Current methods in the analysis of CLIP-Seq data

Kai Fu, Bioinformatics Program,University of California Los Angeles

RNA-binding proteins play important roles in the post-transcriptional regulation. CLIP-Seq has become a popular method to measure genome-wide RNA-protein binding sites. In this review, I describe the general workflow for three types of CLIP-Seq procedures. I also describe the computational and statistical challenges for analyzing CLIP-Seq datasets. Finally, I highlight current computational/statistical methods in the identification of RNA-protein interaction sites.

Categories and Subject Descriptors: H.5.2 [**Computational Biology**]: High-throughput sequencing—*Review / methodology*

Additional Key Words and Phrases: Computational Biology, CLIP-Seq, Statistical methods

ACM Reference Format:

Daniel Pineo, Colin Ware, and Sean Fogarty. 2010. Neural Modeling of Flow Rendering Effectiveness. *ACM Trans. Appl. Percept.* 2, 3, Article 1 (May 2010), 7 pages.
DOI : http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

During the process of transcription and post transcription, RNA synthesis and modifications such as alternative splicing is mainly regulated through the binding of functional proteins. Thus, to understand the fate and the molecular function of RNA, one of the key tasks is to map the binding sites between RNA-binding proteins and RNAs[1]. Regard of interactions between RNA and RNA-binding proteins, interactions between DNA and DNA-binding proteins has been investigated in depth using chromatin immunoprecipitation (ChIP) following by high-throughput sequencing. A large number of transcription factor binding sites have been mapped in various cell lines. These resources provide us a more and more clear view about how the genomic DNA is controlled by the binding of DNA specific proteins.

Inspired by the success of ChIP-Seq, emerging studies are now focusing on the interactions between RNA and RNA-binding proteins. Crosslinking immunoprecipitation followed by high-throughput sequencing (CLIP-Seq) then becomes quite popular in the RNA genomics field [2]. CLIP method was first performed to study interactions between the neuron-specific RNA-binding protein and splicing factor NOVA1 and NOVA2 in Darnell RBs lab at Rockefeller University[3]. In 2008, CLIP was combined with high-throughput sequencing to generate genome-wide protein-RNA interactions maps. This is almost at the same time when ChIP was combined with high-throughput sequencing. Since then a number of RNA-binding protein maps have been generated, especially for the splicing factors. In this paper, I will review CLIP-Seq technology and its applications, focusing on the computational analysis strategy and statistical models and methods that been used.

2. WORKFLOW OF CLIP-SEQ

Based on different library preparation protocols, there are mainly three types of CLIP-Seq: HITS-CLIP, PAR-CLIP and iCLIP (Fig1). HITS-CLIP begins with the in-vivo cross-linking of RNA-protein complexes using ultraviolet light. Upon ultraviolet exposure, covalent bonds are formed between proteins and RNA that are in close interactions. After lysing the cross-linked cells, immunoprecipitation is performed to isolate a certain interested protein[4]. In order to separate RNA-protein complexes

