Published in final edited form as:

Stat Appl Genet Mol Biol. ; 11(3): Article–1. doi:10.1515/1544-6115.1660.

A New Approach for the Joint Analysis of Multiple Chip-Seq Libraries with Application to Histone Modification

John P. Ferguson,
Yale University

Judy H. Cho, and
Yale University

Hongyu Zhao
Yale University

Abstract

Most approaches for analyzing ChIP-Seq data are focused on inferring exact protein binding sites from a single library. However, frequently multiple ChIP-Seq libraries derived from differing cell lines or tissue types from the same individual may be available. In such a situation, a separate analysis for each tissue or cell line may be inefficient. Here, we describe a novel method to analyze such data that intelligently uses the joint information from multiple related ChIP-Seq libraries. We present our method as a two-stage procedure. First, separate single cell line analysis is performed for each cell line. Here, we use a novel mixture regression approach to infer the subset of genes that are most likely to be involved in protein binding in each cell line. In the second step, we combine the separate single cell line analyses using an Empirical Bayes algorithm that implicitly incorporates inter-cell line correlation. We demonstrate the usefulness of our method using both simulated data, as well as real H3K4me3 and H3K27me3 histone methylation libraries.

Keywords

Empirical Bayes; EM; ChIP-Seq; histone methylation

1 Introduction

In recent years, ChIP-Seq has emerged as a popular method to detect protein DNA interactions on a genome-wide scale. As suggested by its name, the procedure involves the successive application of two distinct techniques. The goal of the first phase (or ChIP phase) of the procedure is to isolate fragments of DNA that are associated with DNA-protein interactions. This involves the application of multiple steps. First, the original sample of genomic DNA is cross-linked to the protein of interest and the resulting cross-linked DNA-protein complexes are sheared into small fragments. The next step is known as immunoprecipitation and involves the addition of a specific antibody that is known to bind to the protein under investigation and subsequent purification of the DNA-antibody complexes. Finally the cross links between DNA and proteins are broken (reverse cross-linking) and the DNA is prepared for sequencing (Mardis, 2007). The 'Seq' phase of the procedure is conceptually simpler. It consists of sequencing the purified segments of DNA that resulted from the ChIP step, and the subsequent mapping of the resultant sequence reads to a reference genome. Unfortunately, not all the DNA fragments sequenced will represent genome locations associated with true DNA protein interactions. Even though the antibody may be selected to have high affinity for the protein of interest, it is possible that it will bind to other DNA-binding proteins (in the immunoprecipitation step) and as a result DNA

fragments, associated with non-specific antibody binding, will be sequenced. Even still, one would expect to see an enrichment of sequence reads at sites that associated with interactions of DNA and the protein of interest.

Recently, great attention has been given to ‘peak-calling algorithms’ that attempt to filter out non-specific anti-body binding in ChIP-Seq experiments. Many such algorithms exist for analyzing results from a single sample (see for example Zhang et. al, 2008, Ji et. al, 2008, Valouev et. al., 2008, Zang et. al. 2009 or Rozowsky et. al, 2009). Wilbanks et. al., 2010, gives a comparison of eleven of the most common methods. However, in the situation where libraries exist for multiple related cell lines, these single sample-based algorithms can be sub-optimal since they do not model the inter-cell line correlation that may exist in the binding profiles. Some procedures have been developed that combine two ChipSeq libraries, such as the Hidden Markov Model approach of Xu et. al, 2008, and the non-linear normalization approach of Taslim et. al, 2009, that combine multiple ChIP-ChIP datasets (Wu et. al., 2011), and also that combine ChIP-Seq and ChIP-ChIP data (Choi et. al, 2009). To our knowledge, the only method that proposes to combine more than two ChIP-Seq libraries is the mixture modelling approach of Johannes et. al, 2010. However, only two libraries are compared in the application described in that paper. Here we introduce a novel approach to the problem of combining multiple ChIP-Seq libraries, which probabilistically quantifies the evidence for protein-DNA binding at the gene level, in contrast to the approach taken in Johannes et. al where the binding signals in successive fixed length windows are examined. While considering gene regions instead of fixed window sizes is an unusual feature of our analysis, we found that the correlation between normalized mouse gene expression and log tag-count (from a H3K4me3 histone methylation library) is highest when the log tag-count is calculated over the entire gene region. In addition, our approach allows easy integration of confounding variables such as local GC sequence content, that could bias results if not properly incorporated in an analysis. The probabilistic nature of our method allows easy and natural application of multiple testing corrections such as the false discovery rate. We demonstrate the utility and power of our methodology using both simulated data, and real H3K4me3 (transcriptional activation) and H3K27me3 (transcriptional repression) histone-methylation ChIP-Seq libraries. We have made an R-package, chipem, that implements the methodology we describe in this manuscript, available as part of the Supplementary Material.

2 Methods

Because our method was motivated from the ChIP-Seq analysis of histone methylation data, we will focus on histone methylation libraries in the following discussion. However, the ideas can be applied to infer binding from other types of ChIP-Seq experiments. We assume that ChIP-Seq libraries are available for a number of cell lines of interest. For example, a H3K4me3 library will consist of short sequence reads (commonly referred to as tags) that are associated with genomic positions where a specific antibody binds to H3K4me3 molecules. Although an ideal experiment would only produce tags that are related to binding positions, some reads in any library will inevitably be associated with non-specific antibody binding. These reads do not represent genomic positions where H3K4me3 methylation is found, and can be regarded as noise (Jothi et. al, 2008). For the purpose of inferring binding sites based on ChIP-Seq data, the goal is to distinguish those genes for which there is a combination of specific (signal) and non-specific (noise) antibody binding from those genes where there is only non-specific binding.

