Statistics and Its Interface Volume 8 (2015) 419–436

Statistical issues in binding site identification through CLIP-seq

Xiaowei Chen, Dongjun Chung, Giovanni Stefani,
Frank J. Slack, and Hongyu Zhao*

With the advent and development of CLIP-seq technologies, a growing number of CLIP-seq experiments are being performed to identify the targets of RNA-binding proteins and understand the regulation mechanism of these proteins. Although broad similarities exist between CLIP-seq and ChIP-seq, statistical methods developed to identify binding sites from ChIP-seq data are not directly applicable to CLIP-seq data because of some differences between the two technologies. First, transcript abundance has a large impact on CLIP-seq results, and needs to be accounted for when analyzing CLIP-seq data. Second, mutations near the binding sites from CLIP-seq data offer valuable information that can be incorporated in analysis. Other differences arise from the ability of RNA to form complex secondary structures and from many other technical aspects of the two purification protocols. To date, no systematic studies have been conducted to investigate the general statistical properties of CLIP-seq data, the merits of including RNA-seq as a matching control, and the performance of different binding site identification methods for CLIP-seq data. In this study, we performed a comprehensive evaluation of various statistical issues in using CLIP-seq data to identify RNA-protein binding sites. We demonstrate the value of RNA-seq data in background estimation and peak calling. We show that the large dispersion in CLIP-seq data compared to ChIP-seq data is the main reason for the difficulty in peak calling in the former. Using both real and simulated data, we also show the importance of biological/technical replicates and of combining mutation and peak analysis to accurately identify binding sites from CLIP-seq data.

1. INTRODUCTION

Nucleic-acid (DNA and RNA) binding proteins play a crucial role in gene expression regulation. The identification of the full spectrum of DNA or RNA binding sites of these proteins is necessary to fully characterize their biological roles. ChIP-seq has been widely and successfully used in DNA-binding protein studies and many computational methods have been developed to identify binding sites from ChIP-seq data (Kharchenko, Tolstorukov et al. 2008, Zhang, Liu et al. 2008, Rozowsky, Euskirchen et al. 2009, Kuan, Chung et al. 2011). Similarly, CLIP (crosslinking and immunoprecipitation) coupled with high throughput sequencing (also called HITS-CLIP) facilitates the identification of binding sites for RNA-binding proteins (Ule, Jensen et al. 2005, Wang, Tollervey et al. 2009, Murigneux, Sauliere et al. 2013). For HITS-CLIP, protein and bound RNAs are crosslinked by exposure to UV light. The resulting covalent bonds allow very stringent immunoprecipitation conditions. The isolated RNA fragments are then subjected to reverse transcription and cDNA sequencing. In addition to HITS-CLIP, there are other types of CLIP data generated from different protocols, such as PAR-CLIP (Hafner, Landthaler et al. 2010) and iCLIP (Konig, Zarnack et al. 2011). Because the CLIP data generated from each protocol have their unique features, we will focus on HITS-CLIP in this article, and the term CLIP-seq refers to the sequencing data generated from HITS-CLIP experiments for the rest of the article.

Although CLIP-seq and ChIP-seq data are generated by sequencing RNA and DNA co-purified with proteins, directly applying ChIP-seq data analysis methods to CLIP-seq datasets is neither suitable nor efficient for four main reasons. First, ChIP-seq data have relatively uniform and random distribution of background tags, so the binding regions can be detected with excess tags compared to other genomic regions. In contrast, the tag counts are much more variable in CLIP-seq data because its read count is not only related to protein binding (or affinity of protein binding), but also strongly correlated with the transcript abundance. As a result, some true binding sites may not have high read counts if they fall within transcripts that are expressed at low levels. Therefore, the ChIP-seq binding site identification methods assuming uniform background may perform poorly in the identification of binding sites for CLIP-seq data. Second, the potential binding regions for CLIP-seq are spatially much more restricted than those for ChIP-seq because CLIP-seq only identifies protein binding on coding or non-coding transcribed regions. However, ChIP-seq binding site identification methods assume that binding sites could be identified in any positions on the genome and as a result, the protein binding site prediction using these methods

*Corresponding author.

could be suboptimal for CLIP-seq. Third, mutations (substitutions, deletions, insertions) occur at UV cross-linking sites (usually corresponding to the binding sites) in CLIP-seq data (Kishore, Jaskiewicz et al. 2011, Zhang and Darnell 2011). Unlike formaldehyde-induced cross-linking used in ChIP, UV cross-linking is not reversible by heat. When the protein is digested by proteinase K, incomplete digestion leaves amino acid residues on the RNA, which affects fidelity as reverse transcriptase reads through the fragments. As a result, some mutations may be introduced exactly where the intermolecular contacts have taken place at the binding sites. Hence, cross-linking induced mutations are a unique feature in CLIP-seq, which could potentially improve the spatial resolution of protein binding site identification. Fourth, in ChIP-seq, sequencing of both DNA strands constructs a pair of peaks of reads on forward and reverse strands (Park, 2009) and such peak pairs help pinpointing binding sites. In contrast, CLIP-seq delivers strand specific sequences that result in a single peak of reads on the strand which RNA is transcribed from. Hence, the ChIP-seq binding site identification methods based on peak pair approach are not applicable to CLIP-seq.

Because of the distinct features of CLIP-seq data, CLIP-specific computational methods are required to gain more accurate binding site identification. However, the effects of these unique features of CLIP-seq data on protein binding site identification are not fully investigated yet. To more effectively facilitate the design of binding site identification methods for CLIP-seq, it is necessary to conduct systematic analyses to understand the difference of statistical properties between CLIP-seq and ChIP-seq and explore statistical reasons for their differences.

Meanwhile, using a matching control sample is an effective way to reduce false positives and false negatives in binding site identification. In ChIP-seq, there is a general consensus on the choice of matching control samples and the protocols to generate them (Park 2009, Landt, Marinov et al. 2012). In contrast, although the background levels are highly variable in CLIP-seq data, it is still experimentally challenging to generate appropriate controls (Ule, Jensen et al. 2005, Murigneux, Sauliere et al. 2013). Because CLIP-seq data have been found to be correlated with transcript abundance, RNA-seq data have been proposed as possible control for CLIP-seq (Darnell 2010). However, the advantage of including RNA-seq as control for CLIP-seq data analysis has not been thoroughly studied.

In this study, we conducted a comprehensive investigation of statistical aspects of binding site identifications using CLIP-seq data. First, we considered background modeling for CLIP-seq data. Second, we evaluated the performance of binding site identification methods designed for CLIP-seq. Third, through the analysis of simulated CLIP-seq and ChIP-seq data, we compared the performance in binding site identifications from these two data types, and investigated possible reasons for their different performances. We also studied the merit of including RNA-seq data as matching controls.

2. METHODS

2.1 Data preprocessing

2.1.1 CLIP-seq datasets

To study the statistical aspects of binding site identifications using CLIP-seq data, we generated an in-house CLIP-seq dataset for protein LIN-28 from *C. elegans* with high coverage. We also generated a corresponding RNA-seq dataset. (Details about data preparation/generation can be found in Supplementary Methods section in online supplement http://www.intlpress.com/SII/p/2015/8-4/SII-8-4-CHEN-supplement.pdf.) The RNA-binding protein LIN-28 is an important regulator of proper temporal succession of several developmental events in *C. elegans*. The reasons we chose this pair of CLIP-seq and RNA-seq datasets are two folds. First, LIN-28 is known to recognize a 'GGAG' motif in a sequence-specific manner to regulate target mRNAs or miRNAs (Wilbert, Huelga et al. 2012, Mayr and Heinemann 2013). Compared to CLIP-seq datasets studying Ago or Ago-like proteins that bind to several different motif seeds (Chi, Zang et al. 2009), the relatively unique motif pattern of LIN-28 provided more well-defined ground truth to assess binding site identification performance. Second, since both CLIP-seq and RNA-seq datasets were generated from the same biological samples, these RNA-seq datasets allowed us to evaluate the effect of taking RNA-seq data as a matching control for CLIP-seq data in a more unbiased way. Both CLIP-seq and RNA-seq datasets have two replicates. Since replicates are highly reproducible, we only present the results from CLIP-seq1 and RNA-seq1 for the rest of the article.

Reads from both CLIP-seq and RNA-seq experiments were mapped to the *C. elegans* genome version WS190/ce6 using Novoalign (http://www.novocraft.com/) with parameters '-F ILMFQ -t 85 -l 25 -s 1 -o SAM -r None'. Novoalign was chosen for mapping reads because it can remove adapters at the ends and allow identification of substitutions and small indels in the reads. To exclude ambiguous regions, we only considered reads mapped to exon regions. Since most genes in the *C. elegans* RefSeq database in the UCSC genome browser (http://genome.ucsc.edu/) lack UTR annotation, we extended 200bp at 5'end and 750bp at 3'end based on the known UTR length in Wormbase (http://www.wormbase.org/) and the mapped tags nearby genes (Suppl. Fig. 1). Then the overlapping exon regions were concatenated to generate the target regions for subsequent analysis. Reads mapped to the exons were extracted and summarized for each 150bp non-overlapping window. Since the CLIP-seq data were generated from strand-specific sequencing, it was summarized for the forward and reverse strands separately. Two strands were combined to generate the final count for RNA-seq data because the RNA-seq data was generated from two-stranded sequencing.

2.1.2 ChIP-seq datasets

In this study, we also included a ChIP-seq dataset as a comparison to CLIP-seq data. Because binding site identification has been well studied for ChIP-seq data (Kharchenko, Tolstorukov et al. 2008, Park 2009), we would gain a better understanding of binding site identifications using CLIP-seq data by comparing data features and identification results between CLIP-seq data and ChIP-seq data. For fair comparisons, we selected ChIP-seq data for transcription factor PHA-4 in *C. elegans* at the embryonic stage (Zhong et al., 2010). Aligned read files were downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/geo/) with accession number GSE14545. Reads were mapped using ELAND and only uniquely mapped reads were used. Both ChIP and input samples have two biological replicates and the replicates for each sample were merged. The reads were extended by 200bp to its 3'end and the number of reads was summarized for each non-overlapping 200bp window.

