Computational Analysis of CLIP-seq Data

Michael Uhl* Torsten Houwaart* Gianluca Corrado
Patrick R. Wright Rolf Backofen*†

March 8, 2017

Affiliations: Michael Uhl, Torsten Houwaart, Patrick R. Wright, Rolf Backofen: Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany
Rolf Backofen: Centre for Biological Signalling Studies (BIOSS), University of Freiburg, Freiburg, Germany
Gianluca Corrado: Department of Information Engineering and Computer Science, University of Trento, Italy

Abstract

CLIP-seq experiments are currently the most important means for determining the binding sites of RNA binding proteins on a genome-wide level. The computational analysis can be divided into three steps. In the first pre-processing stage, raw reads have to be trimmed and mapped to the genome. This step has to be specifically adapted for each CLIP-seq protocol. The next step is peak calling, which is required to remove unspecific signals and to determine bona fide protein binding sites on target RNAs. Here, both protocol-specific approaches as well as generic peak callers are available. Despite some peak callers being more widely used, each peak caller has its specific assets and drawbacks, and it might be advantageous to compare the results of several methods.

Although peak calling is often the final step in many CLIP-seq publications, an important follow-up task is the determination of binding models from CLIP-seq data. This is central because CLIP-seq experiments are highly dependent on the transcriptional state of the cell in which the experiment was performed. Thus, relying solely on binding sites determined by CLIP-seq from different cells or conditions can lead to a high false negative rate. This shortcoming can, however, be circumvented by applying models that predict additional putative binding sites.

*These authors contributed equally to this work
†Corresponding author

1 Introduction

The rise of next-generation sequencing (NGS) techniques over the past decade has led to an enormous boost in RNA research thanks to numerous discoveries concerning the fundamental role of RNA in gene regulation [1]. To exert these functions, RNAs in eukaryotic cells can form ribonucleoprotein complexes by interacting with a multitude of RNA-binding proteins (RBPs), allowing for the evolution of complex regulatory networks. Recent studies revealed more than 1500 RBPs in human cells, which emphasizes their fundamental importance for virtually all aspects of post-transcriptional gene regulation (PTGR), including RNA maturation, alteration, transport, stability, and translation [2] [3] [4] [5]. Beside their physiological roles, various diseases have been linked to dysregulated or deficient RNA-binding proteins [6] [7]. Hence, a comprehensive understanding of RNA-based networks is only possible when also considering the contributions of these RBPs. The scientific community is therefore increasingly turning to the characterization of RBP-based regulation. RBPs regulate their target gene(s) by directly binding to the transcribed RNA. Typically, specific sequence motifs are required for binding site recognition, although the relative contributions of RNA sequence, structure and backbone to the binding can differ greatly among RBPs [8] [9]. The majority of RBPs appears to prefer single-stranded regions [4]. There are, however, also many RBPs that prefer structured RNA such as Staufen 1 [10], Roquin [11], or MLE [12]. It is currently unclear to what extent the observed general tendency towards single-stranded RNA regions is caused by biases in the experimental protocols. As RNA molecules generally form extensive secondary structures, it is not surprising that the binding specificity of RBPs also strongly depends on the structural context of their binding sites. Indeed, the importance of binding site accessibility has been shown for many RBPs [13].

The recent development of high-throughput protocols for determining RBP binding sites on a genome-wide scale has greatly influenced the field and opened up new avenues for the investigation of regulatory relationships. Particularly, CLIP-seq (cross-linking and immunoprecipitation followed by next generation sequencing) [14] has become the standard experimental procedure for studying transcriptome-wide RBP binding. Briefly, RBPs are crosslinked to their RNA binding sites, followed by extraction and sequencing of the crosslinked RNA fragments. After mapping of the sequenced fragments, binding regions are identified based on the read profiles and various additional information (e.g. from control experiments or replicates). The process of determining significantly enriched binding regions is also known as peak calling. Subsequently, binding motifs or predictive models can be derived from the identified sites. These can then be employed to identify potential binding sites in yet unreported target sequences.

In this paper, we describe selected tools and pipelines required for a comprehensive bioinformatics analysis of CLIP-seq datasets. We do not intend to give a complete overview of available methods, since there is already plenty of literature available on CLIP-seq data analysis [15] [16] [17]. Rather, we will

concentrate on tools which have proven valuable to us in the past. For these, we will describe important aspects of a comprehensive analysis. A special focus will lie on the process of peak calling, which is the process of recovering bona fide protein binding sites by signal detection and removal of false positives originating from unspecific interactions. To our knowledge, this component of the data analysis is still lacking a more comprehensive discussion in literature, even though it is arguably the most critical part of the whole analysis. We will start with a description of the different CLIP-seq variants available, addressing specific features and (dis)advantages. In the following section on CLIP-seq data analysis, we will describe the different steps of pre-processing, mapping and peak calling in greater detail. The last section considers the task of determining binding models for computationl binding site prediction. Such models are needed to reduce the false negative rates of CLIP-seq experiments which originate from their dependency on the expression of the detected RNA binding sites. Without these models, information from published data cannot be transferred to different cells or conditions.

2 Overview of CLIP-seq variants

In recent years CLIP-seq has become the standard experimental procedure to identify binding sites of RBPs on a transcriptome-wide level. Several variants have been proposed since the introduction of CLIP [18] [19] in 2003 and its first high-throughput sequencing extension HITS-CLIP (high-throughput sequencing of RNA isolated by CLIP) [14] in 2008, each addressing various shortcomings of the previous versions. The most widely used modifications over the last years are PAR-CLIP (photoactivatable-ribonucleoside-enhanced CLIP) [20] and iCLIP (individual-nucleotide CLIP) [21], while recently the eCLIP (enhanced CLIP) protocol [22] was introduced and promoted by the ENCODE consortium. Another protocol termed irCLIP (infrared-CLIP) [23], which has been compared to eCLIP [24] [25], has also been published in 2016. Besides, several specialised modifications for double-strand binding RBPs exist [26] [10] [27]. These protocols add an additional ligation step to the standard protocol in which the two double-strand RNA segments bound by the RBP are connected, leading to chimeric reads that allow for the simultaneous identification of both RNA strand regions. So far, CLIP-seq has been applied in numerous studies on single RBPs. Furthermore, the method has been employed by a study on global mRNA binding preferences [2].

Principle CLIP-seq workflow The principle workflow of a CLIP protocol starts with UV radiation of the cell or tissue culture, which induces covalent crosslinks between RBPs and their bound RNAs. This is followed by immunoprecipitation of the RBP-RNA complexes and partial RNase digestion to narrow down the binding sites to appropriate sequencing and mapping lengths. Further steps aim at stringent purification, including radioactive labeling, recovery by SDS-PAGE, transfer to nitrocellulose membrane to abolish loose RNA frag-

ments, excision and proteinase K treatment to remove the RBP and recover the trimmed RNA fragments. Finally, the fragments are reverse-transcribed and their cDNAs are subjected to deep sequencing. The resulting sequencing data is then analysed to obtain RBP binding sites which can be identified based on the mapped read profiles.

