BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 8 2014, pages 1154–1162
doi:10.1093/bioinformatics/btu003

Gene expression

Advance Access publication January 7, 2014

Integrative analysis of histone ChIP-seq and transcription data using Bayesian mixture models

Hans-Ulrich Klein[1,*,†], Martin Schäfer[2,†], Bo T. Porse[3,4,5], Marie S. Hasemann[3,4,5], Katja Ickstadt[6] and Martin Dugas[1]

[1]Institute of Medical Informatics, University of Münster, D-48149 Münster, [2]Mathematical Institute, Heinrich Heine University, D-40225 Düsseldorf, Germany, [3]The Finsen Laboratory, Rigshospitalet, Faculty of Health Sciences, [4]Biotech Research and Innovation Center (BRIC), [5]Danish Stem Cell Centre (DanStem), Faculty of Health Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark and [6]Faculty of Statistics, TU Dortmund University, D-44221 Dortmund, Germany

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Histone modifications are a key epigenetic mechanism to activate or repress the transcription of genes. Datasets of matched transcription data and histone modification data obtained by ChIP-seq exist, but methods for integrative analysis of both data types are still rare. Here, we present a novel bioinformatics approach to detect genes that show different transcript abundances between two conditions putatively caused by alterations in histone modification.

Results: We introduce a correlation measure for integrative analysis of ChIP-seq and gene transcription data measured by RNA sequencing or microarrays and demonstrate that a proper normalization of ChIP-seq data is crucial. We suggest applying Bayesian mixture models of different types of distributions to further study the distribution of the correlation measure. The implicit classification of the mixture models is used to detect genes with differences between two conditions in both gene transcription and histone modification. The method is applied to different datasets, and its superiority to a naive separate analysis of both data types is demonstrated.

Availability and implementation: R/Bioconductor package epigenomix.

Contact: h.klein@uni-muenster.de

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

Received on August 20, 2013; revised on December 2, 2013; accepted on December 30, 2013

1 INTRODUCTION

Modifications of histone proteins are an epigenetic mechanism to regulate gene transcription and are fundamental to stem cell differentiation as well as to the genesis of cancer (Baylin and Jones, 2011; Dawson and Kouzarides, 2012). Histone modifications can be localized genome wide using chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Furey, 2012; Park, 2009). Recently, several studies used this technique to explore the role of different chromatin states in transcriptional regulation (Cheng *et al*., 2011; Dong *et al*., 2012; Ernst and Kellis, 2010). In cancer research, it is often of interest to detect differences in histone modifications between two conditions, e.g. between cancer and normal cells or between cells carrying a mutation of an epigenetic regulator and wild-type cells. The main focus is usually on epigenetic modifications that occur together with a change in gene transcription, as these modifications are more likely to contribute to the phenotype or cancer development. Hence, in addition to histone ChIP-seq, many studies measure genome-wide RNA abundances from the same samples using expression microarrays or RNA sequencing (RNA-seq). The identification of genes showing differences in both histone modification and gene transcription data is crucial not only due to their potential causative role in cancer but also for identifying putative therapeutic targets for epigenetic drugs. Before presenting our integrative analysis approach to detect such genes using a Bayesian mixture model, we briefly review existing methods for data preprocessing and recent applications of mixture models to integrate genomic data.

Before matching both data types, an appropriate preprocessing is necessary to obtain standardized ChIP-seq and transcription values. Although normalization methods for transcription data, especially expression microarray data, are well established, rigorous preprocessing methods for ChIP-seq data are still an active field of research. When comparing two or more ChIP-seq samples, differences in immunoprecipitation efficiency and in sequencing depth should be accounted for. Many recent methods focus on the estimation of the proportion of background reads and require control samples either derived from input DNA or from ChIP against an unspecific antibody (Diaz *et al*., 2012; Enroth *et al*., 2012; Liang and Keles, 2012; Nair *et al*., 2012). Other methods, especially for comparative analysis, do not rely on control samples; for example, Xu *et al*. (2008) proposed dividing read counts by the total number of reads and thereby account for different sequencing depths. Taslim *et al*. (2009) suggested a non-linear normalization based on local regression. A related regression-based approach has recently been suggested by Shao *et al*. (2012). Song and Smith (2011) and Bao *et al*. (2013) presented models for detecting regions with differences in histone modifications that inherently account for sequencing depth or ChIP efficiency. Some methods originally designed for normalizing RNA-seq data may also be adequate

*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.

(Dillies *et al*., 2012), especially when ChIP-seq data are restricted to genomic regions like genes.

Mixture models have been used in many ways to cluster or classify genes. Recent applications include integrative analyses of various data types (Kirk *et al*., 2012), including gene transcription together with either DNA copy number (van Wieringen and van de Wiel, 2009), DNA methylation (Kormaksson *et al*., 2012) or transcription factor binding activity (Liu *et al*., 2007; Savage *et al*., 2010), as well as assessing concordance between ChIP-chip and ChIP-seq platforms (Schäfer *et al*., 2012). Models may consider a fixed number of components or allow for a flexible and even infinite number of components. The standard techniques for fitting such mixture models are EM algorithms in frequentist statistics (McLachlan *et al*., 2006; Newton *et al*., 2004; Taslim *et al*., 2009; van Wieringen and van de Wiel, 2009) and Markov chain Monte Carlo (MCMC) algorithms for Bayesian mixture models implying a prior distribution on the mixture (Kirk *et al*., 2012; Liu *et al*., 2007; Savage *et al*., 2010; Schäfer *et al*., 2012). Input measures for mixture models are case-specific, such as different log-ratio intensity measurements or transformed *P*-values.

