A Generalized Linear Model for Peak Calling in ChIP-Seq Data

JIALIN XU and YU ZHANG

ABSTRACT

Chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-Seq) has become a routine for detecting genome-wide protein-DNA interaction. The success of ChIP-Seq data analysis highly depends on the quality of peak calling (i.e., to detect peaks of tag counts at a genomic location and evaluate if the peak corresponds to a real protein-DNA interaction event). The challenges in peak calling include (1) how to combine the forward and the reverse strand tag data to improve the power of peak calling and (2) how to account for the variation of tag data observed across different genomic locations. We introduce a new peak calling method based on the generalized linear model (GLMNB) that utilizes negative binomial distribution to model the tag count data and account for the variation of background tags that may randomly bind to the DNA sequence at varying levels due to local genomic structures and sequence contents. We allow local shifting of peaks observed on the forward and the reverse strands, such that at each potential binding site, a binding profile representing the pattern of a real peak signal is fitted to best explain the observed tag data with maximum likelihood. Our method can also detect multiple peaks within a local region if there are multiple binding sites in the region.

Key words: generalized linear model, ChIP-Seq, peak calling

INTRODUCTION

CHROMATIN IMMUNOPRECIPITATION FOLLOWED BY MASSIVELY PARALLEL SEQUENCING (ChIP-Seq) is a high-throughput technique for detecting genome-wide protein-DNA interaction and gene regulation. The success of ChIP-Seq data analysis depends on the quality of distinguishing effective binding signals from background noise. ChIP-Seq data contains short read sequences (20–50bp) called tags, which are mapped to the genome of interest. The observation is the number of tags mapped to each genomic location, called tag count. A unique feature of ChIP-Seq data is that the tag counts are obtained from the two strands of the genome, which are referred to as the forward and the reverse strands. Due to the ChIP-Seq technology, at each protein-DNA binding site, the tag counts observed from the forward strands are located on the left hand side of the binding site, and the tag counts observed from the reverse strand are located on the right hand side. A challenge of peak calling in ChIP-Seq data is therefore how to combine the tag counts from the two strands to increase the power of detecting real protein-DNA interaction sites. In particular, the

accumulation (peaks) of tag counts at real binding sites often form specific shapes, which we refer to as a binding profile, which can help us distinguish real protein binding from random binding. The binding profile is partially determined by the ChIP-Seq technology and by the structure of the proteins of interest. In addition to real peaks, regions in the genome have varying level of random peaks. Tags are more frequently observed at locations with open chromatin structures, and sequence contents also affect the variability of random tag counts (Spyrou et al., 2009; Anders and Huber, 2010). How to best account for the variation of tag counts across the genome and distinguish between real protein–DNA interaction and random peaks is an important problem in peak-calling.

Several peak calling programs have been published since 2008, including MACS (Zhang et al., 2008), SPP (Kharchenko et al., 2008), QuEST (Valouev et al., 2008), and CisGenome (Ji et al., 2008). MACS (Zhang et al., 2008) estimates a global peak shift size from regions with significant fold changes. MACS then shifts and combines all forward and reverse strand tags toward the center by the estimated shift size, and calls peaks on the combined tags using a Poisson model through sliding windows. To account for the local variability of tag counts due to genomic features, MACS estimates a local Poisson parameter as the average tag counts from an up to 10kb neighboring region around each sliding window. Due to the constraint of mean and variance equality in the Poisson distribution, however, MACS is not able to model peak data if the variability of tag counts far exceeds the mean. Also, MACS reports binding regions with highly variable sizes, ranging from 200bp to 7kb. Due to its wide range of sizes of the predicted binding intervals, MACS tends to call only a single peak at regions of clusters of peaks. SPP (Kharchenko et al., 2008) first selects a global peak shift size from a cross-correlation analysis, which maximizes the linear Pearson correlation of the tag counts between forward and reverse strands. SPP then chooses a window size based on the estimated peak shift size. To detect binding, SPP utilizes a sliding window and calls a peak if a binding score is locally maximized, where the score is defined as twice of the difference between the geometric mean and the arithmetic mean of the forward tag counts in upstream window and the reverse tag counts in downstream window. SPP returns a false discovery rate (FDR) (Benjamini and Hochberg, 1995) for each window. Since SPP is not based on any statistical models, it estimates FDR by comparing binding scores in the signal data with the scores in a control data. When there is no real control data, a randomized data set (generated from the signal data) is used. QuEST (Valouev et al., 2008) also estimates a global peak shift size, which is used to build a combined score profile by shifting tags in the forward and the reverse strands toward the center. QuEST then calls a peak if a specific score profile achieves a local maximum. CisGenome (Ji et al., 2008) utilizes a negative binomial distribution rather than a Poisson distribution to call peaks. A peak is called if the observed tag counts within a sliding window (default 100bp) significantly exceeds the expected tag counts from a background distribution. CisGenome estimates the negative binomial parameters from the non-binding regions. It has been demonstrated that the negative binomial model can better fit ChIP-Seq datasets than the Poisson model (Ji et al., 2008). The above programs all estimate a global (and constant) peak shift size for all potential binding regions. They simply merge the forward and the reverse strand tags together before peak calling, and thus may lose power if the estimated peak shift size is inaccurate at some real binding sites. In addition, they do not incorporate any binding pattern information in their algorithm. At a real protein binding site, the tag counts often follow specific spatial patterns specific to the target protein, which provide valuable information for us to best distinguish between a real binding site from spurious peaks caused by means other than the target protein.

In this article, we propose a new negative binomial generalized linear model (GLMNB) that effectively uses information from both strands of the genome and model the background tag level using negative binomial distribution to account for the variation of tag counts along the genome. (Pepke et al. (2009) reported that an accurate strand-specific tag shift can considerably improve summit resolution; we therefore estimate a local peak shift size in each sliding window to combine information from both strands more accurately. Different from most methods, we fit the tag count data in a sliding window (500bp by default) for both forward and reverse strand simultaneously, rather than merging the tag counts. Furthermore, we estimate a binding profile (a pattern for the mean tag counts observed at peak regions) to best separate real peaks from spurious ones. We use simulation and a ChIP-Seq dataset of human transcriptional regulatory protein, hepatocyte nuclear factor 3 α (FoxA1), to demonstrate the performance of GLMNB. We also compare GLMNB with two popular peak callers, SPP and MACS. The two programs were reported to perform the best among many existing peak callers in a comprehensive evaluation study (Willbanks and Facciotti, 2010).