PAR-CLIP PAR-CLIP [20] marked the first successful adaptation of the original protocol, introducing a number of modifications over HITS-CLIP. To increase crosslinking efficiency, cells are additionally supplemented with 4-thiouridine (4SU), and UV radiation is applied at 365 nm instead of 256 nm. Interestingly, these modifications also lead to a high number of thymidine to cytidine transitions in the cDNA at the crosslink sites, which can be exploited in a subsequent mutational analysis for pinpointing the crosslink position, thus basically enabling PAR-CLIP to achieve single-nucleotide resolution. On the other hand, 4SU usage restricts the method to cell cultures and preferential crosslinking to 4SU naturally biases site recovery towards U-containing sites. Also, 4SU exhibits an increased affinity towards G:U base pairing [28], which might influence cellular RNA structure and thus also RBP binding. In addition, RNase T1 digestion leads to a depletion of G-containing sites, due to the enzyme's preferential cleaving after G nucleotides [29]. Another problem is the usage of inducible tagged proteins in the original publication, which can result in the recovery of non-physiological binding events due to overexpression. The last two problems can and have been addressed in subsequent PAR-CLIP versions [29] [30], where the latter one also describes an *in vivo* approach for *C. elegans*.

iCLIP iCLIP [21] has been particularly designed to address a specific problem inherent to HITS-CLIP and PAR-CLIP: during cDNA synthesis, the reverse transcriptase frequently stalls at crosslink sites still containing residual peptides, leading to an estimated loss of over 80 % of cDNA fragments [31]. To solve this issue, the authors developed a two-part cleavable adapter together with an additional circularization and linearization step, allowing for the recovery of both complete and truncated cDNAs. Additionally, random barcodes are used, enabling easy identification and removal of PCR duplicates after mapping. These measures lead to increased efficiency, while single-nucleotide resolution is achieved due to the truncated cDNAs which pinpoint the crosslink position to the reads' 5' ends. Still, as with PAR-CLIP and the original HITS-CLIP, the protocol remains time-intensive (up to 5 days) and error-prone due to its many different steps [24]. Also, a fairly huge amount of starting material (typically 10^6 to 10^8 cells) is required in order to generate a library of sufficient complexity. This often makes successful library preparation difficult. This is especially true for lowly expressed RBPs, RBPs with widespread binding or RBPs with low crosslinking efficiencies and / or antibody affinities.

eCLIP Both eCLIP [22] and irCLIP [23] have been developed to deal with the shortcomings of previous CLIP-seq variants. Particularly, high demands in cell

numbers, many different preparation steps including radioactive reagents and long preparation times frequently result in poor library generation efficiency. In the eCLIP protocol, the inefficient circularization step from iCLIP is exchanged by two separate adapter ligation steps, which results in much higher RNA fragment recovery. This ultimately leads to a significantly improved library complexity. Furthermore less cells are needed. Both aforementioned improvements enable the application of this method on formerly difficult RBPs. Single-nucleotide resolution is achieved the same way as in iCLIP, meaning that the reads' 5' end should mark the crosslink position for the huge majority of reads. Moreover, the autoradiographic visualization step is omitted and different samples can be pooled early in the protocol. This allows for much faster preparation times, but leaving out the autoradiographic step is also a clear drawback since the quality of the IP can no longer be monitored. Another new feature is the inclusion of a size-matched input control (SMInput), which enables efficient background normalization and thus leads to a higher specificity in subsequent binding site identification. For SMInput, 2 % of the pre-immunoprecipitation sample is taken and sequenced together with the immuno-purified sample. It was shown that normalization by SMInput significantly improves authentic binding site recovery, whereas an IgG control, which is frequently employed as a CLIP-seq control, was found unsuitable for this task. The authors also provide a peak calling pipeline called CLIPper [32], which will be discussed in a later section. The described improvements have made eCLIP the method of choice for the ENCODE consortium. So far, the consortium has published eCLIP data for more than 70 diverse RBPs, which underlines its usability and will likely help eCLIP to become more popular in the near future.

irCLIP Compared to eCLIP, irCLIP [23] uses a complementary approach to deal with the described shortcomings of previous CLIP protocols: the circularization step from iCLIP is kept but optimized and applied in a single-tube reaction together with reverse transcription to reduce preparation time. In addition, both circularization and reverse transcription are performed at 60°C using thermostable enzymes to resolve potential RNA secondary structures. It will be interesting to see whether this step also helps to improve binding site recovery in the case of structure-binding RBPs, which might yield low library complexities for other CLIP protocols. irCLIP achieves single-nucleotide resolution analogous to iCLIP and eCLIP. Like eCLIP, irCLIP too skips radioactivity steps, but instead introduces an infrared fluorescent dye to visually check IP quality. It can thus prevent certain IP-related quality issues which can become a problem in the eCLIP protocol, since eCLIP ommits the autoradiographic visualization without substitution. Infrared dye labeling also improves other steps of the protocol, which as with eCLIP results in lesser starting material (typically only 20,000 cells) and overall increased efficiency. On the other hand, working with infrared dyes also requires specialized equipment, such as a gel documentation system with near-infrared capabilities, which might not be highly available or affordable [24]. It remains to be seen which of the two protocols will be ap-

plied more frequently by the field. In any case, future comparisons in recovered binding profiles should help to reveal protocol-specific advantages and biases.

3 Analysis of CLIP-seq data

The analysis of CLIP-seq data usually involves three major steps which will be addressed here. As in many other protocols, the reads first have to be mapped to a reference genome. If the CLIP experiment was performed for a specific RBP, the generated reads should agglomerate in regions to which the RBP binds. To identify these regions, a second step is performed under application of a peak caller. Peak callers are used on the coverage profiles to determine regions that are bound by the RBP with high affinity. Once the peaks are identified, they can be quantified and their statistical significance should be evaluated by comparing them to a control experiment. In a third step, the resulting data can be utilized to find binding motifs and to train binding models, which enable the prediction of novel RBP binding sites on transcripts not present in the CLIP-seq data. The last step is especially important when investigating RBP binding sites in cells or conditions for which no CLIP-seq data is publicly available.

3.1 Preprocessing of raw data and mapping

Most CLIP-seq studies are performed on organisms with well annotated genomes like human, mouse or *C. elegans* [33]. Reads from CLIP-seq experiments performed on these organisms can be mapped to the according reference genome or transcriptome. A major problem regarding the quantification of read data is the reliance of sequencing-based techniques on PCR amplification of the sequence libraries prior to sequencing. Although necessary in order to generate a sufficient amount of sequencing material, the occurrence of some sequences can be artificially boosted in the process because of biases in the PCR protocol, where so-called PCR duplicates are introduced. With the introduction of random barcodes or unique molecular identifiers (UMI) in iCLIP this problem is mitigated, as reads which contain the same random barcodes and map to the same coordinates can be collapsed to unify all PCR duplicates into just one representative. The methodology is not completely flawless though, as it has been shown that during library preparation mutations can be introduced in the random barcodes which can have a big effect on the crosslinking-event counts [12]. Before mapping the reads, these UMIs have to be removed. Tools such as flexbar [34] can be used to accomplish this. If no UMIs were used then tools such as FastUniq can be employed to collapse potential PCR duplicates [35]. Adapters that are used in the amplification steps of the sequences also have to be trimmed from the sequences. Several programs can be used for this, e.g. cutadapt [36], Trim Galore [37], which is based on cutadapt and fastqc, or trimmomatic [38], which is specifically made for Illumina sequencing data.

