BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 8 2010, pages 1000–1006
doi:10.1093/bioinformatics/btq087

Genome analysis Advance Access publication March 5, 2010

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

Frank Johannes[1,*,†], René Wardenaar[1,†], Maria Colomé-Tatché[2], Florence Mousson[3], Petra de Graaf[3], Michal Mokry[4], Victor Guryev[4], H.Th. Marc Timmers[3], Edwin Cuppen[4] and Ritsert C. Jansen[1,5]

[1]Groningen Bioinformatics Centre, University of Groningen, Kerklaan 30, Biologisch Centrum, 9751 NN Haren, The Netherlands, [2]Institute for Theoretical Physics, Leibniz Universität Hannover, Appelstr. 2, D-30167 Hannover, Germany, [3]Department of Physiological Chemistry, University Medical Center Utrecht, Universiteitsweg 100, 3508 AB Utrecht, [4]Hubrecht Institute, KNAW and University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht and [5]Department of Genetics, University Medical Centre Groningen, Hanzeplein 1, 9713 GZ Groningen, The Netherlands

Associate Editor: Alex Bateman

ABSTRACT

Motivation: ChIP-chip and ChIP-seq technologies provide genome-wide measurements of various types of chromatin marks at an unprecedented resolution. With ChIP samples collected from different tissue types and/or individuals, we can now begin to characterize stochastic or systematic changes in epigenetic patterns during development (intra-individual) or at the population level (inter-individual). This requires statistical methods that permit a simultaneous comparison of multiple ChIP samples on a global as well as locus-specific scale. Current analytical approaches are mainly geared toward single sample investigations, and therefore have limited applicability in this comparative setting. This shortcoming presents a bottleneck in biological interpretations of multiple sample data.

Results: To address this limitation, we introduce a parametric classification approach for the simultaneous analysis of two (or more) ChIP samples. We consider several competing models that reflect alternative biological assumptions about the global distribution of the data. Inferences about locus-specific and genome-wide chromatin differences are reached through the estimation of multivariate mixtures. Parameter estimates are obtained using an incremental version of the Expectation–Maximization algorithm (IEM). We demonstrate efficient scalability and application to three very diverse ChIP-chip and ChIP-seq experiments. The proposed approach is evaluated against several published ChIP-chip and ChIP-seq software packages. We recommend its use as a first-pass algorithm to identify candidate regions in the epigenome, possibly followed by some type of second-pass algorithm to fine-tune detected peaks in accordance with biological or technological criteria.

Availability: R source code is available at http://gbic.biol.rug.nl/supplementary/2009/ChromatinProfiles/

Access to Chip-seq data: GEO repository GSE17937

Contact: f.johannes@rug.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Received on November 18, 2009; revised on February 19, 2010; accepted on February 22, 2010

1 INTRODUCTION

Epigenetic modifications such as methylation of DNA or histones are associated with the transcriptional output of genes. Hence, they occupy a central role in genome function. The use of chromatin immunoprecipitation techniques coupled with tiling-arrays (ChIP-chip) or deep sequencing (ChIP-seq) now permit a detailed characterization of various types of chromatin marks on a genome-wide scale. Global surveys of this type have been mainly conducted in the context of a single genome analyses (Suzuki and Bird, 2008). However, to begin to understand the epigenetic basis of biological variation, detailed comparisons of chromatin profiles between different tissues and/or individuals are becoming increasingly important. Such comparative studies can be broadly divided into those that are concerned with intra-individual variation and those concerned with inter-individual variation.

The goal of intra-individual analysis is the detection of key chromatin modifications that are associated with, or possibly drive, normative processes such as tissue development. Meissner *et al.* (2008), for instance, recently compared mammalian cells across several stages of differentiation in order to map the genome-wide changes in DNA methylation and several histone marks that reflect this progression. The focus of inter-individual comparisons, in contrast, rests on the analysis of chromatin differences between individuals for a given tissue (Bock *et al.*, 2008). This is accomplished by either treating each individual as a separate unit of analysis, or by comparing pooled tissue from individuals that belong to a priori well-defined groups. A clear example for this latter type of application comes from oncology studies where cancerous tissues from patients are directly evaluated against normal tissue from healthy individuals (Esteller, 2008).

Apart from clinically relevant inter-individual differences, attention has recently shifted to naturally occurring chromatin variation (Richards, 2008). A recent survey of several *Arabidopsis* ecotypes, for instances, has revealed substantial variation in genome-wide DNA methylation patterns, part of which has been shown to be stably transmitted across generations (Vaughn *et al.*, 2007).

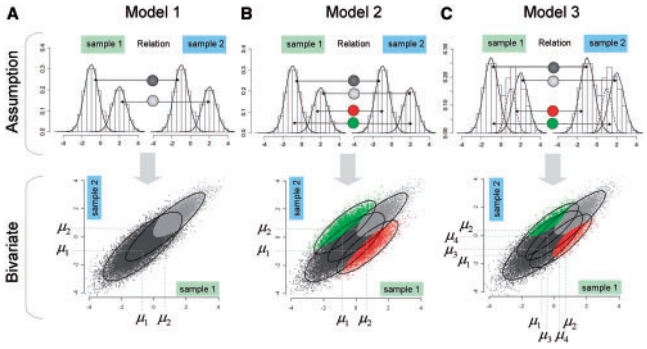

Fig. 1. Conceptual overview of the mixture model approach. We consider three different models for the analysis of chromatin differences between two or more samples. Provided is a visual illustration of the conceptual basis of these models, with a focus on the specific case of two-sample comparisons. (**A**–**C**, top panel) Univariate representation. (A–C, bottom panel) Bivariate representation. The colors represent the expected distribution of the data corresponding to RSE (light gray), RSNE (dark gray), RDE (red: sample 1 > sample 2) and RDE (green: sample 2 > sample 1). The color coding is consistent with Supplementary Figure S1.