2. RESULTS

2.1. Simulation results

We simulated a ChIP-Seq dataset as described in the method section. The simulated data contained 500 peaks distributed in a 300Mb region. We first used the simulated data to estimate the binding profiles from the forward and the reverse strands. We then apply GLMNB to call peaks using a sliding window of 500bp and step size 10bp.

We first examined the p-values output by GLMNB using simulated data from non-binding regions. A non-binding region is defined as a region at least 500bp away from all simulated peaks. Figure 1a shows the

FIG. 1. Quantile-quantile plot of z-scores output by GLMNB (a), SPP (b) and MACS (c) using data from non-binding regions in the simulated data. Scatter plot between FDR in log scale and simulated peak strength called by GLMNB (d), SPP (e), and MACS (f). The simulated data contained 500 peaks randomly distributed in a 300Mb region, and each peak was separated from each other by at least 20kb. Histogram of distance between true peak positions and peaks called by GLMNB (g), SPP (h), and MACS (i) among true positives.

quantile-quantile (QQ) plot of the GLMNB's z-scores compared to a standard normal distribution. We observed that the GLMNB z-scores are approximately normally distributed, with a slight deviation that is likely due to the small tag counts in non-binding regions and also the irregular peak shift parameter θ used in our model. We did not observe strong departure in the QQ plot at the extreme values, suggesting that our p-values by normal approximation are appropriate for peak calling. We further calculated the FDR of GLMNB. FDR is in practice calculated as the expected number of false positives divided by the total number of positives (Efron, 2010). Given that we simulated the peaks, we can also calculate the observed FDR, defined as the observed number of false positive peaks divided by the total number of called peaks. At a 5% FDR threshold, GLMNB called 459 (non-overlapping) peaks, among which 458 were true peaks, yielding an observed FDR 0.2%. A true peak is detected if the predicted binding site is within 1kb to the true binding site. We used 1kb distance because otherwise MACS will miss too many true peaks due to its inaccurate prediction of binding locations.

The observed FDR of GLMNB is less than the expected FDR at 5%, because we assumed independence between sliding windows, but the test statistics of sliding windows are in fact correlated due to the tag counts shared among overlapping windows. Out of the 500 simulated true peaks, 42 peaks were missed by GLMNB at 5% FDR, yielding a 91.6% power. Figure 1d further shows a scatter plot between GLMNB's FDR in log scale and the simulated peak strength. We observed a strong correlation between the two, suggesting the correctness of GLMNB's peak rank. The only false positive peak had a high FDR very close to 5%. Figure 1g shows the histogram of distances from GLMNB's predicted binding sites to the true binding site among the 458 true positives. The average distance is 4bp, which is the smallest among the three programs. The standard deviation is 18bp, which is also the smallest.

Figure 1b shows the QQ plot of SPP's z-scores (converted from FDR) at non-binding regions. At 5% FDR, SPP called 584 peaks, including 483 true positives and 101 false positives. Even though SPP called 25 more true peaks than GLMNB, the number of false positives is much larger, resulting in an observed FDR $101/584 = 17\%$. Since SPP is not based on any statistical models, it assigns a minimum FDR to all top ranked peaks if their scores are stronger than the maximum scores observed in a control data. As a result, SPP's peak rank do not reflect protein binding strength. The histogram of predicted peak distance (mean 5bp and SD 24bp) by SPP to the true binding sites is shown in Figure 1h.

Figure 1c shows the QQ plot of MACS z-scores (converted from p-values) compared with a standard normal distribution. Due to the restriction of MACS program, we were only able to obtain p-values < 0.1 , rather than all p-values in the full range of $[0, 1]$. Further due to MACS automatic peak region expansion procedure, we were not able to restrict the same peak width as used by GLMNB at 500bp. We obtained 13,781 MACS p-values < 0.1 from the non-binding regions, and the sizes of MACS peaks ranged from 400bp to 6kb. As observed in Figure 1c, MACS z-scores from the non-binding regions significantly deviated from the standard normal distribution at large values. That is, the significance output by MACS is greatly inflated in our simulated data. For instance, at a threshold where we expect 30 false positive peaks, the actual number of false positives called by MACS is 234. The inflation of the significance by MACS is likely due to its Poisson model assumption.

MACS reports a predicted binding interval ranging up to thousands of basepairs rather than a single binding site. For fair comparison with GLMNB and SPP, which report binding positions, we used the center of MACS binding interval as the predicted binding site. At 5% FDR, MACS called 374 peaks, among which only 144 peaks were true positives, yielding 29% power. As seen in Figure 1f, MACS FDR and the simulated peak strength are not well correlated, suggesting that the peaks ranked by MACS may not correctly reflect the protein binding strength. For example, some false positive peaks even have much significant p-values than the true positives. Figure 1i shows the histogram of distances from MACS binding sites to the simulated binding site among the true positives. The histogram is flat in the whole range of 2kb, with a standard deviation of 495bp. This is because MACS only reports large binding intervals and cannot pinpoint the exact binding locations, especially when the background is noisy.