We motivate our approach to the above problem by first considering the data from a single ChIP-Seq library. The analysis centers on Y_i that is the log count of the number of tags from this library that are aligned within the region of the i^{th} gene of a set of N genes. We are

interested in the relationships between Y_i , the methylation status of the gene, and other potential covariates. To explore these relationships, consider Figure 1, where Y_i (representing log H3K4me3 tag-count in mouse Naive T cells) is plotted against log gene-length for 21,101 genes. This plot reveals 2 distinct clusters, which likely correspond to the genes being methylated and those not being methylated. Within each cluster, the relationship can be well described via a quadratic model suggesting that separate quadratic regression models can be used to model tag-counts for the 2 clusters.

The local G-C percentage (that is the proportion of guanine and cytosine nucleotides in a particular genomic region) is another variable known to have a significant impact on sequencing read count (Dohm et. al, 2008). For each of the 21,101 genes being surveyed we calculated this proportion for the proximal promoter region – which is defined for a particular gene as the 2000 base pair region centered on the transcription start site. After adjusting for the quadratic relationship with gene length via a regression model, we found that the resulting residuals have an approximately linear relationship with G-C percentage for each of the 2 clusters, as Figure 2 shows for H3K4me3 in Naïve mouse T cells.

2.1 Model and Parameter Estimation When Only a Single Library Is Used

Based on the above observations, we consider the following model to describe the within-cluster relationships between tag-count, log gene length and promoter GC percentage:

$$Y_i = (1 - Z_i)(\beta_0^n + \beta_1^n X_{i1} + \beta_2^n X_{i1}^2) + Z_i(\beta_0^m + \beta_1^m X_{i1} + \beta_2^m X_{i1}^2) + \beta_3 X_{i2} + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. In the above model, Y_i and $Z_i \in \{0,1\}$, represent the log tag-count and the methylation status for gene i , X_{i1} represents the log gene length for gene i , X_{i2} represents GC percentage (defined as above), β_0^n, β_1^n and β_2^n represent the regression coefficients defining the relationship between log gene length and gene-tag-count for the non-methylated cluster, while β_0^m, β_1^m and β_2^m are similarly defined for the methylated cluster. Essentially, we impose two separate quadratic regression models between observed log-count and gene length for the subsets of genes following the two methylation states. A common linear relationship between log GC percentage and log tag-count is assumed for both the methylated and non-methylated subsets of genes, and is represented by the parameter β_3 in the model. Finally,

we define the quantities $\pi_i = I\{Z_i = 1\}$ and $\pi = \frac{1}{N} \sum_{i=1}^N \pi_i$, representing the probability that the i^{th} gene is involved in a methylation event, and the fraction of all genes involved in such events.

2.1.1 Parameter estimation—Our objective is to infer the latent methylation state for each gene, denoted by $Z_i \in \{0,1\}$, from the observed data. This can be accomplished through the EM algorithm (Dempster et. al, 1977). An implementation of the EM algorithm applies two steps iteratively. First, given the current estimates of the parameters $\theta^{old} = (\pi^{old}, \beta_0^{n,old}, \beta_1^{n,old}, \beta_2^{n,old}, \beta_0^{m,old}, \beta_1^{m,old}, \beta_2^{m,old}, \beta_3^{old}, \sigma^{old})$, we compute the conditional expectation of the ‘complete-data’ log likelihood:

$$\begin{aligned}
& E\left(\sum_{i=1}^N Z_i \log \pi\right. \\
& \quad + \frac{Z_i}{2\sigma^2} (Y_i - \beta_0^m - \beta_1^m X_{i1} - \beta_2^m X_{i1}^2 - \beta_3 X_{i2})^2 | \theta \\
& \quad = \theta^{old}, Y_1, \dots, Y_N) \\
& \quad + E\left(\sum_{i=1}^N (1 \right. \\
& \quad \quad - Z_i) \log(1 - \pi) + \frac{(1 - Z_i)}{2\sigma^2} (Y_i - \beta_0^n - \beta_1^n X_{i1} - \beta_2^n X_{i1}^2 - \beta_3 X_{i2})^2 | \theta \\
& \quad \quad = \theta^{old}, Y_1, \dots, Y_N) \\
& \quad - E\left(\frac{n}{2} \log(\sigma^2)\right) \\
& \quad + C | \theta \\
& \quad = \theta^{old}, Y_1, \dots, Y_N)
\end{aligned} \tag{1}$$

where C is a constant, not depending on θ or $\{Z_1, \dots, Z_N\}$. The quantities