2.2 Background estimation

The improvements of purification and sequencing technologies have greatly increased the specificity of DNA/RNA pull-down in ChIP and CLIP experiments. However, a large portion of unwanted nucleotide sequences, which are usually called *background tags*, are still observed in sequencing data. One essential step to improve the sensitivity and specificity in binding site identification is to accurately distinguish protein binding enriched tags from background tags. The distribution of background tags could be statistically estimated using only ChIP-seq/CLIP-seq data or utilizing both ChIP-seq/CLIP-seq data and their matching controls. Currently, the optimal strategy to estimate background in CLIP-seq is still under-investigated, especially for the case that matching control experiments are utilized. Thus, we considered various background models by fitting different statistical models to the CLIP-seq dataset with and without control experiments. We also conducted similar analysis on the ChIP-seq dataset as a comparison.

2.2.1 Background estimation in CLIP-seq data without matching control

Since the majority of the tag-mapped regions in ChIP/CLIP experiments are non-binding regions, we could use the immunoprecipitation experiment itself to estimate the background tag distribution. In the case without control samples, we considered three models for background estimation: *Poisson model*, *negative binomial model* (*gamma-Poisson*) and *beta-binomial model*. In our following discussion, x is used to denote the number of reads in the window under consideration.

Poisson model Poisson model is the first distribution to describe the background tags in ChIP-seq experiments and it is still one of the most widely used statistical models (Zhang, Liu et al. 2008, Rozowsky, Euskirchen et al. 2009, Zang, Schones et al. 2009). In ideal protein binding DNA experiments, it is expected the background tags would be almost uniformly generated from the whole genome. In this scenario, the generation of random background tags could be described as a Poisson process with a single parameter λ_{bg} along the genome (Equation 1). In contrast, in CLIP-seq experiments measuring RNA binding, it is hard to expect that such uniformity assumption holds because the amount of background tags in each region would be related to the abundance of the corresponding RNA in the cells. Despite this, we still applied standard Poisson distribution to see how well (or poorly) the Poisson model fit the CLIP-seq data.

$$P(x \mid \lambda_{bg}) = \frac{\lambda_{bg}^x}{x!} e^{-\lambda_{bg}} \tag{1}$$

Negative binomial/gamma-Poisson model If we use Poisson distribution to model tag counts, we implicitly assume that variance equals to mean in the data. However, this assumption usually does not hold in high throughput immunoprecipitation experiments because of many factors in these experiments, such as non-randomness and biases of sequencing on some regions. Violation to the Poisson assumption is even more severe in other sequencing data types such as RNA-seq because the read counts vary in different regions due to transcript abundance. Such a violation to the Poisson assumption is referred as *overdispersion*. Popular strategy to handle overdispersion in the data is to allow more flexibility in the relationship between mean and variance in the distribution. In *gamma-Poisson model*, it is assumed that Poisson means follow a Gamma prior distribution instead of considering it as a fixed constant. The marginal distribution of gamma-Poisson model follows the negative binomial distribution (equation 2). Thus, negative binomial distribution (equivalent to gamma-Poisson model) could provide better fits to the over-dispersed datasets. In fact, it is commonly used to model background tags in ChIP-seq data (Ji, Jiang et al. 2008).

$$\begin{aligned} &\lambda \sim Gamma(a, b) \\ &(X \mid \lambda) \sim Poisson(\lambda) \\ &let\ \mu = ab, \alpha = 1/a \\ &p(x \mid \mu, \alpha) = \frac{\Gamma(x + \alpha^{-1})}{x!\Gamma(\alpha^{-1})} \left(\frac{\mu}{\mu + \alpha^{-1}}\right)^x \left(\frac{\alpha^{-1}}{\mu + \alpha^{-1}}\right)^{1/\alpha} \\ &\qquad = NegBin(x \mid \mu, \alpha) \end{aligned} \tag{2}$$

Beta-binomial model Instead of using Poisson distribution, uniformly distributed background tags can also be modeled by binomial distribution with constant parameter across the genome. However, the overdispersion problem still remains because a single parameter determines both mean and variance in binomial distribution. Similar to gamma-Poisson model, we can handle overdispersion within the

binomial framework by assuming that the success probability follows a Beta distribution (Skelly, Johansson et al. 2011, Zhou, Xia et al. 2011). This hierarchical beta-binomial model (equation 3) provides more flexibility in modeling the over-dispersed RNA expression data. It has been used to model tag distributions in RNA-seq data (Zhou, Xia et al. 2011).

$$\begin{aligned} &\theta \sim Beta(a,b) \\ &(X \mid \theta) \sim binomial(n,\theta) \\ &\Rightarrow X \sim Bb(a,b,n), \\ &\quad \text{where probability mass function is given as} \\ &p(x \mid a,b) = \binom{n}{x} \frac{B(x+a, n-x+b)}{B(a,b)} \end{aligned} \tag{3}$$

2.2.2 Background estimation in CLIP-seq data with matching control

One effective way to improve background estimation is to collect data from matching immunoprecipitation control experiments (Zhang, Liu et al. 2008, Kuan, Chung et al. 2011). For ChIP-seq data, input samples can be obtained by following the same protocol used to generate ChIP samples except that DNA fragments are purified without immunoprecipitation (Park 2009). Input data essentially reflects contaminated tags pulled down by other proteins and non-random background tags resulting from biases of pulling-down, sequencing and mapping. Similar biases and contamination also exist in CLIP-seq experiments. In principle, the ideal matching control could be constructed by removing the RNA binding protein of interest (for example, as a consequence of genetic mutation), or, alternatively, by using a non-specific antibody instead of one targeting the RNA-binding protein. However, in practice, such control experiments often yield a very limited number of reads to be sequenced and sequencing such a small amount of reads is prone to result in significant biases in protein binding identification. Therefore, they are usually not the preferred controls in CLIP-seq studies (Ule, Jensen et al. 2005, Murigneux, Sauliere et al. 2013). Alternatively, RNA-seq can be considered as a matching control for CLIP-seq because the CLIP-seq tag counts are also influenced by the transcript abundance in a sample, which RNA-seq is designed to measure (Uren, Bahrami-Samani et al. 2012). We also note that in addition to transcript abundance, RNA-seq can also provide local biases information, such as sequencing/mapping biases, which are usually present in CLIP-seq as well. Thus, although RNA-seq is not a direct matching control for CLIP-seq, it may offer valuable information about transcript abundance, regional biases and background tag distributions for CLIP-seq data. So we evaluated its usage as control for CLIP-seq.

Poisson/negative binomial regression model When a matching control dataset is available, regression models are often used to describe the relationship between immunoprecipitation data and matching controls (Kuan, Chung et al. 2011). We considered two models here for background estimation in CLIP-seq data with controls: Poisson regression model and negative binomial regression model (equation 4). Let x_i be the number of reads in the i-th window in CLIP-seq and r_i be the number of reads in the i-th window in the matching control (RNA-seq here).

(4)

$$\log E(X_i \mid R_i = r_i) = \log \mu(r_i) = a + b \log(r_i)$$

$$\text{Poisson regression: } p(x_i \mid r_i, a, b) = \frac{\mu(r_i)^{x_i} e^{-\mu(r_i)}}{x_i!}$$

$$NegBin \text{ regression: } p(x_i \mid r_i, a, b, \alpha)$$

$$= \frac{\Gamma(x_i + \alpha^{-1})}{x_i! \Gamma(\alpha^{-1})} \left(\frac{\mu(r_i)}{\mu(r_i) + \alpha^{-1}} \right)^{x_i} \left(\frac{\alpha^{-1}}{\mu(r_i) + \alpha^{-1}} \right)^{1/\alpha}$$

For all the models described above, parameters were estimated using the maximum likelihood estimation (MLE). Details of parameter estimation for each model can be found in Supplementary Methods (http://www.intlpress.com/SII/p/2015/8-4/SII-8-4-CHEN-supplement.pdf).

2.3 Binding site identification in CLIP-seq real data

Many computational methods have been developed to identify protein-DNA binding sites from ChIP-seq experiments (Zhang, Liu et al. 2008, Rozowsky, Euskirchen et al. 2009, Kuan, Chung et al. 2011). However, these methods cannot be directly applied to CLIP-seq data due to the non-homogeneous background tags, specific study regions (RNA only) and the unique feature of mutation information of CLIP-seq. Because of the valuable information offered by the mutations in CLIP-seq data, binding site identification in CLIP-seq can be improved by incorporating mutation information in addition to considering the tag accumulation patterns (as peak calling). We studied binding site identification in real CLIP-seq data through both mutation analysis and peak calling.

2.3.1 Binding site identification by mutation analysis

Crosslinking-induced mutations (CIMS) can facilitate accurate binding site identification. However, optimal strategies to utilize mutation information have not been fully investigated yet. In order to address this question, we examined the mutation patterns induced by cross-linking in CLIP-seq and also studied the methods to identify crosslinking induced mutations (i.e., to distinguish them from the mutations introduced by sequencing and/or mapping errors) from CLIP-seq. To identify the subtype of the mutations representing crosslinking sites, we summarized and analyzed three types of mutations—substitution, deletion, and insertion, respectively. Furthermore, to show that the sites identified using mutation information are truly related

to protein binding, we applied the same analysis to the corresponding RNA-seq data as a negative control.

To generate the mutation profiles, mutations were clustered if they were at the same position. Less confident mutation calls were pre-filtered using the following criteria because they might be errors or biases due to sequencing. First, because sequencing usually introduces errors on repeated tandem sequences (e.g., region containing a sequence of the same nucleotides, such as TTTT), we extracted the surrounding regions of mutation cluster positions and excluded those on the tandem sequences with at least 5 repeats. Second, some mutations result from PCR biases on particular regions in high-depth CLIP sequencing. To avoid PCR amplification biases, we required mutation clusters containing at least three unique mutations. In this work, unique mutation is defined as the number of reads with unique mutation pattern, in the sense of the length of the mutations (indels can be more than 1bp), the position of the mutation on the read and the strand of the read (applicable for RNA-seq). Third, for mutations longer than 1bp, only the first base was retained.