2.2. Peak calling in real datasets

We next applied GLMNB, SPP, and MACS to a real dataset of FoxA1 ChIP-Seq data. An example of the estimated binding profiles by GLMNB is shown in Figure 2a, where red and green curves represent the smoothed binding profiles for the forward and the reverse strands, respectively, and the vertical lines represent the raw cumulative tag counts. Further shown in Figure 2b is a FoxA1 binding site detected by GLMNB. The

FIG. 2 Binding profiles constructed from FoxA1 ChIP-Seq data and an example of FoxA1 peak called by GLMNB. **(a)** Smoothed (curve) and raw (vertical bars) forward and reverse binding profiles are estimated from FoxA1 ChIP-Seq data, shown in red and green, respectively. **(b)** A FoxA1 peak detected at chr1:199,518,119 (in blue dashed line) with $-\log_{10}(\text{pvalue}) = 22.9$ by GLMNB. All $-\log_{10}(\text{pvalue})$ from adjacent sliding windows are shown in blue connected dots. A blue horizontal line of length $2 \times \theta = 90\text{bp}$ represents the distance between the fitted forward peak (red curve) and the reverse peak (green curve). The y-axis is the tag counts per bin and negative log pvalues with base 10.

observed forward and reverse tag counts per bin (10bp) are plotted by red/green vertical bars in the figure. There are 13 sliding windows around the binding site, whose $-\log_{10}(\text{p-value})$ are illustrated as blue connected dots in the figure. The window centered at chr1:199,518,119 yielded the most significant p-value = $10^{-22.9}$, which is then called as a binding site. The fitted forward and reverse binding profiles by GLMNB is shown in red and green curves, which are located $\theta = 45\text{bp}$ away from the window center to each side. The blue horizontal line of length $2 \times \theta = 90\text{bp}$ marks the width of the predicted binding interval.

At 5% FDR, GLMNB detected 4,008 FoxA1 peaks. Figure 3a shows the ranked FDR in log scale of these peaks. The total number of sliding windows tested by GLMNB was 246,144 after filtration, and thus

FIG. 3. GLMNB peak calling results for FoxA1 ChIP-Seq. **(a)** GLMNB peaks ranked by expected FDR in increasing order. 4,008 GLMNB peaks were called at FDR $\leq 5\%$. **(b)** Histogram of the estimated peak shifting parameter (θ), with mean 44bp and standard deviation 22bp. Matched motif comparison between GLMNB peaks and SPP, MACS peaks for FoxA1 ChIP-Seq dataset. **(c)** Percentage of detected peaks carrying at least one FoxA1-related motifs within 150bp to predicted binding sites by GLMNB, SPP and MACS. **(d)** Histogram of the distance between predicted binding sites and closest FoxA1-related motifs.

with $p\text{-value} < 10^{-3.09}$ the expected number of false positives is 201 of 4,008 calls, yielding an expected 5% FDR. As shown in our simulation study, the FDR estimated by GLMNB is actually conservative, and thus we expect the actual FDR to be less than 5%. One of GLMNB's feature is allowing the peak shifting parameter vary in different windows. Figure 3b shows the distribution of the estimated peak shifts from all FoxA1 peaks. The estimated peak shifts for FoxA1 have mean 44bp and standard deviation 22bp.

FoxA1 binding sites were reported closely related to FoxA1 motifs, including Forkhead motif (FKHR) (Zhang et al., 2008; Matsuzaki et al., 2003), FoxA1/LNCAP, and FoxA1/MCF7 motifs. We show in Figure 3c the histogram of the distance between each detected FoxA1 peak to its closest FoxA1-related motifs, if there is at least one motif within 150bp of the predicted binding site by GLMNB (circle in solid line), SPP (upper triangle in dashed line), and MACS (inverse triangle in dotted line). We also show in Figure 3d the percentage of the detected FoxA1 peaks containing a FoxA1-related motif within 150bp distance. Among the top 4,008 FoxA1 peaks by GLMNB, there were roughly 87.8% to 95% peaks containing at least one FoxA1-related motif. SPP had similar motif percentage (88.1% to 94.7%). MACS peaks, in contrast, can be matched with 85.4% to 89.6% FoxA1-related motifs within 150bp distance, slightly lower than GLMNB and SPP. This is due to the inaccuracy of MACS predicted binding positions. Figure 3d further shows the histogram of distance between GLMNB (circle in solid line), SPP (upper triangle in dashed line), and MACS (inverse triangle in dotted line) predicted binding sites and the closest FoxA1-related motifs. The distances between GLMNB peaks and the closest FoxA1-related motifs were mostly within 100bp, suggesting the high spatial resolution of the predicted binding sites by GLMNB. Both GLMNB and SPP outperformed MACS.

Figure 4a shows the scatter plot of the $-\log_{10}(p\text{-value})$ of peaks detected by GLMNB and MACS. By comparing the peaks between GLMNB (4,008 peaks with $\text{FDR} \leq 5\%$ and tags ≥ 16 per window) and MACS (5,964 peaks with $\text{FDR} \leq 5\%$ and tags ≥ 16 per kb), 3,435 GLMNB peaks matched one-to-one with MACS peaks (black dots), 220 GLMNB peaks matched multiple-to-one with MACS peaks (purple dots), and 353 GLMNB peaks did not match with any MACS peaks (red dots). The 220 "multiple-to-one" GLMNB peaks matched to 98 MACS peaks, which in fact were 98 clusters of peaks. Each cluster included 2 to 6 significant GLMNB binding sites, whereas MACS only called a single binding interval of the entire region. Among the 5,964 MACS peaks, 3,533 matched with GLMNB's, and 2,431 did not match with any

FIG. 4. Comparison of FoxA1 ChIP-Seq results between GLMNB and MACS. (a) Scatter plot of the 4,008 peaks called by GLMNB at $\text{FDR} \leq 5\%$ compared to peaks called by MACS at the same FDR level. (b) A example of multiple GLMNB peaks and multiple SPP peaks matched to one MACS peak region. Both peaks were matched to FoxA1-related motifs. (c) A example of GLMNB peak undetected by MACS. GLMNB called this peak because the forward/reverse tags follow closely to the binding profiles, even though there were not many tags in this region. This peak had a nearby FoxA1-related motif. (d) A example of MACS and SPP peak undetected by GLMNB. No FoxA1-related motif was found.

GLMNB peaks (green dots). SPP called 6,401 peaks at 5% FDR, among which 3,970 matched with 99% of GLMNB peaks, and 2,431 did not match with any GLMNB peaks. The discrepancy of peaks detected by the three programs could be due to two reasons: (1) The information utilized by each program is different; and (2) the FDR control by each method are not exactly accurate. In particular, we suspect that SPP and MACS FDR are inflated, while GLMNB FDR is conservative.