In traditional clustering using mixture models, each cluster is represented by one Gaussian component. The number of clusters is determined by an external criterion or, especially in a Bayesian context, estimated within the model (Frühwirth-Schnatter, 2006). However, often clusters are poorly fitted by a single component; as a remedy, models involving a mixture of mixtures have been introduced in which each cluster is represented by a mixture itself (see Baudry *et al*., 2010; Rousseau and Mengersen, 2011, for frequentist and Bayesian approaches, respectively). In this work, we present a model involving a mixture of mixtures that circumvents overfitting and identifiability problems both by considering a fixed number of clusters suggested by the biological task and by fitting a fixed number of components corresponding to different types of distributions. The latter reduces the number of necessary mixture components and is beneficial for interpretation. Although a similar approach has been applied on ChIP-seq data by Taslim *et al*. (2009), our innovation lies in using a Bayesian framework that compared with the Expectation-Maximization (EM) algorithm holds several benefits, such as the possibility of assessing the uncertainty through credible intervals and less dependence on initial values. Furthermore, we apply our model on data obtained from integrating gene transcription and ChIP-seq datasets. We demonstrate the advantages of our method over the EM approach and over a separate analysis of both datasets on simulated data as well as on actual biological data.

In Section 2, we first describe our data matching and integration approach and subsequently present our Bayesian mixture model. Section 3 gives a description of results obtained on two different datasets and a simulation study. A discussion follows in Section 4 and final conclusions in Section 5.

2 METHODS

2.1 Datasets

The first dataset was derived from hematopoietic stem and progenitor cells (Lineage-, Sca-1+, c-Kit+) from *Cebpa*$^{fl/fl}$ mice (hereafter termed wild-type mice) and *Cebpa*$^{fl/fl}$; *Mx1Cre* mice conditionally deleted for *Cebpa* (hereafter termed knockout mice). Specimens from three *Cebpa* knockout mice and three wild-type mice were hybridized separately on six Affymetrix Mouse Gene 1.0 ST arrays. Transcription data were normalized using the robust multi-array average algorithm (RMA, Irizarry *et al*., 2003) (Gene Expression Omnibus GSE49975). ChIP against histone H3 lysine K4 trimethylation (H3K4me3) was applied to two pools of three wild-type mice and two pools of three knockout mice each. After ChIP, specimens were sequenced on an Illumina Genome Analyzer IIx sequencer producing 36-bp reads (GSE43007). The Burrows–Wheeler aligner (Li and Durbin, 2009) was applied to map the reads against the reference genome (mm10).

The second dataset was taken from a study (Bert *et al*., 2013) where epigenetic differences between a prostate cancer cell line (LNCaP) and normal primary prostate cells (PrEC) were studied. RNA-seq was carried out on an Illumina Genome Analyzer IIx to measure RNA transcription in LNCaP and PrEC cells. The STAR algorithm (Dobin *et al*., 2013) was used to align reads (76 bp, single end) against the human reference genome (hg19) annotated with splice junctions obtained from the GENCODE project (Harrow *et al*., 2012) to improve alignment accuracy. H3K4me3 and histone H3 lysine K27 trimethylation (H3K27me3) were localized in the LNCaP and PReC cells by ChIP-seq (Illumina HiSeq 2000 and Illumina Genome Analyzer IIx resp., 50 bp reads). ChIP-seq reads were aligned against the human reference genome (hg19) using the Burrows–Wheeler aligner (Li and Durbin, 2009). The dataset can be downloaded at Gene Expression Omnibus (GSE38685).

A third dataset was taken from a study (Schenk *et al*., 2012) comparing the effect of all-trans-retinoic acid (ATRA) treatment with the effect of a combined treatment of ATRA and tranylcypromine (TCP) on a leukemic cell line. Details of the dataset and analysis results are given in the Supplementary Material.

A simulation dataset was generated based on the first two knockout mice replicates from the *Cebpa* dataset. We assumed that all genes were neither differential in gene transcription nor in histone modification between the two biological replicates. Differential genes were then simulated by multiplying the transcription value and the ChIP-seq value (calculated as described in Section 2.2) of the first replicate by $(1+c)$ with $c \in \{-1, -0.9\ldots, -0.1, 0.1, \ldots, 1\}$. The 100 genes with a ChIP-seq value above the median of all ChIP-seq values were randomly chosen for each level of factor c leading to 2000 of 21 236 genes with equally directed differences in both data types.

2.2 Data matching and normalization

Data matching is performed at transcript level. The basic idea is to obtain a measure for the abundance of a transcript or group of transcripts from either RNA-seq data or microarray data and then to allocate a ChIP-seq value based on the number of ChIP-seq reads aligned within the genomic region of that transcript. Here, we focus on genomic regions centered at the transcripts' transcriptional start sites (TSSs), as the histone modifications studied here primarily occur at TSSs, with H3K27me3 occurring additionally throughout gene bodies (Dong *et al*., 2012).

2.2.1 RNA-seq data After alignment of RNA-seq reads to the genome by software capable of producing spliced alignments, transcript abundance is estimated using the Cufflinks algorithm (Trapnell *et al*., 2010). Abundances are reported in fragments per kilobase of transcript per million fragments mapped (FPKM) and are scaled via the ratio of the sample's 0.75 quartile to the average 0.75 quartile value across all samples as implemented in the Cuffdiff software (Trapnell *et al*., 2013). Transcripts sharing the same TSS are grouped and their FPKM values are summarized to obtain a single FPKM value for each TSS. Finally, FPKM values are logarithmized. In the following, in case of RNA-seq data, X_i and A_i denote the normalized FPKM values of the two conditions/treatments of interest (e.g. cancer and normal cells) for all transcripts sharing TSS $i = 1, \ldots, n$.

2.2.2 Gene expression microarray data In contrast to RNA-seq or ChIP-seq, gene expression microarray data consist of continuous measurement values that can be directly assigned to transcripts based on the given array design. Several normalization methods for various array platforms exist. We chose RMA (Irizarry *et al.*, 2003) for Affymetrix GeneChips arrays and variance-stabilizing transformation (Lin *et al.*, 2008) for Illumina Bead Chips. Both methods apply a logarithmic or similar transformation on the transcription values. A drawback of the array technology is that, depending on the array design, an array probe may measure several transcripts and that these transcripts may have different TSSs. In case of array data, X_i and A_i denote the normalized transcription values of the two conditions/treatments of interest for all transcripts measured by probe $i = 1, \ldots, n$.