$\pi_i^{new} = P\{Z_i = 1 | \theta^{old}, Y_i = y_i\}$ which equal

$$\begin{aligned}
& \text{where } P_i^m(\theta_{old}) = \pi^{old} \varphi\left(\frac{\frac{P_i^m(\theta_{old})}{P_i^m(\theta_{old}) + P_i^n(\theta_{old})}}{\frac{Y_i - \beta_0^{m,old} - \beta_1^{m,old} X_{i1} - \beta_2^{m,old} X_{i1}^2 - \beta_3^{old} X_{i2}}{\sigma^{old}}}\right) \text{ and} \\
& P_i^n(\theta_{old}) = (1 - \pi^{old}) \varphi\left(\frac{Y_i - \beta_0^{n,old} - \beta_1^{n,old} X_{i1} - \beta_2^{n,old} X_{i1}^2 - \beta_3^{old} X_{i2}}{\sigma_{old}}\right),
\end{aligned} \tag{2}$$

need to be calculated to evaluate (1) (Here φ represents the probability density function for the standard normal distribution). Having performed this computation, we can separately maximize (1) over θ and σ^2 . First, maximizing over θ entails first calculating:

$$\sum_{i=1}^N \frac{\pi_i^{new}}{2\sigma^2} (Y_i - \beta_0^m - \beta_1^m X_{i1} - \beta_2^m X_{i1}^2 - \beta_3 X_{i2})^2 - \sum_{i=1}^N \frac{(1 - \pi_i^{new})}{2\sigma^2} (Y_i - \beta_0^n - \beta_1^n X_{i1} - \beta_2^n X_{i1}^2 - \beta_3 X_{i2})^2 - \frac{n}{2} \ln(\sigma^2) \tag{3}$$

The evaluation of (3) can be thought of as the E-step to the algorithm. The M-step requires finding the values of $(\beta_0^n, \beta_1^n, \beta_2^n, \beta_0^m, \beta_1^m, \beta_2^m, \beta_3, \sigma)$ that maximize (3) with π_i^{new} considered fixed. This can be achieved by identifying (3) with a weighted least squares problem for an ‘augmented’ set of response variables $\mathbf{Y}^{aug} = (Y_1, \dots, Y_N, Y_1, \dots, Y_N)$ and an augmented design matrix $\mathbf{X}^{aug}$ having columns:

$(\mathbf{1}_N, \mathbf{0}_N), (\mathbf{0}_N, \mathbf{1}_N), (\mathbf{0}, X_{11}, \dots, X_{N1}), (\mathbf{0}, X_{11}^2, \dots, X_{N1}^2), (X_{11}, \dots, X_{N1}, \mathbf{0}_N)$ and $(P_1, \dots, P_N, P_1, \dots, P_N)$. However, in application we found that it was necessary to slightly modify the M step to ensure that the two regression lines did not ‘cross over’ for large values of gene length. This was implemented by treating the M step as a quadratic programming problem, with (3) being the criterion function with the linear inequality constraints ensuring that for all possible values of G_i :

$$\widehat{Y}_m = \beta_0^m + \beta_1^m X_{i1} + \beta_2^m X_{i1}^2 + \beta_3 x_{i2} > \widehat{Y}_n = \beta_0^n + \beta_1^n X_{i1} + \beta_2^n X_{i1}^2 + \beta_3 X_{i2}.$$

Let us denote the maximizing values as

$$(\beta_0^{n,new}, \beta_1^{n,new}, \beta_2^{n,new}, \beta_0^{m,new}, \beta_1^{m,new}, \beta_2^{m,new}, \beta_3^{n,new}, \sigma^{new}).$$

Finally we set $\pi^{new} = \frac{1}{N} \sum_{i=1}^N \pi_i^{new}$. The E and M steps described above are iterated until the parameter estimates converge. After convergence, posterior probabilities for each gene being in the methylated cluster, given a tag-count of Y_i can be calculated using

$$P\{Z_i=1|\theta\} = \frac{P_i^m(\theta)}{P_i^m(\theta) + P_i^n(\theta)},$$

where $P_i^m(\theta) = \pi \varphi\left(\frac{Y_i - \hat{\beta}_0^m - \hat{\beta}_1^m X_{i1} - \hat{\beta}_2^m X_{i1}^2 - \hat{\beta}_3^m X_{i2}}{\hat{\sigma}}\right)$ (4)

and $P_i^n(\theta) = (1 - \pi) \varphi\left(\frac{Y_i - \hat{\beta}_0^n - \hat{\beta}_1^n X_{i1} - \hat{\beta}_2^n X_{i1}^2 - \hat{\beta}_3^n X_{i2}}{\hat{\sigma}}\right).$

2.2 Multiple Library Model

Estimated probabilities of the 2 cluster memberships are by-products of this EM algorithm from (2). However, only 1 cell line is taken into account when forming these estimates. In a situation where ChIP-Seq libraries exist for multiple cell lines, strong correlation may exist between the binding status, and as a result the observed log tag-counts, for the various cell lines. For example, Table 1 shows very strong inter-cell line correlations in gene level log tag-count data for the 6 T-cell K4 histone methylation libraries analyzed by Wei et. al, 2009.

In general, suppose that we have K libraries (representing the detection of the same protein in K different tissues or cell lines), and Y_{ik} represents the tag-count in gene region i for tissue k . It is clear that $\{Y_{ij}\}_{j=1}^K$ may contain additional information (even after considering Y_{ik}) about the likelihood of true binding in gene i for cell line k . In this situation, only using the tag-counts for a single library, when estimating the binding probabilities $I_1, \dots, I_K$ is inefficient. As an alternative, we adapt the model considered by Datta and Zhao, 2008, to describe the joint binding patterns of transcription factors for a particular gene.