A few things have to be considered in order to correctly map the trimmed reads to a reference genome. In most cases, this step consumes the most compu-

tational power. The sequences stem from RNA molecules which can be subject to splicing in eukaryotes. The choice of the mapping software depends on prior knowledge about the targets of the RBP and is not independent from the following peak calling step, since the peak caller has to deal with gaps which occur in spliced reads. The reads can either be mapped to the genome or the transcriptome. The advantage of mapping the reads to the transcriptome is that a higher sensitivity can be achieved, but it also comes at the cost of limiting the analysis to known transcripts. Since RBP binding sites can be located in introns (especially in the case of splicing regulators), mapping only to exonic parts would lead to the exclusion of these sites. Mapping to exons also leads to a depletion of sites spanning exon borders, since the read parts are often too short to be mapped to their corresponding exons with sufficient quality. All these issues have to be considered in order to choose a meaningful mapping strategy. A layered procedure of first mapping strictly to the transcriptome and afterwards mapping the remaining reads to the genome is often used and might work best in such cases. A wide range of mapping algorithms originally developed for RNA-seq are available. To list a few good choices for this task, TopHat [39], GSnap [40] and segemehl [41] fulfill the aforementioned requirements and are widely used, but also STAR [42] should be mentioned, which is the mapper of choice in the eCLIP pipeline used by the ENCODE consortium. Of course this list is not comprehensive and many other good choices exist. Benchmarking and isolating the best program for this task go beyond the scope of this review and can be found elsewhere [43, 44, 45].

3.2 Methods for peak calling

The next task after mapping reads to a reference genome or transcriptome is to extract authentic binding sites from the mapped read profiles. Many reads stem from unspecific binding and thus have to be discarded, which is done in the process of *peak calling*. This task can typically be divided into two parts: one first extracts potentially interesting peaks based on peak shape or height and then filters the resulting peaks such that only sites enriched over a certain threshold or background are kept. The first part usually results in a huge number of initial sites, including many false positive predictions. The second part therefore should incorporate additional experimental information like read profiles from replicates, controls, or RNA-seq samples in order to increase the signal-to-noise ratio. Information on underlying transcript abundances is particularly important to peak calling on CLIP-seq data, since transcript amounts differ between transcripts from different loci, and thus directly influence the peak heights found in the read profiles. Therefore one cannot be sure if e.g. a high peak corresponds to a strong binding site or if this is just the result of the underlying transcript being highly expressed in the observed cell type or condition. A correction for transcript abundance is therefore of fundamental importance in CLIP-seq peak calling. Interestingly, correction for transcript abundance has been shown to significantly improve peak calling results even in the case of external RNA-seq data [46]. Ideally however, one should choose a CLIP-specific control for

background correction, which also incorporates protocol-intrinsic biases. The authors of eCLIP [22] e.g. showed that using a pre-immunoprecipitation control (as described in the eCLIP section) led to a significant enrichment of true binding sites, whereas an IgG control, which is frequently used in CLIP, was not suitable for background correction. In general, controls that produce low complexity libraries and thus poor coverage of the underlying transcriptome should be avoided. Besides using controls, results from different replicates can be intersected to further increase specificity. In order to assign significance values to peaks, it is also important to find a suitable probability distribution for modeling the underlying read counts. In the following, some prominent CLIP-seq peak callers which have been used by our group will be discussed in more detail.

Piranha Piranha [46] is a CLIP-seq peak caller which can be applied to all available CLIP-seq as well as RIP-seq datasets in order to identify significant peaks. It was the first generic CLIP-seq peak caller developed, i.e. it does not depend on certain CLIP variant properties in order to call peaks, as opposed to PARalyzer [47], which relies on PAR-CLIP data, or CIMS [48] and CITS [49], which were developed for HITS-CLIP. Based on the mapped reads as input, Piranha first divides the genome into non-overlapping bins of a user-defined size and counts the number of read starts falling into each bin. Piranha assumes that the read starts define the site where the crosslink events take place. Bins with zero counts are discarded, and the counts of the remaining bins are then used to fit a probability distribution. Covariates, e.g. in the form of reads from RNA-seq or a CLIP-seq control experiment, can be supplied to correct for different transcript abundances or protocol biases. In the case of covariates, Piranha uses a zero-truncated negative binomial regression for fitting the read counts together with the supplied covariate data. If no covariates are given, the user has the choice between four different distributions. However, the zero-truncated negative binomial distribution is set as default and recommended, as it was shown to have the best fit on a collection of over 100 CLIP-seq datasets. Since Piranha assumes that most read-covered sites represent background binding, the fitted distributions essentially model background probabilities. Therefore, the p-value of a given bin corresponds to the probability of the site being background. By default, Piranha reports p-values corrected for multiple testing using the Benjamini-Hochberg method [50] with a default threshold of 0.05. As for the bin size, the authors suggest the size to be adapted to the depth of coverage and the CLIP-seq variant used. This is of course not intuitive, especially for novice users. According to the authors, a good starting point for RIP-seq is 100, while e.g. for iCLIP, one could start with low sizes (e.g. 5 nt) and then depending on the amount of noise in the dataset gradually increase the window size. Either way, having to deal with manually adjustable bin sizes is a clear drawback of Piranha. In addition, it lacks support for the integration of replicate information, although one could still do a manual intersection by calling peaks on all replicates separately and merging the results afterwards.

PARalyzer PARalyzer [51] is a computational tool[1] for the discovery of crosslinking sites from PAR-CLIP sequencing data. In the PAR-CLIP protocol the protein crosslinking is boosted by additionally culturing the cells with a photoreactive ribonucleoside analogue, usually 4SU. The crosslink product of 4SU is known to have a preferential base pairing to guanine (G) instead of adenine (A), resulting in thymine (T) to cytosine (C) conversions in PCR-amplified cDNA.

The rationale of PARalyzer is to examine the pattern of T to C conversions in order to spot, with high confidence, RNA-protein interaction sites. A kernel-density-based classifier is used to characterize crosslinked regions, identified by T to C conversions (the signal), against not crosslinked ones, characterized by the absence of T to C conversions (the background).