After pre-filtering, we distinguished potential crosslinking-induced mutation sites from sequencing errors using the two methods detailed below. For the first method, we ranked the mutation positions by the number of unique mutations present at the position. This is because the number of unique mutations may be a more robust measure than the total number of mutations. We call this method *rule based approach* in the following discussion. For the second method, mutation sites were ranked by the p-values from the hypothesis testing whether the proportion of reads with mutation in the position is significantly higher than that in the whole genome (equation 5). The reported p-values were adjusted for multiple testing using the Benjamini-Hochberg (BH) method (Benjamini and Hochberg 1995). This method is called *binomial test* in the following.

$$p\text{-}value(a \mid y, p) = \sum_{x \geq a} \binom{y}{x} p^x (1-p)^{y-x} \quad \text{where } p = \frac{\#\text{ of mutation type}}{\#\text{ of reads * read length}} \tag{5}$$

and a is the number of mutations at the position and y is the total number of reads mapped to that position.

We inferred which of the three mutation types is the primary one for crosslinking induced mutations in CLIP and evaluated two ranking methods for their ability to detect the known GGAG motif in the *de novo* motif analysis. Specifically, we extracted sequences with 15bp up and downstream of each mutation position using the UCSC genome browser. We then identified the *de novo* motifs from the top 500 mutations based on each ranking method (for the binomial test, we also required BH adjusted P-value ≤ 0.05) using the MEME algorithm (Bailey, Boden et al. 2009) with parameters -mod zoops -nmotifs 3 -minw 4 -maxw 8 -dna -maxsize 500000. Finally, to see the enrichment levels of motifs in each ranking method, we searched the motif identified with MEME in mutation sequences using the FIMO algorithm (Grant, Bailey et al. 2011) with parameters –output-pthresh 5e-3 –motif 1 –norc –max-stored-scores 500000.

2.3.2 Binding site identification by peak analysis

Although mutation analysis provides a *de novo* method to identify binding sites in CLIP-seq, it still cannot substitute for the whole-region peak analysis. The mutations are estimated to be induced at crosslinking sites 8% to 20% of the time depending on the protein binding factors (Zhang and Darnell 2011), so the sensitivity would be low if only mutations are used to identify binding sites. Hence, it is necessary to also consider peak analysis to increase sensitivity in protein binding site identification using CLIP-seq data. Since CLIP-seq data have the features of significantly non-uniform distribution of background tags and high correlation with RNA-seq data, we compared the binding site identification results with and without using RNA-seq as a matching control in peak analysis. These are called *one sample* (CLIP-seq only) and *two sample* (CLIP-seq VS RNA-seq) *studies*, respectively, in our following discussion. Such comparisons can directly assess the effectiveness of each method and the usefulness of including RNA-seq in CLIP-seq data analysis. Considering the background tag features of CLIP-seq, we evaluated three methods (one one-sample study and two two-sample studies) for CLIP-seq peak analysis.

One-sample study Since negative binomial distribution can well capture the overdispersion in sequencing data and the nonrandom distribution of background tags, we used a simple negative binomial test to call the peaks from the CLIP-seq dataset for one-sample analysis. Because a single distribution is used to model background tag distribution across different genomic regions, those regions having higher tag counts are more likely to be selected by this method.

Two-sample study We considered two methods to include RNA-seq data as matching controls to model background tag distributions. With these control data, the nonrandomness and non-uniformity of background tags may be incorporated by allowing distinct means and variances for different regions. We used a dynamic Poisson method to estimate local Poisson parameters from RNA-seq control for each CLIP-seq region, and a negative binomial regression method to capture the relationship between CLIP-seq and RNA-seq globally. We also considered the overdispersion of CLIP-seq data given the RNA-seq data. These two methods represent local and global estimation methods of mean and variability of background for CLIP-seq data, respectively. Because the negative binomial regression method was discussed in 'Background estimation' of the Method section (equation 4), we only describe the dynamic Poisson model here.

In the dynamic Poisson model, background Poisson mean in each window is locally estimated using the read counts in nearby windows of control samples. Note that for RNA-seq data, transcript abundance is usually summarized by exons or genes and a window may belong to more than one gene or exon region in our processed data. So we chose to use the maximum parameter from genes, exons and surrounding regions as the parameter for each window in dynamic Poisson model, as shown in equation 6.

$$
\begin{aligned}
& p(x_i \mid \lambda_i) = \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i}, \\
& \lambda_i = \max\Big[\max_{j=1,\ldots,J}(\lambda_{gj}), \max_{k=1,\ldots,K}(\lambda_{ek}), \max_{l=1,\ldots,L}(\lambda_{sl}) \Big]
\end{aligned} \tag{6}
$$

where x_i is the number of reads in the i-th window; λ_i is the Poisson parameter estimated from RNA-seq data and used to calculate p-value for the i-th window, normalized based on the total read count ratio in exon regions of CLIP-seq and RNA-seq; λ_{gj} is the parameter for gene j that the i-th window belongs to; λ_{ek} is the parameter for exon k that the i-th window belongs to; and λ_{sl} is the parameter from surrounding region l of the i-th window. The surrounding region is defined as the windows on the exon island that the i-th window under study belongs to, and the exon island is defined as the non-overlapping and concatenated exon regions on the genome. We chose λ_i to be the maximum value among all these parameters to control false positive peak identifications.

After we estimated parameters associated with the background model, the statistical evidence for peaks is calculated as $P(X > x_i \mid \Theta)$, where x_i is the read count in the i-th window and Θ are the parameter estimates. Finally, all the windows were ranked based on their p-values and the top peaks were considered for comparisons.

Assessment of peak calling methods Reliable ground truth for binding sites is essential to accurately assess the performance of peak calling methods for CLIP-seq data. In practice, true binding sites are rarely available for CLIP-seq data and the most straightforward alternative is to examine the enrichment of known motifs in the binding sites predicted by each method. However, such an approach is not satisfactory for LIN-28 because its short known motif 'GGAG' occurs all over the genome. This could be even more problematic when the resolution of binding site identification is low, as in the case of peak detection analysis, because there is a higher probability that an identified peak region contains this short motif just by chance when the region becomes wider. As a result, the occurrence of the LIN-28 motif itself is insufficient to be used as a validation for binding sites and additional criteria are required to obtain more reliable ground truth for binding sites.

Fortunately, for CLIP-seq data, the crosslinking induced mutations provide valuable information on 'gold standard' binding sites that could be used to assess the peak calling methods. Therefore, we define the 'true' binding sites to be the positions that 'GGAG' motif (more precisely, the motif identified by MEME algorithm from the top ranked mutations) was detected within 15bp up and downstream of high confident mutations. We chose this window size because most of the 'GGAG' motif was detected within 15bp from the binding sites identified by the mutation analysis. Then, the performance of binding site identification was assessed based on the proportion of top peak regions (extended by 100bp to both sides) containing these binding sites. In addition, since some random 'GGAG' motif would also be detected on high confident mutation clusters just by chance, we further randomly inserted the same number of selected high confident mutations into the regions of study and calculated the proportion of top peak regions overlapping with randomly inserted 'GGAG' motif. The final enrichment score of true binding sites for each method is defined as

$$
EnrichScore = p_{robust} - p_{random} \tag{7}
$$

where p_{robust} is the proportion of peak regions with 'true' binding sites and p_{random} is the proportion of peak regions with random motif. Finally, in order to exclude those possible artificial binding sites identified by peak analysis, we only considered the windows with more than 10 tag counts in both CLIP-seq and normalized RNA-seq data and window length equal to 150bp.

2.4 Simulation analysis

To investigate how the choice of peak calling methods (with or without RNA-seq) affects binding site identification in CLIP-seq and to examine why similar methods show different performance between CLIP-seq and ChIP-seq experiments, we performed simulation for both CLIP-seq and ChIP-seq data. To generate more realistic simulated datasets, we first fitted negative binomial regression models of each real CLIP-seq and ChIP-seq dataset on the corresponding RNA-seq data and input data, respectively. Then, we generated background tags of CLIP-seq/ChIP-seq dataset based on real RNA-seq/input controls using the negative binomial regression models with estimated regression coefficients (a,b) and dispersion parameters α. Finally, we randomly spiked in 2,000 binding sites to the simulated datasets. Here, the signal strength λ_1 for these spiked-in binding events was generated based on the estimated strength for 'true' binding sites in real CLIP-seq data. Specifically, by considering the non-uniformity of binding affinity strengths along the genome, we randomly simulated affinity values from the empirical distribution of λ_1 as follows. First, we calculated the ratio of CLIP-seq tag count versus the RNA-seq tag count for each window in regions containing high confident mutations and other background regions (i.e., regions that do not contain high confident mutations), denoted as $ratio_{binding}$ and $ratio_{background}$, respectively. Then, we derived the empirical distribution of λ_1 by estimating the q-th quantile of distribution of affinity

strength parameter λ_1 as the q-th quantile of $ratio_{binding}$ divided by the q-th quantile of $ratio_{background}$. After we obtained the empirical distribution of λ_1, the smallest 5% and the largest 5% were trimmed to avoid extreme affinity strength. Finally, the strengths of spiked-in binding sites were randomly sampled from this trimmed empirical distribution of λ_1, denoted as $\hat{F}(\lambda_1)$ (equation 8). To further exclude ambiguous regions, the affinity strength estimation, the generation of spiked-in binding sites and the analysis of simulation results were implemented using only the windows of size 150bp with CLIP-seq (original or simulated) and/or normalized RNA-seq count larger than 10. To avoid confounding effects of binding strength in the comparison between ChIP-seq and CLIP-seq data, binding affinity of the ChIP-seq data was also simulated from the empirical distribution of λ_1 estimated from the CLIP-seq data. Simulations were iterated 10 times for each scenario.

$$
\begin{aligned}
&\lambda_{1i} \sim \hat{F}(\lambda_1) \\
&Z_i \sim Bernoulli(\pi_1) \\
&\begin{cases} \text{if } \mathrm{Z_i} = 0, & X_i \sim NegBin(e^{a+b\log(r)}, \alpha) \\ \text{if } \mathrm{Z_i} = 1, & X_i \sim NegBin(\lambda_{1i} e^{a+b\log(r)}, \alpha) \end{cases}
\end{aligned} \tag{8}
$$

Performance of peak calling methods on CLIP-seq and ChIP-seq data For the performance comparison of peak calling methods, we first simulated a spiked-in dataset of which affinity strengths were sampled from the empirical distribution of λ_1 as described in the previous section to mimic realistic immunoprecipitation datasets. In addition, we also simulated a dataset with extremely high constant affinity strength ($\lambda_1' = 5 * \lambda_{1,50th \text{ quantile}}$) to see the performance of different methods when there is much clearer separation between binding and non-binding regions. We considered the same set of methods (negative binomial, dynamic Poisson and negative binomial regression) for both CLIP-seq and ChIP-seq data. For ChIP-seq data, MACS (Zhang, Liu et al. 2008) and MOSAiCS (Kuan, Chung et al. 2011) were used as implementations of the dynamic Poisson method and the negative binomial regression method, respectively. The performance of peak calling methods was assessed by both the receiver operating characteristic (ROC) curves and the proportion of true binding sites identified among the top ranking peaks predicted by each peak calling method.