2.2.1 Model formulation—We assume that ChIP-Seq libraries exist for K cell lines. For a particular gene, i , consider the random vector $(Z_{i1}, \dots, Z_{iK})$, where $Z_{ik} = 1$ if there is a binding event in the region of the gene for cell line k , and 0 otherwise. Consider the proportion,

$I_1, \dots, I_K$ of all genes that show a $(I_1, \dots, I_K)$ pattern of binding over the K cell lines. Here each $I_k \in \{0, 1\}$ ($1 \leq k \leq K$), where I_k indicates the event of binding $\{I_k = 1\}$, or no binding $\{I_k = 0\}$, in cell line k . These proportions can be used as prior probabilities of a particular gene having a certain binding pattern. Let us write $f_{I_1, \dots, I_K}(\cdot | X)$ to denote the density function for the log tag-counts, $Y_{i1}, \dots, Y_{iK}$, given that the true binding pattern for gene i is $Z_{i1} = I_1, \dots, Z_{iK} = I_K$ and covariates X_i (Recall the covariates we consider are gene length and GC percentage). It follows that a posterior probability can be constructed as:

$$\pi_{i, I_1, \dots, I_K} = P\{Z_{i1} = I_1, \dots, Z_{iK} = I_K | Y_{i1}, \dots, Y_{iK}\} = \frac{\pi_{I_1, \dots, I_K} f_{I_1, \dots, I_K}(Y_{i1}, \dots, Y_{iK} | X_i)}{\sum_{\mathbf{a} \in \{0, 1\}^K} \pi_{\mathbf{a}} f_{\mathbf{a}}(Y_{i1}, \dots, Y_{iK} | X_i)}. \quad (5)$$

To find the posterior probability of binding in cell line k , we simply calculate the marginal of this joint probability. In the following section, we describe an empirical Bayes algorithm

to estimate $I_1, \dots, I_K$. Once estimates for the global prior probabilities are obtained, (5) can be used to generate the posterior probabilities for each configuration for each gene. We now describe the algorithm we use in more detail.

2.2.2 Initial estimate of $\pi_{I_1, \dots, I_K}$ —Recall that the per-cell line analysis from the initial EM algorithm (described in Section 2.1) generates an estimated probability, $\widehat{\pi}_{ik}$, that gene i is methylated in cell line k . This gives an estimate for the probability that gene i has

configuration $(I_1, \dots, I_K)$ as $\widehat{\pi}_{i,I_1, \dots, I_K}^{(0)} = \prod_{k=1}^K \widehat{\pi}_{i,k}^{I_k} (1 - \widehat{\pi}_{i,k})^{1-I_k}$ based upon the naïve assumption of independence of the event of methylation in each cell line. From this, an initial estimate of the proportion of genes that have configuration $(I_1, \dots, I_K)$ is

$$\widehat{\pi}_{I_1, \dots, I_K}^{(0)} = \frac{1}{N} \sum_{i \leq N} \widehat{\pi}_{i,I_1, \dots, I_K}^{(0)}.$$

To estimate the density $f_{I_1, \dots, I_K}(\cdot | X_i)$ we assume that given the true binding pattern $(I_1, \dots, I_K)$, the tag-counts for a given gene in the K cell lines are independent, so that f can be factorized as a product of its marginal densities:

$$f_{I_1, \dots, I_K}(\cdot | X_i) = \prod_{k=1}^K f_{k, I_k}(Y_k | X_i).$$

The marginal densities for cell line k , $f_{k,0}$ and $f_{k,1}$ are estimated via the initial EM algorithm described above (recall this algorithm estimates 2 regression lines and hence 2 corresponding densities for each cell line). Note that different intercept terms are estimated in these regression models for each of these K libraries. In this way, the method automatically adjusts for differing sequencing depths across the K libraries.

2.2.3 Iterative scheme to update the estimates for $\pi_{I_1, \dots, I_K}$ —To update the estimates we apply two steps iteratively:

I Update posterior probabilities (on iterative step $t+1$): Given the current estimates,

$\widehat{\pi}_{I_1, \dots, I_K}^{(t)}$ for $\pi_{I_1, \dots, I_K}$, we use formula (5), to get estimates for $\pi_{i,I_1, \dots, I_K}^{(t+1)}$ via

$$\pi_{i,I_1, \dots, I_K}^{(t+1)} = P^{(t)}\{Z_{i1}=I_1, \dots, Z_{iK}=I_K | Y_{i1}, \dots, Y_{iK}, X_i\} = \frac{\pi_{I_1, \dots, I_K}^{(t)} \widehat{f}_{I_1, \dots, I_K}(Y_{i1}, \dots, Y_{iK} | X_i)}{\sum_{\mathbf{a} \in \{0,1\}^K} \pi_{a_1, \dots, a_K}^{(t)} \widehat{f}_{a_1, \dots, a_K}(Y_{i1}, \dots, Y_{iK} | X_i)}$$

II Update estimates of $I_1, \dots, I_K$: First we calculate an average of the posterior probabilities