Class-specific densities (one for the signal and one for the background) are assessed by employing a Gaussian kernel density estimator that, for each T nucleotide, considers the number of T to C conversions and the number of non T to C conversions in the aligned reads. For each T nucleotide in the RNA sequence, the number of T to C conversions occurring in that position is represented using a Gaussian distribuion with fixed variance. The distribution is peaked on the T nucleotide and the variance distributes the signal over the neighbouring nucleotides. The function in green, shown in Figure 1A, is the sum of all the individual Gaussian distributions that indicate T to C conversions and represents the signal. The background (red function in Figure 1B) is estimated by summing all the Gaussian contributions of T nucleotides that have not turned into C nucleotides instead of the T to C conversions. After estimating the class-specific densities, the interaction sites are defined by the nucleotides for which the density estimate of the signal (T to C conversions) is greater than the one for the background (non T to C conversions) (Figure 1C).

CLIPper To distinguish peak regions from non-peak regions, the CLIPper software [52] utilizes different statistical measures. CLIPper is intended for calling CLIP-seq peaks on known genes only and therefore requires annotation. It provides annotations for a few genome assemblies, i.e. hg19, mm9, mm10, and ce10. For other species the user has to provide the annotation. The program defines sections on the genome where reads agglomerate and identifies peaks on the read profiles. A threshold is defined based on the amount of reads in each section, the amount of reads in the vicinity of the section, and the amount of reads in the gene. The threshold specifies the minimum amount of reads necessary within this region to be deemed statistically significant. This procedure makes sure the false positive rate of peaks is controlled. By default, CLIPper then fits a spline function to the read profile and defines regions which are above the threshold and those that are in between local minima of the fitted spline as peaks. For these peaks, a p-value is calculated with the amount of reads in the peak region X being modeled as $X \sim \text{Poisson}(1 + \text{reads\_in\_gene} \cdot \frac{\text{peak\_length}}{\text{gene\_length}})$. This procedure assigns p-values to all peaks which in turn can be corrected for multiple testing

[1] PARalyzer is available at `https://ohlerlab.mdc-berlin.de/software/PARalyzer_85/`

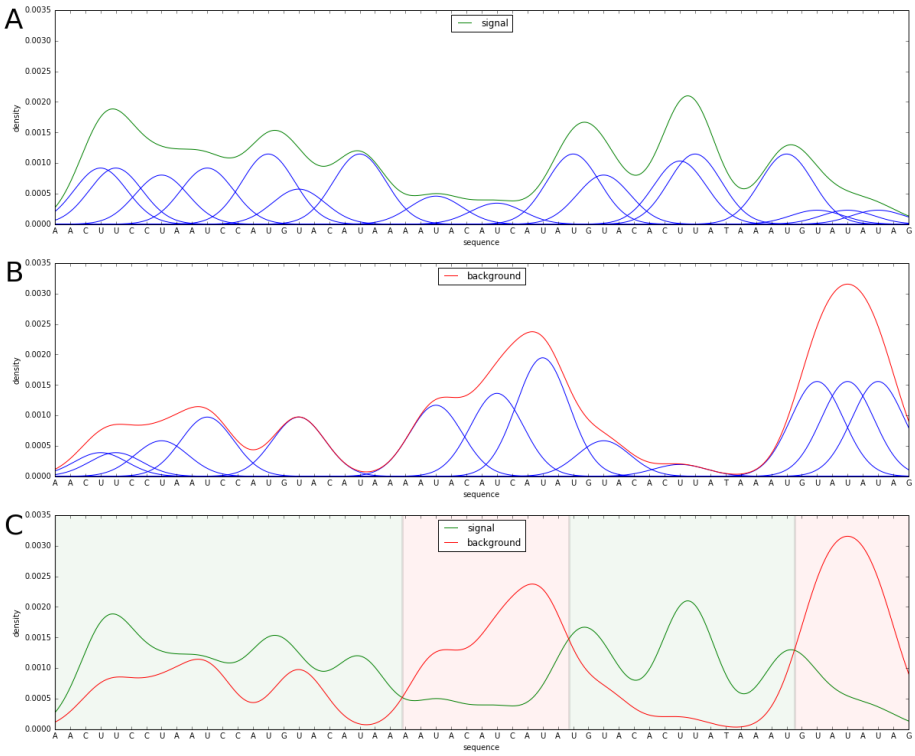

Figure 1: Crosslink site indentification with PARalyzer on a synthetic example. Class-specific densities for both the signal and the background are estimated using a Gaussian kernel density estimator. (A) Density estimation of the signal (green function). For each T nucleotide a Gaussian with fixed variance is peaked representing the number of T to C conversions occurring in the position, normalized by the total number of T to C conversion in the associated read group. The normalized sum of all their Gaussian functions is the signal. (B) Density estimation of the background (red function). The estimation is based on the number of T nucleotides that have not turned into Cs. (C) After estimating the class-specific densities, the interaction sites are defined by the nucleotides where the density estimate of the signal (T to C conversions, green line) is greater than the one for the background (non T to C conversions, red line).

with respect to all tested peaks using the Benjamini-Hochberg procedure [50]. The local maxima in the fitted splines are explicitly highlighted (in the resulting BED file) because these positions are the best candidates for where the analysed RBP binds to. In the eCLIP pipeline that is used by the ENCODE consortium [53] the peaks are annotated qualitatively after their identification. Each CLIP experiment dataset can be compared to one control dataset. For each peak a log2-fold-change is calculated based on the mapped reads within the peak region

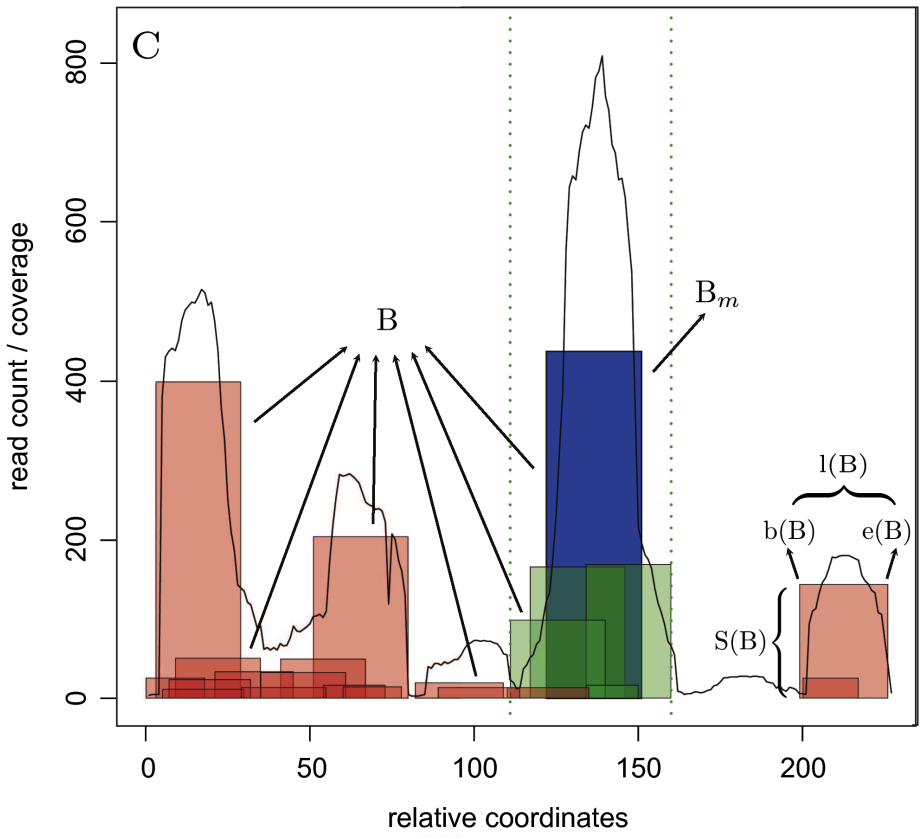

Figure 2: The figure shows a specific cluster (C) of blocks (B) and their attributes e(B), b(B), l(B), S(B). The first B_m selected for this C is shown in blue, while overlapping blocks that reach at least into the middle of B_m are green. The dotted green lines show the borders of the first peak. All B not used for the definition of the boundaries of the first peak are red. The density of sequencing reads at nucleotide resolution is shown as black line.