This latter observation suggests that many of the experimental populations that are derived from these ecotypes and used for genetic mapping experiments, such as F_2-intercrosses or Recombinant Inbred Lines (RILs), could segregate a source of epigenetic variation that impacts many commonly studied complex traits, independent of DNA sequence polymorphisms (Johannes *et al.*, 2008, 2009). In light of this possibility, we have argued for a genome-wide characterization of available parental lines to identify those that are most divergent on the level of chromatin (Johannes *et al.*, 2008). Such lines can be used to obtain epigenetically informative line crosses, which could become the basis for the dissection of the epigenetic architecture underlying continuous trait variation.

The above intra- and inter-individual applications share a common problem: they require a methodological approach that permits a meaningful partition of the epigenome into regions with differentially enrichment (RDE), regions with shared enrichment (RSE) and regions with shared non-enrichment (RSNE) for a chromatin mark of interest (Fig. 1 and Supplementary Fig. S1). In many practical situations, researchers are interested in comparing a small number of ChIP samples, possibly as low as two, each sample being presented by one tiling array or sequencing experiment. Small numbers of samples are common because of financial constraints, or because they represent pilot studies. Most available software packages for the analysis of ChIP-chip or ChIP-seq data have been designed for single genome (i.e. single ChIP sample) investigations and therefore have limited application in this setting (Supplementary Tables S1 and S2).

In an attempt to overcome these limitations, we introduce a parametric classification method based on multivariate mixture models. The approach permits a direct comparison of two (or more) ChIP samples (Fig. 1). Its unique feature lies in the characterization of the global distribution of the data in accordance with several plausible biological models. At the same time, it provides probabilistic estimates of locus-specific chromatin states. To ease computational demands, we implement an incremental version of the Expectation–Maximization (IEM) algorithm. We illustrate our approach in the context of two published ChIP-chip experiments and one unpublished ChIP-seq experiment. We highlight the utility of our approach in terms of the biology of each example dataset. We also evaluate our approach against several published ChIP-chip and ChIP-seq software packages.

2 METHODS

2.1 The parametric classification approach

For clarity, we focus on the special case of a two-sample comparison throughout, although extensions to additional dimensions can be implemented. The samples should represent two different natural or experimental units (e.g. tissue 1 versus tissue 2, genotype 1 versus genotype 2, cancer patients versus control cases). Let x_j and y_j denote the signals of the two samples recorded for the j-th genomic region. In practice, region j represents an arbitrary unit of analysis. For ChIP-chip data, for instance, this can be a single tile on the array or the average signal taken over a larger sequence, such as a promoter, or gene. For ChIP-seq data, we first divide the genome into equally sized bins and take 'read' counts within each one of these bins. Hence, in this case, region j represents one or several of these bins and its associated counts represent the 'signal', which after normalization over several experiments can be treated as a continuous variable (see below).

When Input data is available, x_j and y_j are obtained by calculating the signal difference $g(\mathrm{IP}_j)-g(\mathrm{Input}_j)$, where $g(\cdot)$ denotes some suitable normalization. IP_j is the signal belonging to the immunoprecipitate and Input_j that to the total DNA (Supplementary Fig. S1). Note that the IP and Input material can come from the DNA of a single individual, or from pooled DNA of multiple individuals representing a larger group.

Based on the expected distributional properties of ChIP data (Fig. 1 and Supplementary Fig. S1), we let the probability of the random pair $\mathbf{w}_j=(x_j,y_j)$ be given by a four-component mixture model

$$f(\mathbf{w}_j;\vec{\Psi})=\sum_{i=1}^{4}\lambda_i f_i(\mathbf{w}_j,\vec{\theta}_i) \qquad (1)$$

where $\vec{\Psi}=(\lambda_1,\ldots,\lambda_4,\vec{\theta}_1,\ldots,\vec{\theta}_4)$ is a vector of parameters assumed a priori to be distinct. Note that the component indexing corresponds to the components associated with RSNE (dark gray), RSE (light gray), RDE (red) and RDE (green), respectively (Fig. 1 and Supplementary Fig. S1), and that the mixing weights must meet the constraint $\sum_i\lambda_i=1$. For the present application of a two-sample comparison, we take the i-th component density to be bivariate normal; that is, $\mathbf{w}_j|\vec{\theta}_i\sim\mathrm{BVN}(\vec{\mu}_i,\Sigma_i)$ with $(j=1,\ldots,n),(i=1,\ldots,4)$, and where $\vec{\mu}_i$ and $\mathbf{\Sigma}_i$ are the vector of the component means and the component variance–covariance matrix, respectively. Based on (1), we consider three competing models, a Null model (Model 1), a Full-switching model (Model 2) and a Flexible-switching model (Model 3), each of which is detailed below.

2.1.1 Null model (Model 1) Model 1 (Fig. 1A) assumes that there are effectively no chromatin differences between the two samples, implying that $\lambda_3=\lambda_4=0$. In this conceptualization, the signals of the enriched and non-enriched components are 'stable' across conditions and thus highly correlated (i.e. both samples tend to show enrichment or non-enrichment for region j). This is equivalent to assuming that the two samples are technical replicates of each other. The null model is given by

$$f(\mathbf{w}_j;\vec{\Psi})=\lambda_1 f_1(\mathbf{w}_j;\vec{\mu}_1,\Sigma_1)+\lambda_2 f_2(\mathbf{w}_j;\vec{\mu}_2,\Sigma_2) \qquad (2)$$

where $\vec{\mu}_1=(\mu_1,\mu_1)$ and $\vec{\mu}_2=(\mu_2,\mu_2)$. The component means are constrained to be equal, and the variance–covariance matrices are allowed to be different (heteroscedastic).