$\pi_{I_1, \dots, I_K}^{(t+1)} = \frac{1}{N} \sum_{i=1}^N \pi_{i,I_1, \dots, I_K}^{(t+1)}$. Consider now the 2^K contingency table, with $(I_1 \dots I_K)$ entry equal to $\sum_{i=1}^N \pi_{i,I_1, \dots, I_K}^{(t+1)}$. This contingency table counts the expected number of genes that have each

possible binding configuration according to the probabilities $\{\pi_{i,I_1,\dots,I_K}^{(t+1)}\}_{i \leq N, (I_1,\dots,I_K) \in \{0,1\}^K}$. $\pi_{I_1,\dots,I_K}^{*(t+1)}$ is the estimated probability for cell $(I_1,\dots,I_K)$ according to the saturated log-linear model for this 2^K table. However, if a conditional independence structure exists between the events $\{Z_{i1}=I\}, \dots, \{Z_{iK}=I\}$, the estimates for these probabilities can be improved via fitting a reduced log-linear model. We apply a forward selection algorithm, in an attempt to find the log-linear model having the lowest BIC. The updated probabilities $\pi_{I_1,\dots,I_K}^{(t+1)}$ are then estimated via this log linear model.

Note that if we used the saturated log linear model (instead of estimating $I_1,\dots,I_K$ using the log linear model with minimum BIC), steps **I** and **II** could be viewed as the E and the M steps of an EM algorithm approach to estimate the configuration probabilities, $\{I_1,\dots,I_K\} \in \{0,1\}^K$. The selection of the model with the lowest BIC implies a modification of the M step. However, we expect model selection to help when estimating the probabilities for the less frequent configurations, since the estimates from the saturated model may have high variances if no model structure is assumed.

2.2.4 Stopping rule—The probabilities $I_1,\dots,I_K$ are iteratively re-estimated until the difference in ratio between the estimates from two successive steps is below a cut-off threshold. Let us denote these final estimated probabilities as $I_1,\dots,I_K$. This implies that the

algorithm will terminate if $\pi_{I_1,\dots,I_K}^{(t)} \sim \pi_{I_1,\dots,I_K}^{(t+1)} \sim \frac{1}{N} \sum_{i=1}^N \pi_{i,I_1,\dots,I_K}^{(t)}$. More details about the stopping rule for both single and multiple cell line algorithms are given in Appendix A of the Supplementary material.

3 Results

3.1 Data

We apply the methodology developed to H3K4me3 and H3K27me3 ChIP-Seq libraries for 6 T-cell subsets (Naïve, Th1, Th2, Th17, iTreg and nTreg) in mouse. This data set was previously collected and analyzed in Wei et al. The mm8 assembly for the mouse genome was downloaded from (<http://genome.ucsc.edu/>) and was used to annotate gene regions. Again, following the analysis in Wei et al., we computed the count of all mapped ChIP-Seq reads that lie in each gene region (defined as starting at 1000 base pairs upstream of the annotated transcription start site and ending at the transcription end site), counting reads that map to a unique location only once. These counts were then used to infer the binding patterns over the 6 cell lines by the method that we outline in Section 2.

3.2 Single Cell Line Based Analysis

Having calculated log tag-count scores for all 21,101 genes and 12 ChIP-Seq libraries, we fitted the two (cluster - specific) regression lines via the EM algorithm as described in Section 2. In Figures 3a. and 3b. (Appendix B, Supplementary material), we show the resulting plots for each of the libraries. The clustering of the genes is certainly more apparent in the K4 cell lines than in the K27 lines. This is to be expected because it is known that K4 modifications tend to happen in gene promoter regions, while K27 modifications are more evenly spread across the gene region and can be intergenic (see Wei et al, 2009). This more localized pattern for the K4 modification is reflected in the more defined clusters found from the K4 libraries.

3.3 Multiple Cell Line Analysis

As described in the Section 2, we apply an additional iterative algorithm (separately for the K4 and K27 libraries) which uses the correlation in tag-counts between the various libraries to better predict the joint modification pattern over all cell lines. Interestingly, and perhaps not surprisingly in light of the similarity of the cell lines, the majority of genes, 97.1% for K4 cell lines and 94.7% for K27 cell lines are predicted to be methylated over all 6 cell lines, or are predicted to show no modifications over any cell line. Tables 2 and 3 display the patterns that are most common for both K4 and K27 libraries. The third most common pattern is consistent for the K4 and K27 cell lines and seems to represent genes which are expressed in Naïve and nTreg cell lines, but not in the other cell lines.

One way that we can test the validity of these inferred patterns is by examining ‘key cytokine genes’ that should be expressed in some but not all of the cell lines. For instance, the gene *IFNG* represents the inflammatory cytokine Interferon Gamma that should be expressed only in Th1 cell lines and not in the other cell lines. Since K4 methylation is associated with gene expression, we would expect to see a K4 signal in this cell line alone. Table 3 lists the modification patterns predicted for some of these genes in both the cases when cell line sharing is and is not used. The only surprise below is perhaps the predicted absence of any methylation modifications for the *IL17* gene. Immune system theory would suggest that there should be a K4 modification for *IL17* in Th17 cells and there is evidence in the data that there is a K4 modification for *IL17* genes. For instance, the estimated probability of a K4 modification in Th17 cells (from the initial EM algorithm) is 0.9987.