2.2.3 ChIP-seq data ChIP-seq values are calculated for a given TSS i by counting the number of sequenced fragments lying within the genomic region $\mathcal{R}_i$ of width w centered at the TSS. Let Y_i and B_i denote the number of reads overlapping $\mathcal{R}_i$ in the two conditions. Reads are expanded toward the 3′-end to the mean DNA fragment length (here 200 bp or 350 bp) before calculating Y_i and B_i. The size w of $\mathcal{R}_i$ must be chosen depending on the studied histone modification. For many histone modifications, an appropriate size is known (Hebenstreit *et al.*, 2011), and we show that the choice of w is not crucial for our approach. Alternatively, a proper choice of $\mathcal{R}_i$ can be obtained from a peak detection algorithm applied to the ChIP-seq data. In case of RNA-seq data, X_i and A_i can be directly matched to Y_i and B_i. However, array probes may be located at exons associated with more than one transcript, which may have different TSSs. In such cases, we define $\mathcal{R}_i$ as the union of the regions derived for all single transcripts measured by probe i. So $\mathcal{R}_i$ may consist of more than one genomic interval, and its size may differ for different probes.

To account for different total number of reads and different ChIP efficiency, quantile normalization (Bolstad *et al.*, 2003) is applied to the ChIP-seq values. The ChIP-seq values are ordered, so that $\mathrm{Y}_{\iota_1} \leq \ldots \leq \mathrm{Y}_{\iota_n}$ and $\mathrm{B}_{\tau_1} \leq \ldots \leq \mathrm{B}_{\tau_n}$. The normalized values are then defined as $Y_{\iota_i} = B_{\tau_i} := (\mathrm{Y}_{\iota_i} + \mathrm{B}_{\tau_i})/2$.

2.2.4 Data integration After data normalization and matching, a correlation value Z inspired by the externally centered correlation coefficient (Schäfer *et al.*, 2009, 2012) is calculated by multiplying the standardized difference of transcription values with the standardized difference of ChIP-seq values:

$$Z_i = \frac{X_i - A_i}{\sigma_{XA}} \frac{Y_i - B_i}{\sigma_{YB}}, \quad i = 1, \ldots, n \tag{1}$$

The variances of the differences $\sigma^2_{XA} = 1/(n-1)\sum_{i=1}^{n}(X_i - A_i)^2$ and $\sigma^2_{YB} = 1/(n-1)\sum_{i=1}^{n}(Y_i - B_i)^2$ are calculated across all transcripts. If a transcript shows equally directed differences in transcription and histone modification data, the corresponding Z value is positive, whereas it is negative in case of unequally directed differences. If a transcript has differences in only one of both data types or has no differences at all, Z is expected to be close to zero. Thus, for an activating histone modification, the distribution of Z is expected to be slanted toward positive values.

If biological replicates are available, the average is calculated after matching and normalization and then used in Equation (1), i.e. $Y_i = 1/m\sum_{j=1}^{m} Y_{ij}$ if m ChIP-seq replicates are available and Y_{ij} is the quantile normalized ChIP-seq value of the j-th replicate for transcript i.

2.3 Bayesian mixture model

In the following, the distribution of Z will be studied to detect transcripts with alterations and to explore the association between histone modification and transcript levels. As motivated in the previous section, we view the distribution of Z as consisting of three clusters: A large probability mass centered around zero corresponding to transcripts displaying differences between the two conditions/treatments of interest in none or just one of both data types, a smaller probability mass on the positive axis corresponding to transcripts displaying equally directed differences between the two conditions in transcription and histone modification data and an also smaller probability mass on the negative axis corresponding to transcripts displaying unequally directed differences between the two conditions in both data types. Measure Equation (1) aggregates the information necessary to distinguish between these three clusters of transcripts. The difficulty lies in discriminating Z values whose deviation from zero is small enough to be explained by random variability from those values that represent transcripts displaying (either equally or unequally directed) differences between the two investigated conditions in both transcription and histone modification data.

A classic approach to analyze the distribution of Z that reflects this conception would be a three-component mixture model, i.e. a model in which each cluster is represented by one component. However, a cluster may be represented by more than one component in a mixture model when convenient for achieving a good fit and classification. In an investigation of an histone modification via ChIP-chip and ChIP-seq measurements (Schäfer *et al.*, 2012), a measure similar to Equation (1) was used in a mixture model with eight Gaussian components representing three clusters, concluding that less components, whereas preferable for interpretation, do not provide enough flexibility to fit the histogram of the Z values. Here, we want to propose a more general modeling approach that potentially reduces the number of necessary mixture components for analyzing three clusters, e.g. when considerable probability mass in the tails of the distribution constitute a challenge to standard models.

We take up the idea of Taslim *et al.* (2009) to fit a mixture of not only normal distributions, as in traditional model-based clustering, but also both normal and exponential distributions. Although a number of normal components represent the center of the distribution, the probability mass in the distributions' tails is covered by two exponential distributions, one of which is mirrored at zero. By using the exponential distributions, the number of components can be considerably reduced, as each exponential component potentially replaces several normal components. In determining the number of normal components representing the center of the distribution, besides the fit, one criterion is to achieve a clean classification in the sense that the resulting classification should produce three contiguous domains of values.

2.3.1 Model Formally, Z is assumed to be a random variable and $Z_1, \ldots, Z_n$ to be an independent and identically distributed random sample of Z. The mixture model for the distribution F of Z with density f is defined as follows:

$$\begin{aligned} Z_i|\lambda, \mu, \sigma^2 &\sim F(\lambda, \mu, \sigma^2) \\ f(z_i|\lambda, \mu, \sigma^2) &= \pi_1 \cdot h(-z_i|\lambda_1) \\ &\quad + \sum_{k=2}^{K-1} \pi_k \cdot g(z_i|\mu_k, \sigma_k^2) \\ &\quad + \pi_K \cdot h(z_i|\lambda_K), \quad i = 1, \ldots, n \end{aligned}$$

with mixture proportions $\pi_k, k = 1, \ldots, K$ and $\sum_{k=1}^{K} \pi_k = 1$. The g denotes the density of the normal and h the density of the exponential distribution. Inverse gamma distributions are assigned to the variances of the normal distributions, σ_k^2, $k = 2, \ldots, K-1$, whereas gamma distributions are assumed for the parameters λ_1 and λ_K of the two exponential distributions:

$$\begin{aligned} 1/\sigma_k^2 &\sim \mathrm{Gamma}(a_{\sigma_k}, b_{\sigma_k}) \text{ for } k = 2, \ldots, K-1 \\ \lambda_k &\sim \mathrm{Gamma}(a_{\lambda_k}, b_{\lambda_k}) \text{ for } k = 1, K \end{aligned}$$