for the experiment in comparison with the control. Furthermore, a p-value is determined for each peak using a χ^2 test or Fisher's exact test using the mapped and total reads of the experiment and control. Given that eCLIP pinpoints the crosslink positions to the read starts, it is surprising that CLIPper does not take advantage of this information, instead considering the full-length reads for peak calling.

Block-based peak calling A recent CLIP-seq study on the transcriptome-wide binding sites of the bacterial RBPs Hfq and CsrA in the human pathogen *Salmonella enterica* [54] introduced an experimental procedure including paired-end signal and background libraries in triplicates. The library preparation for the background data solely differs in the fact that no UV induced cross linking is performed. The identification of significant peaks in the signal data was performed based on a three step procedure. In the first step, the blockbuster algorithm [55] subdivides the pooled signal sequencing data into clusters (C) of blocks (B). A block is fundamentally a pile of similar reads which is characterized by its beginning position b(B), its ending position e(B), its size S(B) and its length l(B) (see Figure 2). Since the block boundaries set in this initial structure do not provide appropriate peak boundaries, overlapping blocks are joined into peaks in the second step of the procedure. The biggest block (B_m) is selected from C if $S(B_m) \geq S(C)*0.01$. Then, all B overlapping with B_m are selected and removed from C. The peak boundaries are extended using all blocks that overlap with at least half of B_m and also fulfill $S(B) \geq S(B_m)*0.1$. The leftmost and rightmost coordinates of the remaining B are set as final boundaries and the procedure restarts by selecting the next B_m. In the last step, the DESeq2 algorithm [56] assesses the statistical significance of each peak based on the individual amount of reads counted for each of the peaks in the signal and background libraries. The final output is a p-value sorted list.

Summary and Comparison A meaningful and fair comparison of the different peak callers is problematic. On the one hand each tool incorporates several parameters which change the behaviour of each peak caller significantly. On the other hand no datasets of absolute truth exist on which the different tools can be benchmarked. Tools that work only with very specific protocols because they rely on signatures in the data that are introduced in these protocols can not be fairly compared. In the following discussion, PARalyzer was not considered because its method of finding peaks is specific to PAR-CLIP data and not applicable to other CLIP-seq methods. For the other tools (CLIPper, Piranha and block-based peak calling) specific filtering steps were undertaken as explained in the following. To give a quantitative measure for the comparison of the different peak callers we propose a genomic position based metric. One position corresponds to one nucleotide in the reference genome of the investigated organism. Each position that is assigned to a peak by at least one peak caller is evaluated on whether it is also within a peak region defined by the other tools.
The problem of peak calling can be considered as two distinct steps. The first step consists of defining regions of interest solely based on the fact that one or more signal libraries show an agglomeration of reads in these areas. The second step is a statistical evaluation of these regions of interest where both the signal and the background libraries are taken into account. The two peak callers CLIPper and Piranha can perform a statistical analysis on just signal libraries and report a p-value for the peaks they find. The block-based method can only perform the first step of finding read-enriched areas and relies on other

programs, i.e. DESeq2, to do the statistical analysis. If no replicate or control samples are available, CLIPper's or Piranha's built-in functionality to estimate the background distribution is obviously the only possibility to assign significances to peaks in the read profiles. In theory this is also possible for the block-based method, but has not yet been implemented. Piranha offers the possibility to add covariate datasets to improve estimation of the background. CLIPper itself does not offer this capability, yet the significance of the regions can be reevaluated with the scripts used in the eCLIP pipeline of the ENCODE consortium [53]. Both Piranha and CLIPper cannot handle replicate samples which are essential for the estimation of the technical or biological variance in the experiments, which is why splitting up the two tasks as mentioned above is advised when replicates are available. In the following paragraphs the three tools Piranha, CLIPper, and blockbuster based peak calling are compared with each other on one example dataset to illustrate similarities and some distinguishing features between the tools.
For this example the human RBP Histone Stem-Loop-Binding Protein (SLBP) was chosen. An eCLIP experiment [22] was recently published for this RBP and is available on the ENCODE consortium website [2]. For the analysis we utilized the files that provide the already mapped reads to reference genome hg19. The second-in-pair reads which should contain the cross link position at their 5' ends have an average length of approximately 38 nucleotides. SLBP targets histone protein mRNAs and has a well known stem-loop binding motif [57]. One target of SLBP are transcripts of HIST2H2AC with the aforementioned stem-loop motif in its 3' UTR. In Figure 3 eCLIP read profiles for this one target site are depicted. The 3' end of gene HIST2H2AC lies in the region 149.858800 mb – 149.858910 mb (see Figure 3 tracks 1-3). The read coverage and the read start coverage (tracks 4,5) are the determining signals for the three different peak callers CLIPper (track 6), extended blockbuster (track 7) and Piranha (track 8). Comparing the coverage tracks with its size matched input counterparts for this region (track 4,5 red), it can be safely stated that this region is targeted by SLBP. This one example already clearly illustrates some of the three tools' major differences. Where CLIPper and the block-based approach follow the overall read coverage, Piranha is more aligned with the read start distribution. It should be noted that in this example the stem-loop motif starts right after the peak that was called by Piranha (a more detailed discussion of this issue can be found in the Conclusion section).

