D} > \tilde{q}_u + \#\mathbf{D} < \tilde{q}_l\right)$$

and

$$\text{FPR} \mid \tilde{\mathbf{q}} \approx \#\mathbf{D}^* > \tilde{q}_u + \#\mathbf{D}^* < \tilde{q}_l,$$

respectively, where and $\tilde{\mathbf{q}} = (\tilde{q}_u, \tilde{q}_l)$. Values for $\tilde{q}_l$ and $\tilde{q}_u$ can then be chosen to control the FDR or FPR at a desired level.

6) Supplemental Figures

Figure S1

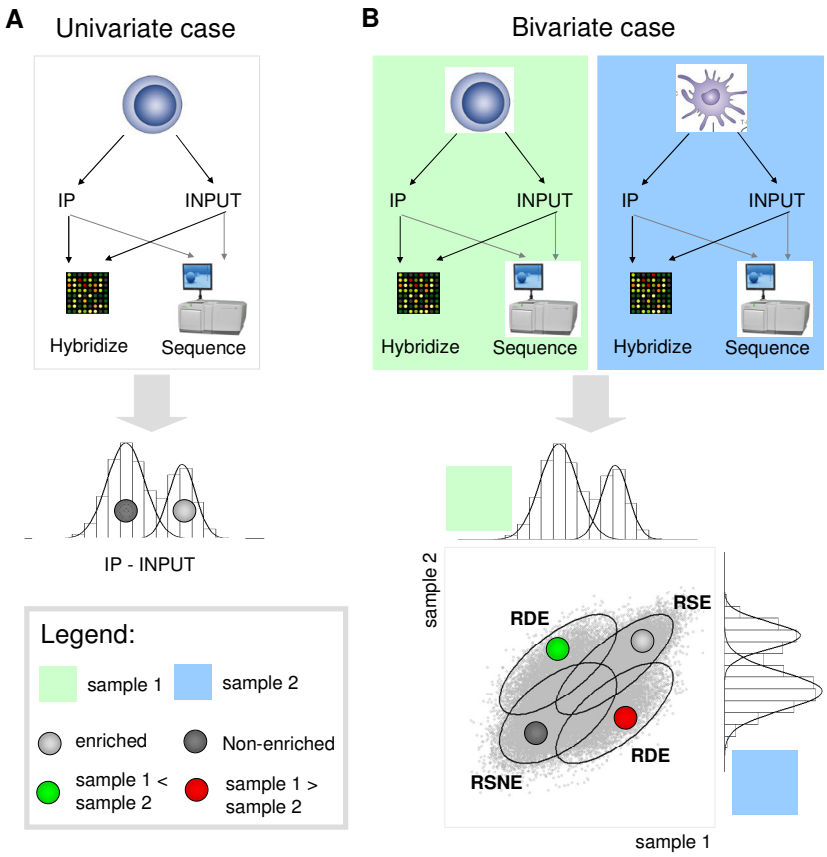


Figure S1. Data generation and distributions.
Panel A shows the familiar univariate case: ChIP-chip or ChIP-seq data is obtained from a single tissue (e.g. stem cells), possibly by pooling from several biological replicates. The INPUT-normalized signal distribution reveals the characteristic bimodal shape, with one distribution consisting of signals from enriched regions (light-grey) and the other of non-enriched regions (dark-grey). Several mixture modeling approaches have been used for the analysis of this type of univariate data. In the bivariate (or multivariate) case (Panel B), univariate methods are no longer appropriate. When comparing chromatin profiles of two samples, the goal is to distinguish Regions with Differential Enrichment (**RDE**, green and red) from Regions with Shared Enrichment (**RSE**, light-grey), and Regions with Shared Non-Enrichment (**RSNE**, dark-grey). In absence of differential enrichment, as it should be the case with two technical replicates, the signals of the two samples should highly correlate and cluster along the diagonal of the bivariate scatter plot. Systematic off-diagonal departures are evidence for chromatin differences, and raise the need to account for them in a modeling context (see **Figure 1** in the main text).

