Bioinformatics, 31(12), 2015, 1889–1896
doi: 10.1093/bioinformatics/btv094
Advance Access Publication Date: 13 February 2015
Original Paper

OXFORD

Genome analysis

A novel statistical method for quantitative comparison of multiple ChIP-seq datasets

Li Chen[1], Chi Wang[2], Zhaohui S. Qin[3,4] and Hao Wu[3,*]

[1]Department of Mathematics and Computer Science, Atlanta, GA 30322, USA, [2]Department of Biostatistics and Markey Cancer Center, University of Kentucky, Lexington, KY 40536, USA, [3]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA and [4]Department of Biomedical Informatics, Emory University, Atlanta, GA 30322, USA

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on November 12, 2014; revised on January 26, 2015; accepted on February 10, 2015

Abstract

Motivation: ChIP-seq is a powerful technology to measure the protein binding or histone modification strength in the whole genome scale. Although there are a number of methods available for single ChIP-seq data analysis (e.g. 'peak detection'), rigorous statistical method for quantitative comparison of multiple ChIP-seq datasets with the considerations of data from control experiment, signal to noise ratios, biological variations and multiple-factor experimental designs is underdeveloped.

Results: In this work, we develop a statistical method to perform quantitative comparison of multiple ChIP-seq datasets and detect genomic regions showing differential protein binding or histone modification. We first detect peaks from all datasets and then union them to form a single set of candidate regions. The read counts from IP experiment at the candidate regions are assumed to follow Poisson distribution. The underlying Poisson rates are modeled as an experiment-specific function of artifacts and biological signals. We then obtain the estimated biological signals and compare them through the hypothesis testing procedure in a linear model framework. Simulations and real data analyses demonstrate that the proposed method provides more accurate and robust results compared with existing ones.

Availability and implementation: An R software package ChIPComp is freely available at http://web1.sph.emory.edu/users/hwu30/software/ChIPComp.html.

Contact: hao.wu@emory.edu

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

1 Introduction

Coupling chromatin immunoprecipitation (ChIP) and next-generation sequencing (seq), ChIP-seq is a powerful technology for profiling protein bindings or histone modifications in the whole genome scale. Since the introduction of the technology (Johnson *et al.*, 2007), a large number of experiments were conducted to create genome-wide profiles for many DNA-binding proteins and different types of histone modifications under various biological contexts, for example, by large national consortiums such as ENCODE (Thomas *et al.*, 2006) and modENCOD (Celniker *et al.*, 2009).

The main goal of analyzing data from a single ChIP-seq experiment is to detect protein binding or histone modification regions, often referred to as 'peaks'. The raw data produced from ChIP-seq experiments are many short DNA segments called 'reads'. After aligning the reads to the reference genome, genomic regions with unusually large number of reads clustered are often deemed peaks. In recent years, a number of methods and software tools are developed for peak detection. Two benchmark studies have also been conducted to compare different peak calling methods (Laajala *et al.*, 2009; Wilbanks and Facciotti, 2010). With the continuous reduction of sequencing costs and the rapid accumulation of public data, it is

now a common practice to compare data from different ChIP-seq experiments, for example, to compare the binding of certain protein under different biological conditions. Such analysis provides important information for studying the dynamics of epigenetic regulations. Results from the analysis can be further associated with other data, such as gene expressions, to better understand the gene regulation mechanisms.

The comparisons of ChIP-seq data have been widely performed. The most straightforward method is the 'overlapping analysis', which is to compare the peaks called from different experiments and defines 'common peaks' or 'unique peaks', then represents them by Venn diagram (Chen *et al.*, 2008). This method, however, is highly dependent on the thresholds used for calling peaks. Genomic regions barely over the threshold in one sample but under the threshold in the other will be declared as unique peaks even if the quantitative difference is small. Moreover, it completely ignores the quantitative differences of peaks, that is, genomic regions being peaks in both samples will be deemed common peaks, even if the quantitative difference is large. Due to these reasons, quantitative comparison is more desirable to compare ChIP-seq datasets.

The quantitative comparison of ChIP-seq can be performed by comparing the read counts among different experiments, which is similar to RNA-seq differential expression (DE) analysis. However, it is a more complicated problem due to several reasons. First, the data from the IP experiments are affected by the genomic background, such as chromatin structures and DNA sequence. These backgrounds are non-uniform across the genome, and could be highly variable across different experiments. The backgrounds, measured by control experiments, need to be taken into account in quantitative comparison of multiple ChIP-seq datasets. Another complication arises from the different signal to noise ratios (SNRs) of the experiments. Many technical or biological artifacts contribute to SNRs. For example, sample with less binding sites will have taller peaks because reads are allocated into narrower genomic regions. Moreover, different SNRs may result from differences of antibody qualities, experimental protocols or lab technician skills, etc. Therefore, correctly accounting for SNRs is important in quantitative comparison of ChIP-seq. In addition, considerations for biological variance and experimental designs remain, similar to that in DE analysis of RNA-seq.