Effect of dispersion parameters Although ChIP-seq and CLIP-seq technologies share many similarities, there is a critical difference in their tag generating processes for both binding site regions and background regions. Specifically, ChIP-seq tags are relatively uniformly distributed over the genome whereas CLIP-seq tags are strongly associated with transcript abundance in a sample. Statistically, such a difference in tag generating process can be described as significantly larger dispersion parameters for CLIP-seq compared to those for ChIP-seq data. Hence, to examine the effect of dispersion parameters on binding site identification, we simulated ChIP-seq and CLIP-seq datasets on which the dispersion parameters were switched while all the other parameters remained the same as in the previous simulation settings. The effect of dispersion parameters was assessed by comparing the ROC curves and the proportion of true binding sites identified in the top ranking peaks between the datasets of original and switched dispersion parameters.

Effect of biological/technical replicates As ChIP-seq protocol becomes mature, two or three biological replicates are often considered to be sufficient to obtain reliable identification of binding sites in ChIP-seq. For example, the ENCODE Consortium sets two replicates as a standard to implement ChIP-seq experiments (Landt, Marinov et al. 2012). In contrast, the necessary number of replicates for CLIP-seq has not been established in the literature. Thus, we evaluated the importance of biological/technical replicates in CLIP-seq studies by simulating datasets with various numbers of replicates (1, 2, 4, 6, 8, 10, 12, 14, and 16). For each number of replicate, the tag count values averaged over the CLIP-seq replicates was considered as the estimator for CLIP-seq tag count in each window. This simulation would give us an idea about the number of replicates required in CLIP-seq to obtain reliable binding site identification by peak analysis.

3. RESULTS

3.1 Background estimation

In this section, we considered which model would provide the best fit for CLIP-seq background data. In addition, we also examined whether RNA-seq can be used as a suitable control for CLIP-seq data and whether binding site identification could be improved by including RNA-seq data.

We first examined the relationship between CLIP-seq and RNA-seq data by comparing it with that between ChIP-seq and input data. Figures 1A and 1B show that the immunoprecipitation datasets are well correlated with corresponding controls for both ChIP-seq and CLIP-seq experiments. The correlation was 0.773 for ChIP-seq versus input (Fig. 1A), 0.455 for CLIP-seq versus RNA-seq (Fig. 1B), 0.803 among CLIP-seq replicates (Suppl. Fig. 2A) and 0.719 among RNA-seq replicates (Suppl. Fig. 2B), respectively. Such a high correlation between immunoprecipitation and input datasets is mainly due to the large amount of the background tags present in protein-binding experiments. This implies that input data and RNA-seq data can capture background information of ChIP-seq and CLIP-seq data, respectively. Compared to ChIP-seq, correlation is weaker in the CLIP-seq dataset and the correlation between ChIP-seq and input is as high as those between replicates of CLIP-seq or RNA-seq.

Figure 1A also shows that in ChIP-seq data, there are mainly two well-separated clusters. The larger cluster with lower ChIP-seq counts represents the background windows while the smaller cluster with higher ChIP counts represents

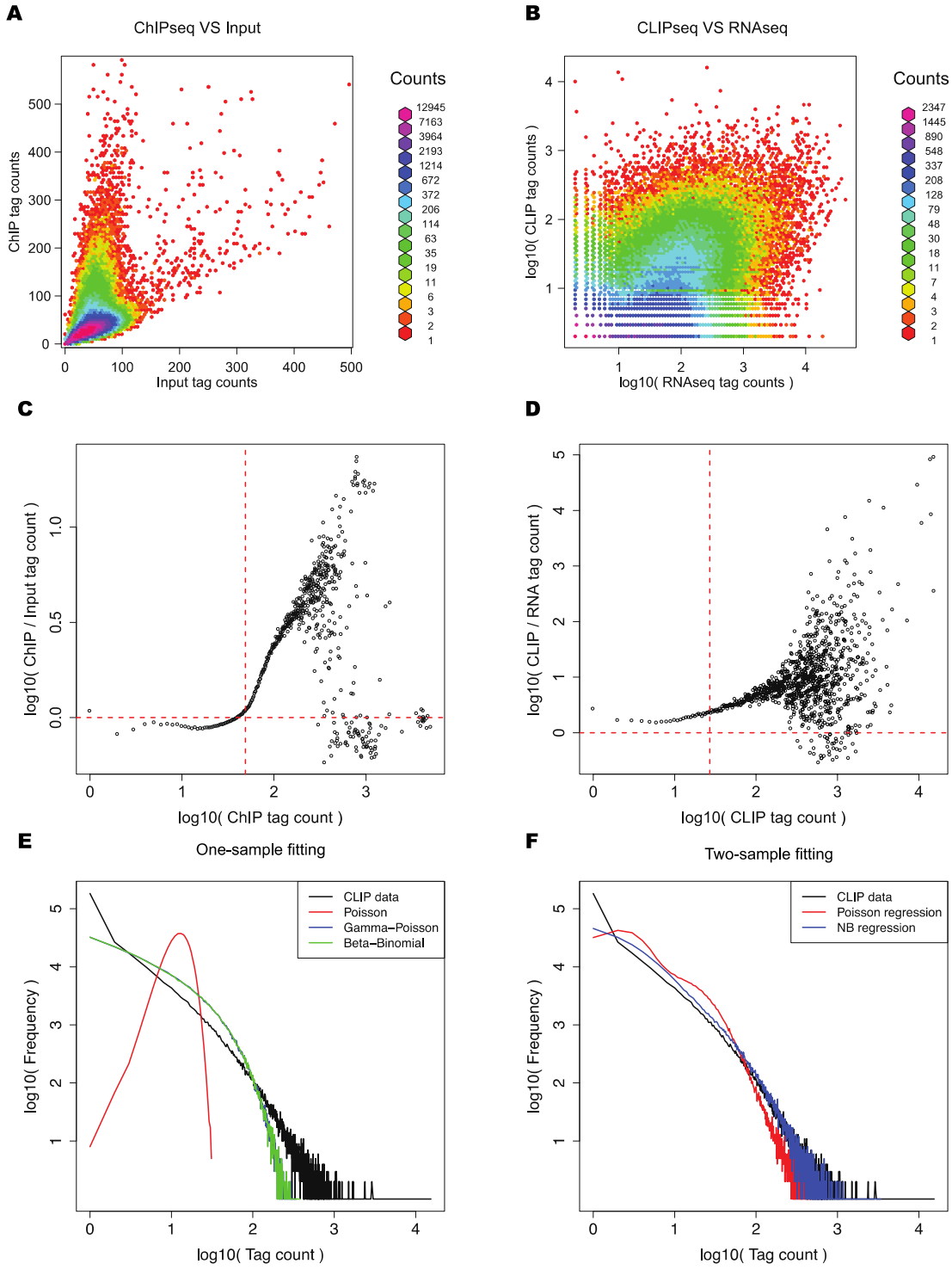


Figure 1. Background estimation and model fitting. (A) ChIP-seq tag count versus input tag count (windows with tag count $\leq$ 500). (B) CLIP-seq tag count versus RNA-seq tag count (log10 transformed counts are presented for windows with tag count $>$ 0). (C) ChIP-seq tag count versus ratio of ChIP-seq over input tag count. (D) CLIP-seq tag count versus ratio of CLIP-seq over RNA-seq tag count. In (C) and (D), log10-transformed counts are presented in x-axis and the red vertical lines indicate 90% quantile of ChIP-seq or CLIP-seq tag count. (E) CLIP-seq one-sample model fitting without RNA-seq as a matching control, using Poisson, gamma-Poisson (negative binomial) and Beta-binomial models. (F) CLIP-seq two-sample model fitting with RNA-seq as a matching control, using Poisson regression and negative binomial regression on log-transformed RNA-seq tag counts.

the potential binding sites of the protein. In contrast, Figure 1B shows that there is no such clear separation between backgrounds and binding sites for CLIP-seq vs. RNA-seq. This observation is consistent with the relatively lower correlation between CLIP-seq data and RNA-seq data (0.455). In summary, compared to input control which captures the background information of ChIP-seq data nicely, RNA-seq data provides relatively limited information on the background of CLIP-seq data. However, the statistically significant positive relationship between CLIP-seq and RNA-seq counts suggests that RNA-seq data still offer valuable information and may be used as a matching control for CLIP-seq data.

In ChIP-seq data, in the majority (∼90%) of windows with lower ChIP-seq tag counts, the counts were comparable to those in input data (Fig. 1C), which implies the background regions. In the remaining 10% of windows with higher ChIP-seq tag counts, the ratio of ChIP-seq vs. input was significantly higher than the majority of windows, which provides strong evidence for binding. In contrast, in CLIP-seq data, the ratio of immunoprecipitation over control globally increases as CLIP-seq tag counts increase (Fig. 1D) and this might imply that windows with high CLIP-seq tag counts might not correspond to binding regions. Moreover, large variability of the ratio of CLIP-seq vs. RNA-seq in the windows with high CLIP-seq counts hints that the one-sample study (CLIP only) may not be able to accurately identify the binding sites.

We fitted background distribution using the Poisson, negative binomial (gamma-Poisson), and beta-binomial models for the CLIP-seq datasets without RNA-seq data, and using Poisson regression and negative binomial regression models for the CLIP-seq datasets with RNA-seq data as controls. Figure 1E shows the results for one-sample model fitting. As expected, the results indicate that the background could be significantly underestimated if we use the Poisson model (BIC = 17,526,877; Table S1), of which mean and variance are assumed to be identical. This implies that the Poisson model is not suitable for CLIP-seq data with large variance due to transcript abundance. In contrast, both negative binomial (BIC = 2,289,642) and beta-binomial (BIC = 2,289,658) models fit the background almost equally well by appropriately accounting for the overdispersion. These results are consistent with the statistical theory that beta-binomial model is asymptotically equivalent to gamma-Poisson (negative binomial) model when we have large total counts and small success probability (Skellam 1948). This is actually the case for CLIP-seq and RNA-seq data analyzed here because we do have large total counts since both data sets are very deeply sequenced and the proportion for each transcript is very small as *C.elegans* has about 20,000 genes.