In the FoxA1 study, we found that MACS tended to call peaks in larger sizes due to its automatic peak interval expansion procedure. This is why often we found multiple binding sites by GLMNB and SPP that matched to just a single MACS interval. A desirable feature of GLMNB is its capability to call nearby peaks. As shown in Figure 4b, a 1,270bp binding interval (orange horizontal segment) was reported by MACS between chr17:70,966,887-70,968,157, with an extremely small p-value 10^{-100} and a motif (orange star) 56bp to the right of the interval center. It is however obvious that the region contained two binding sites, with a stronger peak on the left and a weaker peak on the right. GLMNB called two peaks in this region, one centered at 70,967,317 (left blue circle) with $\theta = 59$ bp, and the other centered at 70,967,747 (right blue circle) with $\theta = 44$ bp. The p-values for the two peaks were $10^{-18.1}$ and $10^{-8.6}$, respectively. GLMNB is capable of capturing multiple peaks within a local region, because the method is a model based approach that fits the data with a specific binding profile. Both GLMNB peaks can be matched to two FoxA1-related motifs (blue stars, 33bp and 17bp to the left of the predicted binding sites, respectively). SPP was also able to call two peaks in this region. Two SPP peaks (purple upper triangles) were found, one at 70,967,302 with a motif (left purple star) 13bp to the left, the other at 70,967,760 with a motif (right purple star) 25bp to the left. As an extreme example (data not shown), GLMNB detected 6 peaks in a 5kb region, with p-values ranging from 10^{-5} to $10^{-24.1}$, on chr20:51,913,000-51,918,000. All six GLMNB peaks except one contained at least one FoxA1-related motifs. SPP identified seven peaks in the same region, six of which is within 30bp of GLMNB peaks. The one that does not match to GLMNB peaks does not contain any FoxA1-related motif. In contrast, MACS called the entire region as a single binding interval of 7,278bp with p-value = $10^{-13.5}$. This relatively large p-value does not reflect the real binding strength, as the region was ranked 3,551th by MACS, whereas the strongest GLMNB peak among the six was ranked 129th, and the strongest SPP peak among the seven was ranked 74th. Figure 4c shows another example of binding sites called by GLMNB and SPP that is missed by MACS. GLMNB called a peak at chr17:56,970,872 with p-value = $10^{-4.12}$, which can be matched with a motif 1bp on its left. SPP called a peak at chr17:56,970,852 with FDR = 0.2%, which can be matched with the same FoxA1-related motif but 19bp on its right.

Figure 4d shows an example of the peaks called by both SPP and MACS, but missed by GLMNB. Notice that the tag counts spread across a 1kb region fairly uniformly. The size of this region is much wider than a typical binding region for FoxA1, and there were no clear pattern of the forward and the reverse tags distributed within the region. As a result, GLMNB did not report this site as a significant FoxA1 binding site, but treating it as a spurious peak due to potential open chromatin structures and background noise. SPP reported a binding site at chr1:196,770,446 with FDR = 1%. MACS also reported a binding interval starting at chr1:196,769,941 with p-value = $10^{-14.9}$. However, there is no FoxA1-related motif within 500bp of this region. In practice, it is not uncommon to observe such local accumulation of tag counts due to mechanisms other than protein binding. A comparison of the tag counts in this region with a control sample would help determining its true binding status.

3. DISCUSSION

One feature of our GLMNB is the utilization of negative binomial distribution. Negative binomial distribution allows the background level of tag counts, β_0 in our model, and an dispersion parameter α , to vary across different genomic regions (Ji et al., 2008). The flexibility of using these two parameters makes GLMNB a better model for the ChIP-Seq data than a Poisson based model. Using a negative binomial model also allows us to properly account for the effect of biological variability (Spyrou et al., 2009); Anders and Huber, 2010). ChIP-Seq background tags are frequently unevenly distributed, as they depend on the chromatin structures and the sequence contents. By fitting a likelihood function to the data and obtaining its maximum likelihood estimator for β_0 and α within each sliding window, GLMNB can most efficiently and flexibly account for the effects of local genomic features. Furthermore, GLMNB fits the tag data by a binding profile, which is estimated from highly tag-enriched regions. As a result, GLMNB can detect refined protein binding sites that follow a particular spatial pattern rather than simply relying on the total tag counts.

Another key feature of GLMNB is its local peak shifting parameter estimation. All previous peak calling programs estimate a global peak shifting size from highly tag-enriched regions across the genome, and then use the parameter to merge forward/reverse strand tags together before peak calling. This strategy ignores the variation of the lengths of DNA sequence protected by the binding protein. For example, a wide interval of DNA sequence may be resulted from sonication in a ChIP-Seq experiment, if the binding protein forms a complex in large size. Similarly, a narrow interval of DNA sequence may be resulted if a protein only partially binds to a site. If a peak shifting parameter is not accurately estimated, the tags on the forward and reverse strands may not be correctly merged, and thus reducing power in peak calling and/or reducing the accuracy for pinpointing the binding positions. GLMNB estimates the peak shifting parameter along with other model parameters simultaneously within each sliding window. This strategy not only provides a more accurate peak shift estimate based on local tags, but also properly combines peak strength from the two strands of the genome.

As demonstrated in the FoxA1 data, GLMNB was able to detect 220 peaks that were clustered within 98 local regions along the genome. Rather than reporting a single long genomic interval containing multiple unspecified binding sites, as did by MACS, GLMNB pinpointed the binding locations of each site within a peak cluster. The output of GLMNB includes the predicted binding location, the local peak shifting parameter that measures the size of the protein binding, the statistical significance of the peak. Also, GLMNB can output the fitted spatial distribution of mean tag counts at the predicted binding sites.