2.1.2 Full-switching model (Model 2) Model 2 (Fig. 1B) starts from the assumption that differential chromatin states exist between the two samples.

The model is

$$f(\mathbf{w}_j;\vec{\Psi})=\lambda_1 f_1(\mathbf{w}_j;\vec{\mu}_1,\Sigma_1)+\lambda_2 f_2(\mathbf{w}_j;\vec{\mu}_2,\Sigma_2)$$
$$+\lambda_3 f_3(\mathbf{w}_j;\vec{\mu}_3,\Sigma_3)+\lambda_4 f_4(\mathbf{w}_j;\vec{\mu}_4,\Sigma_4) \quad (3)$$

where, as above, $\vec{\mu}_1$, $\vec{\mu}_2$, $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are the means and covariance matrices of the RSNE and RSE component and $\vec{\mu}_3=(\mu_2,\mu_1)$, $\vec{\mu}_4=(\mu_1,\mu_2)$, $\Sigma_3=\Sigma_4$ are the means and covariance matrices of the two RDE components. From (3) and Figure 1B, it becomes clear that the differentially enriched components are governed by different combinations of the means of the enriched and non-enriched component. This constraint has the direct interpretation that chromatin changes that may have arisen in one sample compared with the other sample can be defined by 'full' transitions from enrichment to non-enrichment or vice versa. We assume that the two RDE components are characterized by a homoscedastic variance–covariance structure.

2.1.3 Flexible-switching model (Model 3) Model 2 represents only a special case of a more flexible model in which chromatin transitions can either be more subtle or more severe. This implies that the means of the RDE components can differ from those of the RSNE and RSE components. This flexibility is provided by allowing the RDE components to be governed by a separate set of means μ_3 and μ_4 that may or may not equal the means of the enriched and non-enriched signals. The form of this type of model is as in (3) but now with symmetrical constraint $\vec{\mu}_3=(\mu_4,\mu_3)$ and $\vec{\mu}_4=(\mu_3,\mu_4)$, where most likely but not necessarily $\mu_3\neq\mu_1$ and $\mu_4\neq\mu_2$.

2.1.4 Estimation and classification Parameter estimation is carried out in a maximum likelihood framework. In this case, the maximum likelihood estimates (MLEs) must fulfill $\nabla\log L(\vec{\Psi}|\mathbf{w})=0$, where $\mathbf{w}=(\mathbf{w}_{j=1},\mathbf{w}_{j=2},\ldots,\mathbf{w}_n)$ and $L(\vec{\Psi}|\mathbf{w})=\prod_j f(\mathbf{w}_j;\vec{\Psi})$ is the likelihood function. Solutions can be obtained iteratively via the EM algorithm (Dempster *et al*., 1977; McLachlan and Peel, 2000) subject to the specific mean and covariance constraints of each model. In genome-wide ChIP-chip and ChIP-seq applications, the dimension n can be large. This can pose serious computational difficulties when applying any global classification approach. To account for this, we implemented an incremental updating strategy for the EM algorithm (IEM), which considerably speeds up convergence (McLachlan and Peel, 2000). We also provide approximations to the standard errors (SE) of the final parameter estimates by calculating the so-called empirical information matrix $I_e(\vec{\Psi};\mathbf{w})$ (Meilijson, 1989) evaluated at the MLEs as

$$I_e(\hat{\vec{\Psi}};\mathbf{w})=\sum_{j=1}^{n}\nabla\log L(\hat{\vec{\Psi}}|\mathbf{w}_j)\nabla^T\log L(\hat{\vec{\Psi}}|\mathbf{w}_j) \quad (4)$$

Once the parameter estimates, $\hat{\vec{\Psi}}$, have been obtained, each $\mathbf{w}_j$ can be given a probable component membership via its posterior density $\tau_i(\mathbf{w}_j;\hat{\vec{\Psi}})=\hat{\lambda}_i f_i(\mathbf{w}_j;\vec{\theta}_i)/f(\mathbf{w}_j;\hat{\vec{\Psi}})$. The highest posterior density of region j corresponds to its most probable membership. This probability assignment provides a quantitative measure of the locus-specific chromatin status, which is an attractive feature given the fact that chromatin measures over a defined region are likely not discrete.

2.1.5 Criteria for model comparisons The parametric clustering approach outlined above lends itself to a global evaluation of the best fitting model to the observed data. This is similar to testing a single hypothesis about the entire genome. For these model comparisons, we utilize the commonly used Akaike Information criterion (AIC; Akaike, 1974), as extensive simulations (data not shown) indicate that this criterion appears to be very reliable in the context of our modeling framework. The AIC is defined as $\mathrm{AIC}=2k-2\ln(L)$, where k denotes the number of parameters in the model and L is the likelihood value of the model. The model with the lowest AIC is favored.