Quantitative comparison of ChIP-seq (often referred to as 'differential binding' problem) has gained some interests recently, and several methods have been proposed for two-condition comparison. There are two methods take the approach to model the differences of normalized read counts from two IP experiments: ChIPDiff (Xu *et al.*, 2008) applies hidden Markov model on the differences to identify differential histone modification regions, and DIME (Taslim *et al.*, 2009, 2011) uses a finite exponential-normal mixture model on the differences to detect differential binding sites. However, neither the control experiment nor the biological replicates are considered in these methods. Moreover, these methods do not account for SNRs and cannot be easily extended for multiple condition comparison. MAnorm (Shao *et al.*, 2012) and ChIPnorm (Nair *et al.*, 2012) consider different SNRs. Both methods normalize the data before comparison: MAnorm performs normalization based on MA-plot, and ChIPnorm uses quantile normalization. But again, neither considers control data at the normalization step, and these methods cannot be easily extended to handle more complicated experimental designs.

There are two software packages provide functionalities to consider the control data: DBChIP (Liang and Keleş, 2012) and DiffBind (Stark and Brown, 2013). Both methods directly apply existing methods and software package developed for RNA-seq DE analysis. They start from a list of candidate regions which are unions of peaks called for each individual experiment. These regions are then treated like genes, and RNA-seq DE methods are directly applied for comparison. To account for the control experiment, the software provide option to subtract the normalized control counts from IP counts, then round the differences and use them as inputs for the software. There are several problems with this approach. First, the underlying assumption of the methods is that the background noise and biological signals are additive, which is not always true based on our real data observation (details in later section). Second, these methods don't consider the SNRs from different experiments. Finally, most RNA-seq DE methods are developed based on negative binomial distribution assumption of the gene counts. Subtracting control from IP counts then rounding will likely to violate that model assumption, which lead to incorrect statistical inferences.

In this work, we develop a comprehensive and rigorous statistical method, named 'ChIPComp', to perform quantitative comparison of multiple ChIP-seq data from experiments with narrow peaks, including data for most of the protein binding, some histone modifications if the modification regions are narrow, and the DNase-seq experiments. ChIPComp takes into consideration of (i) genomic background measured by the control data; (ii) SNRs in different experiments; (iii) biological variances from the replicates and (iv) multiple-factor experimental designs. We demonstrate using simulations and real data analyses that ChIPComp provides more accurate and robust results compared with existing methods.

2 Methods

We use a two-step procedure for the quantitative comparison of ChIP-seq datasets. In the first step, we apply existing peak calling algorithm to each individual dataset to identify peaks. We then obtain the union of peaks called from all datasets as the candidate regions for quantitative comparison. Since the first step peak calling method is well developed, we will only present the method for quantitative comparison in this section.

2.1 The data model

Suppose there are D datasets and N candidate regions. For candidate region i ($i = 1, 2, \ldots, N$) in dataset j ($j = 1, 2, \ldots, D$), let Y_{ij} be the observed IP counts. We assume that Y_{ij} follow a Poisson distribution with underlying rate μ_{ij}, which is a function of the background λ_{ij} and ChIP signal S_{ij}, e.g. $\mu_{ij} = f(\lambda_{ij}, S_{ij})$. Here λ_{ij} represents the background signals caused by technical or biological artifacts. The observed read counts from the control experiment can be considered as realizations of the backgrounds, and can be used for estimating λ_{ij} (details of the estimation procedure is presented in later section). S_{ij} represents the non-control-related signals in the IP sample. Further, we assume that $S_{ij} = b_j s_{ij}$, where b_j is a constant representing the SNR from dataset j, and s_{ij} measures the relative biological signals (e.g. protein binding or histone modification strength up to a constant).

Now consider a set of general, multiple-factor experiments with design matrix $\mathbf{X}$. At candidate region i, the logarithm of the relative biological signals are assumed to be from a linear model:

$$\log(s_{ij}) = \mathbf{x}_j \boldsymbol{\beta}_i + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma_i^2)$$

where $\mathbf{x}_j$ is the jth row of $\mathbf{X}$, and $\boldsymbol{\beta}_i$ is a vector of coefficient for the ith candidate region. ϵ_{ij} is a random term accounting for the

variations among biological replicates. Putting all pieces together, we have the following model for data at the candidate regions:

$$\begin{aligned} Y_{ij}|\mu_{ij} &\sim \text{Poisson}(\mu_{ij}) \\ \mu_{ij} &= f(\lambda_{ij}, b_j s_{ij}) \\ \log(s_{ij}) &= \mathbf{x}_j \boldsymbol{\beta}_i + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma_i^2) \end{aligned} \quad (1)$$

Under this setup, the quantitative comparison for factor k at candidate region i can be achieved by testing: $H_0 : \beta_{ik} = 0$.

2.2 Estimate the background signal from control data

Obtaining good estimates of the background signal λ_{ij} is the crucial first step. Some existing methods (e.g. DBChIP or DiffBind) simply treat the counts from control experiment as background signals. However, the background noises are generated by artifacts such as chromatin structures and DNA sequence contexts, therefore, the noises fluctuate in genomic regions much wider than the peaks. Using the read counts at peaks regions only to estimate background is inaccurate and has large variances. The spatial correlations of the read counts from control experiment can be utilized to obtain better background estimates. Here we adopt the smoothing technique used in MACS (Zhang *et al.*, 2008) to obtain estimated background, denoted by $\hat{\lambda}_{ij}$. Once $\hat{\lambda}_{ij}$'s are obtained, we treat them as known and constant for the rest of the procedures.