These are the respective conjugate prior distributions and thus allow the application of a Gibbs sampler. The classification of transcripts is carried out by means of an allocation variable T that is given a categorical

distribution. Correspondingly, $T_1, \ldots, T_n$ are the classifications for all transcripts $i = 1, \ldots, n$. The mixture proportions are following a Dirichlet distribution,

$$T_i \sim \text{Categorical}(\pi_1, \ldots, \pi_K)$$
$$\pi_1, \ldots, \pi_K \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \quad (2)$$

The use of the prior in Equation (2) generally favors a low number of non-empty components, which we prefer for the sake of interpretability (Frühwirth-Schnatter, 2011). Because the assignment of transcripts to the components depends essentially on the mass parameter α, we estimate it as part of the model instead of choosing a fixed value. Thus, we also assign a distribution to α (Ishwaran and Zarepour, 2000),

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha).$$

The model is fitted via MCMC methods using a Gibbs sampler (Ishwaran and James, 2001; Neal, 2000, for details of our implementation see Supplementary Material), and model estimates are obtained by averaging across the posterior distributions. This mixture model is closely related to Dirichlet process models in which the potentially infinite number of components is approximated by an effective finite mixture model with an upper bound on the non-empty components (Ishwaran and James, 2001; Ishwaran and Zarepour, 2000, one application being Kirk *et al.*, 2012). Depending on the application, the upper bound is usually large, e.g. on the order of several hundreds. In contrast, we prefer a small number of components for the sake of classification and easy interpretation while still achieving three contiguous classes as well as a good fit by non-standard mixing of different distributions. We found in preliminary analyses that four normal components were sufficient to achieve a good overall fit.

2.3.2 Prior distributions The normal components represent the Z values equal to or near zero and are thus given a fixed mean $\mu_j = 0$ and a small variance a priori. The parameter of the exponential components is assigned a small value a priori, resulting in a large mean, which ensures that they represent the tails of the distribution while avoiding label switching between the components that correspond to distinct clusters. Specifically, we choose $a_{\sigma_{k,0}} = b_{\sigma_{k,0}} = 10$ for $k = 2, \ldots, K-1$, $a_{\lambda_{k,0}} = b_{\lambda_{k,0}} = 0.001$ for $k = 1, K$ and $a_{\alpha_0} = b_{\alpha_0} = 1$ as prior values.

3 RESULTS

3.1 *Cebpa* knockout dataset

Normalized transcription values were averaged across the three wild-type and three knockout samples for each probe annotated with at least one transcript. This resulted in $i = 1, \ldots, 21\,236$ transcription values X_i and A_i for both conditions. In all, 7163 probes were assigned to multiple transcripts with different TSSs according to the ENSEMBL data base (Flicek *et al.*, 2013). ChIP-seq values Y_i and B_i were calculated by counting the number of reads overlapping the promoter regions $\mathcal{R}_i$. Based on work by Hebenstreit *et al.* (2011) and on visual inspection of the read distribution at TSSs, we chose a promoter width of $w = 6\,000$. After quantile normalization, ChIP-seq values were averaged across the two knockout and wild-type replicates. To assess the effect of different choices of w, we calculated the Pearson correlation ρ between $X_i - A_i$ and $Y_i - B_i$ for different choices of w ranging from 100 to 20 000 bp (Fig. 1A). The choice of w seems not to be crucial, unless it is chosen too small. Furthermore, the correlation for probes with a unique TSS ($\rho = 0.200$) does not differ remarkably from probes with multiple TSSs ($\rho = 0.203$), indicating that the aggregation of reads from multiple TSSs is suitable.

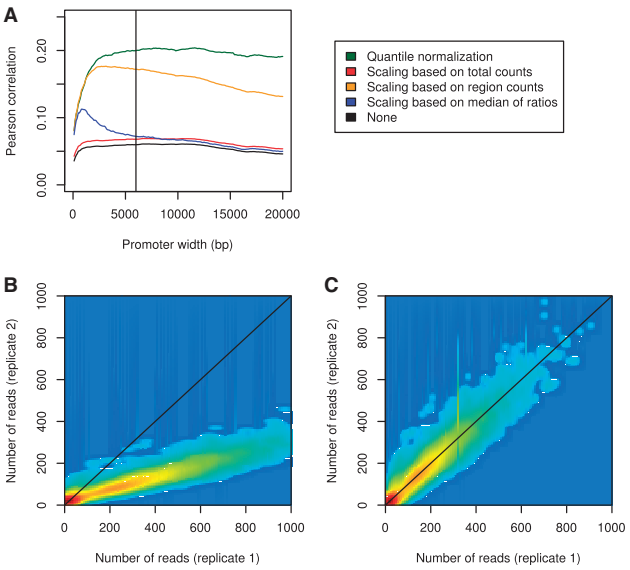

Fig. 1. Quantile normalization of ChIP-seq values. Figure (**A**) shows the Pearson correlation between differences in gene transcription and differences in ChIP-seq values for different promoter sizes (x-axes) and different ChIP-seq normalization approaches. The black vertical line indicates the promoter width of $w = 6\,000$ chosen for this dataset. Quantile normalization performed superior compared with linear scaling based on the total number of reads as used by many peak detection algorithms, e.g. MACS (Zhang *et al.*, 2008), or compared with scaling based on the median of count ratios as proposed by Anders and Huber (2010) for RNA-seq data (see Supplementary Data for details on the different normalization methods). Scatter plot (**B**) between the number of reads observed within the promoters ($w = 6\,000$) in the first and the second *Cebpa* knockout replicate sample clearly shows need for normalization. Figure (**C**) shows the scatterplot after quantile normalization

Figure 1A also shows that quantile normalization of the ChIP-seq data leads to a higher correlation between the differences in the transcription data and the differences in the ChIP-seq data than the other studied normalization methods. A scatter plot between the ChIP-seq values of the two *Cebpa* knockout replicates (Fig. 1B) shows larger read counts in the promoter regions for replicate 1, although replicate 2 had more mapped reads in total (Supplementary Table S1) indicating different ChIP efficiency between both replicates. Figure 1C shows the ChIP-seq values after quantile normalization.