There were several default parameters used by GLMNB: the bin size, the window size, the minimum tag counts in a window, and the step size of sliding windows. We evaluated the impact of these parameters to the performance of our method using the simulated data. We first tested bin size of 5bp, 10bp (default), and 20bp, respectively, which yielded almost the same power and the number of false positives. In practice, a larger bin size allows a better fit to the model, but it may reduce the mapping resolution. We next tested the cutoff value of the minimum tag counts within a window. With a cutoff of 5, 8 (default value), and 10 tags, we again obtained almost the same results in the simulated data. We suggest using a cutoff of 8bp, as it can effectively filter out most regions with potentially no signals, and at the same time reduces the computation time. We further tested the window size of 500bp (default) and 1000bp, respectively. Again we did not observe changes of the performance by GLMNB. In fact, given that our estimated binding profiles are in fixed sizes, increasing the window size itself will not greatly affect the performance. Finally, the step size of sliding windows is an important parameter. If the step size is too large, GLMNB can easily miss a true binding site. This is true for all methods utilizing sliding windows.

One future extension of GLMNB is to incorporate control datasets into peak calling. Currently, GLMNB constructs a background model using the ChIP-Seq signal data only. Even though GLMNB captures the local background variation of tag occurrence via a negative binomial model with overdispersion, the comparison between a signal data and a control data can further improve the modeling of background tag variation, and thus improve the power and the specificity of peak calling. For example, GLMNB was not able to call a peak in Figure 4d, because the tags spread widely in the region that were not significantly different from the background. If a control data at the same location shows a much fewer number of tags, it would help GLMNB tell the real binding status at the location. One straightforward approach of incorporating the control data is to include the control tag counts as additional data in the window. We fit a binding profile to the control data with a different set of parameters. We then call peaks by comparing the parameters for the signal data versus the parameters for the control data. Another interesting extension of the model would be to allow the width of the binding profiles to vary. For example, if the tag counts spread in a wider interval, but their spatial distribution confirms to an estimated peak shape after proper scaling, then it may still suggest a valid protein binding site.

4. METHODS

4.1. Dataset

ChIP-Seq data for FoxA1 were obtained from MACS (Zhang et al., 2008). The ChIP sample contains about 3.9 million uniquely mapped tags with length of 36bp.

4.2. Generate binding profiles

GLMNB requires input of tag counts, generated from a ChIP-Seq experiment and mapped to a reference genome, including its genomic location and the strand orientation (forward and reverse). The tag counts in

forward and reverse strands usually form two approximately bell shaped peaks on the two sides of a real binding site. The distance between the two peaks is determined by the length of the DNA sequence protected by the binding protein. We first build a binding profile for the forward and the reverse strands, separately, using the data from highly likely protein binding regions. In particular, we construct a raw forward (and reverse) binding profile by aligning all non-overlapping windows of $w_0 = 1,000\text{bp}$ that contain forward (and reverse) tag counts exceeding $n_0 = 20$. The windows are aligned at the peak center within each window. Note that the peak center is not necessarily the center of the window. For example, if most tags are located at the right side of a window, the peak center is defined as the center position of those tags forming the peak.

From the aligned windows, we use a kernel regression estimator with Gaussian kernel (Eubank, 1999) to construct a smoothed forward (and reverse) binding profile $\hat{m}_F(t)(\hat{m}_R(t))$. The bandwidth h of the estimator is set at 10bp . As a result, we obtain a binding profile for each strand, representing the smoothed and double differentiable shapes of real binding peaks. As shown in Figure 2a, red and green curves represent the smooth binding profiles for the forward and the reverse strands, respectively, and the vertical lines represent the raw profile.

4.3. Peak shift parameter

We use a sliding window to scan through the genome for peak calling. Within a window, we fit the tag counts data using the estimated binding profiles. Assuming that the center of the window is the true binding site, we expect to observe two peaks of tag counts, one on each strand of the chromosome, and we assume that the distance from each peak to the window center is the same (this is our definition of the unknown binding site relative to the observed peaks on two strands). Let $\theta (\geq 0)$ denote the distance from each peak on a strand to the window center. We call θ a peak shifting parameter. Within each window, we shift the forward (and the reverse) strand profile to the left (and right) of the window center by θ basepairs. The shifted forward and reverse strand profiles can be written as $\hat{m}_F(t + \theta)$ and $\hat{m}_R(t - \theta)$. Instead of fixing θ for all windows, we treat θ as an unknown parameter specific to each sliding window, and we estimate θ using the data within each window.

To fit a model between the tag counts and the estimated binding profiles, we first partition the window into non-overlapping bins of size b_1 , and we sum over the tag counts within each bin. This is to ensure that there are enough tags within each bin and also to reduce computation. Correspondingly, we convert the binding profiles according to the bin coordinates. Let t_i denote the positions of the bin boundaries, the values of the binding profiles within each bin $[t_i, t_{i+1}]$, for the forward and the reverse strands, respectively, are

$$\begin{aligned} x_i^F(\theta) &= \int_{t_i}^{t_{i+1}} \hat{m}_F(t + \theta) dt \simeq \hat{m}_F\left(\frac{t_i + t_{i+1}}{2} + \theta\right) \times b_1 \doteq \hat{m}_F(dm_i + \theta) \times b_1 \\ x_i^R(\theta) &= \int_{t_i}^{t_{i+1}} \hat{m}_R(t - \theta) dt \simeq \hat{m}_R\left(\frac{t_i + t_{i+1}}{2} - \theta\right) \times b_1 \doteq \hat{m}_R(dm_i - \theta) \times b_1 \end{aligned} \quad (1)$$

where dm_i denotes the center of bin $[t_i, t_{i+1}]$.

4.4. Negative binomial generalized linear model

We construct a negative binomial generalized linear model (GLMNB) for the observed tag counts (response), using the shifted binding profiles as a predictor. We choose the sliding window size as $w_1 = 500\text{bp}$, and the bin size as $b_1 = 10\text{bp}$. Let y_i^F and y_i^R , $i = 1, \dots, n$ denote the forward and the reverse tag counts within the i th bin, respectively, and $n = \text{ceiling}(w_1/b_1)$ denote the number of bins within a sliding window. We write the response variable of our GLMNB as a vector

$$\mathbf{y} = (y_1^F, \dots, y_n^F, y_1^R, \dots, y_n^R)^T \quad (2)$$

where the forward and the reverse tag counts are concatenated. Similarly, we write the predictor variable (binned values of the binding profiles) of our GLMNB as

$$\mathbf{x}(\theta) = (x_1^F(\theta), \dots, x_n^F(\theta), x_1^R(\theta), \dots, x_n^R(\theta))^T \quad (3)$$

where $x_i^F(\theta)$ and $x_i^R(\theta)$ are calculated by formula (1). Note that $\mathbf{x}(\theta)$ is a function of the unknown peak shifting parameter θ . For notation convenience, we write $\mathbf{x}(\theta)$ as $\mathbf{x}$ and $x_i(\theta)$ as x_i hereafter. We will further skip the superscripts ‘‘F’’ and ‘‘R’’ in $\mathbf{y}$ and $\mathbf{x}$.