2.3 Model the IP-background relationship

The most important component for the proposed data model is to characterize the relationship of IP and background signals, which is the f function. The approaches taken by DBChIP and DiffBind, e.g. subtracting the normalized control data, implicitly assume that the IP signal is the sum of the background and biological signals, or $\mu_{ij} = \lambda_{ij} + s_{ij}$. Another possible solution for quantitative comparison is to put the IP and background data into a 2×2 table at each candidate region, and then use χ^2 or Fisher's exact test for hypothesis testing. The underlying assumption for such approach is that the background and biological signals are multiplicative, e.g. $\mu_{ij} = \lambda_{ij} \times s_{ij}$.

To discover the true IP-background relationship, we obtain several public ChIP-seq datasets from ENCODE project (a description of the data is provided in the Section 4) and perform exploratory analyses. For peaks in an experiment, we obtain the read counts from IP experiments and estimate backgrounds from control, then plot the IP counts versus backgrounds in the logarithm scale.

Figure 1 shows such scatterplots from two ChIP-seq dataset: H3K27 acetylation (H3K27ac) in K562 cell line and RNA polymerase II (PolII) binding in HelaS3 cell line. These figures reveal several important aspects for the IP-background relationship. First, the IP and background signals are positively correlated, as expected. Second, the IP-background relationship is neither additive nor multiplicative. The relationship is non-linear in the log scale. Finally, the IP-background relationship is different in different datasets, demonstrated by the different slopes of the scatterplots in two datasets. This emphasizes the importance of building individual background model for each dataset separately.

Based on these observations, we use a smooth function to model the IP-background relationship in logarithm scale. The IP-background response function in dataset j is described by the following model:

$$\log \mu_{ij} = g_j(\log \lambda_{ij}) + \log S_{ij} = g_j(\log \lambda_{ij}) + \log b_j + \log s_{ij}$$

Here g_j is a experimental specific smooth function. This model assumes that at a candidate region, the IP signal is the sum of background-related noise (which is a smooth function of $\log \lambda_{ij}$), SNR and biological signal in the logarithm scale.

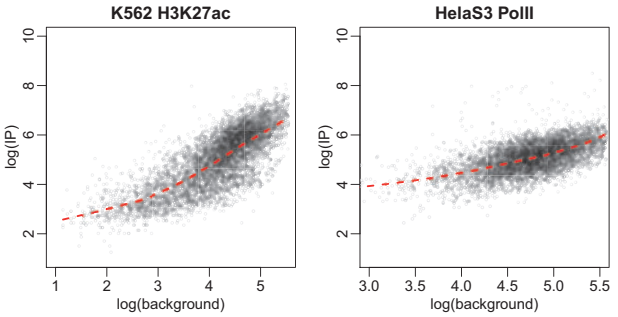

Fig. 1. Scatterplots of IP counts versus estimated background signals from the peak regions, in logarithm scale. The red dashed line is the result from cubic smoothing spline fitting (Color version of this figure is available at *Bioinformatics* online.)

2.4 The final model

Plugging in the IP-background model, the data model as described in Eq. 1 can be written as:

$$\begin{aligned} Y_{ij}|\mu_{ij} &\sim \text{Poisson}(\mu_{ij}) \\ \log \mu_{ij} &= g_j(\log \lambda_{ij}) + \log b_j + \log s_{ij} \\ \log s_{ij} &= \mathbf{x}_j \boldsymbol{\beta}_i + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma_i^2) \end{aligned} \quad (2)$$

This model implies that at a candidate region in an experiment, the underlying rates for the read counts are from a lognormal distribution. The mean of the distribution depends on the genomic background, the SNR for the experiment, and the true biological signal. In this model, the observed data are Y_{ij} from IP experiment. Background λ_{ij} can be estimated from the control experiment data, and g_j can be estimated from IP-background model. β are parameters of interests that one wants to make inference about.

2.5 The procedures for quantitative comparison

In our approach for quantitative comparison of ChIP-seq datasets, the quantities to be compared across experiments are the log biological signals $\log s_{ij}$. For each dataset, we first obtain the signals by removing the estimates of background-related noises and SNRs from observed read counts from IP experiment, and then perform statistical tests. The differential protein binding or histone modifications regions are defined based on test results. The rest of this section provides detailed descriptions of the estimation and hypothesis testing procedures.

2.5.1 Estimating g_j and b_j

At a candidate region, the IP signals from different experiments might exhibit quantitative differences. These differences could be due to differences in biological signals, or simply because different experiments have different backgrounds, SNRs or the IP-background responses. In order for data from different experiments to be comparable, a proper baseline is needed for normalization. A common normalization approach for ChIP-seq comparison uses the total number of reads under the peaks for adjustment. However, this approach only works for correcting technical artifacts. Biological differences such as different number of peaks cannot be

corrected by this approach. For example, even if the total numbers of reads from all peaks are identical in two conditions, the peak height can still be different due to different number of peaks.