2.1.6 Annotation-based genome partitioning The proposed mixture modeling approach can also flexibly incorporate information about specific sequence contexts. To illustrate this, suppose each signal $\mathbf{w}_j$ ($j=1,\ldots,n$), can be assigned, a priori, to one of m mutually exclusive annotation sets $S_1,S_2,\ldots,S_m$, corresponding to genes, promoters, transposable elements, CpG islands, etc. Let $n_1,n_2,\ldots,n_m$ denote the respective sizes of these sets. Define $\Theta(\mathbf{w}_j;S_k)$ to be an indicator function performing this assignment, with $\Theta(\mathbf{w}_j;S_k)=1$ if $\mathbf{w}_j\in S_k$ and 0 otherwise. Thus, the 'composite' mixture density of $\mathbf{w}_j$ is given by

$$f_C(\mathbf{w}_j;\vec{\Psi})=\sum_{k=1}^{m}\Theta(\mathbf{w}_j;S_k)\frac{n_{S_k}}{n}f_{S_k}(\mathbf{w}_j;\vec{\Psi}_{S_k}). \quad (5)$$

In this way, the density contributions are simply weighted by a constant that is proportional to the relative size of the set. Accordingly, the logarithm of the likelihood function can be written as

$$\log L_C(\vec{\Psi}|\mathbf{w})=\mathcal{A}+\sum_{k=1}^{m}\log L_{S_k}(\vec{\Psi}_{S_k}|\mathbf{w}_{S_k}), \quad (6)$$

where $\mathcal{A}=-n\log n+\sum_{k=1}^{m}n_{S_k}\log(n_{S_k})$. Hence, in the case where the parameters are uniquely indexed across annotation sets $S_1,S_2,\ldots,S_m$, the estimation of the likelihood function can be performed independently for each set, and then compared with an analysis where no such a prior categorization was performed. This partitioning strategy can also be useful if certain sequences are known to affect the signal distribution due to technical reasons, such as cross-hybridization, or amplification and sequencing biases. Formal model comparison procedure, such as the AIC, can be employed to assess if a separate treatment of different sequence contexts is warranted. For the sake of clarity, we will restrict ourselves to the special case of no a prior partitioning. However, we do provide a full example of its implementation in the Supplementary Material.

2.2 Data description and normalization

2.2.1 ChIP-chip data We consider two publicly available ChIP-chip datasets. The first ChIP-chip experiment compares the genome-wide methylation profiles of a wild-type *Arabidopsis* plant to that of a methylation maintenance mutant (Penterman *et al*., 2007). ChIP samples were hybridized to NimbleGen tiling arrays consisting of 382 178 probes with an average length of 45–85 bp. The second ChIP-chip dataset compares genome-wide promoter methylation between mouse embryonic germ cells and sperm cells (Farthing *et al*., 2008). This experiment used mouse NimbleGen promoter tiling arrays. Arrays were composed of 1.5 kb promoter regions for a set of 26 275 mouse genes. On average, tiled regions contained 15 probes (50 bp length) separated by a 50 bp gap. We refer the reader to the original publications for details concerning sample preparation.

In the ChIP-chip *Arabidopsis* experiment, we calculated the signal ratios for the j-th tile on the array as $x_j=\log_2(\mathrm{IP}_{j,wt}/\mathrm{Input}_{j,wt})$ and $y_j=\log_2(\mathrm{IP}_{j,mt}/\mathrm{Input}_{j,mt})$, where *wt* and *mt* denote the wild-type and mutant genotypes, respectively. Each genotype was represented by a single array. For the ChIP-chip mouse data, dye-swaps were available for each of the two cell types, embryonic germ cells (*EG*) and sperm cells (*SP*). Hence, the signal ratios were taken as $x_j=\log_2(\mathrm{IP}_{j,EG}/\mathrm{Input}_{j,EG})$ and $y_j=\log_2(\mathrm{IP}_{j,SP}/\mathrm{Input}_{j,SP})$, where the IP_j and Input_j, in this case, represent the average of the swap of Cy3 and the Cy5 signal. Due to the constraints imposed on the component means of the mixture model, we required that the signal distribution of both samples be brought to a common scale, and therefore applied quantile normalization using standard software (Gentleman *et al*., 2004).

2.2.2 ChIP-seq data The unpublished ChIP-seq dataset investigates the differences in the distribution of TATA-binding protein (TBP) across the human genome between BTAF1 and GAPDH knockdown HeLa cells derived from human cervix carcinoma. A protocol outlining sample preparation and library construction is provided in the Supplementary Material. Overall, the ChIP-seq data considered here consists of three independent deep-sequencing experiments (using SOLiD technology), two involving the

immunoprecipitate (pull-down) of the GAPDH and BTAF1 knockdown samples (IP_B and IP_G, respectively), as well as one common Input control sample (IC). The total number of mapped reads for each of these experiments were $IP_G = 2798922$, $IP_B = 5064143$ and $IC = 4296061$. Consistent with previous approaches for ChIP-seq analysis (Ji *et al*., 2008), we partitioned the genome into bins of size 1 kb, with each bin containing the total read counts in that particular region. Since there was a 1.5- to 2-fold difference in the total number of reads between these three experiments, we normalized the read counts as follows: let $K_{j,p}$ denote the count for the j-th bin and p-th experiment ($p = 1, 2, 3$), and take the average total count across the three experiments as $\overline{K} = \frac{1}{3}(\sum_p \sum_j K_{j,p})$. Hence, a normalized count can be given by

$$K'_{j,p} = \frac{K_{j,p}}{\sum_j K_{j,p}} \cdot \overline{K} \tag{7}$$

Bins with a value of zero in all three deep-sequencing experiments were removed from further analysis. Finally, we express the Input-normalized signal as $x_j = \log_2\left\{(1+K'_{j,\mathrm{IP}_B})/(1+K'_{j,\mathrm{IC}})\right\}$ and $y_j = \log_2\left\{(1+K'_{j,\mathrm{IP}_G})/(1+K'_{j,\mathrm{IC}})\right\}$, followed by quantile normalization of the two 'signal' vectors. As already pointed out by others (Kharchenko *et al*., 2008), read adjustments within and across experiments invariably result in continuous rather than count data. This justifies the use of continuous density functions in the context of multiple-sample ChIP-seq experiments. Nonetheless, care should be taken when applying the above normalization as the linearity assumption in the scaling of reads may break down when read counts in one or several samples are either too low (i.e. near background) or too high (i.e. near saturation). This consideration deserves further research.