After normalization, Z values were calculated and the model from Section 2.3.1 was fitted to the data. Figure 2A and B show the model fit and Figure 2C the classifications obtained from 100 000 iterations using every 10th iteration after a burn in of 2000 iterations. Parameter estimations are given in Table 1 and corresponding trace plots in Supplementary Figure S1. Classifications were obtained by calculating the mode of T_i across iterations. The distribution of Z had more probability mass at the right tail than at the left tail, as reflected by the estimated weights $\hat{\pi}_j$ of the mixture components. The weight of the negative component $\hat{\pi}_1 = 0.002$ was negligible, whereas $\hat{\pi}_6 = 0.039$ was estimated for the positive component's weight.

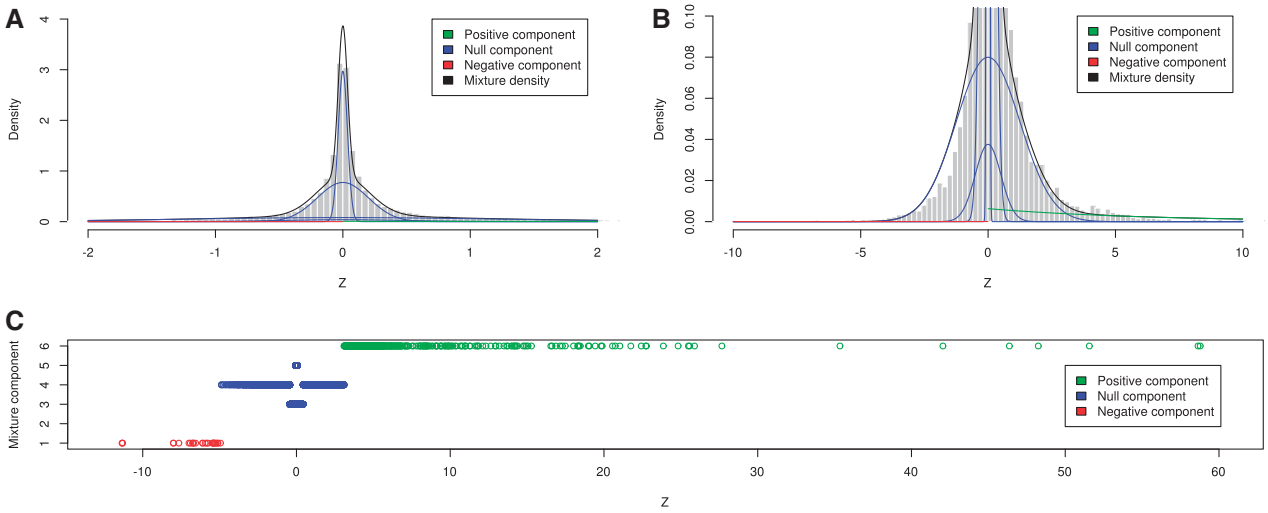

Fig. 2. Model fit and classification for the *Cebpa* knockout dataset. The gray histograms in (**A**) and (**B**) show the empirical distribution of Z at different scales. The histograms are overlaid with the mixture density and the densities of the single components. In (**C**) the observed values z_i are plotted against their classification in the mixture model. One transcript classified to component 6 with $z_i \approx 107$ was omitted. More transcripts were classified into the positive than into the negative component indicating a positive correlation between H3K4me3 modifications and gene transcription. No transcripts were classified to component 2: it is completely dominated by component 4 and thus plays a role with respect to the fit of the mixture distribution to the distribution of Z but not for classification

Table 1. Estimated parameters of the Bayesian mixture model for the Cebpa knockout dataset

	Negative	Null				Positive
	1	2	3	4	5	6
$\hat{\sigma}^2$	–	0.255 [0.197, 0.326]	0.041 [0.036, 0.046]	1.572 [1.442, 1.717]	0.001 [0.001, 0.001]	–
$\hat{\lambda}$	0.025 [0.016, 0.039]	–	–	–	–	0.162 [0.145, 0.181]
$\hat{\pi}$	0.002 [0.001, 0.002]	0.048 [0.036, 0.059]	0.390 [0.376, 0.406]	0.251 [0.236, 0.266]	0.270 [0.254, 0.286]	0.039 [0.034, 0.045]
		0.959 [0.953, 0.965]				
Number of genes	27	0	8524	4948	7279	458
		20751				

Note: The 95% credibility intervals are given in square brackets. The curly brackets summarize the weights π, or, respectively, the number of classified transcripts for all four null components.

Only 27 transcripts were classified to the negative component, 20 751 to the null components and 458 to the positive component (Supplementary Spreadsheet S1). Hence, the majority of transcripts did not show differences in both data types. However, transcripts with differences in both data types clearly revealed that an increase (decrease) in H3K4me3 is associated with an increase (decrease) in gene transcription. Among the 458 transcripts in the positive cluster, we found several genes that have previously been implicated in hematopoietic stem cell biology (*Mecom*, *Kit*) and/or acute myeloid leukemia (*Hoxa9*, *Meis1*), highlighting the functions of *Cebpa* in normal and malignant hematopoiesis. Moreover, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses (Dennis *et al.*, 2003) identified *acute myeloid leukemia* as significantly enriched (False discovery rate (FDR) <0.05).