We fit the observed tag counts using the shifted binding profiles via a variable overdispersion GLMNB, which allows more flexibility on estimating the dispersion parameter (Hardin and Hilbe, 2007) in order to fit biological variability in ChIP-Seq data. In particular,

$$\log(\mu_i) = \beta_0 + \beta_1 x_i + \delta_i$$

where μ_i denotes the mean of y_i , δ_i is an individual unobserved effect in the conditional Poisson mean and $\exp(\delta_i)$ is gamma noise with mean 1 and variance α (Hardin and Hilbe, 2007). Therefore, the log likelihood function can be written as

$$\begin{aligned} \ell(\mu; y, \alpha) = & \sum_{i=1}^{2n} \left\{ y_i \log(\alpha \exp(x_i \beta_1 + \beta_0)) - \left(y_i + \frac{1}{\alpha} \right) \log(1 + \alpha \exp(x_i \beta_1 + \beta_0)) \right. \\ & \left. + \log \Gamma(y_i + 1/\alpha) - \log \Gamma(y_i + 1) - \log \Gamma(1/\alpha) \right\} \end{aligned} \quad (4)$$

Here, the four parameters are the background coefficient β_0 , the effect (or the fitness) β_1 of the binding profile x_i to the observed tag counts, the peak shifting parameter θ , and the dispersion parameter α .

We use Newton-Raphson method to obtain the maximum likelihood estimators (MLE) for parameters $\omega = (\beta_1, \theta, \beta_0, \alpha)$. In particular, starting from a random value of $\omega^{(0)}$, we iteratively update ω by

$$\omega^{(r)} = \omega^{(r-1)} - \{ \ell''(\omega^{(r-1)}) \}^{-1} \ell'(\omega^{(r-1)}) \quad (5)$$

for $r=1, 2, \dots$ until convergence. Please refer to the *Appendix* for the gradient vector ℓ' and the observed Hessian matrix ℓ'' . The standard errors of the MLE for the individuals parameters in ω are calculated as the square root of the diagonal elements of the inverse observed Hessian matrix.

In our GLMNB, we set the initial values of $\beta_1^{(0)}$ and $\alpha^{(0)}$ as the MLE calculated from all tag counts in the current chromosome under the null hypothesis of no protein binding. We set $\theta^{(0)}$ as half of the median distance between the two peaks on forward/reverse strands observed in the top 100 windows (ranked by total tag counts). We further set $\beta_1^{(0)}$ as zero.

4.5. Peak calling

Our interest is to identify peaks that follow the shapes specified by the estimated binding profiles. We therefore test the following hypothesis

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 > 0$$

According to Hardin and Hilbe (2007), the distribution of $\hat{\beta}_1$ under the null hypothesis of no protein binding should be asymptotically

$$\hat{\beta}_1 \xrightarrow{D} N(\beta_1, \text{Var}(\beta_1)) \quad (6)$$

We therefore calculate the p-value of the test using z-score $z = \frac{\hat{\beta}_1}{\sqrt{\text{Var}(\beta_1)}}$. Although in our case the regularity condition may be violated because of the added θ in x , we still observed close to normal distribution of the z-scores in the simulated non-binding data. As discussed in Fan et al. (2001), even if the regularity conditions do not hold in some cases, an asymptotic normal approximation would still be reasonable. Alternatively, the users can empirically estimate a null distribution for the z-scores from a control data (or from a randomized tag count data) to calculate p-values.

We apply the above model to test all sliding windows across the genome for peak calling. Since the null distribution of our test statistic is only known asymptotically with enough tag counts, we filtered out all windows with tag counts less than cutoff = 8 on either the forward or the reverse strand. That is, we do not test in windows with very few tag counts. This filtration may reduce the power for detecting very weak binding sites with fewer tags, but it can significantly reduce the number of tests and increase the computation speed. To adjust for multiple comparisons, we calculate FDR assuming independence between tests. For example, in FoxA1 ChIP data, we called 4,008 peaks from 246,144 sliding windows with expected FDR $\leq 5\%$ at a p-value threshold of $10^{-3.09}$. Since our sliding windows are overlapping, however, our FDR control is conservative.

Note that our hypothesis test on β_1 requires the tag counts within a region follow the shape specified by the estimated binding profile. The shape of the peaks and their relative positions on the two strands of the chromosome are the main targets of our peak calling. For example, if a peak on the forward strand is on the right side of a peak on the reverse strand, our test will not call it as a binding site. If one is interested in calling peaks with enough tag counts, regardless of their shapes and orientations, then an alternative solution is to test whether β_0 is significantly above some background value. A multivariate joint test of both β_0 and β_1 can also be applied. In practice, however, frequently the tags are not uniformly distributed across the genome even if there are no protein binding. Certain genomic regions have more tags on average than other regions due to its chromatin structure and sequence contents. This is the motivation of our GLMNB method, which requires a specific peak shape and orientation, and is robust to local random variation. Control data can be further added to the model for a better control of false positives.