We make a crucial assumption that there exists a subset of all candidate regions, where the averages of logarithm biological signals are identical in all datasets conditional on the background signals. This is a similar assumption used by MAnorm, and by some methods for gene expression microarray data normalization where the expressions of house keeping genes are assumed constant across conditions. Denoted such set by A, $A \in \{1, 2, \ldots, N\}$. By default, A is chosen as the common peaks from all datasets, or can be specified by user.

Further, we define a new function $g_j'(\log\lambda_{ij}) = g_j(\log\lambda_{ij}) + \log b_j$ to absorb the SNR into the background noise function. We take the following approach to estimate g_j' functions. For each individual dataset, we first obtain the IP counts (Y_{ij}) and estimated background signals ($\hat{\lambda}_{ij}$) for all peaks in A. Next, a cubic smoothing spline is fitted for $\log Y_{ij}$ versus $\log\hat{\lambda}_{ij}$. The fitted spline function is deemed $\hat{g}_j'$.

2.5.2 Hypothesis testing

The hierarchical model in Eq. 2 essentially describes the data as log-normal-Poisson compound distribution. The hypothesis testing can be performed using either likelihood ratio or Wald-based test. However either method requires numerical integration to obtain the marginal likelihood of β, which are computationally too intensive to be practically useful given large number of candidate regions. Further, with limited number of biological replicates, it is desirable to borrow information across different candidate regions to improve the estimation of biological variances and hence statistical inference, similar to that in many other high-throughput data analysis methods (Smyth, 2005; Wu *et al.*, 2013; Feng *et al.*, 2014). Such information sharing is usually achieved by adding another hierarchy in the model, for example, imposing a parametric distribution on the biological variances (σ_i^2). That will further increase the complexity of the model and make the model fitting more difficult. To overcome these difficulties, we use following approximate procedures to fit the model and perform hypothesis testing at each candidate region.

We first obtain $\widehat{\log(s_{ij})}$ as

$$\widehat{\log s_{ij}} = \log(Y_{ij} + c_0) - \hat{g}_j'(\log\hat{\lambda}_{ij})$$

Here c_0 is a small constant (0.5) added to the IP counts to avoid having $Y_{ij} = 0$. The estimated $\widehat{\log s_{ij}}$ can be viewed as 'normalized relative log fold changes'. They are quantities representing log fold changes between IP signals and background noises. They are further normalized to remove SNRs, and are values relative to the average $\log s_{ij}$'s from peaks in A and with similar background. Under our model assumptions, these quantities are directly comparable across datasets.

We then fit linear regression of $\widehat{\log s_{ij}}$ on $\mathbf{X}$, and obtain the estimates for coefficient β and residual variances σ_i^2. To overcome the small sample size problem, we apply existing variance shrinkage method developed for microarray analyses (Smyth, 2005) to obtain the shrunk estimates of σ_i^2, denoted by $\sim \sigma^{\sim 2}_i$. For statistical inference, an approximate estimate of the variances of $\widehat{\boldsymbol{\beta}}$ with consideration of the read counts can be derived as:

$$\widehat{var(\boldsymbol{\beta})} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\hat{\Sigma}\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}.$$

Here Σ is a diagonal matrix with the diagonal elements being $var(\widehat{\log s_{ij}})$. The detailed derivation of $\hat{\Sigma}$ is provided in the Supplementary Materials. In a nutshell, Σ takes into consideration of both the biological variances σ_i^2 and the uncertainty of $\log(s_{ij})$ point estimation affected by the read count Y_{ij}.

Hypothesis testing of $H_0 : \beta_{ik} = 0$ can be perform via Wald test, with the test statistics being $t = \widehat{\beta_{ik}}/\widehat{var(\beta_{ij})}$. The test statistics approximately follows normal distribution under null hypothesis. P-values and false discovery rate (FDR) can be obtained using canonical procedures (Benjamini and Hochberg, 1995).

In real data analyses, however, we found that the results from the Wald test are often influenced by the read counts, because candidate region with larger counts have greater power to detect differences. At these regions, the statistical significance is usually greater, e.g. with smaller P-values, even when the effect size is small. This is undesirable since the statistical significance doesn't necessarily imply biological significance. To overcome this problem, we provide an alternative approach in Bayesian framework. Assuming a non-informative prior on β_{ik}, e.g. $P(\beta_{ik}) \propto 1$, the following posterior probabilities are used to rank candidate regions:

$$Pr(|\beta_{ik}| > c | Y_{ij}, \hat{g}, \hat{\lambda}_{ij})) \quad (3)$$

Here c is a user specified threshold. In two-group comparison case, c represents the log fold change of biological signals. Under the normality assumption of $\hat{\beta}_{ik}$, the posterior probability can be obtained from normal p-values and used to rank the candidate regions. We find that this procedure often provides better results in real data analyses.