Figure 1F indicates that incorporating RNA-seq using regression models can improve background estimation. Among all the regression models under consideration, negative binomial regression with log-transformed RNA-seq tag counts (BIC = 2,040,052) provides the best model fit. We also obtained similar conclusions for ChIP-seq data (Suppl. Fig. 2C and D) and this is consistent with the knowledge that ChIP-seq and CLIP-seq have similar data generating processes. However, closer inspection of the fitted regression lines of immunoprecipitation data vs. control data (Suppl. Fig. 2E and F) still indicates that ChIP-seq shows better separation of binding sites and backgrounds when the controls are incorporated in background estimation.

In summary, our correlation, ratio trend, and model fitting analyses for both ChIP-seq and CLIP-seq data indicate that (1) although RNA-seq is not a perfect control for CLIP-seq data compared to input data for ChIP-seq data, it is still clearly beneficial to incorporate corresponding RNA-seq datasets in the analysis of CLIP-seq data, because it can improve the background fitting of CLIP-seq and help identify the binding sites in the regions with relatively low CLIP-seq counts; (2) negative binomial regression models on log-transformed RNA-seq tag counts provide the best model fits to the background tags of CLIP-seq data.

3.2 Binding site identification in real data

3.2.1 Binding site identification by mutation analysis

To identify binding sites from CLIP-seq, we started with the unique feature of CLIP-seq—crosslinking induced mutations—to identify binding sites. In order to characterize subtypes of mutations related to binding events, we examined mutation frequencies for each mutation type (substitution, deletion, and insertion) in CLIP-seq data. We also applied the same analysis to RNA-seq data in order to estimate the mutation frequencies when there are no crosslinking effects. The UV crosslinking process has changed mutation profiles by increasing the total number of mutations, especially the proportion of deletions among three mutation types (Table 1). In non-crosslinked RNA-seq, there was about equal numbers of deletions (3.1% of total mutations) and insertions (3.9%). In contrast, there was a clear preference for deletions (5.8% for CLIP-seq1 and 28.4% for CLIP-seq2) compared to insertions (0.6% for CLIP-seq1 and 2.1% for CLIP-seq2) in CLIP-seq data. Furthermore, the proportion of deletions increased most rapidly among three mutation subtypes as CLIP-seq is more deeply sequenced (Table 1). This suggests that deletions are particularly enriched by crosslinking effects.

We further evaluated the preference for different mutation types by examining the mutation clusters. Mutations were clustered by their positions in exon regions and filtered by requiring at least 3 unique mutations in each cluster and tandem nucleotides with a length shorter than 5bp at surrounding sequences. The mutation clusters showed preferential enrichment for deletions as well (Table 1). The proportions of deletions in clustered data (3.5% and 6.1%) from two CLIP-seq replicates were more comparable than those in the total counts of mutations (5.8% and 28.4%). This

Table 1. Mutation statistics in CLIP-seq and RNA-seq. Numbers in the parenthesis are the proportions of each type of mutation for each dataset

		All Mutations		
Dataset	**Mapped reads**	**Substitutions**	**Deletions**	**Insertions**
CLIP-seq1	5,087,544	1,223,405 (0.937)	75,400 (0.058)	7,272 (0.006)
CLIP-seq2	156,886,622	18,493,981 (0.695)	7,563,806 (0.284)	566,672 (0.021)
RNA-seq	26,467,641	1,175,157 (0.931)	38,519 (0.031)	49,116 (0.039)
		Mutation clusters in exon regions **Clusters with unique mutations > 2 & tandem nt < 5**		
Dataset		**Substitutions**	**Deletions**	**Insertions**
CLIP-seq1		24,467 (0.964)	897 (0.035)	27 (0.001)
CLIP-seq2		436,193 (0.937)	28,342 (0.061)	1,015 (0.002)
RNA-seq		37,219 (0.986)	195 (0.005)	319 (0.008)

suggests that higher sequencing depth not only improves the sensitivity of mutation cluster detection (more clusters detected), but also provides increased evidence for deletion cluster identification (clusters have higher numbers of mutations).

To further evaluate the mutation types induced by crosslinking, we performed de novo motif searching by MEME for the top 500 mutations ranked by binomial tests. The 'GGAG' motif was clearly identified near deletion sites with a significant E-value equal to 2.4e-56 (Fig. 2A and Suppl. Fig. 3A), which further suggests that deletion is specifically induced by UV crosslinking. However, interestingly, 'GGAG' was also identified from the substitution sites as the second most significant motif (E-value 2.2e-36) from MEME in CLIP-seq1 (Fig. 2A) and as the most significant motif (E-value 9.4e-51) in CLIP-seq2 (Suppl. Fig. 3A). Hence, although enrichment of substitution clusters was not as obvious as enrichment of deletion clusters in general, they do harbor a significant amount of crosslinking induced mutations other than natural variations and sequencing errors. A clear motif identified from substitutions in CLIP-seq with deeper sequencing depth also implies the importance of sequencing depth for a higher confidence calling of crosslinking induced mutations.

Because mutations in different regions were called with different levels of statistical significance, we further investigated the enrichment level of 'GGAG' motif in mutations called at various confidence levels. Figure 2B shows that when we ranked mutations using binomial tests, there was much higher enrichment of motif containing sites at higher confidence levels (35% to 45%), which suggests that higher confident mutations are more likely corresponding to binding sites. Specifically, both deletions and substitutions with higher confidence (about top 1,000 mutations) presented 35%–45% enrichment of the motif although substitutions were overwhelmed by sequencing errors and/or natural variations at lower confident levels. In contrast, if we ranked mutations using the rule based approach, there was no more motif enrichment in highly ranked mutation regions. These results suggest that the binomial test may better capture signals from crosslinking than the rule based approach. In Figure 2C, non-crosslinked RNA-seq showed only random motif enrichment about 20% at all confident levels and this indicates that the sites identified from CLIP-seq using mutation information are more likely to be related to protein binding.

Finally, we assessed the resolution of mutation analysis using the distance between deletions and the 'GGAG' motif identified within 600bp flanking regions of deletions. Figure 2D indicates that the motif is mostly enriched near mutations (estimated resolution: ∼30–40bp), which suggests that mutations could identify protein-binding sites with high resolution.

3.2.2 Binding site identification by peak analysis

High-resolution mutation analysis provided us a list of high confident putative binding sites for protein LIN-28. We selected the top 897 deletion and 1,000 substitution sites containing 'GGAG' motif within the 30bp flanking sequences from the binomial test mutation analysis as the ground truth for binding sites to assess the performance of peak calling methods. We first mapped these high confident binding sites to corresponding windows on the genome and examined the CLIP-seq tag counts and the ratios of CLIP-seq vs. RNA-seq tag counts in these windows (Fig. 3A). In contrast to ChIP-seq data of which binding sites are located on high immunoprecipitation count and ratio regions and well separable from background regions, in CLIP-seq, there is no clear separation of the mutation-defined binding sites from the background, although there is a tendency towards higher ratio regions. This suggests that identifying binding regions by peak analysis could be more challenging for CLIP-seq because peak analysis usually finds the binding sites based on CLIP-seq tag count or the ratio between CLIP-seq and control counts.

We applied the negative binomial model, the dynamic Poisson model, and the negative binomial regression model to identify binding sites. These three methods represent one-sample analysis, two-sample analysis with local parameter estimation, and two-sample analysis with global parameter