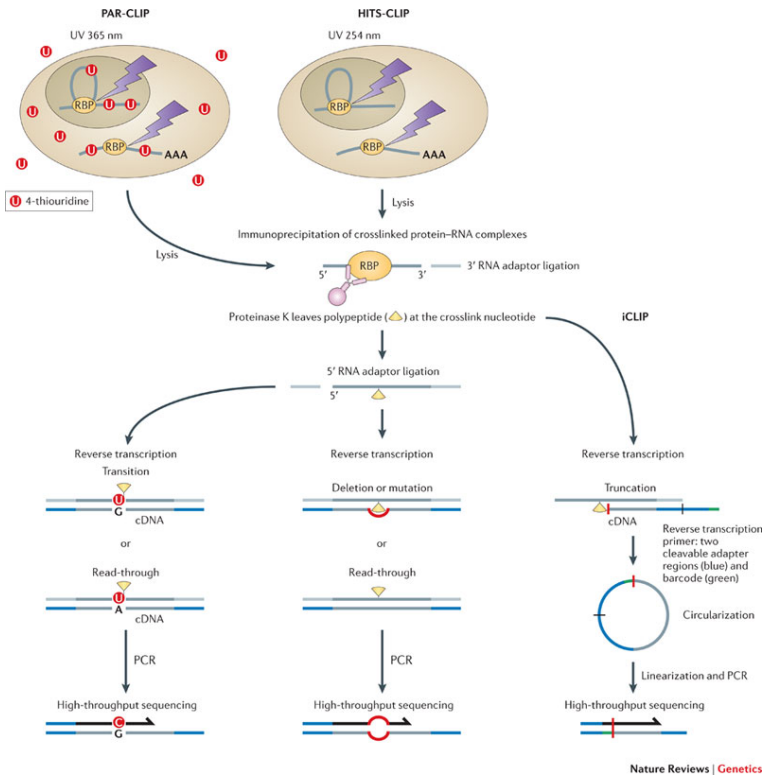

Fig. 1. Workflow of HITS-Seq, PAR-CLIP and iCLIP

from free RNA, gel electrophoresis and membrane transfer are used. After this, proteinase K digestion is used to remove protein from the RNA-protein complexes, leaving a peptide at the cross-link site. This cross-link site is then regarded as the binding sites between RNA and RNA-binding proteins. After reverse transcription of the target RNA to cDNA, high-throughput sequencing is used to generate genome-wide RNA-protein binding sites. It is important to note that during the process of reverse transcription, due to the biochemical property of RNA-peptide complex, some of the reversed cDNAs are truncated and stopped at the start site of RNA-protein binding sites. At the same time, about 20-30 percent of RNAs can be fully reversed but leaving a crosslinking induced mutation sites (CIMS) at the binding sites between RNA and peptides.

PAR-CLIP and iCLIP have a similar experimental strategy with HITS-CLIP expect for the following: In the experiment of PAR-CLIP (Photoactivatable Ribonucleoside-enhanced CLIP), cells are fed with 4-thiouridine, which becomes incorporated into newly transcribed RNA. This is because 4-thiouridine could enhance the covalent bonds between RNA and RNA-binding proteins. As a result, PAR-CLIP uses ultraviolet light at a wavelength of 365 nm to crosslink RNA and proteins while both HITS-CLIP (High-throughput sequencing CLIP) and iCLIP (individual CLIP) use ultraviolet light at a wavelength of 254 nm [5]. Moreover, using of 4-thiouridine causes a U to G transition in the RNA-protein binding sites. iCLIP is designed to capture cDNAs that truncate at the peptide that remains at the crosslinked nucleotide after proteinase K digestion. In order to do this, the 5 adaptor is added after reverse transcription and circularized cDNAs are generated.

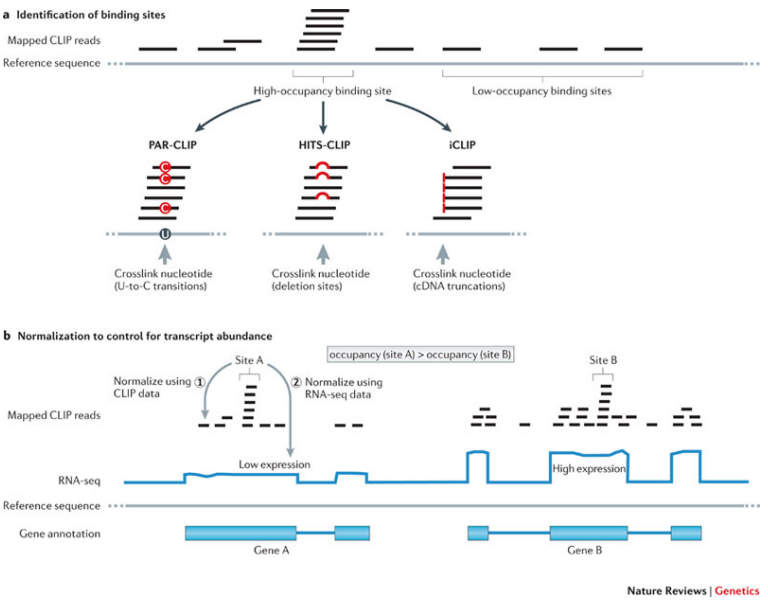

Fig. 2. Schematic to identify RNA-protein binding sites and normalize by RNA abundance

3. COMPUTATIONAL FRAMEWORK AND STATISTICAL CHALLENGES FOR CLIP-SEQ

CLIP-Seq generates hundreds of millions of reads that require highly effective computational framework and statistical methods. The first thing after obtaining sequenced reads is mapping. Secondly, enriched reads that reflect truly RNA-Protein binding sites should be inferred and identified. Thirdly, downstream analysis such as motif finding could be performed in order to deduce general regulatory principles. Figure 2 shows a general schematic to identify RNA-protein binding sites and normalization by RNA abundance.