3 RESULTS

3.1 *Arabidopsis* genome-wide methylation data

As a first data example, we consider the experiment by Penterman *et al*. (2007). This study compared the methylation profile of a wild-type *Arabidopsis* plant to that of a *ros1-3; dml2-1; dml3-1* triple mutant. All three genes are members of the DEMETER (DME) family, which is known to be involved in the removal of methylcytosine from DNA, and in establishing genomic imprinting patterns in the endosperm. It was hypothesized that a disruption of the DME genes leads to locus-specific hypermethylation. To identify candidate loci that are targeted by DME enzymes, the authors resorted to a genome-wide ChIP-chip approach.

Supplementary Figure S2 shows a univariate plot of the quantile-normalized signal distribution of the original data, as well as a bivariate plot of the wild-type array signal against that of the triple mutant. In both cases, the data structure visually complies with our expectations. We applied our models to this data treating individual probes on the tiling array as separate units of analysis. The results show that Model 2 provides the best overall fit (Fig. 2A). This confirms that the triple mutation leads to non-negligible changes in methylation levels. As anticipated, we estimate that ~0.35% ($SE = 8.48 \times 10^{-3}$) of all tested regions had become hypermethylated in the mutant relative to the wild-type (Fig. 2A), but also note a 0.15% reduction in methylation ($SE = 2.38 \times 10^{-3}$). This locus-specific hypomethylation in the mutant is unexpected and suggests an additional role of the DME genes in *de novo* methylation; alternatively, it reflects the presence of genomic mechanisms that counteract pervasive hypermethylation through epigenetically mediated compensatory pathways. Apart from the differentially methylated regions, we also estimate that 35.4% ($SE = 1.7 \times 10^{-2}$) of the *Arabidopsis* genome is methylated in both genotypes, a number that is roughly consistent with previous reports (Zilberman *et al*., 2007).

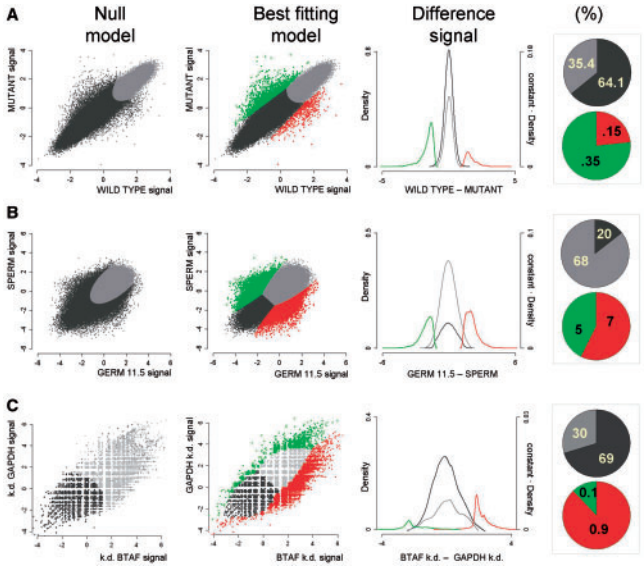

Fig. 2. Mixture modeling results for the three example datasets. (**A**–**C**) The results for the *Arabidopsis*, mouse and human data examples, respectively. The color coding in these figures is consistent with the notation introduced in Figure 1. In each panel, the first plot shows the mixture model solution for the null model. The second plot provides the solution for the best fitting model; the third plot, is a univariate version of the best fitting model, which is obtained by calculating the signal difference between the two samples (i.e. sample 1 and sample 2), while keeping track of the most likely classification. This univariate plot reveals the potential difficulty in using univariate methods to de-convolute the RSE distribution (light gray) from the RSNE distribution (dark gray). The fourth plot (pie-charts) shows a summary of the genome-wide classification results in terms of percentages.

3.2 Mouse promoter methylation data

Cell differentiation processes are accompanied by specific changes in DNA methylation patterns in gene promoters (Reik, 2007). This reprogramming is necessary to return the cell to a pluripotent state. How many of these changes are taking place and their level of tissue specificity is currently under intense investigation. In order to understand promoter methylation changes, Farthing *et al*. (2008) carried out a genome-wide promoter analysis of DNA methylation in mouse embryonic stem (ES) cells, embryonic germ (EG) cells, sperm, trophoblast stem (TS) cells and primary embryonic fibroblasts (pMEFs) using ChIP-chip promoter tiling arrays. For illustrative purposes, we focus on a comparison between EG and sperm cells.

Supplementary Figure S2 shows the univariate and bivariate plots of the genome-wide mouse promoter methylome data. By visual inspection, the bivariate plot of the methylation signals of EG cells (x-axis) against those of the sperm cells (y-axis) suggests the presence of RDE, with data points scattering along off-diagonal components. Interestingly, applying our approach to this data, we find that Model 3 provides the best fit (Fig. 2B). As the means of the RDE components (red and green) are between those of the RSE and the RSNE components, we conclude that the methylation differences between the tissues are, at least globally, more subtle than under a

model that posits full transitions between the methylation states. It is tempting to speculate that such subtle global methylation changes are a common feature during tissue development, and are sufficient to contribute to the establishment of new cell identities.

On a genome-wide scale, we estimate that promoter methylation in EG cells is higher compared with sperm cells. This observation contradicts the patterns seen for particular candidate loci that transition from a hypomethylated to a hypermethylated state during the process of differentiation/specialization (Farthing *et al*., 2008). Using a classification based on maximum posterior loadings, we find that a total of ∼12% (SE = 6.04×10^{-4}) of the measured regions (i.e. tiles on the array) are differentially methylated between the two tissues.