The above procedures are developed for data with biological replicates. When replicate is unavailable in the comparison, ChIPComp will use the difference in the estimated biological signals between two conditions, e.g. $\widehat{\log(s_{i1})} - \widehat{\log(s_{i2})}$, to rank the candidate regions.

3 Implementation

The proposed method is implemented in an R package `ChIPComp`, which is currently available at http://web1.sph.emory.edu/users/hwu30/software/ChIPComp.html, and being prepared to submit to Bioconductor (Gentleman *et al.*, 2004). The function takes detected peaks from all datasets and the aligned sequence files as inputs, and reports a list of genomic regions showing differential binding or histone modification, with estimated P-values and FDRs.

4 Results

4.1 Data description

Both simulation and real data analysis results are based on a number of public ChIP-seq datasets. We obtain several public ChIP-seq datasets generated by ENCODE consortium Thomas *et al.* (2006), including three cell lines (HelaS3, GM12878 and HUVEC) for RNA polymerase II binding (PolII), and three cell lines (H1, K562 and HelaS3) for H3K27 acetylation (H3K27ac). Both the aligned sequence files (aligned to human reference genome build hg19, in BAM format) and peak calling results are obtained from ENCODE.

4.2 Simulation

We first perform several simulation studies, based on parameters estimated from real data, to evaluate the performance of ChIPComp. All simulations are for two-condition comparison, with 10,000 candidate regions, and 20% of these regions are true differential regions. Data are simulated based on two different data generative models: the proposed data model as described in Eq. 2; and an additive model where the underlying IP rates (μ_{ij}) are sums of

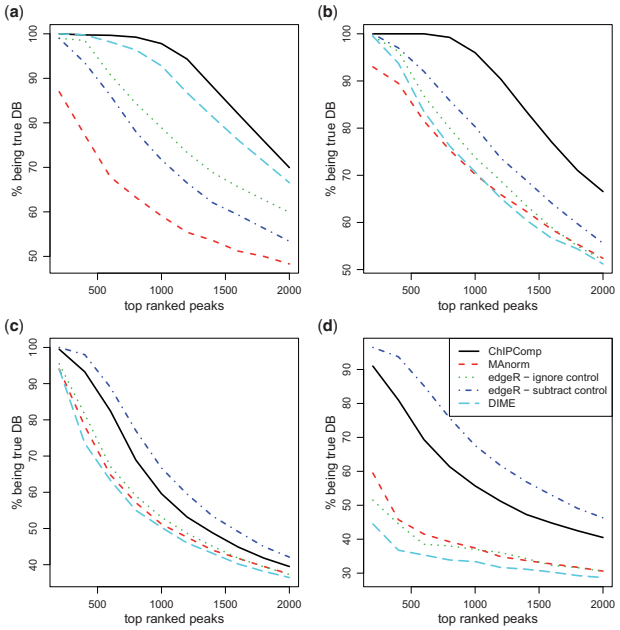

Fig. 2. Comparison of differential peak detection accuracies from simulations. The proportions of true discovery among top-ranked candidate regions is plotted against the number of top-ranked regions. (a) and (b) data are generated based on the proposed model. (c) and (d) data are generated based on the additive model. (a) and (c) are based on H3K27ac. (b) and (d) are based on PolII (Color version of this figure is available at *Bioinformatics* online.)

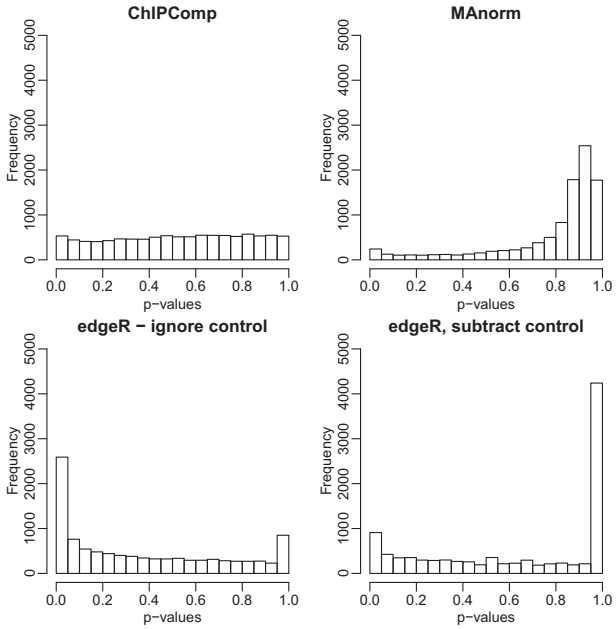

Fig. 3. Histogram of *P*-values reported from different methods, based on null model that there's no differential regions. The data are generated from the proposed model

background (λ_{ij}) and biological signals (s_{ij}). The additive model is the underlying assumption of DBChIP and DiffBind. We include the additive model in order to demonstrate that the proposed model work fine even under this assumption.

For the simulation results shown here, the simulation parameters are sampled from the real data estimates from H3K27ac and PolII. Parameters include background rate (λ_{ij}), biological signals (s_{ij}), and the IP-background relationship (g'_j) for the proposed model. For non differential candidate regions, the biological signals are made identical for two conditions. For differential regions, we randomly sample biological signals from real data for two conditions independently, so that they are different.