As stated above, a fair comparison is difficult to achieve when the tools are very flexible with different parameter settings. For the following more general analysis the tools were called with standard parameters where possible. The other parameter settings are best guesses as a thorough evaluation of these settings is beyond the scope of this review. To achieve an even higher parity in the evaluation of the peak callers, the normalization and the statistical analysis were done with the same pipeline. For CLIPper the ENCODE consortium offers peak files in a BED-like format where the signal library is normalized with a size

[2]Datasets available at `https://www.encodeproject.org/experiments/ENCSR483NOP/`

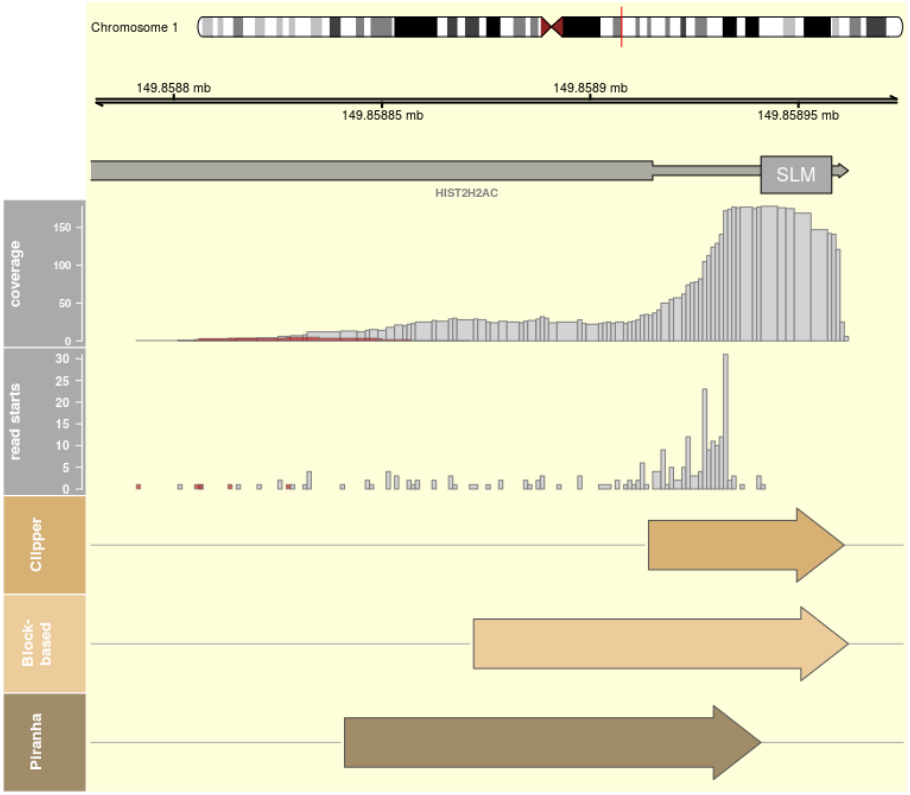

Figure 3: Comparison of called peaks on data stemming from an eCLIP experiment of the RBP SLBP. Tracks from top to bottom. (1) and (2) location on chromosome 1. (3) gene HIST2H2AC (thick) with 3' UTR (thin) and stem-loop motif region (SLM), (4) read coverage (SMI coverage in red), (5) read start coverage (SMI coverage in red), (6) peaks called by CLIPper, (7) peaks called by extended blockbuster, (8) peaks called by Piranha. The figure was generated in the R environment [58] with gviz [59].

matched input library (SMI). The peak boundaries in these BED files were taken as input for the second step of the peak analysis: counting the reads of the signal and the input library that fall into each peak region and evaluating the fold change in the region with DESeq2. Piranha was called with the signal library only to define the peak boundaries and the normalization with the SMI was done thereafter with the same pipeline as for CLIPper and the block-based approach. Piranha can be used with covariates that should normalize the results, but the results of this analysis did not allow for the filtering steps that were applied afterwards as the output of Piranha in this mode was not verbose enough. Furthermore it has to be mentioned that Piranha was called with a bin size of 20 (-z 20) and the merging of bins was disabled (-u 0). Piranha calculates

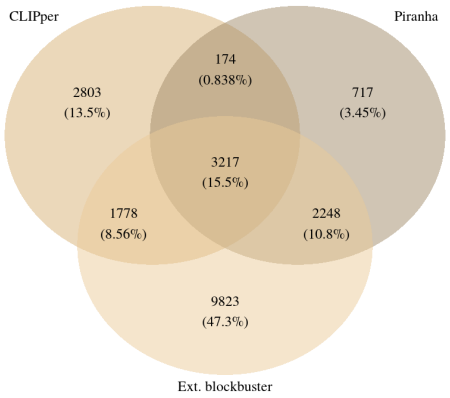

Figure 4: Venn diagram of genomic positions contained in peak regions defined by the three peak callers.

Tools	Number of peaks	Average peak length	Positions in peaks
Piranha	116 (312)	53.79 (20.00)	6240
CLIPper	135 (180)	58.05 (43.54)	7837
Block-based	146 (180)	115.89 (97.97)	17635

Table 1: General statistics of called peaks for the three methods (in brackets: adjacent peaks not merged).

p-values for merged bins counter-intuitively such that results with merged bins were inconsistent because the implicit output filtering of peaks relies on these p-values. In the block-based approach blockbuster was called with a minimum block height of 10 (-minBlockHeight 10), the blocks were extended as described and the resulting regions were again evaluated with DESeq2. Afterwards the identified peak regions were filtered such that only those peaks were kept that had a normalized fold change of at least 2 when comparing signal library to SMI.

The genomic position based overlap between the different peak callers is depicted in Figure 4 and the overall distribution of the peaks determined by the different tools is shown in Table 1. The block-based approach is the most inclusive as it generates the biggest total number of positions in peaks. The number of peaks is quite similar for all three peak callers with the block-based approach generating the longest peaks. Piranha generates many small peaks that are adjacent to each other as expected due to disabled bin merging. Only 11.3 % of peak positions generated by Piranha are exclusive to that tool, for CLIPper and extended blockbuster this percentage is much higher (35.2 % and 57.6 % respectively). This shows that there are significant differences between

Tools	Pros	Cons	Observations
Piranha	• models background • fast	• p-value for merged bin counterintuitive • fixed bin width • no replicates	• calls peaks on read starts • takes read ends instead of starts for minus strand
CLIPper	• models background • dynamic peak width	• slow • needs specific annotation • no replicates	• calls peaks only on known transcripts • broad peaks
Block-based	• dynamic peak width • fast • supports replicates	• does not model background	• relies on blockbuster and DESeq2 • broad peaks • peaks can overlap

Table 2: Observed assets and drawbacks of the described CLIP-seq peak callers.

the tools and it might be worth applying the different tools to the same dataset to find significant regions and subsequently motifs. In any case, an in-depth knowledge of the tools is advisable and the most appropriate tools should be chosen based on the given wet-lab protocol. Table 2 gives an overview of the three tools, adressing strengths, weaknesses and some general observations we gathered during this analysis.