Figure S2

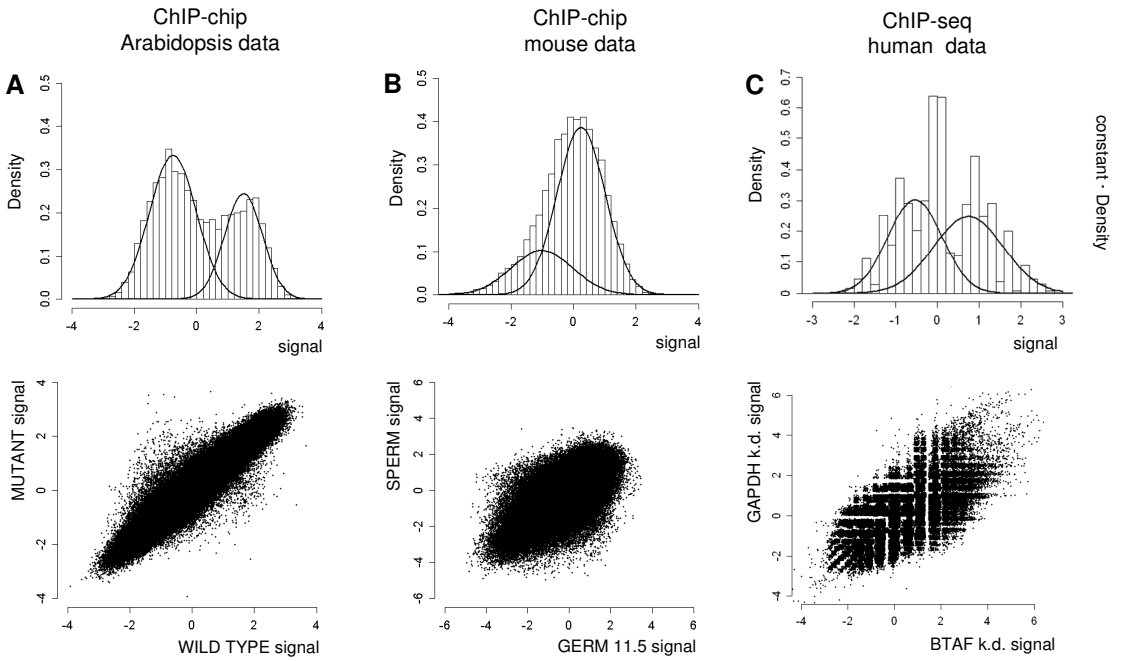


Figure S2. Univariate and bivariate distributions of three example datasets.
Shown are the univariate and bivariate distributions of three example datasets. Panel A: ChIP-chip *Arabidopsis* data (see main text) comparing genome-wide DNA methylation profiles between a wild type plant (*x*-axis) with that of a mutant plant (*y*-axis). Panel B: Mouse ChIP-chip data (see main text) comparing genome-wide promoter methylation in different cell types. For illustrative purposes we focus on the comparison of germ cells (*x*-axis) with sperm cells (*y*-axis). Panel C: Human genome-wide ChIP-seq data comparing TBP distribution in an experimental cell line with a knock-down for TBP-associated factor BTAF1 (*x*-axis) with that of control cells with GAPDH knock-down (*y*-axis). In each data example, the univariate distributions (top) refer to quantile normalized signals, and it therefore suffices to show the distribution of only one sample. It should become clear that for the ChIP-chip data examples the distributions are largely consistent with the conceptual framework set out in **Figure S1 and Figure 1**.

Figure S3

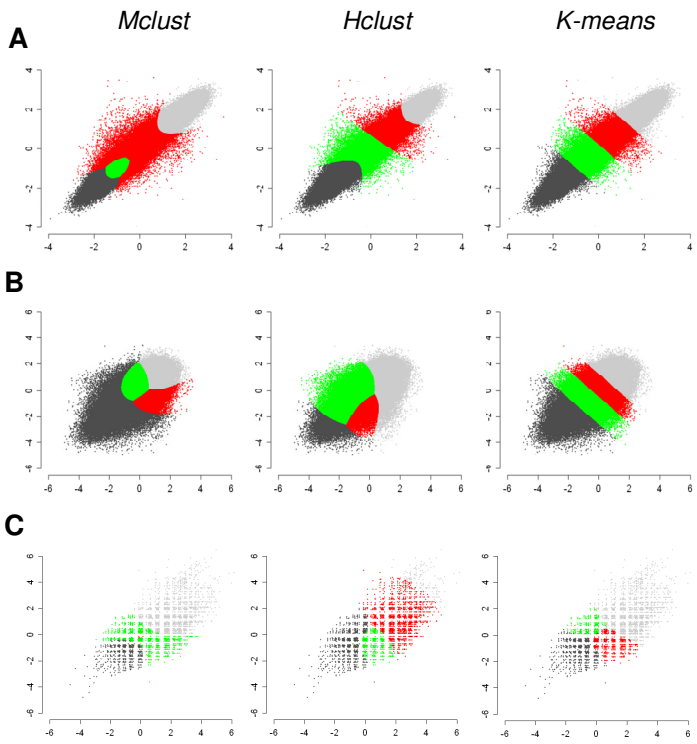


Figure S3: Classification results of the three data examples using alternative clustering methods.
Shown are the classification results of the three real data examples discussed in the text using alternative clustering methods. (A) *Arabidopsis* methylation data, (B) Mouse promoter methylation data, (C) Human basal transcription factor data. The results appear to vary quite substantially across methods. This may be attributable to the relalively small **RDE** components that are difficult to capture without sensible constraints in the classification procedure, especially when sub-sampling of the data is involved (see text).