To study the reproducibility of the results, the dataset was split into two datasets each consisting of one microarray and one ChIP-seq sample from the knockout and wild-type collective, respectively. Thus, each dataset consists of a single biological replicate. The model fits and classification results obtained by applying the model to each set of replicates separately are shown in Figure 3. The contingency table (Table 2) of the classification results shows that fewer transcripts were classified to the positive component in each of the two separate analyses compared with the analysis of the full dataset. This may be due to smaller variances σ^2_{XA} and σ^2_{YB} in the full dataset caused by averaging the

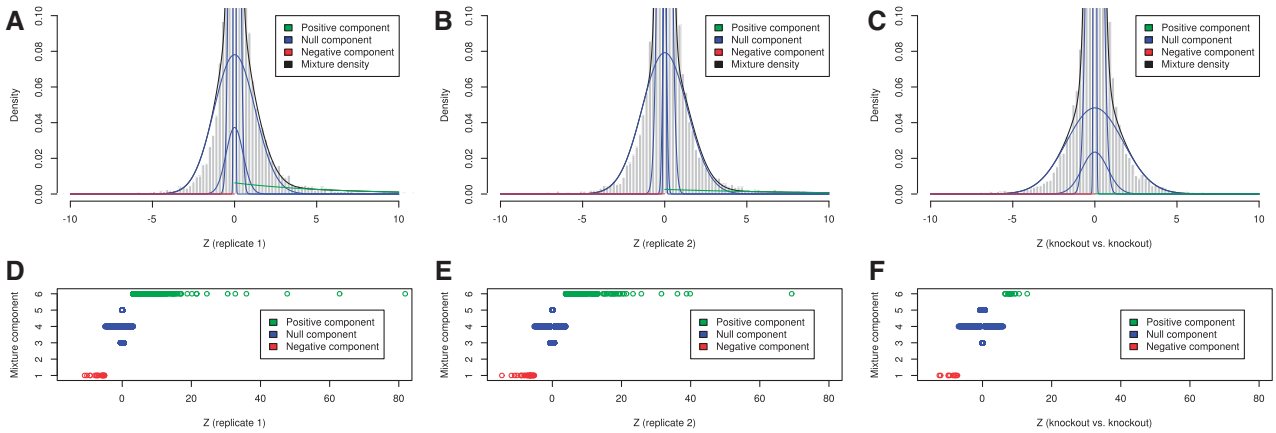

Fig. 3. Model fits and classifications for subsets of the *Cebpa* knockout dataset. (**A**) and (**B**) show the distribution of Z and the mixture model's fit when comparing the first (second) *Cebpa* knockout replicate to the first (second) wild-type replicate. The corresponding classifications are given in (**D**) and (**E**), respectively. (**C**) and (**F**) present the mixture model of the comparison of the first to the second *Cebpa* knockout replicate

Table 2. Classification results of both replicates from the Cebpa knockout dataset

		Replicate 2			
		Negative	Null	Positive	Sum
Replicate 1	Negative	1	18	1	20
	Null	28	20635	159	20822
	Positive	3	287	104	394
	Sum	32	20940	264	

Note: Shown is the number of transcripts classified into the negative, null and positive components when applying the Bayesian mixture model separately to the first and second set of biological replicates.

replicates. Furthermore, only one transcript was classified to the negative component in both analyses, whereas 104 transcripts were classified to the positive component twice.

The specificity of our approach was assessed by comparing the first *Cebpa* knockout replicate to the second knockout replicate. In this setup, no biological differences exist and all transcripts should be classified to the null components. We observed small weights $\hat{\pi}_1 = 0.001$ and $\hat{\pi}_6 = 0.004$ of the two exponential components (Fig. 3C). Only 11 (18) of 21 236 transcripts were assigned to the negative (positive) component (Fig. 3F). On the whole, these findings speak in favor of reproducibility of the presented results.

3.2 Prostate cancer dataset

Normalized transcription values of LNCaP and PrEC cells were calculated for 113 663 (groups of) transcripts that share the same TSS based on transcript annotation from the GENCODE project (Harrow *et al.*, 2012). In all, 46 657 of these showed transcript abundances in LNCaP or PrEC and were used for model fitting and classification. For both H3K4me3 and H3K27me3 histone modifications, a promoter width of $w = 3\,000$ was chosen. We chose a smaller promoter size than for the *Cebpa* knockout dataset on the basis of the observed correlation between differences in ChIP-seq values and transcription values for different choices of w (Supplementary Fig. S2) and due to the fact that RNA-seq allows to distinguish between different transcripts of the same gene that may have their TSSs in proximity. The observed correlation of $\rho = 0.289$ for H3K4me3 was higher than in the *Cebpa* knockout dataset, indicating larger differences or a larger fraction of genes showing epigenetic and transcriptional differences when comparing prostate cancer cells with normal cells. Interestingly, although it is known that H3K27me3 occurs at the TSS but also throughout the gene body (Ernst and Kellis, 2010; Hebenstreit *et al.*, 2011), a larger window width w covering larger parts of the gene body did not lead to a stronger negative correlation (Supplementary Fig. S2b). Recently, Dong *et al.* (2012) also chose windows close to the TSS to predict transcription based on the occurrence of H3K27me3. We observed $\rho = -0.197$ for H3K27me3, and for both histone modifications, quantile normalization performed superior in terms of maximizing the absolute correlation (Supplementary Fig. S2), albeit the need for normalization was not as evident as in the *Cebpa* knockout dataset.

Applying our Bayesian mixture model to the integrated transcription and H3K4me3 data lead to 3526 transcripts classified to the positive and 272 transcripts classified to the negative component (Supplementary Spreadsheet S2). Consistently, we observed a small weight of $\hat{\pi}_1 = 0.007$ for the negative component compared with the weight of the positive component $\hat{\pi}_6 = 0.107$ (Supplementary Fig. S3A–C and Supplementary Table S4). *Focal adhesion*, *axon guidance* and *pathways in cancer* were significantly (FDR < 0.05) enriched KEGG pathways (Dennis *et al.*, 2003). Axon guidance genes were recently reported to be involved in pancreatic carcinogenesis (Biankin *et al.*, 2012) and may also play a role in other cancers (Mehlen *et al.*, 2011). Alternative promoter usage was observed only for 8 genes involving 20 transcripts, whereas mostly all transcripts of a gene that were classified to component 6 were consistently upregulated or downregulated. *TPM1* and *SMTN* were among these eight genes and are involved in actin cytoskeleton development and stabilization.