Since the differential analyses of ChIP-seq data is often used as a hypothesis generating tool, the goal is to have as many true positives as possible in the top-ranked candidate regions. We compare the proportions of true positives (i.e. true discovery rate, or TDR) in the top-ranked regions from different methods. The methods in comparison include the ChIPComp using the posterior probability in Eq. 3, MAnorm, edgeR with and without subtracting controls, and DIME. Both DBChIP and DiffBind require aligned read files as inputs, which pose difficulties in simulations. Since both methods are based on existing RNA-seq DE detection methods, we use edgeR to approximate their performances in simulations.

Figure 2 compares the TDR curves of differential peak detection from different methods in several simulation scenarios. Figure 2(a) and (b) shows the results when data are generated from the proposed model. In these cases, ChIPComp provides the best performance among the methods in comparison, and the gain of accuracy could be significant. It also shows that when data are generated from proposed model, subtracting control does not necessarily provide better results, that is, edgeR with and without subtracting control perform similarly. Figure 2(c) and (d) shows the results when data are generated from the additive model. Not surprisingly, edgeR with subtracting control provides the best results. However, ChIPComp performs the second best and significantly outperforms all other method. These simulation results show good performance of ChIPComp. The results from the additive model further demonstrate its robustness.

Another commonly used distribution assumption for sequencing read counts is negative binomial distribution, or the Gamma–Poisson compound distribution. The difference is that it assumes the underlying Poisson rate follows Gamma distribution instead of lognormal. We perform additional simulations when data are generated from negative binomial distribution (results shown in Supplementary Fig. S3). The TDR curves show that the proposed method is robust to that distribution assumption, and ChIPComp still performs the best overall.

Furthermore, we perform an additional 'null' simulation when there are no differential peaks. The data are generated from the proposed model using the same settings as the previous simulation. Because there are no differential peaks, the result *P*-values should follow uniform distribution. Figure 3 shows the histogram of *P*-values from different method, Results from edgeR ignoring control reports many false positives. *P*-values from MAnorm and edgeR subtracting control are heavily skewed toward 1 and tend to be over-conservative (number of false positives under different *P*-value threshold are shown in Supplementary Table S1). Overall, ChIPComp provides the most uniform *P*-value distribution, which again indicates that the statistical inference will be the most accurate. Similar simulation is conducted when data are generated from additive model (results shown in Supplementary Figure S1 and Table S2). Again, p-values from ChIPComp are more uniform compared than others when data are generated from additive model.

We further investigate the FDR for different estimated models. For each candidate region, one minus the posterior probability obtained from Eq. 3 can be viewed as local FDR. Based on the connection between the local FDR and the classical global FDR (Efron,

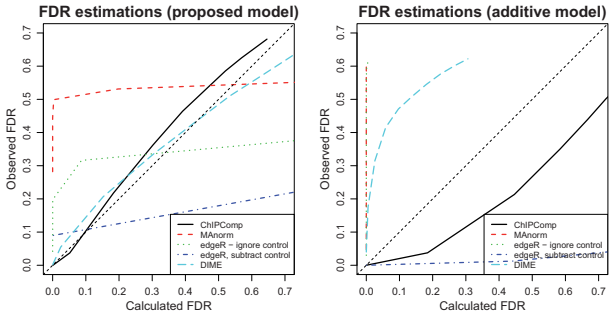

Fig. 4. Comparison of FDR estimations from different methods, based on simulation. *X*-axis shows the FDR reported from different methods, and *y*-axis shows the observed FDR (Color version of this figure is available at *Bioinformatics* online.)

2004), the local FDR can be converted to the global FDR. Figure 4 shows the comparison of global FDR estimates from different methods, when data are generated from the proposed and additive model. When data are from the proposed model, ChIPComp provides accurate FDR estimation. DIME performs well too but all other methods have poor performances. From other methods, the estimations of the FDR in the top-ranked regions are too liberal and give overly optimistic results. When data are generated from additive models, none of the methods provide very accurate FDR estimation, but ChIPComp still has the best performance relatively.

All simulation results demonstrate that ChIPComp is more accurate and robust compared to existing methods. It provides better ranking and statistical inference in detecting differential peaks. It is also fairly robust against model mis-specification, for example, when data are from additive model.

4.3 Real data results

We further evaluate the performance of ChIPComp in several real datasets. The analyses are based on two-condition comparisons. Since the gold standards for quantitative differences between ChIP-seq data are not available, we utilize other data to create 'silver standard' to compare different methods. It was known that PolII binding and H3K27ac are positively correlated with gene expressions. We obtain the gene expression data from RNA-seq experiments for these samples (also from ENCODE), and then use them to create silver standard for comparison. To be specific, in a two-condition comparison, we first perform differentially expression (DE) analyses on the RNA-seq data using edgeR. Genes with FDR less than 0.01 are deemed DE, with FDR greater than 0.2 are deemed non-DE, and the rest are deemed unknown. Next, we keep candidate regions that are within 1000 base pairs of the transcriptional start sites (TSS) of a gene. Finally, a region will be deemed differential or non-differential for the protein binding or histone modification between two conditions if its corresponding gene is DE or non-DE.