However, while the prior probability of the pattern (0, 0, 0, 0, 0, 0) (no modifications for any of the 6 cell lines) is 0.497, the probability of the pattern (0, 0, 0, 1, 0, 0) (which represents a modification only in the Th17 cell line) is only 0.00002 (or in other words the model expects 4 genes to show such a pattern of methylation across the 6 cell lines). Since this pattern is so unlikely, exceptional evidence in favor of a sole modification in the Th17 cell line is required to conclude that this is the most likely pattern. While for the *IL17* gene, this prediction may seem counter-intuitive, this strategy requiring greater evidence to detect less likely patterns protects against false positives and, as we shall show later in our simulation study minimizes the overall classification error.

3.4 GO Enrichment Analysis

As an additional practical validation, we looked for biological process GO groups that were most associated with certain interesting gene sets chosen via our algorithm using DAVID bioinformatics software (Dennis et al., 2003 and Huang et al., 2009). The gene sets we investigated for enrichment were comprised of genes for which the algorithm predicts a true histone modification in only one of the 6 cell lines. The most associated GO group in each case should be related to the immune system. We report these GO groups in Table 4 below. Note that there were no genes showing a cross cell-line pattern of modification only in the Naïve cell line, or only in the nTreg cell lines, and only 1 showing modification only in the Th17 cell line so these cell lines have been removed from the table.

3.5 Simulation Study

In order to demonstrate the improvement gained through ‘sharing’ the inter-cell line information, we performed simulations to assess the prediction error when either single libraries are analyzed separately or multiple libraries are jointly analyzed. More specifically, for each simulation and for each of the 6 K4 libraries, we simulate the tag-count for the 21,101 transcripts considered in our original model. The 6 log tag-counts for a particular gene, are drawn assuming the true methylation configuration for the gene is the most likely posterior configuration according to our model. Suppose then, for a particular gene this

pattern is (0, 1, 0, 0, 0, 0) indicating that there is a histone modification associated with the gene only in the Th1 cell line but not in the other cell lines. We simulate the log tag-counts for this gene, using the 'upper' regression line (which relates expected log tag-count to GC percentage and gene length for the subset of methylated genes) for the Th1 cell line and the 'lower' regression line (relating expected log tag-count to GC percentage and gene length for the subset of un-methylated genes) for the other 5 cell lines.

Having simulated the data set, we re-estimate the regression lines via the EM algorithm. For each gene, the marginal probabilities that are calculated in the initial EM algorithm are used to estimate probabilities for the joint methylation configuration by taking their product. We then select the configuration with the largest such product probability as the 'predicted configuration'. In addition, for each gene we re-estimate the methylation configuration from the simulated data using the iterative algorithm (Section 2.2) that shares information across the 6 cell lines. 95% confidence bounds for the error rate, that is the expected proportion of genes which are assigned incorrect methylation configurations, indicate that the multiple cell line algorithm (CI: 0.0133 ± 0.0002) is performing better than the single cell line algorithm (CI: 0.456 ± 0.009). The 95% confidence bounds are calculated using a standard error estimated from 100 simulations. Figure 3 displays a ROC curve that also suggests that the Multi-Cell line method has superior performance regardless of how we estimate the configuration (see the caption above the figure for more information). Note that the good classification error of the multi-cell line algorithm is driven by the high specificity. In other words, because of the predominance of the configurations: (0,0,0,0,0,0) and (1,1,1,1,1,1), the algorithm has highest power to detect these patterns.

4 Discussion

An important property of our single cell line analysis is that the predictions for methylation are made at the transcript or gene level. The total tag-count for each gene under investigation needs to be supplied as input to the algorithm. However, the initial input to the algorithm could be tag-counts over other regions of DNA. For instance, we could have instead used tag-counts over the promoter region when looking at H3K4me3 methylation. In fact, it is well known that H3K4me3 signals appear predominantly in gene promoter region. This suggests that we should define our 'methylation score' for a particular gene via the tag-count over the gene's promoter region, rather than over the entire gene region. However, we found that the correlation between gene expression and this tag-count score is highest when the methylation score is calculated over the entire gene region which is why we define our score this way.

For each gene, the multiple cell line algorithm reports posterior probabilities for all possible binding configurations across the analyzed cell lines. This is done via an Empirical Bayes approach, where the global prior probabilities $j_1 \dots j_K$ are iteratively estimated using all genes and all cell lines. This 'sharing' of information alleviates the need for multiple testing corrections. For example, if we desire to detect genes which have histone modifications in the Th1 cell line, the false discovery rate among the subset of genes having a predicted marginal probability of methylation of at least 0.8, should be less than 20%. To see this, suppose we identify all genes (indexed by g), such that the marginal probability of methylation is greater than q , as methylated. Let's say there are n_q such genes. The Posterior expectation of the false discovery rate is then given as:

$$\begin{aligned}
& n_q E(FDR_q | \mathbf{Y}, \mathbf{X}) \\
&= \sum_{g: P\{g \text{ methylated} | \mathbf{Y}, \mathbf{X}\} > q} E(I\{\text{g not methylated}\} | \mathbf{Y}, \mathbf{X}) = \sum_{g: P\{g \text{ methylated} | \mathbf{Y}, \mathbf{X}\} > q} P(I\{\text{g not methylated}\} | \mathbf{Y}, \mathbf{X}) \leq \sum_{g: P\{g \text{ methylated} | \mathbf{Y}, \mathbf{X}\} > q} \\
& - q = n_q(1 - q)
\end{aligned}$$

Most ChIP-Seq software packages allow the user to incorporate a control library into the analysis. If such control data is available, our method can easily be modified to accommodate the extra information. For example, suppose that the total tag-counts are T and C for the ‘test’ and ‘control’ library, for which T_i and C_i reads align to the region of gene i . One can then apply the techniques that we have developed to the transformed tag-counts,

$\tilde{T}_i = T_i - \frac{T}{C} C_i + h$, where h is a constant added so that the transformed count $\tilde{T}_i$ is larger than 0 for every gene. Incorporating a control library in this way should significantly reduce the number of falsely detected binding events.