For the repressive H3K27me3 mark, we obtained an estimated weight of $\hat{\pi}_1 = 0.060$ ($\hat{\pi}_6 = 0.015$) and 2 098 (436) transcripts that were classified to the negative (positive) component (Supplementary Fig. S3D–F, Supplementary Table S5 and Supplementary Spreadsheet S3). The same three pathways derived from the H3K4me3 model were also significant (FDR < 0.05) when the pathway analysis was applied to the transcripts classified as negative by the model fitted with the H3K27me3 data. In all, 976 of the 2098 transcripts of the negative component were classified as positive by the model based on the H3K4me3 data, reflecting interactions between occurrences of active H3K4me3 and repressive H3K27me3 marks.

3.3 Simulated dataset

The simulated dataset contained 2000 of 21 236 transcripts with equally directed differences in both data types. Our approach classified 1656 of these transcripts into the positive component. A total of 15 transcripts were falsely classified to the positive component, and 344 transcripts were falsely classified to a null or negative component. In summary, a sensitivity of 0.828 and a specificity of 0.999 were observed. We compared our approach with a naive separate analysis of both data types. The differences $X_i - A_i$ of the gene transcription and $Y_i - B_i$ of the ChIP-seq values were calculated, and a threshold t was chosen. All k transcripts with $|X_i - A_i| \geq t$ were considered as differentially transcribed, and the k transcripts with the largest absolute differences $|Y_i - B_i|$ were considered as differentially histone modified. A Receiver-Operating Characteristic curve was plotted for varying thresholds t and compared with the results obtained from our integrative method (Supplementary Fig. S6). At the same level of specificity, the naive approach achieved a smaller sensitivity of 0.683, probably due to a loss of information caused by the separate analyses. Supplementary Table S7 shows that especially for moderate differences, our approach achieved a gain in sensitivity of $\sim$0.3.

4 DISCUSSION

The presented results are consistent with the literature (Zhang *et al.*, 2009) with respect to the association between gene transcription and histone modifications. Results from the simulated dataset and from comparing two biological replicates from the *Cebpa* dataset give evidence for a high specificity when detecting transcripts with differences in both datasets. Furthermore, the simulation study indicates that a reasonable sensitivity may be achieved. This is substantiated by the good reproducibility of the classification results when splitting the *Cebpa* dataset into two datasets of sample size one.

Using distributions of different types is beneficial for classifying transcripts in our Bayesian mixture model approach. In combination with a small fixed number of components, it helps to avoid label switching problems that often occur for mixture models with a large number of normal components, of which a considerable proportion often remains empty. In preliminary analyses, a mixture model using exclusively normal components had to use 15 components to achieve both a good overall fit and three contiguous classes where the proposed new model only needs six.

When fitting the mixture model with the EM algorithm (like in Taslim *et al.*, 2009), the results depended on the set of initial values due to convergence in different local maxima (see Supplementary Fig. S10). It was shown that some initial values did not result in reasonable classifications, although they led to a higher likelihood. Although in our Bayesian framework in combination with MCMC algorithms for model fitting, informative prior distributions had to be chosen to ensure a sensible classification, we argue that this explicit model-based approach is preferable, at least when it is consistent across several datasets like in our analysis. Moreover, unlike the EM algorithm, the Bayesian framework allows for several possible extensions of the model, such as incorporation of interactions and spatial correlation between the transcripts.

Compared with naive separate analyses of both datasets, our approach demonstrated to be superior in the simulation study. This is probably due to the fact that the classification is based on information from both datasets aggregated by our score given in Equation (1) and underlines the need for novel integrative methods for studies incorporating more than one genomic dataset. Model-based classification may consume more computing time but has the advantage that no threshold has to be chosen, in contrast to naive approaches based on *P*-value or fold-change rankings.

We consider our approach as an appropriate framework also for other classification tasks when integrating two types of 'omics' data, helping to cope with sophisticated distributions when the number of clusters is known. As alternatives to the proposed model, a model with a variable number of components or a different number of components in the mixture may be considered. However, a fixed number of components, whereas less flexible in terms of model fit, holds more benefits in terms of interpretability and is often preferred when classification is the major goal of the analysis, and the number of clusters is obvious due to the underlying biological question. Kormaksson *et al.* (2012), e.g. use a two-component normal mixture model to classify probe sets with respect to low or high methylation, whereas Broet and Richardson (2006) and Wei and Pan (2008) fit three-component models to classify loci into being unmodified or being subject to loss or gain of genetic material. Although our model was validated by application to distinct datasets, we do not rule out the possibility that for datasets of different structure, a different number of components might be optimal. If considered appropriate for another application, the framework can be easily adapted to provide for a variable and unlimited number of clusters and/or components, e.g. based on the finite approximation to a Dirichlet process mixture model (Ishwaran and James, 2001). In such models, overfitting might be an issue and sufficient sparsity should be ensured for better interpretability (Frühwirth-Schnatter, 2011). Of course, other distributions may also be incorporated in the model.

Alternatively, one might consider to approximate the distribution of each factor in Equation (1) by a normal distribution, and hence of Z_i by a normal-product distribution. We assume that this works well for datasets that lead to roughly equally formed positive and negative tails in the Z distribution, when the focus is on the fit of the distribution. Our model additionally offers the possibility of classification by representing classes by distinct components of the mixture model and is flexible enough for

one tail of the Z distribution to contain considerably more probability mass than the other one.

When larger sample sizes are available, uncertainty could be modeled across samples for each locus, introducing an additional model layer. An adapted measure Z could be defined similarly as was done in Schäfer *et al*. (2009) in a frequentist context, focusing on the summands in an externally centered correlation coefficient, which could be assigned an, e.g. normal distribution.

5 CONCLUSION

We propose quantile normalization for ChIP-seq data and a novel Bayesian mixture approach involving a mixture of mixtures and distributions of different type (normal and exponential) to classify transcripts based on a new measure for the correlation between histone modifications and gene transcription. This integrative analysis was able to detect transcripts for which alterations in transcript abundances and histone modification exist between two different conditions in several datasets, including different histone modification and transcription data from either microarrays or RNA-seq. We assessed the sensitivity and specificity of our approach based on simulated data and on biological replicates and showed its superiority toward naive separate analyses. Given the fact that modern studies are often not limited to one type of 'omics' data, the presented method is a useful and important tool for the integrated analysis of epigenetic and transcription data.