3.1 Mapping the sequence reads

Currently, there are a number of famous alignment tools such as Bowtie, BWA and SOAP for mapping high-throughput sequencing reads to the reference genome. However, in terms of CLIP-Seq, mapping procedure should be different. This is because CLIP-Seq includes both immature RNAs and mature RNAs. If the RNA-binding proteins bind to mature RNAs, the mapping procedure needs to consider exon-exon junctions, while if the RNA-binding proteins bind to immature RNAs, algorithms like Burrows-Welch Transform can be used. Therefore, mapping of CLIP-Seq reads should ideally include either the use of splicing-aware algorithms such as TopHat or direct alignment to processed transcripts. In addition, crosslink-induced point mutations should be considered during the process of reverse transcription of RNA into cDNA in CLIP-Seq experiment. This point mutation could be substitution, insertion or deletion. To capture these mutations, algorithms allow gapped alignments should be used.

3.2 Identification of RNA-Protein binding sites

Based on the property of CLIP-Seq experiment, there are mainly two different types of computational methods to identify RNA-protein binding sites. The idea of first method is to identify enriched CLIP-Seq clusters of reads along the genome. These enriched clusters of reads are then regarded as RNA-Protein interaction regions. The second method uses the characteristic of crosslink-induced point mu-

tations. The goal of the method is thus to identify point mutations in the CLIP-Seq reads, and then regard these point mutation sites as RNA-Protein interaction sites. Naturally, this method identifies RNA-Protein binding sites at a single base pair resolution. Details of the statistical methods used in these two methods will be reviewed in section four [6][7].

3.3 Downstream analysis of RNA-Protein binding sites

Once the binding sites between RNA and proteins are identified, enriched motifs could be checked using motif finding algorithms. Currently there are a number of motif discovery algorithms. For example, MEME uses expectation-maximization algorithm to fit a two-component mixture model to the sequence data, while AlignACE uses gibbs sampler to perform a leave-one-out sampling strategy in order to find the enriched motifs given sets of DNA sequences. These enriched motifs inferred from CLIP-Seq interaction sites suggest the consensus sequence of the specific protein in the recognition of RNA sequences. It is worth noting that not all the interactions identified from CLIP-Seq will be functional important. Thus, in order to identify functional interactions between RNA and protein, the binding sites inferred from CLIP-Seq can be integrated with other genome-wide data sets that provide functional information. For example, the integration of CLIP-Seq and RNA-Seq can be used to generate position-dependent splicing regulation by RNA-binding proteins.

4. STATISTICAL METHODS TO IDENTIFY RNA-PROTEIN BINDING SITES

4.1 Site identification in high-throughput RNA-protein interaction data

All the CLIP-Seq reads are mapped to the reference genome and a junction database. Then the mapped reads are binned based on the nucleotide at which they begin. Choice of bin size should dependent on depth of the sequencing coverage. In this paper, the authors use 200bp as bin size. Let y_i be the count of the number of reads which start in the ith bin. Each bin optionally has an associated vector of covariates, which denoted as x_i. A covariate measures some property that varies in parallel with the CLIP-Seq read counts, but need not be count data. The types of covariates used in the paper are mappability and transcript abundance.

Mappability measures how many locations within the bin start sequences of length equal to the read length, which are not duplicated elsewhere in the genome and can be uniquely mapped back. Unlike ChIP-Seq, which we can consider different genomic regions have equally amount of DNA, in CLIP-Seq, RNAs have a very different level of abundance. It is entirely possible that two genomic regions of different RNAs may have same CLIP-Seq read counts while the expression of the RNAs vary a lot. In this way, in order to correctly infer the statistical significance for RNA-protein binding sites, expression level of RNAs should be considered.

Naturally, binomial distribution could be used to model the probability of read counts falling into a certain bin. Let n be the total number of reads, k be the number of reads of a certain bin and p be the probability of a read falling in a bin, the probability of getting exactly k reads in that bin will be:

$$f(k;n,p) = Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{1}$$

When the n is sufficiently large and p is sufficiently small, Poisson distribution can be used as an approximation of the binomial distribution. There is a rule of thumb stating that the Poisson distribution is a good approximation of the binomial distribution if n is at least 20 and p is smaller than or equal to 0.05, and an excellent approximation if n 100 and np 10. In CLIP-Seq or other high-throughput platform of sequencing, n is usually very large and p is usually very small, thus we can approximate

binomial distribution to Poisson distribution:

$$F_{Binomial}(k; n, p) \approx F_{Poisson}(k; \lambda = np) \tag{2}$$

Then the probability of getting k reads in the bin will be:

$$f(k; \lambda) = Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3}$$

Because both the mean and the variance of a poisson distribution is the same, it is very convenient for us to calculate the probability of getting k read counts using poisson distribution. This poisson distribution based approach are widely used in the modeling of high-throughput sequencing read counts. However, recent research reveal that read counts from high-throughput immunoprecipitation experiments are Poisson over-dispersed. The negative binomial distribution is a more appropriate approach to model the read counts distribution. Negative binomial is a distribution that models the trial of observing number k of success until a predefined number r of failures has occurred. In each trial, the probability of success is p and of failure is 1-p. The probability of getting k reads within a bin then becomes:

$$f(k; r, p) = Pr(X = k) = \binom{k + r - 1}{k} p^k (1 - p)^r \tag{4}$$

In CLIP-Seq, the authors found that zero-truncated negative binomial (ZTNB) is even better than negative binomial in the modeling of read counts distribution. In practice, zero-truncated negative binomial can be more efficient since it only uses bins with one or more reads mapping instead of all bins to perform the significant test. The zero-truncated negative binomial has the following log-likelihood function:

$$\Gamma(\mu|\alpha, y) = \Gamma_{NB}(\mu|\alpha, y) - \sum_{i=0}^{n} ln(1 - 1(1 + \alpha\mu)^{-\frac{1}{\alpha}}) \tag{5}$$

where μ is the (un-truncated) mean, α is the dispersion parameter and $\Gamma_{NB}(\mu|\alpha, y)$ is the log-likelihood of the non-adjusted negative binomial that is

$$\Gamma_{NB}(\mu|y, \alpha) = \sum_{i=0}^{n} y_i ln(\frac{\alpha\mu}{1 + \alpha\mu} - \alpha^{-1} ln(1 + \alpha\mu) + ln\Gamma(y_i + \alpha^{-1}) - ln\Gamma(y_i + 1) - ln\Gamma(\alpha^{-1}) \tag{6}$$

With the information of mappability and transcript abundance, the parameter μ is replaced with a vector of $\overrightarrow{\mu}$, where each $\mu_i = exp(\overrightarrow{\mu}^T \overrightarrow{x_i})$ The model is fitted using a Newton−Raphson algorithm for the estimation of the regression parameters and a dispersion dampening algorithm for estimating α. An example of calculating is described in Figure 3.

Each bin is then assigned a p-value to indicate the statistical significance of the read counts enrichment. The smaller the P-value, the more unlikely the read count in the bin is given the fit distribution. As a result, the Genomic regions with a p-value under the cutoff are then regarded as the RNA-protein binding regions.

4.2 Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data

In the standard CLIP procedure, the residual amino-acid-RNA complex can impose a potential obstacle for reverse transcriptase to read through when RNA fragments are converted into cDNAs. This

Y	X_1	X_2	$\exp(\vec{\beta}^T\vec{x_i}) = \mu$	P-value
1	4	10	$\exp(0.17\times 4+0.02\times 10)=2.41$	$1-\sum_{j=1}^{1}\Pr(y_i=j\|\mu=2.41,\ \alpha=2)=0.71$
3	2	37	$\exp(0.17\times 2+0.02\times 37)=2.94$	$1-\sum_{j=1}^{3}\Pr(y_i=j\|\mu=2.94,\ \alpha=2)=0.45$
250	5	30	$\exp(0.17\times 5+0.02\times 30)=4.26$	$1-\sum_{j=1}^{250}\Pr(y_i=j\|\mu=4.26,\ \alpha=2)<8.63\times 10^{-14}$ ***
5	4	17	$\exp(0.17\times 4+0.02\times 17)=2.77$	$1-\sum_{j=1}^{5}\Pr(y_i=j\|\mu=2.77,\ \alpha=2)=0.27$
7	7	13	$\exp(0.17\times 6+0.02\times 13)=3.60$	$1-\sum_{j=1}^{7}\Pr(y_i=j\|\mu=3.60,\ \alpha=2)=0.24$
300	10	180	$\exp(0.17\times 10+0.02\times 180)=200.34$	$1-\sum_{j=1}^{300}\Pr(y_i=j\|\mu=200.34,\ \alpha=2)=0.23$

Fig. 3. Example of calculating P-values for bins using the ZTNBR with two covariates: mappability (X1, in arbitrary units) and transcript abundance (X2, in reads mapped from RNA-seq control), assuming the model has already been fit with β= 0.17, 0.02 and α = 2

obstacle causes three different types of cDNAs as a result: truncated cDNA at the crosslink site, correct read been read through or read with an error at the crosslink site. These crosslink mutation sites (CIMS) can be recovered by mapping sequence reads to the reference genome that allow for deletions, insertions or substitutions. More importantly, identification of these CIMSs reveals RNA-protein interaction sites at a single base pair resolution. The question of finding RNA-protein interaction sites that becomes figuring out where these statistically significant CIMSs are.