4.6. Data simulation

We simulated a background ChIP-seq data using a negative binomial distribution with the size parameter 0.0042 and the probability of success parameter 0.57, which roughly corresponds to 9 million tags mapped to the whole genome on each strand. We further simulated 500 ChIP-Seq peaks under another negative binomial distribution, with the mean values given by the FoxA1 profile shown in Figure 2a multiplied by a simulated peak strength, 2^γ , where γ is defined as a signal fold change relative to the profile, and the probability of success parameter of the negative binomial distribution is $p = 0.5$. The fold change parameter γ is generated from a standard normal distribution. Most values of 2^γ ranged between [0.125, 8], which is slightly wider than the estimated coefficients, $\hat{\beta}_1$, from the FoxA1 peaks. The peak shift for each simulated peak region is also generated from a normal distribution with mean 50 and standard deviation 10. We finally merged the simulated ChIP-Seq background data and the simulated ChIP-Seq peak data by adding the tag counts in both data together at each position.

4.7. Motif identification

We used HOMER (Heinz et al. (2010)) to identify motifs at the predicted FoxA1 binding sites. We let HOMER to search for *de novo* motifs within 150bp around each predicted binding site for GLMNB, SPP, and MACS (using the center of the reported peak interval), separately. There are three FoxA1-related motifs, FKHR motif (CTGTTTAC) FoxA1/LNCAP (WAAGTAAACA) and FoxA1/MCF7 (WAAGTAAACA). The core sequence of FKHR motif is actually the reverse complement of the other two.

5. APPENDIX

The gradient vector of the log likelihood function of GLMNB is

$$\begin{aligned}\frac{\partial \ell}{\partial \omega} &= \left(\frac{\partial \ell}{\partial \beta_1}, \frac{\partial \ell}{\partial \theta}, \frac{\partial \ell}{\partial \beta_0}, \frac{\partial \ell}{\partial \alpha} \right)^T \\ \frac{\partial \ell}{\partial \beta_1} &= \sum_{i=1}^{2n} \left\{ y_i x_i - \left(y_i + \frac{1}{\alpha} \right) x_i \frac{\alpha \exp(x_i \beta_1 + \beta_0)}{1 + \alpha \exp(x_i \beta_1 + \beta_0)} \right\} \\ \frac{\partial \ell}{\partial \theta} &= \sum_{i=1}^{2n} \left\{ y_i \left(\frac{\partial x_i}{\partial \theta} \right) \beta_1 - \left(y_i + \frac{1}{\alpha} \right) \frac{\alpha \exp(x_i \beta_1 + \beta_0)}{1 + \alpha \exp(x_i \beta_1 + \beta_0)} \left(\frac{\partial x_i}{\partial \theta} \right) \beta_1 \right\} \\ \frac{\partial \ell}{\partial \beta_0} &= \sum_{i=1}^{2n} \left\{ y_i - \left(y_i + \frac{1}{\alpha} \right) \frac{\alpha \exp(x_i \beta_1 + \beta_0)}{1 + \alpha \exp(x_i \beta_1 + \beta_0)} \right\} \\ \frac{\partial \ell}{\partial \alpha} &= \sum_{i=1}^{2n} \left\{ \frac{y_i}{\alpha} + \frac{1}{\alpha^2} \log(1 + \alpha \exp(x_i \beta_1 + \beta_0)) \right. \\ &\quad \left. - \left(y_i + \frac{1}{\alpha} \right) \frac{\exp(x_i \beta_1 + \beta_0)}{1 + \alpha \exp(x_i \beta_1 + \beta_0)} + \psi_0 \left(y_i + \frac{1}{\alpha} \right) (-\alpha^{-2}) - \psi_0 \left(\frac{1}{\alpha} \right) (-\alpha^{-2}) \right\}\end{aligned}$$

where $\psi_0(t) = \frac{\partial \Gamma(t)}{\partial t}$

The observed Hessian matrix of GLMNB is:

$$\frac{\partial^2 \ell}{\partial \omega^2} = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \beta_1^2} & \frac{\partial^2 \ell}{\partial \beta_1 \partial \theta} & \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ell}{\partial \beta_1 \partial \alpha} \\ \frac{\partial^2 \ell}{\partial \theta \partial \beta_1} & \frac{\partial^2 \ell}{\partial \theta^2} & \frac{\partial^2 \ell}{\partial \theta \partial \beta_0} & \frac{\partial^2 \ell}{\partial \theta \partial \alpha} \\ \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 \ell}{\partial \beta_0 \partial \theta} & \frac{\partial^2 \ell}{\partial \beta_0^2} & \frac{\partial^2 \ell}{\partial \beta_0 \partial \alpha} \\ \frac{\partial^2 \ell}{\partial \alpha \partial \beta_1} & \frac{\partial^2 \ell}{\partial \alpha \partial \theta} & \frac{\partial^2 \ell}{\partial \alpha \partial \beta_0} & \frac{\partial^2 \ell}{\partial \alpha^2} \end{pmatrix} \quad (7)$$