5 Conclusions

In this paper, we have proposed a new method for analyzing ChIP-Seq data over multiple cell lines tissues or even treatments, and demonstrated its application with real histone modification data. As shown in the Section 3, analyzing the libraries jointly increases both the sensitivity and specificity of finding particular joint binding patterns over the set of cell lines. An additional more subtle advantage of a joint analysis is that it leads to more accurate single cell line inference. For instance, a ‘marginal’ probability for there being a binding event in a particular gene in a particular cell line can be calculated by marginalizing the joint distribution (which is the collection of probabilities for all possible binding patterns for that gene) over all the other cell lines. Calculating marginal cell line probabilities of binding in this way naturally incorporates correlation over the different cell lines and so is superior to a corresponding probability calculated by analyzing each cell line in isolation.

In addition to demonstrating the increased power of the joint approach, we demonstrate the application of our method using real H3K4me3 and H3K27me3 histone methylation libraries. The sets of genes predicted to show certain binding patterns make real biological sense in that they are associated with the pathways we expect. As a secondary validation step, we examine how our model classifies the binding pattern for key-cytokine genes where we have a-priori knowledge for the expected binding pattern. Our model classifies most of these genes correctly, but has a tendency to ‘shrink’ towards the more prevalent binding patterns compared to separate single-cell line analyses. As we explain in Section 3, this is actually a desirable feature of any Empirical or Hierarchical Bayes approach and helps our algorithm achieve better sensitivity and specificity compared to separate single cell line analyses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Author Notes: This work was supported in part by the following grants and granting agencies: NIH R01 GM59507, NSF DMS 1106738, U01 DK062422, RC1DK086800, and CCFA. We used the computational facilities at the Yale University Biomedical High Performance Computing Center for our analyses. In addition, we would like to thank two anonymous reviewers whose comments have greatly improved the quality of the manuscript.

References

- Choi A, Nesvizhskii I, Ghosh D, Qin ZS. Hierarchical hidden Markov model with application to joint analysis of ChIP-chip and ChIP-Seq data. *Bioinformatics*. 2009; 25(no. 14):1715–1721. [PubMed: 19447789]
- Datta D, Zhao H. Statistical methods to infer cooperative binding among transcription factors in *Saccharomyces cerevisiae*. *Bioinformatics*. 2008; 24(no. 4):545. [PubMed: 17989095]
- Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*. 1977; 39(no. 1):1–38.
- Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*. 2003; 4(5):P3. [PubMed: 12734009]
- Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*. 2008; 36(no. 16):e105. [PubMed: 18660515]
- Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc*. 2009; 4(1):44–57. [PubMed: 19131956]
- Ji H, Jiang H, Ma W, Johnson D, Myers R, et al. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature Biotechnol*. 2009; 26:1293–1300. [PubMed: 18978777]
- Johannes F, Wardenaar R, Colome-Tatche M, Mousson FI, De Graaf P, Morky MI, Guryev V, Timmers H, Cuppen E, Jansen R. Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-Seq. *Bioinformatics (Oxford, England)*. 2010; 26(no. 8):1000–1006.
- Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucl Acids Res*. Sep 1; 2008 36(no. 16):5221–5231. [PubMed: 18684996]
- Mardis ER. ChIP-Seq: welcome to the new frontier. *Nat Meth*. 2008; 4(no.8):613–614.
- Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. PeakSeq enables systematic scoring of ChIP-Seq experiments relative to controls. *Nature Biotechnology*. 2009; 27:66–75.
- Taslim C, et al. Comparative study on ChIP-Seq data: normalization and binding pattern characterization. *Bioinformatics (Oxford, England)*. 2009; 25(no. 18):2334–2340.
- Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods*. 2008; 5:829–834. [PubMed: 19160518]
- Wei G, Wei L, Zhu J, Zang C, Hu-Li J, Yao Z, Cui K, et al. Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity*. 2008; 30(no. 1):155–167. [PubMed: 19144320]
- Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in ChIP-Seq peak detection. *PLoS One*. 2010; 5:e11471. [PubMed: 20628599]
- Wu H, Ji H. JAMIE: joint analysis of multiple ChIP-chip experiments. *Bioinformatics*. 2010; 26:1864–1870. [PubMed: 20551135]
- Xu H, Wei C, Lin F, Sung W. A HMM approach to genome-wide identification of differential histone modification sites from ChIP-Seq data. *Bioinformatics*. 2008; 24(no. 20):2344–2349. [PubMed: 18667444]
- Zang C, Schones DE, Zeng C, Cui K, Zhao K, Wei Qun P. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*. 2009; 25(no. 15): 1952–1958. [PubMed: 19505939]
- Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, Bernstein B, Nussbaum C, et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*. 2008; 9(no.9):R137. [PubMed: 18798982]

Figure 1. Scatter Plot of Gene Length vs Tag-Count

A scatter plot of gene length vs the total H3K4me3 tag-count for each gene (on the log scale) for Naïve mouse T cells. 2 clusters are clearly identifiable. These 2 clusters most likely represent genes that have true H3K4me3 binding, and those that do not. Points are colored either light or dark grey according to whether the probability calculated in equation (4) is above 0.5 or not.