3.3 Human basal transcription factor data

The last data example illustrates the application of our method to the more recent ChIP-seq technology. The goal of the ChIP-seq study was to investigate the distribution of the TBP across the human genome. TBP is the DNA-binding subunit of the basal transcription factor TFIID as well as for other complexes of RNA polymerase I and II (Sharp, 1992). The BTAF1 ATPase forms a stable complex with TBP and regulates its activity in pol II transcription. We hypthesized that BTAF1 is involved in the mobilization of TBP from promoter and non-promoter sites (Pereira *et al*., 2003). To test this hypothesis, TBP ChIP samples were prepared from human HeLa cervix carcinoma cells after knockdown of BTAF1 expression and compared with HeLa cells with a control knockdown of GAPDH. GAPDH is a cytosolic enzyme that participates in glycolysis, and its inactivation is not expected to affect the genomic distribution of TBP, and acts as negative control. ChIP samples were sequenced using SOLiD technology along with the Input samples. The data were normalized as described above.

Based on previous work in yeast (Li *et al*., 2000), we expected the effect of the BTAF1 knockdown to result in an overall reduction in binding of TBP to promoters and redistribution of TBP to the many A/T-rich, non-promoter sequences in the human genome. At the same time, TBP enrichment in the GAPDH knockdown condition should remain unchanged. Under the assumption that TBP removal is efficient, we anticipated substantial RDE between the BTAF1 and the GAPDH knockdown samples, and limited RSE. Fitting our mixture models to these data, we find that Model 2 provides the best fit (Fig. 2C). From these mixture results, we highlight two important findings: first, using a classification based on maximum posterior loadings, we discovered a substantial amount (30%; SE = 6.4×10^{-3}) of RSE (light gray), suggesting that the BTAF1 knockdown does not affect a large percentage of sequences harboring TBP despite BTAF1 inactivation. These RSE corresponded mostly (96.13%) to non-promoter sequences. The second important observation relates to the fact that RDE are more prevalent in the direction of BTAF1 than in the direction of GAPDH (0.9%; SE = 1.43×10^{-3} versus 0.1% SE = 6.79×10^{-4}). This indicates that RDE are predominantly defined by increased TBP levels in the BTAF1 relative to the GAPDH knockdown in this system. Interestingly, a disproportionally high number of sequences contained in the 0.9% of RDE for which BTAF1 shows more TBP binding compared with GAPDH (red), corresponds to annotated promoters (50.09%). This trend becomes even more pronounced when we use a more stringent classification cutoff (Table 1).

Table 1. Overlap between detected RDE bins and promoter regions

	RDE overlapping promoters	
	BTAF1 > GAPDH	GAPDH > BTAF1
Model 2	3152/4440 (71%)	29/947 (3%)
CisGenome	3591/5076 (70%)	34/585 (6%)
SISSRs	1273/1538 (83%)	28/166 (17%)

We distinguish two types of RDE, BTAF1 > GAPDH and GAPDH > BTAF1. Listed are the number of detected bins that map to annotated promoters over the total detected number within each of the two RDE categories (at FDR = 0.05). Promoter regions were defined as −2 kb to +500 bp relative to the TSS.

We have verified this observation using ChIP-PCR experiments on a limited set of genomic regions (data not shown). These findings point toward a primary role of BTAF1 in the mobilization (or removal) of TBP *from* promoter sites, which explains the overaccumulation of TBP at these locations in the BTAF1 knockdown background. This proposal agrees well with our recent finding that the BTAF1 ortholog in yeast, Mot1p, mostly resides on promoter regions as shown by genome-wide ChIP-chip (van Werven *et al*., 2008). More detailed bioinformatic and experimental work will provide additional insight into the specific promoter contexts underlying the detected RDE.

3.4 Comparison with alternative approaches

3.4.1 Problem statement We compared our approach with four representative methods for ChIP-chip and ChIP-seq analysis (ChIPmix, ChIPmonk, CisGenome, SISSRs). Specific criteria for this selection are detailed in the Supplementary Material (see also Supplementary Tables S1 and S2). The goal was to test the performance of each approach in detecting RDE using the example datasets. It is worth noting that there were three major obstacles to a direct comparison.

First, nearly all available methods are specifically designed for single ChIP sample analysis (Input-normalized or unnormalized). In this case, the detection of signal differences between experiments (e.g. BTAF1 versus GAPDH) had to be done in an *ad hoc* fashion. This forced us to analyze each experiment separately. RDE were then defined, informally, as those regions for which enrichment was detected in one experiment but not in the other, and vice versa.

The second obstacle is the use of inconsistent definitions of peak regions between methods. It was therefore important to find a common unit of analysis so that total RDE counts were comparable. In the case of the ChIP-chip data, a convenient choice was the mouse promoter methylation example, because any number of peak regions detected within a promoter could be reported in terms of promoter units; that is, the presence of one or more differential peak regions resulted in calling the entire promoter RDE. In the human ChIP-seq data example, we used the imposed 1 kb bin structure as a common unit of references (i.e. given differential peaks anywhere within a bin, the whole bin was reported as a RDE).