where

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \beta_1^2} &= \sum_{i=1}^{2n} \left(- \left(y_i + \frac{1}{\alpha} \right) x_i \frac{\alpha \exp(x_i \beta_1 + \beta_0) x_i}{(1 + \alpha \exp(x_i \beta_1 + \beta_0))^2} \right) \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \theta} &= \sum_{i=1}^{2n} \left(y_i \frac{\partial x_i}{\partial \theta} - \left(y_i + \frac{1}{\alpha} \right) \frac{\partial x_i}{\partial \theta} \frac{\alpha \exp(x_i \beta_1 + \beta_0)}{1 + \alpha \exp(x_i \beta_1 + \beta_0)} \left(1 + \frac{x_i \beta_1}{1 + \alpha \exp(x_i \beta_1 + \beta_0)} \right) \right) \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} &= \sum_{i=1}^{2n} \left(- \left(y_i + \frac{1}{\alpha} \right) x_i \frac{\alpha \exp(x_i \beta_1 + \beta_0)}{(1 + \alpha \exp(x_i \beta_1 + \beta_0))^2} \right) \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \alpha} &= \sum_{i=1}^{2n} \left(- x_i \frac{\exp(x_i \beta_1 + \beta_0)}{1 + \alpha \exp(x_i \beta_1 + \beta_0)} \left(- \frac{1}{\alpha} + \frac{y_i + \frac{1}{\alpha}}{1 + \alpha \exp(x_i \beta_1 + \beta_0)} \right) \right) \\ \frac{\partial^2 \ell}{\partial \theta^2} &= \sum_{i=1}^{2n} \left(y_i \beta_1 \frac{\partial^2 x_i}{\partial \theta^2} - \left(y_i + \frac{1}{\alpha} \right) \beta_1 \left(\frac{\alpha \exp(x_i \beta_1 + \beta_0) \beta_1 \left(\frac{\partial x_i}{\partial \theta} \right)^2}{[1 + \alpha \exp(x_i \beta_1 + \beta_0)]^2} \right. \right. \\ &\quad \left. \left. + \frac{\alpha \exp(x_i \beta_1 + \beta_0)}{1 + \alpha \exp(x_i \beta_1 + \beta_0)} \frac{\partial^2 x_i}{\partial \theta^2} \right) \right) \\ \frac{\partial^2 \ell}{\partial \theta \partial \beta_0} &= \sum_{i=1}^{2n} \left(- \left(y_i + \frac{1}{\alpha} \right) \frac{\alpha \exp(x_i \beta_1 + \beta_0) \beta_1 \frac{\partial x_i}{\partial \theta}}{(1 + \alpha \exp(x_i \beta_1 + \beta_0))^2} \right) \\ \frac{\partial^2 \ell}{\partial \theta \partial \alpha} &= \sum_{i=1}^{2n} \left(\frac{\exp(x_i \beta_1 + \beta_0) \beta_1 \frac{\partial x_i}{\partial \theta}}{\alpha (1 + \alpha \exp(x_i \beta_1 + \beta_0))} - \frac{(y_i + \frac{1}{\alpha}) \exp(x_i \beta_1 + \beta_0) \beta_1 \frac{\partial x_i}{\partial \theta}}{(1 + \alpha \exp(x_i \beta_1 + \beta_0))^2} \right) \\ \frac{\partial^2 \ell}{\partial \beta_0^2} &= \sum_{i=1}^{2n} \left(- \left(y_i + \frac{1}{\alpha} \right) \frac{\alpha \exp(x_i \beta_1 + \beta_0)}{(1 + \alpha \exp(x_i \beta_1 + \beta_0))^2} \right) \\ \frac{\partial^2 \ell}{\partial \beta_0 \partial \alpha} &= \sum_{i=1}^{2n} \left(\frac{\exp(x_i \beta_1 + \beta_0)}{1 + \alpha \exp(x_i \beta_1 + \beta_0)} \left(\frac{1}{\alpha} - \frac{y_i + \frac{1}{\alpha}}{1 + \alpha \exp(x_i \beta_1 + \beta_0)} \right) \right) \\ \frac{\partial^2 \ell}{\partial \alpha^2} &= \sum_{i=1}^{2n} \left[\frac{-y_i}{\alpha^2} - \frac{2}{\alpha^3} \log(1 + \alpha \exp(x_i \beta_1 + \beta_0)) + \frac{2 \exp(x_i \beta_1 + \beta_0)}{\alpha^2 (1 + \alpha \exp(x_i \beta_1 + \beta_0))} \right. \\ &\quad \left. + \frac{(y_i + \frac{1}{\alpha}) \exp^2(x_i \beta_1 + \beta_0)}{(1 + \alpha \exp(x_i \beta_1 + \beta_0))^2} + \psi'_0 \left(y_i + \frac{1}{\alpha} \right) \frac{1}{\alpha^4} + \psi_0 \left(y_i + \frac{1}{\alpha} \right) \frac{2}{\alpha^3} \right. \\ &\quad \left. - \psi'_0 \left(\frac{1}{\alpha} \right) \frac{1}{\alpha^4} - \psi_0 \left(\frac{1}{\alpha} \right) \frac{2}{\alpha^3} \right] \end{aligned}$$

ACKNOWLEDGMENTS

This research was supported by the NIH (grant R01 DK065806).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Anders, S., and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome Biol.*, 11, R106.
- Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 57, 289–300.
- Efron, B. 2010. *Large-Scale Inference*. Cambridge University Press, New York.
- Eubank, R.L. 1999. *Nonparametric Regression and Spline Smoothing*, 2nd ed. Marcel Dekker, New York.
- Fan, J., Zhang, C., and Zhang, J. 2001. Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Stat.* 29, 153–193.
- Hardin, J.W., and Hilbe, J.M. 2007. *Generalized Linear Models and Extensions*, 2nd ed. STATA.
- Heinz, S., Benner, C., Spann, N., et al. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol. Cell.* 38 576–589.
- Ji, H., Jiang, H., Ma, W., et al. 2008. An integrated software system for analyzing chip-chip and chip-seq data. *Nat. Biotechnol.* 26, 1293–1300.
- Kharchenko, P.V., Tolstorukov, M.Y., and Park, P.J. 2008. Design and analysis of chip-seq experiments for dna-binding proteins. *Nat. Biotechnol.* 26, 1351–1359.
- Matsuzaki, H., Daitoku, H., Hatta, M., et al. 2003. Insulin-induced phosphorylation of fkhr(foxo1) targets to proteasomal degradation. *Proc. Natl. Acad. Sci. USA* 100, 11285–11290.
- Pepke, S., Wold, S., and Mortazavi, A. 2009. Computation for chip-seq and rna-seq studies. *Nat. Methods* 6, S22–S32.
- Spyrou, C., Stark, R., Smith, M.L., et al. 2009. Bayespeak: Bayesian analysis of chip-seq data. *BMC Bioinform.* 10, 299.
- Valouev, A., Johnson, D.S., Sundquist, A., et al. 2008. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nat. Methods* 5, 829–834.
- Willbanks, E.G., and Facciotti, M.T. 2010. Evaluation of algorithm performance in chip-seq peak detection. *PLoS ONE* 5, e11471.
- Zhang, Y., Liu, T., Meyer, C.A., et al. 2008. Model-based analysis of chip-seq (MACS). *Genome Biol.* 9, R137.1–R137.9.

Address correspondence to:

Dr. Yu Zhang

Department of Statistics

The Pennsylvania State University

325 Thomas Building

University, PA 16802

E-mail: yuzhang@stat.psu.edu