Since there are three cell lines (HelaS3, GM12878 and HUVEC) for PolII and another three cell lines (H1, K562 and HelaS3) for H3K27ac, we perform following pairwise two-condition comparisons: HelaS3 versus K562, H1 versus K562, H1 versus K562 for H3K27ac; HelaS3 versus HUVEC, GM12878 versus HelaS3, GM12878 versus HUVEC for PolII. The performance of the proposed method is compared with MAnorm, DBChIP ignoring or subtracting control, and DIME. DiffBind essentially uses the same algorithm as DBChIP (apply existing RNA-seq DE methods), so they are not included in the comparison. We use $c = 1$ in ChIPComp

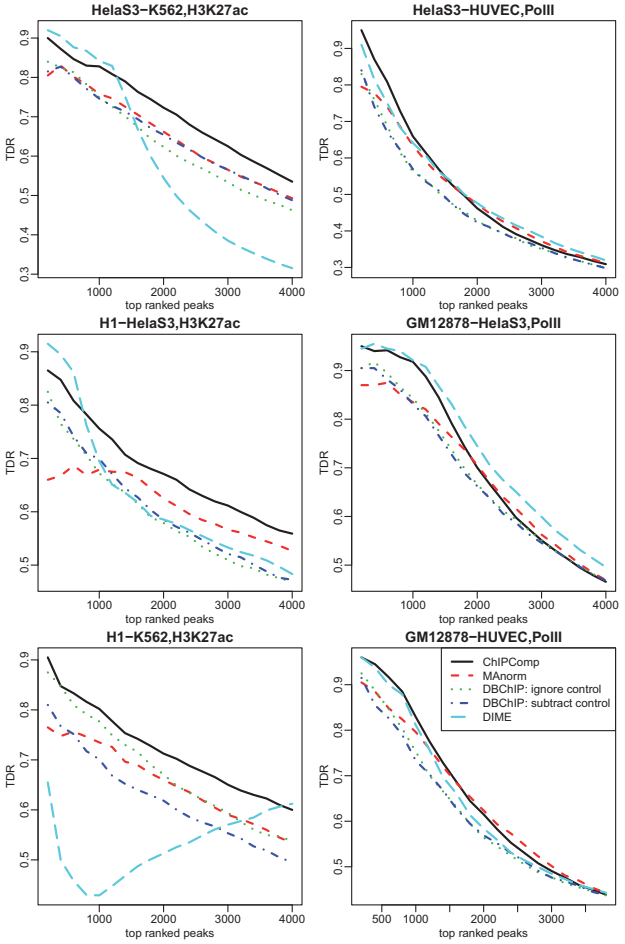

Fig. 5. Comparison of differential peak accuracies from real datasets. All results are for two-condition comparisons on different histone modification or protein binding, as marked in figure titles (Color version of this figure is available at *Bioinformatics* online.)

to generate and rank the differential binding regions for the results presented below.

Figure 5 shows the detection accuracies from all the comparisons. For H3K27ac comparisons, ChIPComp performs best except that DIME slightly outperforms at a small number of top peaks in HelaS3 versus K562 and H1 versus HelaS3. However, DIME fails badly in the H1 versus K562 comparison. In practice, we found that DIME is sometimes unstable, perhaps due to the convergence problem in EM algorithm. Compared with other methods, gains of detection accuracies from ChIPComp is usually over 10%. For comparisons in PolII binding data, ChIPComp and DIME are the best performers across three cell lines. Overall, these real data analyses demonstrate that ChIPComp provides the most accurate and robust results compared with other methods. In addition, we notice that subtracting control, based on the assumption of additive model, does not necessarily provide better performance than ignoring control from the results of DBChIP. In H1 versus K562 comparison for H3K27ac data, ignoring control actually provides much better performance than subtracting control. This is consistent with the results from simulation studies, and further demonstrates that simply subtracting control from IP is not an optimal way to use the data from control experiment in quantitative comparison of ChIP-seq data.

Although the posterior probability threshold c could have some impacts on the performance of ChIPComp, the overall performance

of ChIPComp remains stable with reasonable choice of c value. Since the default c value is 1, we try using different c values (0.5 and 2), and obtain similar TDR curves as using default c value (Supplementary Figures S4 and S5). We also plot the TDR curves by using p-values from hypothesis test instead of the posterior probabilities to rank peaks, and find similar results (Supplementary Figure S6).

In addition, we examine the FDR estimation accuracies from all methods in real datasets, using gene expression as gold standard. We plot the observed versus reported FDR from different methods (Supplementary Figure S7). In general, none of the methods provide very accurate FDR estimation, but ChIPComp still has the best performance overall.

Furthermore, we generate the ROC curves and use area under the curve (AUC) as another criteria to compare the performance of different methods (Supplementary Figure S8). Overall ChIPComp has the highest AUC value.