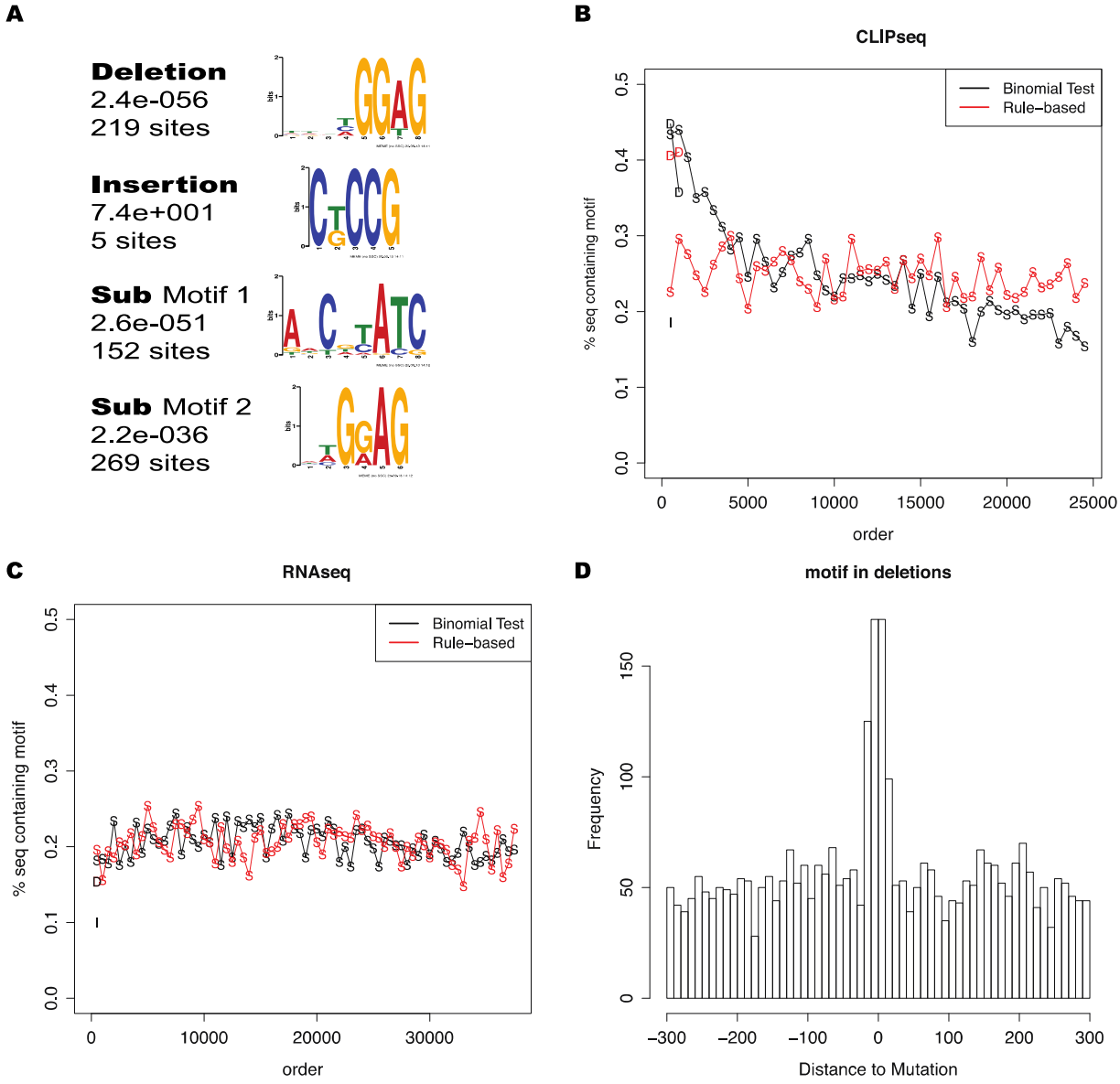


Figure 2. Binding site identification by mutation analysis. (A) Motif identified by MEME for each mutation type of CLIP-seq (the first two motifs are shown for substitutions). (B) Enrichment of motif in the ordered mutations of CLIP-seq, identified by the binomial test (black) and the rule based approach (red) for each substitution, deletion and insertion, marked as S, D and I respectively. The interval is 500 mutations. (C) Enrichment of motif in the ordered mutations of RNA-seq. (D) Motif enrichment according to the distance from the position of the motif to the mutations, which represents the resolution of binding site identification by mutation analysis.

estimation, respectively. As expected, the top 1,500 putative binding sites identified by the negative binomial model corresponded to posing a hard threshold on CLIP-seq tag counts (Fig. 3B). There was a large overlap between the top 1,500 binding sites identified by the dynamic Poisson model and the negative binomial regression (Fig. 3C, D), as they both used information from RNA-seq as controls. With RNA-seq data as a control, we could identify about 20%∼40% novel binding sites that are "true" binding sites defined by high confidence mutations but were missed by one-sample negative binomial method. This again suggests the importance of using RNA-seq control in CLIP-seq peak analysis. We found that in the dynamic Poisson model, due to its local enrichment feature, some windows with high CLIP-seq counts and high ratios could be excluded from the top ranking peaks if they come from high abundant transcripts and their tag counts are not significantly higher than tag counts of other windows in the same transcript/exon.

The top binding sites from peak analysis (Fig. 3B, C and D) only cover a subset of the high confident mutation-

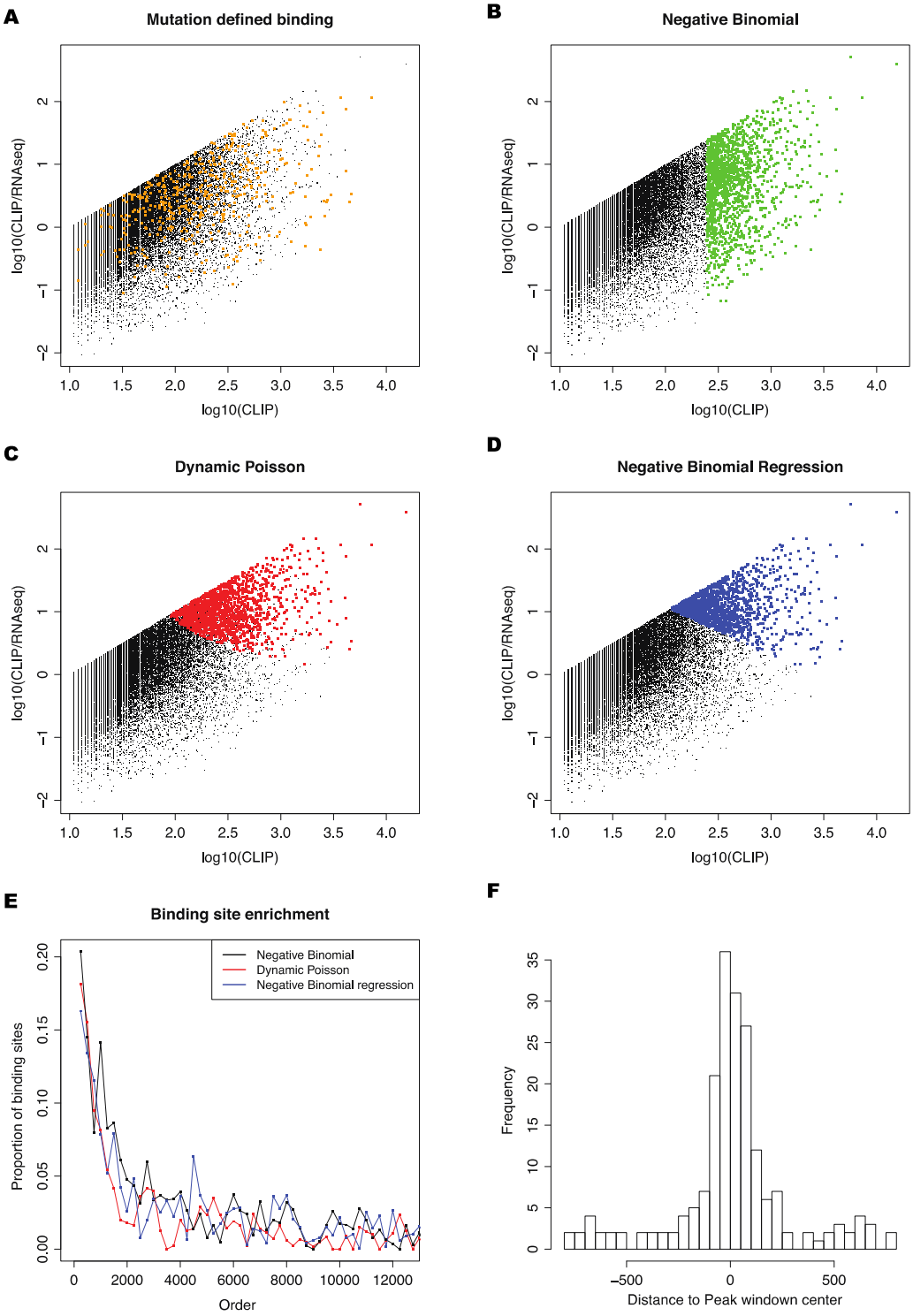


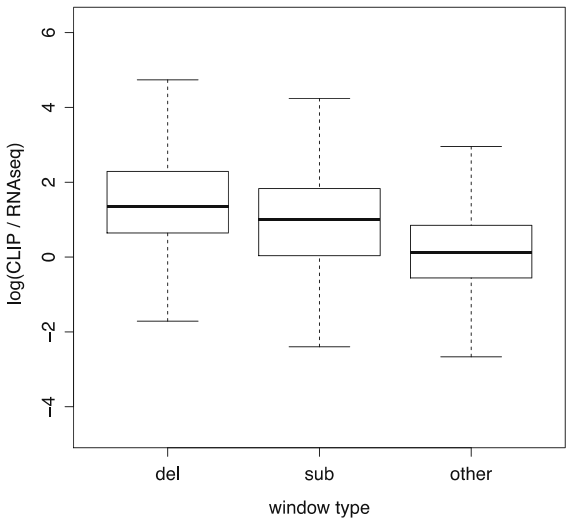
Figure 3. Binding site identification by peak analysis. (A) Distribution of binding sites defined by high confidence mutations on the scatter plot of ratio of CLIP-seq over RNA-seq count vs. CLIP-seq count. (B) Top 1,500 peak regions identified by negative binomial method. (C) Top 1,500 peak regions identified by dynamic Poisson method. (D) Top 1,500 peak regions identified by negative binomial regression method. (E) Binding site enrichment in ordered peak regions identified by each of three methods (each interval contains 250 peaks). (F) Binding site enrichment according to the distance from the position of the motif to peak window centers for top 1,500 peaks (only results for negative binomial regression method shown), which represents the resolution of binding site identification by peak analysis.

defined binding sites shown in Figure 3A. This is also evident by the almost identical binding site enrichment levels observed in the top ranking peaks identified by the three methods (Fig. 3E). Weak separation between binding sites and backgrounds in CLIP-seq count and ratio of CLIP-seq vs. RNA-seq counts resulted in indistinguishable performance of peak analysis methods and made it challenging to identify binding sites with low CLIP-seq count and/or ratio values. However, Figures 3E and 3F reveal that the peak analysis with/without RNA-seq control should still be included as an essential part of CLIP-seq data analysis. Figure 3E indicates that high quality binding sites are clearly enriched in the top peak lists identified by each of the three methods, especially for the top 1,500 to 2,000 windows predicted by each method. Hence, the peak analysis will enable us to detect binding sites of a strong signal missed from the mutation analysis that can identify only about 8∼20% of true binding sites. Figure 3F indicates that peak analysis can identify positions of binding sites quite accurately (resolution ∼300bp) although the resolution is still significantly lower than that of mutation analysis (∼30–40bp).

3.3 Simulations

The peak analysis using real CLIP-seq data showed that there is only weak separation between binding sites and background regions and the peak analysis methods identifying binding sites successfully in ChIP-seq could identify only a small subset of binding sites in CLIP-seq. In this section, we conducted a series of simulation studies to systematically investigate the differences in binding site identification between CLIP-seq and ChIP-seq and the factors contributing to these differences.