The third obstacle relates to the choice of an appropriate false discovery rate (FDR) for calling significant RDE on a genome-wide scale. Ideally, this should be based on a known (or estimated) null distribution of no enrichment differences between ChIP samples. However, with methods where ChIP samples are analyzed separately, such FDR calculations are not available. Our parametric

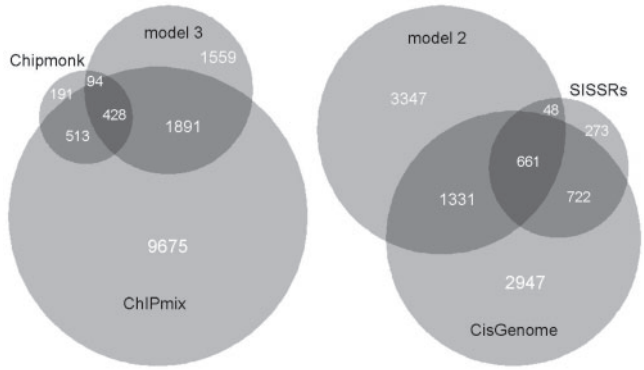

Fig. 3. Comparison of methods for RDE mapping. Left: comparison of ChIPmonk, ChIPmix and Model 3 in detecting RDE in the mouse promoter methylation data (counts are in terms of numbers of promoters). Right: comparison of CisGenome, SISSRs and Model 2 in detecting RDE in the human transcription factor data (counts are in terms of numbers of 1 kb bins).

classification approach elegantly circumvents this problem by providing an estimated null distribution as part of the analysis. This can be directly used for FDR determination (Supplementary Material). In the absence of a more formal approach, we adopted the following strategy: whenever ChIP samples had to be analyzed separately (ChIPmix, CisGenome, SISSRs), a FDR for finding enrichment peaks *within* the sample was applied. In this context, we imposed no further formal requirement to determine enrichment differences *between* experiments. In contrast, when differences between ChIP samples could be assessed directly (ChIPmonk, Model 2, Model 3), an appropriate FDR was chosen. A stringent FDR cutoff of 0.05 was applied throughout.

3.4.2 ChIP-chip methods comparison We applied ChIPmix (Martin-Magniette *et al*., 2008), ChIPmonk (Andrews, 2007) and our best fitting model (Model 3) to the mouse promoter methylation data (Fig. 3). ChIPmix differed notably from the other two methods in the total number of promoters detected as RDE (12507, 1226, 3972, for ChIPmix, ChIPmonk and Model 3, respectively). Only 3% of the RDE detected by ChIPmix overlapped with the other two methods. Better percentages were found for ChIPmonk (34% overlap) and Model 3 (11% overlap). The relatively high overlap percentage for ChIPmonk suggests high specificity. However, this advantage is probably offset by low sensitivity: 1891 promoters overlapping between ChIPmix and Model 3 went undetected with ChIPmonk. This could relate to the fact that ChIPmonk uses a sliding window approach where the signal of several consecutive tiles are pooled. ChIPmix and Model 3, on the other hand, use individual tiles in the analysis. This can lead to the detection of promoters that are subject to single tile differences. Analyses of other ChIP-chip datasets indicate that such localized differences can indeed lead to functional consequences on downstream gene expression (data not shown). Importantly, 99.67% of the 9675 RDE promoters that appear to be specific to ChIPmix were formally classified as RSE by Model 3. We can attribute this to the fact that the RDE classification with ChIPmix was performed *ad hoc*, whereas Model 3 explicitly modeled the null distribution of no methylation differences between ChIP samples. Similar arguments can be given for the unique overlap (513 promoters) between ChIPmonk and ChIPmix, 98% and 2% of which were classified as RSE and RSNE by Model 3, respectively.

3.4.3 ChIP-seq methods comparison We applied SISSRs (Jothi *et al*., 2008), CisGenome (Ji *et al*., 2008) and our best fitting model (Model 2) to the human transcription factor data (Fig. 3). CisGenome and Model 2 detected a comparable number of bins as RDE (5661 and 5387, respectively), whereas SISSRs detected relatively few (1704). Overall 661 RDE bins overlapped between the three methods, resulting in an overlap percentage of 38%, 11%, and 12% for SISSRs, CisGenome and Model 2, respectively. Similar to ChIPmonk above, SISSRs may suffer from low sensitivity with a substantial number of overlapping bins (1331) between CisGenome and Model 2, escaping detection with this method. Again, it is noteworthy that among the 722 bins that overlap exclusively between SISSRs and CisGenome, 30% were identified as RSE and 70% RSNE by Model 2.

It is perhaps most instructive to relate the performance of these three ChIP-seq methods back to our biological expectations of the BTAF1 knockdown system. As discussed above, a disruption of BTAF1 leads to an overaccumulation of TBP at promoters relative to the GAPDH condition. Hence, we expect RDE to be in the direction of BTAF1 rather than GAPDH (i.e. BTAF1 > GAPDH rather than GAPDH > BTAF1). At a FDR = 0.05, Model 2 provides the clearest support for this expectation with 71% of the bins classified as RDE (BTAF1 > GAPDH) mapping to promoter regions, compared with only 3% of the bins classified as RDE (GAPDH > BTAF1). The other methods showed a similar, but less pronounced trend (Table 1).

3.4.4 Common clustering approaches Clustering methods are common in high-dimensional classification problems, particularly in the analysis of microarray gene expression studies (Quackenbush, 2001). We therefore compared our approach with three popular clustering algorithms, hierarchical clustering, K-means partitioning and the model-based multivariate normal clustering as implemented through the *Mclust* R package. The details of this analysis are provided in the Supplementary Material. Our general conclusions highlight the poor performance of these alternative approaches in the context of ChIP-chip and ChIP-seq data, particularly with regard to detecting RDE. We attribute this to the lack of biologically meaningful constraints in these clustering procedures, which can prevent the identification of smaller clusters corresponding to RDE. Furthermore, the size of typical ChIP-chip or ChIP-seq datasets makes these methods computational infeasible.