Funding: Deutsche Forschungsgemeinschaft (Research Training Group 1032/2 'Statistical Modeling' and SCHW 1508/3-1); German Cancer Aid (110495); EuGESMA COST Action (BM0801).

Conflict of Interest: none declared.

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol*., **11**, R106.

Bao,Y. *et al*. (2013) Accounting for immunoprecipitation efficiencies in the statistical analysis of ChIP-seq data. *BMC Bioinformatics*, **14**, 169.

Baudry,J. *et al*. (2010) Combining mixture components for clustering. *J. Comput. Graph. Stat*., **19**, 332–353.

Baylin,S.B. and Jones,P.A. (2011) A decade of exploring the cancer epigenome - biological and translational implications. *Nat. Rev. Cancer*, **11**, 726–734.

Bert,S.A. *et al*. (2013) Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer Cell*, **23**, 9–22.

Biankin,A.V. *et al*. (2012) Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*, **491**, 399–405.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Broet,P. and Richardson,S. (2006) Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics*, **22**, 911–918.

Cheng,C. *et al*. (2011) A statistical framework for modeling gene expression using chromatin features and application to modencode datasets. *Genome Biol*., **12**, R15.

Dawson,M.A. and Kouzarides,T. (2012) Cancer epigenetics: from mechanism to therapy. *Cell*, **150**, 12–27.

Dennis,G. *et al*. (2003) David: database for annotation, visualization, and integrated discovery. *Genome Biol*., **4**, P3.

Diaz,A. *et al*. (2012) Normalization, bias correction, and peak calling for ChIP-seq. *Stat. Appl. Genet. Mol. Biol*., **11**, Article 9.

Dillies,M.-A. *et al*. (2012) A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform*., **14**, 671–683.

Dobin,A. *et al*. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Dong,X. *et al*. (2012) Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol*., **13**, R53.

Enroth,S. *et al*. (2012) A strand specific high resolution normalization method for chip-sequencing data employing multiple experimental control measurements. *Algorithms Mol. Biol*., **7**, 2.

Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol*., **28**, 817–825.

Flicek,P. *et al*. (2013) Ensembl 2013. *Nucleic Acids Res*., **41**, D48–D55.

Frühwirth-Schnatter,S. (2006) *Finite Mixture and Markov Switching Models*. Springer, Berlin.

Frühwirth-Schnatter,S. (2011) Dealing with label switching under model uncertainty. In: Mengersen,K.L., Robert,C.P. and Titterington,D.M. (eds) *Mixture Estimation and Applications*. Wiley, Chichester, pp. 193–218.

Furey,T.S. (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet*., **13**, 840–852.

Harrow,J. *et al*. (2012) Gencode: the reference human genome annotation for the encode project. *Genome Res*, **22**, 1760–1774.

Hebenstreit,D. *et al*. (2011) EpiChIP: gene-by-gene quantification of epigenetic modification levels. *Nucleic Acids Res*., **39**, e27.

Irizarry,R.A. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Ishwaran,H. and James,L.F. (2001) Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc*., **96**, 161–173.

Ishwaran,H. and Zarepour,M. (2000) Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, **87**, 371–390.

Kirk,P. *et al*. (2012) Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, **28**, 3290–3297.

Kormaksson,M. *et al*. (2012) Integrative model-based clustering of microarray methylation and expression data. *Ann. Appl. Stat*., **6**, 1327–1347.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Liang,K. and Keles,S. (2012) Normalization of ChIP-seq data with control. *BMC Bioinformatics*, **13**, 199.

Lin,S.M. *et al*. (2008) Model-based variance-stabilizing transformation for illumina microarray data. *Nucleic Acids Res*., **36**, e11.

Liu,X. *et al*. (2007) Bayesian hierarchical model for transcriptional module discovery by jointly modeling gene expression and ChIP-chip data. *BMC Bioinformatics*, **8**, 283.

McLachlan,G. *et al*. (2006) A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, **22**, 1608–1615.

Mehlen,P. *et al*. (2011) Novel roles for slits and netrins: axon guidance cues as anticancer targets? *Nat. Rev. Cancer*, **11**, 188–197.

Nair,N.U. *et al*. (2012) ChIPnorm: a statistical method for normalizing and identifying differential regions in histone modification ChIP-seq libraries. *PLoS One*, **7**, e39573.

Neal,R. (2000) Markov chain sampling methods. *J. Comput. Graph. Stat*., **9**, 249–265.

Newton,M. *et al*. (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.

Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet*., **10**, 669–680.

Rousseau,J. and Mengersen,K. (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Series B*, **73**, 689–710.

Savage,R. *et al*. (2010) Discovering transcriptional modules by Bayesian data integration. *Bioinformatics*, **26**, i158–i167.

Schäfer,M. *et al*. (2009) Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities. *Bioinformatics*, **25**, 3228–3235.

Schäfer,M. *et al*. (2012) Integrative analyses for Omics data: a Bayesian mixture model to assess the concordance of ChIP-chip and ChIP-seq measurements. *J. Environ. Sci. Health A*, **75**, 461–470.

Schenk,T. *et al*. (2012) Inhibition of the LSD1 (KDM1A) demethylase reactivates the all-trans-retinoic acid differentiation pathway in acute myeloid leukemia. *Nat. Med*., **18**, 605–611.

Shao,Z. *et al.* (2012) Manorm: a robust model for quantitative comparison of ChIP-seq data sets. *Genome Biol.*, **13**, R16.
Song,Q. and Smith,A.D. (2011) Identifying dispersed epigenomic domains from ChIP-seq data. *Bioinformatics*, **27**, 870–871.
Taslim,C. *et al.* (2009) Comparative study on ChIP-seq data: normalization and binding pattern characterization. *Bioinformatics*, **25**, 2334–2340.
Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
Trapnell,C. *et al.* (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, **31**, 46–53.
van Wieringen,W. and van de Wiel,M. (2009) Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, **65**, 19–29.
Wei,P. and Pan,W. (2008) Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics*, **24**, 404–411.
Xu,H. *et al.* (2008) An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics*, **24**, 2344–2349.
Zhang,X. *et al.* (2009) Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*. *Genome Biol.*, **10**, R62.
Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-seq (MACS). *Genome Biol.*, **9**, R137.