In this paper, the authors are interested in two alternative splicing factors, Nova and Ago. Because splicing factors mainly bind to immature RNA, sequencing reads were mapped back to the reference genome instead of the exon-exon junctions. Novoalign, which performs exhaustive searches of hits tolerating substitutions, small insertions and deletions, is used as the mapping program. Unambiguous mapping to the genome with $\leq$2 substitutions, insertions or deletions in $\geq$25 nt is required. To remove potential duplicates caused by PCR amplifications, reads with the same start genomic positions are only kept once.

The authors assume that crosslinking-induced mutations would occur at specific sites and would be reproducibly detected in multiple CLIP tags, whereas technical errors should map to random positions without reproducibility. For substitutions, known single nucleotide polymorphisms or RNA editing sites are excluded since these sites may also cause the mutation clusters. To distinguish crosslinking-induced mutations from sequencing or alignment errors, permutation based statistical test is used. First, for each genomic position, the total number of overlapping unique tags k and the number of unique tags with particular types of mutations m is calculated. Now the question is whether the observed mutation rate m/k for each site is significantly larger than one would expect from random or not. Second, in the permutation procedure, each mutation was planted into a randomly picked CLIP tag with the same offset of mutation observed from the original read. A null distribution of m, given k, was estimated empirically. Third, to estimate the false discovery rate, cumulative number of clustered mutation sites with k tags at the position and $\geq$m tags with mutations c[m,k], and the corresponding cumulative number of permuted mutation sites c_0[m,k] is counted. By definition, the FDR of observing $\geq$m tags with mutations given a total of k tags is c_0[m,k]/c[m,k]. Mutation sites with FDR under the cutoff are then regarded as the RNA-protein binding sites.

5. SUMMARY

CLIP-Seq identifies RNA-protein binding sites at a genome-wide scale. Up to date, there are mainly two strategies to analyze and infer RNA-protein interactions from CLIP-Seq. The first strategy is consistent with that used in ChIP-Seq peak finding, which use statistical distribution to model the significance of read counts within a certain bin size. Another strategy takes advantage of the CLIP-

Seq procedure, which uses the information of crosslink-induced mutation sites to infer the RNA-protein binding sites. In the first method, all the uniquely mapped CLIP-Seq reads provide the information about RNA-protein binding sites, while the second method only uses a proportion of mapped reads. Moreover, the resolution of the first method depends on the bin size, which always span to hundreds of base pairs. On the contrary, the second method identifies RNA-protein binding sites at a single base pair resolution. In addition, read counts based approach works on HITS-CLIP, PAR-CLIP and iCLIP, while mutation sites based approach is specifically designed for HITS-CLIP.

REFERENCES

Darnell R B. HITSCLIP: panoramic views of proteinRNA regulation in living cells[J]. Wiley Interdisciplinary Reviews: RNA, 2010, 1(2): 266-286.

Knig J, Zarnack K, Luscombe N M, et al. ProteinRNA interactions: new genomic technologies and perspectives[J]. Nature Reviews Genetics, 2012, 13(2): 77-83.

Ule J, Jensen K B, Ruggiu M, et al. CLIP identifies Nova-regulated RNA networks in the brain[J]. Science, 2003, 302(5648): 1212-1215.

Ule J, Jensen K, Mele A, et al. CLIP: a method for identifying proteinRNA interaction sites in living cells[J]. Methods, 2005, 37(4): 376-386.

Hafner M, Landthaler M, Burger L, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP[J]. Cell, 2010, 141(1): 129-141.

Uren P J, Bahrami-Samani E, Burns S C, et al. Site identification in high-throughput RNAprotein interaction data[J]. Bioinformatics, 2012, 28(23): 3013-3020.

Zhang C, Darnell R B. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data[J]. Nature biotechnology, 2011, 29(7): 607-614.

Chen B, Yun J, Kim M S, et al. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis[J]. Genome biology, 2014, 15(1): R18.

Kishore S, Jaskiewicz L, Burger L, et al. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins[J]. Nature methods, 2011, 8(7): 559-564.