In Background Estimation section, the negative binomial regression model had the best fit for CLIP-seq/ChIP-seq when taking RNA-seq/input as a covariate. Thus, we simulated CLIP-seq/ChIP-seq datasets from the negative binomial regression model based on real RNA-seq/input. To mimic real binding sites, binding affinity (λ_1) was simulated from the empirical distribution of λ_1 that was essentially estimated as the ratio of CLIP-seq count vs. RNA-seq count in the regions with high confidence mutations when the ratio in background regions was used as a baseline. Figure 4 shows the distribution of the ratio of CLIP-seq count vs. RNA-seq count for (1) binding regions defined by deletion, (2) binding regions defined by substitution, and (3) other background regions. Although signal enrichment of binding sites is relatively weak compared to background regions, it is clear that ratio of CLIP-seq over RNA-seq counts is higher for both deletion- and substitution-defined binding sites compared to background regions, which again implies that taking RNA-seq as a control is helpful for peak analysis. The binding affinity (λ_1) estimated from real CLIP-seq ranged from 1.77 to 3.71 (Table in Fig. 4) and this empirical distribution of λ_1 was used to generate simulated data of both CLIP-seq and ChIP-seq to make the results more



Quan	10%	20%	30%	40%	50%	60%	70%	80%	90%
Lamda1	1.77	2.125	2.386	2.515	2.75	2.871	3.117	3.433	3.711

Figure 4. Estimation of binding affinity (lamda1,λ_1) from binding sites identified by mutation analysis in CLIP-seq. Boxplots show the distribution of log-transformed ratio of CLIP-seq count vs. RNA-seq count for top deletion-defined binding sites, top substitution-defined binding sites and other background regions. The table provided the 10%∼90% quantiles of estimated λ_1 values.

comparable between them. Note that our simulation was based on the negative binomial regression model and did not consider exon relationships in genes. As a result, the results would favor peak analysis using the negative binomial regression method. However, this will not significantly affect the comparison of peak analysis performance between CLIP-seq and ChIP-seq data.

3.3.1 Performance of peak calling methods on CLIP-seq and ChIP-seq

We first considered a set of simulated CLIP-seq and ChIP-seq datasets which best mimic the realistic data by randomly generating binding affinity from the empirical distribution of λ_1. Hence, signal enrichment of binding sites compared to background regions is very close to those in real CLIP-seq data. As in real CLIP-seq data, there is no clear separation between binding regions and background regions in simulated CLIP-seq data (Suppl. Figs. 5A), while ChIP-seq showed much better separation (Suppl. Figs. 4A). To compare the performance of peak calling methods on CLIP-seq and ChIP-seq, three calling methods (negative binomial, dynamic Poisson and negative binomial regression) were applied to simulated datasets. Figure 5A indicates that for ChIP-seq, the negative binomial regression method clearly outperformed the other two methods. The area under the curve (AUC) for negative binomial regression was 0.953 and above 0.85 for the other two methods (Table 2). In contrast,

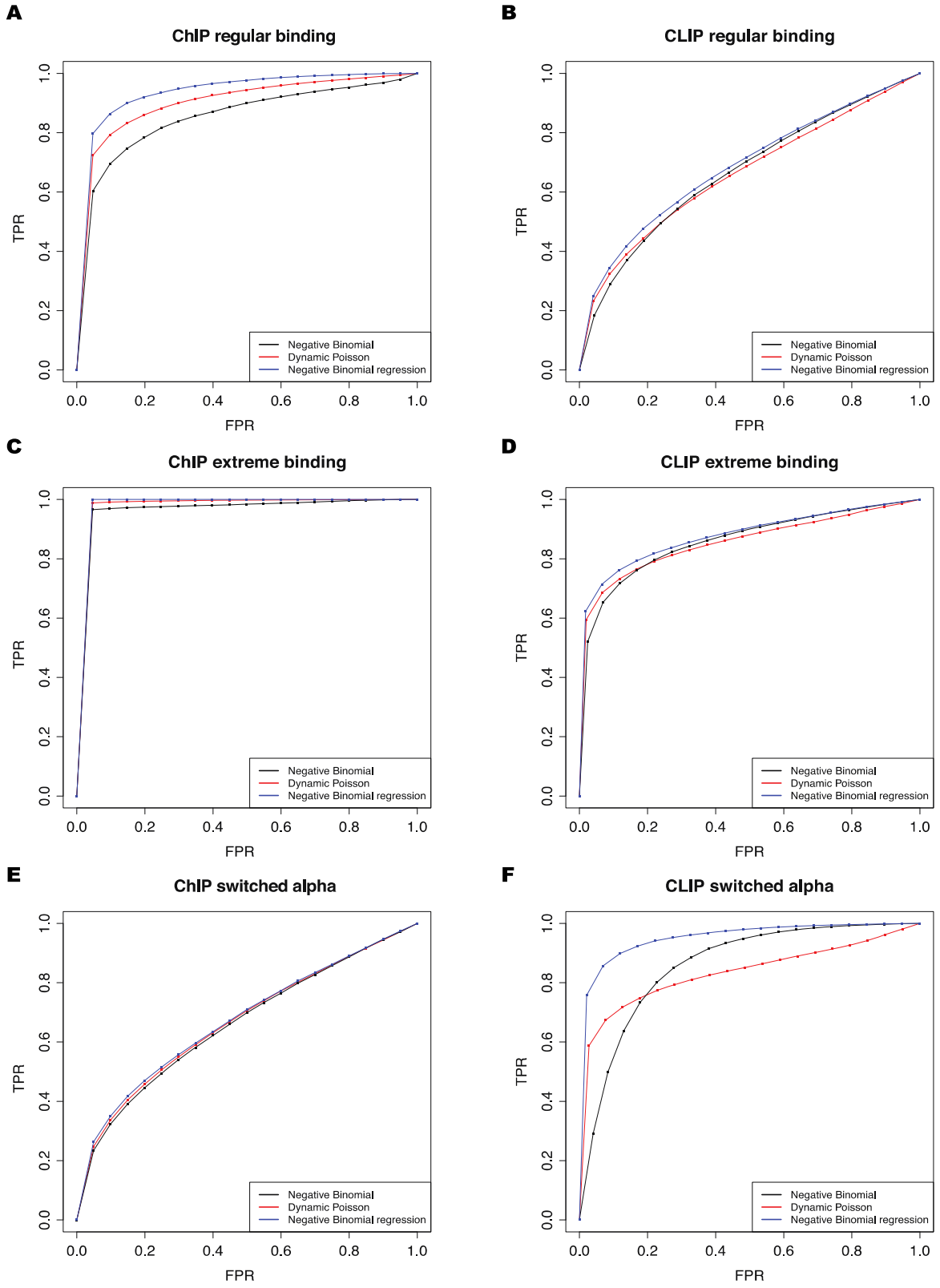


Figure 5. Performance (ROC curves) of peak calling methods on simulated datasets with regular binding affinity (A and B) and extremely strong binding affinity (C and D) and for the case that dispersion parameters for ChIP and CLIP are switched (E and F). A, C, and E are ChIP-seq results and B, D, and F are CLIP-seq results.

Table 2. Dispersion parameters and AUC calculation for simulated datasets

Dataset	Dispersion parameters (α)	Method	Regular binding (AUC)	Extreme binding (AUC)	α switched (AUC)
CLIP-seq	1.529	Negative Binom	0.667	0.864	0.859
		Dynamic Poisson	0.663	0.860	0.839
		NB regression	0.688	0.882	0.957
ChIP-seq	0.115	Negative Binom	0.868	0.985	0.653
		Dynamic Poisson	0.915	0.997	0.657
		NB regression	0.953	1.000	0.662

these three methods performed comparably for CLIP-seq datasets (Fig. 5B) and their AUC values (∼0.65) were much lower than those of ChIP-seq datasets (Table 2).

To see how peak analysis would perform for proteins with extremely high binding affinity, we simulated datasets of which binding affinity is five times stronger, i.e., $5 * \text{median}(\lambda_1) = 13.75$. Figure 5C shows that in ChIP-seq, all three methods identified most binding sites as their top ranking ones and their AUC values were close to 1 (Table 2). In contrast, AUC values were still only about 0.85 for CLIP-seq (Fig. 5D, Table 2), which is close to that of ChIP-seq datasets with moderate affinity. In summary, although the data generating processes of CLIP-seq and ChIP-seq are quite similar, the peak calling methods perform significantly worse for CLIP-seq compared to ChIP-seq data. Moreover, for CLIP-seq, even extremely high binding affinity cannot make the binding sites totally distinguishable from background regions.

3.3.2 Effect of dispersion parameters

As most of the parameters for data simulation were controlled to be the same between simulated CLIP-seq and ChIP-seq data, we found that the major difference between RNA and DNA-binding can be formulated as the differences in dispersion parameters of the negative binomial distribution between CLIP-seq (1.529) and ChIP-seq (0.115) (Table 2). In order to see if the dispersion could statistically explain the differences in peak analysis results between ChIP-seq and CLIP-seq, we simulated CLIP-seq and ChIP-seq data by switching dispersion parameters between them. As expected, after dispersion parameters were switched, the peak analysis results showed exactly opposite patterns (Fig. 5E and F). Specifically, the AUCs for ChIP-seq with CLIP-seq dispersion parameter were only about 0.65 while the AUCs for CLIP-seq with ChIP-seq dispersion parameter were above 0.85 (Table 2). This suggests that the large variance/dispersion in CLIP-seq data might be the major reason for peak analysis challenge.

3.3.3 Effect of biological/technical replicates

Considering the large variation for transcript abundance, the large variance/dispersion in both CLIP-seq and RNA-seq datasets is inevitable. One way to reduce variation is to increase the number of either biological or technical replicates. To examine how many biological/technical replicates are necessary to achieve satisfactory results in peak analysis, we simulated CLIP-seq data with different replicate numbers. Figure 6 shows that the statistical power to identify true binding sites increases significantly as we increase the number of replicates, in the case of large dispersion parameters in CLIP-seq. Under the simulation settings we considered here, there is no significant improvement in performance when more than 10 replicates were used. We could achieve an AUC higher than 0.9 with 8 replicates with the negative binomial regression method in peak analysis. We note that this AUC level corresponds to that for only one replicate in ChIP-seq and this implies that larger numbers of replicates are needed for CLIP-seq in order to handle the larger dispersion compared to ChIP-seq.

4. DISCUSSION

As CLIP-seq is emerging as a major technique for the study of RNA-protein interactions *in vivo*, it becomes more critical to develop appropriate statistical methods to infer binding sites through the analysis of CLIP-seq data. In this article, we reported a thorough investigation of various factors affecting binding site identification in CLIP-seq using both real data and simulation studies. Our results suggest that the negative binomial regression using RNA-seq as controls may be a preferred approach for modeling CLIP-seq data, and the mutation analysis offers valuable information to improve specificity and resolution in binding site identification. We found that larger dispersion could be one of the main sources of difficulties in binding site identification for CLIP-seq, which points to possible directions to improve binding site identification for CLIP-seq. We also showed that in order to address such a large dispersion issue in CLIP-seq and achieve higher sensitivity and specificity in binding site identification, it is important to have a larger number of biological/technical replicates than ChIP-seq.