Figure S4

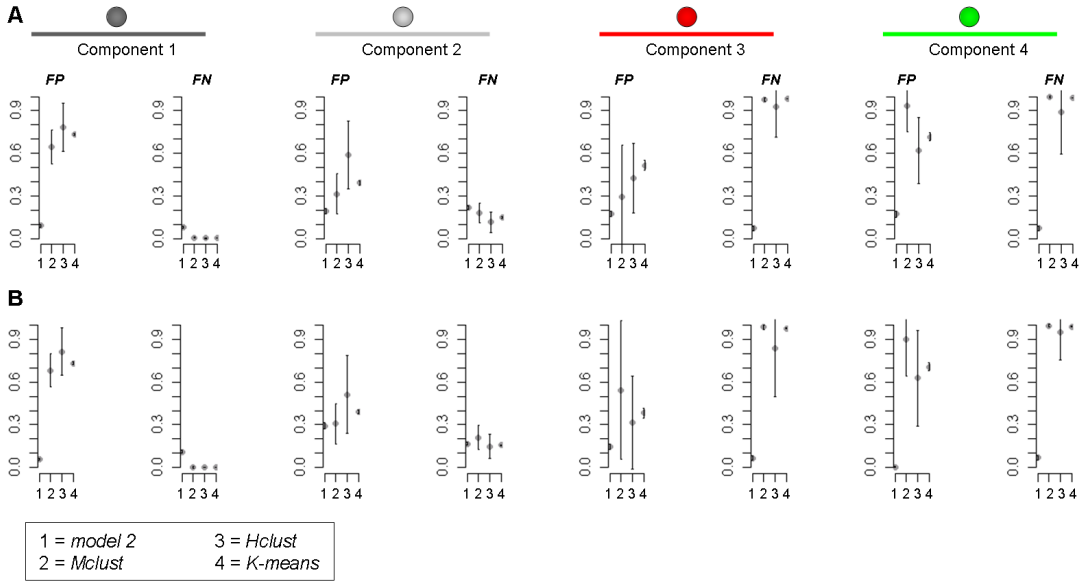


Figure S4: Simulation comparison with other methods.
Shown are the simulations results for two types of datasets: (**A**) the data is generated directly from model 2 and hence all the distributional assumptions are met; (**B**) the distributional assumptions of model 2 are deliberately violated by letting data points be drawn from a bivariate skewed-normal mixture for **RDE** components 3 (red) and 4 (green). For each these two datasets the classification performance of *model 2* (**1**), *Mclust* (**2**), *hclust* (**3**) and *K-means* (**4**) (*x*-axis) are compared over 50 independent simulation replicates (see text). From these simulation replicates we estimated the expected false positive (FP) and false negative rate (FN) (*y*-axis, ranging from 0 to 1) as well as standard errors (error bars). The results show that *model 2* provides consistently low FP and FN compared to other methods, even in the situation considered here where the model assumptions are violated. Noticeably, *Mclust* (**2**) and *hclust* (**3**) show large variation across simulation runs, particularly for **RDE** components 3 and 4. This may be attributable to the small mixing proportions of these components and a result of the sub-sampling that is required to implement these methods for large datasets.

6) References

Fraley,C. and Raftery,A. (2007) MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. Technical Report 504 University of Washington, Department of Statistics.

Fraley,C. and Raftery,A. Mclust: Model-Based Clustering/Normal Mixture Modeling. R package version 3.1-10.

Houseman,E.A. et al. (2008) Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics*, **9**, 365.

Marjoram,P. et al. (2006) Cluster analysis for DNA methylation profiles having a detection threshold. *BMC Bioinformatics*, **7**, 361.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet*, **2**, 418-427.

R Development Core Team R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Royce TE, Rozowsky JS, Gerstein MB. Assessing the need for sequence-based

normalization in tiling microarray experiments. *Bioinformatics.* 2007, **23**:988-997.

Siegmund,K.D. et al. (2004) A comparison of cluster analysis methods using DNA methylation data. *Bioinformatics*, **20**, 1896-1904.

Takai, D. and Jones, P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA*, **99**, 3740-3745.

Wuertz,D. fMultivariate: Multivariate Market Analysis. R package version 270.74.