3.3 Postprocessing

The purpose of peak calling is to reduce the false positive rate and provide a set of high affinity binding sites. Albeit peak calling corrects for differences in expression levels to some extent, the results of a CLIP-seq experiment will still be highly dependent on the expression state of the cells in which the experiment was performed. This implies that the problem of false negatives remains since binding sites in lowly expressed genes or genes that are not expressed at all cannot be detected in a CLIP-seq experiment. Consider Figure 5, where we display the read starts of a CLIP-seq experiment on an artificial genomic locus. Due to unspecific binding, reads can be detected outside of true binding sites. Most of these reads are discarded by peak calling. However, the false negatives, i.e. the binding sites which are not covered by reads from the experiment,

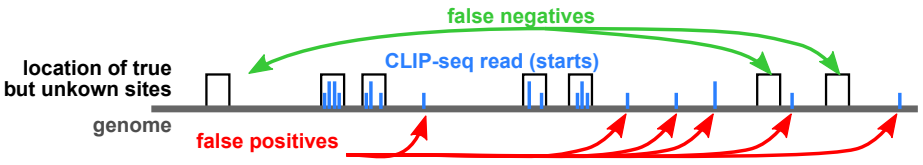

Figure 5: False positives and negatives for a CLIP-seq experiment with respect to true but unknown binding sites.

cannot be found by the analysis described so far. This is a problem when using published CLIP-seq data to analyse cell lines or tissues different from the cell lines that were used to produce the CLIP-seq data. Even for same cell lines there can be considerable variances in expression profiles and thus also differences in recovered binding sites. To give one example, Maticzka et al. [60] re-analysed data from an AGO knockdown by Schmitter et al. [61] using more recent CLIP-seq data [29]. Schmitter et al. showed that genes up-regulated in an AGO knockdown are enriched with putative miRNA-binding sites, consistent with a direct regulation by miRNAs in the wild type. However, one may be inclined to perform an analysis using published CLIP-seq data from the same cell line (which exists), instead of *in silico* seed-based miRNA-binding site prediction, as it was done by Schmitter et al. in the original publication. Surprisingly, Maticzka et al. showed that CLIP binding sites are not enriched in the up-regulated genes, probably due to the low expression of the miRNA-regulated genes in the wild type.

Another example is the work in [62], which shows that publicly available data can be more or less useless (or even harmful by leading to wrong biological conclusions) when only the peak-called sites are used. The group was studying the tumor suppressor *ANXA7*, which is alternatively spliced in glioblastoma compared to normal tissue. They did several experiments to show that an RNA-binding protein, namely the splice factor PTBP1, is involved. Firstly, they showed that *ANXA7* is alternatively spliced. Secondly, they searched for differentially expressed splice factors (again between glioblastoma and normal tissue) and showed that PTBP1 is the only such factor. Thirdly, they did an RNA immunoprecipitation with PTBP1, finding that PTBP1 coprecipitates *ANXA7* RNA. The final step would have been to determine binding sites from a publicly available CLIP-seq dataset, which exists [63]. In this publication a set of binding sites was determined by peak calling. However, as shown in Figure 6B, there are no called binding sites in the vicinity of the alternatively spliced gene, which would lead one to wrongly conclude that there are no binding sites in this transcript region.

Thus, to overcome these kinds of problems and to make publicly available CLIP-seq data usable for a wider community, one has to predict these missing binding sites. Of course, these predictions have to be accompanied by additional experimental approaches to verify them. The general approach for predicting binding sites is to learn a model from the sites detected by a CLIP-seq experi-

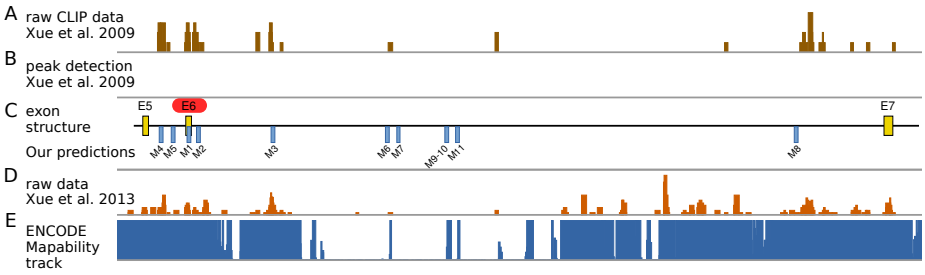

Figure 6: Exon structure of *ANXA7* with the alternative exon E6, which is differentially spliced in glioblastoma. Since exon E6 is repressed by PTBP1, one would expect binding sites of PTBP1 to the left and right of exon E6 [63]. Albeit the raw data from the publicly available CLIP-seq experiment [63] shows some reads in that region (A), no binding sites from the CLIP-seq experiment (as determined by peak calling) can be found (B). We predicted ten binding sites with GraphProt (C). Nine out of the ten predicted sites could be validated by mutation experiments [62]. Track (D) shows the raw read data of a newer CLIP-seq experiment [64]. Some reads accumulate around our predicted binding sites. However, as shown in the mappability track (E), the predicted sites M6,M7, M9-10 and M11 cannot be identified by the CLIP-seq experiment since the reads cannot be uniquely mapped in that region.

ment, and to use this model to determine missing binding sites. In the following we will focus on two approaches for binding site identification most commonly used in CLIP-seq data analysis.

Affinity-based approaches The first types are affinity-based approaches, which try to learn a model that estimates the affinity of the RNA-binding protein P for a specific sequence s. In more detail, consider the binding reaction of a protein P to an RNA sequence s at equilibrium. Then the affinity can be determined by

$$K_a(s) = \frac{[P\text{-}s]}{[P][s]} = \frac{k_{\text{on}}}{k_{\text{off}}} = e^{-\Delta G/RT} \tag{1}$$

where k_{on} (resp. k_{off}) is the rate of association (resp. dissociation), and ΔG is the free energy of binding. $[P\text{-}s], [P]$ and $[s]$ are the concentrations of the protein-sequence complex, the protein, and the sequence, respectively. Now given a set of sequences $\{s_1, \ldots, s_n\}$ that are bound by P, let $\{[P\text{-}s_1], \ldots, [P\text{-}s_n]\}$ be the associated counts indicating how often the sequence s_i occurs as a binding site of protein P. The purpose of motif finding tools is to determine parameters Θ for their models such that the associated score $S_\Theta(s)$ for a sequence s is a good estimate for the affinity, i.e. that $S_\Theta(s) \approx K_a(s)$. If we had enough data and knew the concentration $[s]$ of unbound s for each sequence, then the following score

$$S_\Theta(s) = \frac{[P\text{-}s_i]}{[s_i]\sum_{j=1}^{n}[P\text{-}s_j]}$$

provides such an estimate for the (relative) affinity. However, there are two caveats. First, $[s_i]$ is usually unknown, and is thus often estimated from the background distribution of sequences. Secondly, datasets are usually too small to provide a reliable estimate for $S_\Theta(s)$ for all sequences s. Hence, these scores are often approximated by assuming fixed-length motifs and independent contributions of each position. This basically assumes that the free energy contribution for each base is additive. Since the affinity is related to the free energy as described by equation (1), additivity in the free energy contribution translates to multiplicity in the score. Examples of these types of models are position weight matrices (PWM) [65], as used by the popular MEME tool [66], or position-specific affinity matrices (PSAM) [67].