Our results suggest that it is critical to use both mutation and peak analysis to accurately identify binding sites in CLIP-seq data and these two methods cannot substitute each other. On one hand, peak analysis provides higher sensitivity for binding site identification but suffers from lower resolution (about 300bp) and specificity. On the other hand, mutation analysis identifies binding sites with high resolution (∼30–40bp) and high specificity but suffers from lower sensitivity as mutations are only induced near 8–20% of

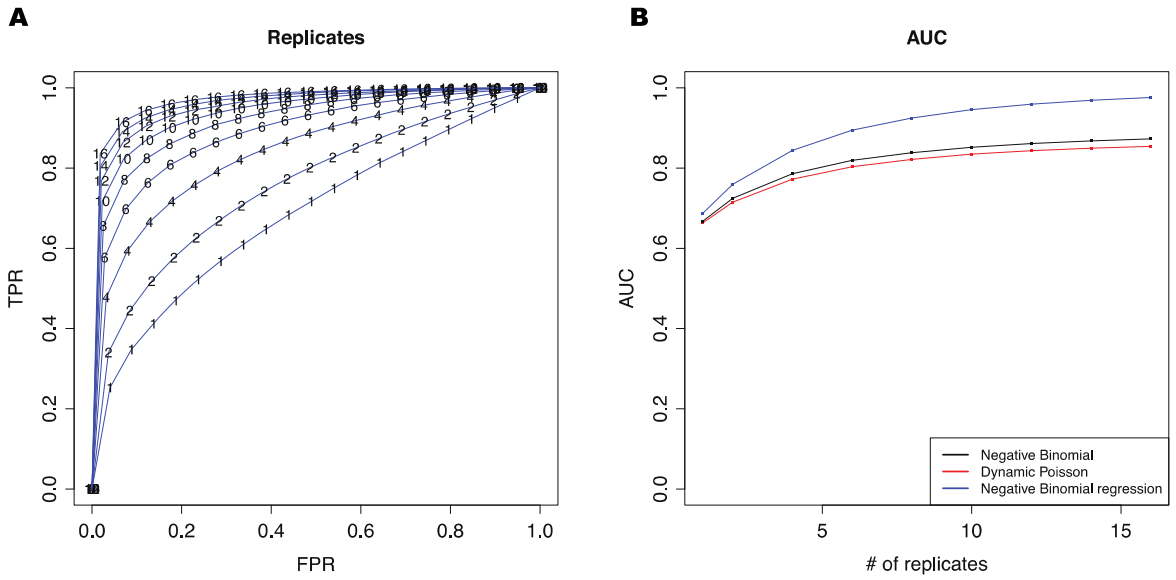


Figure 6. Effect of number of replicates on binding site identification in CLIP-seq. (A) ROC curves for negative binomial regression methods in simulated datasets with 1, 2, 4, 6, 8, 10, 12, 14 and 16 replicates. (B) AUC for different number of replicates on negative binomial, dynamic Poisson and negative binomial regression methods.

the binding sites. Meanwhile, in LIN-28 protein study, if we combined both peak analysis (top 1,500 sites in negative binomial regression model) and mutation analysis (1897 high confident deletions/substitutions), we could identify 447 common binding sites. Such a combined approach improved the enrichment of motif among identified binding sites (50.6% of these overlapped sites were with GGAG motif on mutation positions) compared to mutation analysis alone (42.3%). These observations imply that integration of peak analysis and mutation analysis might improve binding site identification and achieve a good balance among sensitivity, specificity, and resolution.

As CLIP-seq tag counts are partially dependent on transcript abundance, RNA-seq data might be treated as a matching control to infer binding sites. However, RNA-seq is not specifically designed as a matching control for CLIP-seq and there has been no convincing evidence for the benefits of its inclusion. In our study, we systematically investigated this issue by fitting the CLIP-seq data with or without RNA-seq control and calling binding sites with one-sample or two-sample methods. We found that it might be preferable to use RNA-seq as a control for CLIP-seq. First, RNA-seq tag counts are correlated with CLIP-seq, which could provide the information about background noises coming from different transcript abundance and/or other local biases. Second, regression models including RNA-seq as a matching control improved estimation of distribution of background tags. Third, two-sample analysis may identify some potentially true binding sites that are missed by one-sample analysis.

In our study, we have ignored the isoform issue by concatenating overlapped exon regions. However, unlike the whole genome study performed by ChIP-seq, the analysis of CLIP-seq data must take into account the existence of transcript isoforms generated by alternative splicing. It is therefore of interest to investigate the effect of the differential expression of transcript isoforms on binding site identification and to integrate the inference of the isoform abundance to the binding site identification.

ACKNOWLEDGEMENTS

We thank Yale University Biomedical High Performance Computing Center (YHPC) for data storage and computation runs. XC was supported by the Lo Graduate Fellowship for Excellence in Stem Cell Research and a fellowship from the China Scholars Council. DC was supported by NIH P01 CA154295. GS was supported by the State of Connecticut Grant 12-SCA-Yale-02. FJS was supported by GM064701. HZ was supported by NIH R01 GM59507. YHPC was funded by NIH RR19895. There is no conflict of interest.

Received 13 December 2013

REFERENCES

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). "MEME SUITE: tools for motif discovery and searching." *Nucleic. Acids. Res.* **37**(Web Server issue): W202–208.

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society Series B (Methodological)* 289–300. MR1325392

Chi, S. W., Zang, J. B., Mele, A., and Darnell, R. B. (2009). "Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps." *Nature* **460**(7254): 479–486.

Darnell, R. B. (2010). "HITS-CLIP: panoramic views of protein-RNA regulation in living cells." *Wiley Interdiscip Rev RNA* **1**(2): 266–286.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). "FIMO: scanning for occurrences of a given motif." *Bioinformatics* **27**(7): 1017–1018.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). "PAR-CliP—a method to identify transcriptome-wide the binding sites of RNA binding proteins." *J. Vis. Exp.* (41).

Ji, H., Jiang, H., Ma, W., Johnson, D. S., Myers, R. M., and Wong, W. H. (2008). "An integrated software system for analyzing ChIP-chip and ChIP-seq data." *Nat. Biotechnol.* **26**(11): 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**(12): 1351–1359.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (2011). "A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins." *Nat. Methods* **8**(7): 559–564.

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2011). "iCLIP-transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution." *J. Vis. Exp.* (50).

Kuan, P. F., Chung, D., Pan, G., Thomson, J. A., Stewart, R., and S. Keleş (2011). "A statistical framework for the analysis of ChIP-Seq data." *Journal of the American Statistical Association* **106**(495): 891–903. MR2894745

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. (2012). "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." *Genome Res.* **22**(9): 1813–1831.

Mayr, F. and Heinemann, U. (2013). "Mechanisms of Lin28-mediated miRNA and mRNA regulation—a structural and functional perspective." *Int. J. Mol. Sci.* **14**(8): 16532–16553.

Murigneux, V., Sauliere, J., Roest Crollius, H., and Le Hir, H. (2013). "Transcriptome-wide identification of RNA binding sites by CLIP-seq." *Methods* **63**(1): 32–40.

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

Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. B. (2009). "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls." *Nat. Biotechnol.* **27**(1): 66–75.

Skellam, J. G. (1948). "A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials." *Journal of the Royal Statistical Society. Series B (Methodological)* **10**(2): 4. MR0028539

Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J. M. (2011). "A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data." *Genome Res.* **21**(10): 1728–1737.

Ule, J., Jensen, K., Mele, A., and Darnell, R. B. (2005). "CLIP: a method for identifying protein-RNA interaction sites in living cells." *Methods* **37**(4): 376–386.

Uren, P. J., Bahrami-Samani, E., Burns, S. C., Qiao, M., Karginov, F. V., Hodges, E., Hannon, G. J., Sanford, J. R., Penalva, L. O., and Smith, A. D. (2012). "Site identification in high-throughput RNA-protein interaction data." *Bioinformatics* **28**(23): 3013–3020.

Wang, Z., Tollervey, J., Briese, M., Turner, D., and Ule, J. (2009). "CLIP: construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo." *Methods* **48**(3): 287–293.

Wilbert, M. L., Huelga, S. C., Kapeli, K., Stark, T. J., Liang, T. Y., Chen, S. X., Yan, B. Y., Nathanson, J. L., Hutt, K. R., Lovci, M. T., Kazan, H., Vu, A. Q., Massirer, K. B., Morris, Q., Hoon, S., and Yeo, G. W. (2012). "LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance." *Mol. Cell* **48**(2): 195–206.

Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). "A clustering approach for identification of enriched domains from histone modification ChIP-Seq data." *Bioinformatics* **25**(15): 1952–1958.

Zhang, C. and Darnell, R. B. (2011). "Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data." *Nat. Biotechnol.* **29**(7): 607–614.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). "Model-based analysis of ChIP-Seq (MACS)." *Genome Biol.* **9**(9): R137.

Zhou, Y. H., Xia, K., and Wright, F. A. (2011). "A powerful and flexible approach to the analysis of RNA sequence count data." *Bioinformatics* **27**(19): 2672–2678.

Xiaowei Chen
Program in Computational Biology and Bioinformatics
Yale University
New Haven, CT
USA

Dongjun Chung
Department of Biostatistics
Yale School of Public Health
New Haven, CT
USA

Giovanni Stefani
Department of Molecular, Cellular and Developmental Biology
Yale University
New Haven, CT
USA

Centre for Integrative Biology
University of Trento
Trento
Italy

Frank J. Slack
Department of Molecular, Cellular and Developmental Biology
Yale University
New Haven, CT
USA

Hongyu Zhao
Program in Computational Biology and Bioinformatics
Yale University
New Haven, CT
USA

Department of Biostatistics
Yale School of Public Health
New Haven, CT
USA

Department of Genetics
Yale School of Medicine
New Haven, CT
USA
E-mail address: hongyu.zhao@yale.edu