Figure 2. Scatter Plot of GC Percentage vs Adjusted Tag-Count

A scatter plot showing the linear relationship between GC percentage and adjusted log tag-count for a H3K4me3 library in Naive mouse T-cells. Separate quadratic regressions were fit to the set of genes with a posterior probability of methylation greater than 0.5, and to the set of genes with a posterior probability of methylation less than 0.5. The vector of adjusted log tagcounts is the concatenation of the residuals from these two regressions.

Figure 3. ROC Curve Comparing the Single and Multi-Cell Line Algorithm

Here sensitivity is defined as the proportion of configurations $(I_1 \dots I_6)$ where $I_l = I_k$ for some l and k that are correctly identified, and specificity as the proportion of the configurations $(I_1 \dots I_6)$ having $I_l = I_k$ for all $l, k \in \{1, \dots, 6\}$ that are correctly identified. Each point on the ROC curve corresponds to a differing threshold, c , for concluding that $I_l = I_k$ for l, k based on the posterior probability $P\{I_l = I_k \mid \mathbf{Y}, \mathbf{X}\}$. 95% confidence point-wise confidence bounds for sensitivity (given the estimated specificity) are also displayed.

Table 1

Inter cell line correlation in gene region log tag-counts

	K4naive	K4Th1	K4Th2	K4Th17	K4iTreg	K4nTreg
K4naive	1	0.924	0.919	0.942	0.891	0.941
K4Th1	0.924	1	0.966	0.961	0.956	0.911
K4Th2	0.919	0.966	1	0.953	0.945	0.926
K4Th17	0.942	0.961	0.953	1	0.938	0.912
K4iTreg	0.891	0.956	0.945	0.938	1	0.885
K4nTreg	0.941	0.911	0.926	0.912	0.885	1

Table 2

Common K4 modification patterns over 6 CD4+ T cell subtypes

Here the first 6 columns specify a pattern of H3K4me3 histone modifications across CD4+ T cell lines, and the final column the number of genes that are predicted to show such a pattern.

K4naive	K4Th1	K4Th2	K4Th17	K4Treg	K4nTreg	Counts
0	0	0	0	0	0	9985
1	1	1	1	1	1	9536
0	1	1	1	1	0	336
0	0	0	0	1	0	40
0	1	0	0	0	0	29
0	1	1	1	1	1	29
1	1	1	1	1	0	27
0	0	0	1	1	0	18
0	1	1	1	0	0	18
0	1	1	0	0	0	16

Table 3

Common K27 modification patterns

The first 6 columns specify a pattern of H3K27me3 Histone modification across 6 CD4+ T cell lines, and the final column the number of genes that are predicted to show such a pattern.

K27naive	K27Th1	K27Th2	K27Th17	K27iTreg	K27nTreg	Counts
0	0	0	0	0	0	14136
1	1	1	1	1	1	4903
1	0	0	0	0	1	364
1	1	1	1	0	1	151
0	1	1	1	0	0	111
1	1	1	1	0	0	105
1	0	1	1	1	1	83
1	0	0	0	1	1	62
0	0	1	1	1	1	47
0	0	1	1	1	0	29

Table 4

K4 methylation patterns inferred from joint analysis

For each of 8 key immune system genes, we give the predicted binding patterns over the 6 cell lines, from analyzing each cell line independently (labelled by A) and from jointly analyzing all cell lines together (labelled by B).

	K4naive	K4Th1	K4Th2	K4Th17	K4iTreg	K4nTreg
<i>Ifng A</i>	0	1	0	0	0	0
<i>Ifng B</i>	0	1	0	0	0	0
<i>Il4 A</i>	0	0	1	0	0	0
<i>Il4 B</i>	0	0	1	0	0	0
<i>Foxp3 A</i>	1	0	0	0	1	1
<i>Foxp3 B</i>	1	0	0	0	1	1
<i>Il17 A</i>	0	0	0	0	0	0
<i>Il17 B</i>	0	0	0	1	0	0
<i>Il21 A</i>	0	0	0	1	0	0
<i>Il21 B</i>	0	0	0	1	0	0
<i>Il10 A</i>	0	1	1	0	0	0
<i>Il10 B</i>	0	1	1	0	0	0
<i>Rorc A</i>	0	0	0	1	1	0
<i>Rorc B</i>	0	0	0	1	1	0
<i>Tbx21 A</i>	1	1	1	1	1	1
<i>Tbx21 B</i>	0	1	0	0	1	0
<i>Gata3 A</i>	1	1	1	1	1	1
<i>Gata3 B</i>	1	1	1	1	1	1

Table 5
Go groups associated with gene subsets found by algorithm

Unique Cell line	Number genes	Most Associated biological process GO group
Th1	18	positive regulation of cytokine production
Th2	7	Immune Response
iTreg	27	Immune Response