4 DISCUSSION

We presented a parametric classification method for genome-wide comparisons of chromatin profiles between multiple ChIP samples. For simplicity, we focused our discussion on the special case of a two-sample comparison. Extensions to additional dimensions are possible. The number of components that have to be considered to capture the RDE structure between additional conditions increases at a rate 2^d, where d denotes the number of conditions involved in the comparison. While the interpretation of the classification results becomes increasingly complex, the actual modeling benefits greatly from the imposed parameter constraints so that the number of parameters to be estimated increases at a much slower rate.

The proposed approach provides a conceptual framework for the analysis of ChIP-chip and ChIP-seq data. Future work should consider the use of alternative density functions, such as the multivariate *t*-distribution (Peel and McLachlan, 2000) for ChIP-chip data, and some type multivariate discrete distribution for ChIP-seq. The key issue will be to explore the trade-off between model robustness on the one hand, and sensitivity–specificity on the other.

Another interesting extension of our method is to reformulate the mixture approach in the context of a hidden Markov modeling (HMM) framework. In this case, the RDE, RSE and RSNE components can be viewed as hidden Markov states and the mixture densities as the so-called emission probabilities. With ChIP-chip and ChIP-seq measurements collected from the same biological samples, it may even be possible to integrate these two data sources within a single HMM analysis, as previously attempted by Choi *et al*. (2009). While an HMM approach could be appealing, we note that its implementation would forfeit the, arguably, useful annotation-based genome partitioning scheme outlined above. Additionally, it still remains to be seen whether the Markov property of conditional independence between neighboring chromatin states is a valid assumption for this type of data.

ACKNOWLEDGEMENTS

We thank three anonymous reviewers for making this a better manuscript; W. Reik, S. Andrews for the mouse promoter methylation data and very helpful comments; V. Colot and W. Krijnen for stimulating discussions; H. Stunnenberg (Radboud University Nijmegen) for the SL30 antibody against human TBP.

Funding: The authors were supported by grants from the Netherlands Proteomics Centre (to F.M. and H.Th.M.T.); Netherlands Organization for Scientific Research (NWO-CW TOP 700.57.302 to P.dG and H.Th.M.T.); European Union (EUTRACC LSHG-CT-2006−037445 to F.M., P.dG and H.Th.M.T.); Human Frontier Science Program (grant LT-0860/2005 to F.M.); Horizon Breakthrough grant, Netherlands Organization for Scientific Research (NWO, 2009 to F.J.).

Conflict of interest: none declared.

REFERENCES

Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, **19**, 716–723.

Andrews,S. (2007) ChIPmonk: software for viewing and analysing ChIP-on-chip data. *BMC Syst. Biol.*, **1** (Suppl. 1), P80.

Bock,C. *et al*. (2008) Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Res.*, **36**, e55.

Choi,H. *et al*. (2009) Hierarchical hidden Markov model with application to joint analysis of ChIP-chip and ChIP-seq data. *Bioinformatics*, **15**, 1715–1721.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B.*, **39**, 1–38.

Esteller,M. (2008) Epigenetics in cancer. *N. Engl. J. Med*., **358**, 1148–1159.

Farthing,C.R. *et al*. (2008) Global Mapping of DNA Methylation in Mouse Promoters Reveals Epigenetic Reprogramming of Pluripotency Genes. *PLoS Genet.*, **4**, e1000116.

Gentleman,R. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Ji,H. *et al*. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.

Johannes,F. *et al*. (2008) Epigenome dynamics: a quantitative genetics perspective. *Nat. Rev. Genet.*, **9**, 883–890.

Johannes,F. *et al*. (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet.*, **5**, e1000530.

Jothi,R. *et al*. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, **36**, 5221–5231.

Kharchenko,P.V. *et al*. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.

Li,X.Y. *et al*. (2000) Distinct classes of yeast promoters revealed by differential TAF recruitment. *Science*, **288**, 1242–1244.

Martin-Magniette,M. *et al*. (2008) ChIPmix: mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics*, **24**, i181–i186.

McLachlan,G.J. and Peel,D. (2000) *Finite Mixture Models*. John Wiley and Sons, p. 303.

Meilijson,I. (1989) A fast improvement to the EM algorithm on its own terms. *J. R. Stat. Soc. B.*, **51**, 127–138.

Meissner,A. *et al*. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.

Mikkelsen,T.S. *et al*. (2008) Dissecting direct reprogramming through integrative genomic analysis. *Nature*, **454**, 49–55.

Peel,D. and McLachlan,G.J. (2000) Robust mixture modelling using the t distribution. *Stat. Comput.*, **10**, 339–348.

Penterman,J. *et al*. (2007) DNA demethylation in the Arabidopsis genome. *Proc. Natl Acad. Sci. USA*, **104**, 6752–6757.

Pereira,L.A. *et al*. (2003) Roles for BTAF1 and Mot1p in dynamics of TATA-binding protein and regulation of RNA polymerase II transcription. *Gene*, **315**, 1–13.

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

Reik,W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, **447**, 425–432.

Richards,E. (2008) Population epigenetics. *Curr. Opin. Genet. Dev.*, **18**, 221–226.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Sharp,P.A. (1992) TATA-binding protein is a classless factor. *Cell*, **68**, 819–821.

Suzuki,M.M. and Bird,A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.*, **9**, 465–476.

Vaughn,M.W. *et al*. (2007) Epigenetic natural variation in Arabidopsis thaliana. *PLoS Biol.*, **5**, e174.

Vermeulen,M. *et al*. (2007) Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell*, **131**, 58–69.

van Werven,F.J. *et al*. (2008) Cooperative action of NC2 and Mot1p to regulate TATA-binding protein function across the genome. *Genes Dev.*, **22**, 2359–2369.

Zilberman,D. *et al*. (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.*, **39**, 61–69.