5 Discussion

In this work, we develop a novel statistical method to perform quantitative comparisons of multiple ChIP-seq datasets and detect differential protein binding or histone modification regions. Statistical methods of differential analysis for other sequencing data such as RNA-seq have been well developed. The comparison of ChIP-seq data, however, is more complicated because of different background noises and SNRs in distinct experiments. Existing methods either ignore the data from control experiments (such as MAnorm or DIME), or directly apply RNA-seq methods without proper normalization (such as DBChIP or DiffBind). The proposed method describes the data by a rigorous statistical model with the considerations of control data, SNRs, biological variations, and general experimental designs. Statistical test procedures are developed for detecting differential regions. Simulation and real data analyses results demonstrate that ChIPComp provides more accurate and robust results compared with existing methods.

The essence of the method is to extract biological signals from different experiments and then compare. The process involves estimating and removing biological and technical artifacts, and normalization of the biological signals. To ensure that the estimated biological signals are comparable across different experiments, proper references are needed for normalization and put the biological signals in a common baseline. In that regard, the proposed method relies on two important assumptions. First, the ChIP-seq datasets in comparison need to have a non-trivial number of common peaks. In fact, when there are very few common peaks among datasets, a simple overlapping analysis of the peak will be adequate. Second, it is assumed that there's no global difference in the true biological signals for the common peaks across all datasets, which is the same assumption used by MAnorm. This assumption provides a common baseline for different datasets for comparison. Similar assumption has been used in DE analysis for many years: a majority of the genes show no DE.

The hypothesis test is performed based on the log biological signals, which is derived based on log counts. When the counts at candidate regions are very small, this procedure could bring some biases and high variance. To overcome that, we added a small constant in the counts to 'squeeze' the lower end of the log count distribution, and carefully derived the variances for estimated parameters to take the raw counts into consideration. A similar approach has been proposed in a recently developed RNA-seq DE method, voom (Law *et al.*, 2013), and proved to have good performance.

The proposed method describes the count data from replicated samples through a lognormal-Poisson model. More often, these data are described by negative binomial, which is a Gamma–Poisson compound distribution. In our model, the underlying Poisson rate is assumed to follow a lognormal, instead of Gamma distribution. This is mainly motivated by methodological convenience. However, when the shape parameter in Gamma distribution is reasonably large, the Gamma and lognormal distributions become very similar. Simulation results show that the results from ChIPComp is robust and still provides good results when the data are from negative binomial. So we believe that our method will perform well in real data settings.

The method is specifically designed for comparison of ChIP-seq with short peaks, including most of the protein binding data, some histone modification data and DNase-seq. For histones modification data with long peaks/blocks such as H3K9me3, the method is not directly applicable. However the problems presented in those data are similar: consideration of backgrounds, different SNRs, biological variances, etc. To design method working for the quantitative comparison of data with long peaks is our research plan in the near future.

Acknowledgement

The authors acknowledge Dr. Zhijin Wu at Brown University for valuable comments and suggestions.

Funding

H.W. is partially supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR000454. L.C. and Z.S.Q. are partially supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R01 HG005119 and National Institute of General Medical Sciences of the National Institutes of Health under Award Number P01 GM085354.

Conflict of Interest: none declared.

References

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)*, **57**, 289–300.

Celniker,S.E. *et al.* (2009) Unlocking the secrets of the genome. *Nature,* **459**, 927–930.

Chen,X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell,* **133**, 1106–1117.

Efron,B. (2004) Large-scale simultaneous hypothesis testing. *J. Am. Stat. Assoc.,* **99**, 96–104.

Feng,H., Conneely,K.N. and Wu,H. (2014) A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.,* **42**, e69.

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

Johnson,D. *et al.* (2007) Genome-wide mapping of in vivo protein–DNA interactions. *Science,* **316**, 1497.

Laajala,T.D. *et al.* (2009) A practical comparison of methods for detecting transcription factor binding sites in chip-seq experiments. *BMC Genomics,* **10**, 618.

Law,C.W. *et al.* (2013) Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol 2014,* **15**, R29.

Liang,K. and Keleş,S. (2012) Detecting differential binding of transcription factors with chip-seq. *Bioinformatics,* **28**, 121–122.

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.

Shao,Z. *et al.* (2012) Manorm: a robust model for quantitative comparison of chip-seq data sets. *Genome Biol.,* **13**, R16.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 3.

Stark,R. and Brown,G. (2014) DiffBind: Differential Binding Analysis of ChIP-Seq peak data. R package version 1.10.2.2014.

Taslim,C. *et al.* (2009) Comparative study on chip-seq data: normalization and binding pattern characterization. *Bioinformatics,* **25**, 2334–2340.

Taslim,C., Huang,T. and Lin,S. (2011) Dime: R-package for identifying differential chip-seq based on an ensemble of mixture models. *Bioinformatics,* **27**, 1569–1570.

Thomas,D. *et al.* (2006) The encode project at uc santa cruz. *Nucleic Acids Res.,* **35**, D663.

Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in chip-seq peak detection. *PLoS One,* **5**, e11471.

Wu,H., Wang,C. and Wu,Z. (2013) A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics,* **14**, 232–243.

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,Y. *et al.* (2008) Model-based analysis of chip-seq (macs). *Genome Biol.,* **9**, R137.