Given fixed-length motifs, the question is how to score binding sites that are longer than the motif size. Early approaches used the sum of the different subsequences, however, this does not take the concentration of the protein and the effect of binding on the concentration into account. A better approach is to model the occupancy of the sequence by the protein. For a sequence s, the occupancy $N(s)$ is the probability that s is bound by P:

$$N(s) = \frac{[P\text{-}s]}{[P\text{-}s] + [s]} = \frac{[P\text{-}s]\frac{[P]}{[P\text{-}s]}}{[P\text{-}s]\frac{[P]}{[P\text{-}s]} + [s]\frac{[P]}{[P\text{-}s]}} = \frac{[P]}{[P] + K_d(s)},$$

where $K_d(s) = K_a(s)^{-1}$ is the dissociation constant. Assuming that the protein concentration $[P]$ is small compared to $K_d(s)$ to ensure an efficient regulatory scheme [68], one yields

$$N(s) \approx \frac{[P]}{K_d(s)} = [P]K_a(s).$$

Thus, for the small k-mers recognized by the models explained above, the occupancy can be estimated from the score $S_\Theta(s) \approx K_a(s)$. For larger sequences, one can determine the occupancy as the probability that at least one k-mer of the sequence binds. This can be done using a "noisy OR" function [69] by calculating this probability as 1 minus the probability that none of the k-mers bind. RNAcontext [70] is a recent approach for learning sequence and structure preferences for RNA-binding proteins which directly estimates the occupancy of the k-mers using a logistic regression formulation of the occupancy term.

However, it is already known that pure sequence-based models are not good for modeling binding sites on RNAs due to the disregard of secondary structure. Examples of early models that take secondary structure into account are BioBayesNet [71] for DNA and MEMERIS [72] for RNA. More recently, RNAcontext presented a more integrated approach, where the occupancy for

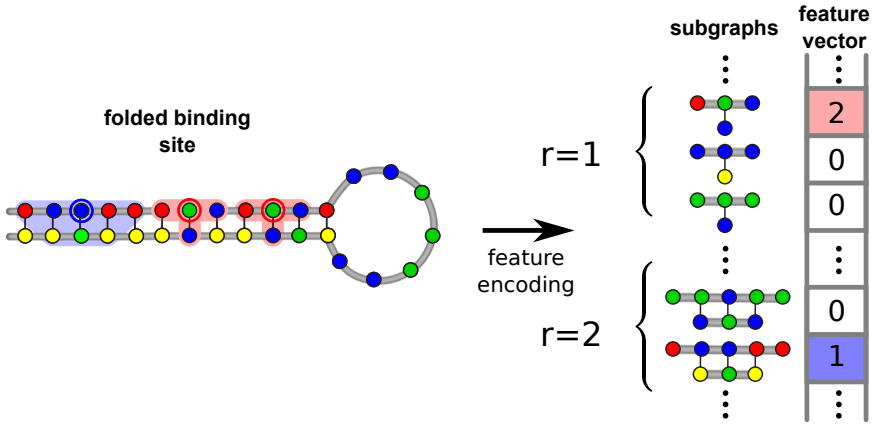

Figure 7: A folded binding site in graph representation (left) and its associated feature vector (right). The red shaded areas on the left indicate two subgraphs of radius (r) 1 with the centre indicated by a red circle (two occurences). The blue shaded area is an example of a subgraph with r=2 (one occurence). Again, the blue circle indicates the central node of the subgraph.

a k-mer s is determined by taking a sequence contribution $N^{\text{seq}}(s)$, which is interpreted as the occupancy of the sequence s in its optimal context, and multiplying it with a contribution of the structural context $N^{\text{struct}}(s, p)$. Here, p is the matrix that assigns to each position a distribution of possible structural states such as being in a hairpin, internal or multi loop, or being in a stem. p is calculated from s using SFOLD [73].

Classification- and Regression-based Approaches Another type of approach is not based on a physical model but considers the problem of determining binding sites as a classification or regression problem. As a classification task, in contrast to the previous approaches, one needs a positive set (i.e. the regions determined by the peak caller) and a negative set, which are sites that are *not* bound by protein. Since the latter is usually not available, the set of negative instances has to be generated, e.g. by shuffling the true regions on the genome. The idea is to determine features that differentiate the binding sites from the non-binding sites. Oversimplifying, when using k-mers, one would try to determine k-mers that are highly enriched in the positive data and depleted in the negative data. However, a simple k-mer approach would not work due to the complexity of the task. Instead, advanced machine learning approaches have to be used.

One example for such an approach is GraphProt [60], which uses sequence- and structure-based features for that purpose. The binding site together with a collection of different near optimal foldings is encoded as a graph. Then, GraphProt considers small subgraphs that are determined by two different parameters, namely radius r and distance d, as features. Starting from each node

of the graph, the radius defines how many edges can be visited to determine the subgraph. The distance parameter d includes as features all possible pairs of subgraphs determined by the radius r that have an edge distance of exactly d. Thus, these features can be considered as the upgrade from sequence k-mers with gaps to graphs. The number of occurrences of these subgraphs are stored into a huge but sparse feature vector (see Figure 7). These feature vectors for each binding and non-binding site are then used by a support vector machine as input to discriminate positive from negative sites. If quantitative binding data is available, support vector regression instead of support vector classification can be used.

Another example for a regression-based approach is iONMF [74], which uses orthogonal matrix factorization to determine a model for the strength of binding sites. iONMF basically uses as features the probability of each position around the binding site to be double-stranded, the number of occurrences for all possible 4-mers in a region around the binding sites, the region type, the GO annotation of the RNA, and the CLIP-seq counts for a collection of proteins different from the one investigated as possible features. The idea for training a model is to determine a coefficient matrix for a linear regression task. I.e. multiplying these coefficients with the values for the features listed above should approximate the CLIP-seq counts of the actual experiment as well as possible, using the determined values for all features. Once this is achieved, new binding sites can be scored by determining the feature values and multiplying them with the coefficients. However, due to the large number of features, one would immediately run into overfitting problems. Omitting a lot of details, iONMF introduces a new approach for orthogonal matrix factorization. The idea is to yield a low-rank approximation of the feature matrices by determining modular projection of the original data matrices, yielding an effective regularization by avoiding multicollinearity between feature vectors.

4 Conclusion

CLIP-seq is currently one of the most important means to determine binding sites of RNA-binding proteins on a genome-wide level. Since peak height alone is not a good measure of significance, we advise preparing signal and background CLIP-seq libraries in replicates. This enables highly specific removal of background noise from the signal data under application of statistical modeling.

The computational analysis of CLIP-seq data requires three steps, which have to be adapted to the specificities of the CLIP protocols to different extents. The first and most protocol-specific step is the preprocessing of the raw data. Sequenced reads have to be trimmed and mapped to the genome or the transcriptome. What exactly has to be trimmed depends on the adapter sequences, as well as barcode sequences for PCR duplicate removal and de-multiplexing. For the mapping part, several widely-applied tools exist which can also handle splice-sensitive mapping.

The second, and one of the most important steps, is peak calling, which

determines high confidence binding sites by removing signals corresponding to unspecific binding. Here, both protocol-specific and generic peak callers exist. However, as shown in the comparison of different peak callers, results can vary drastically. This can even hold within individual tools when they are run using slightly different parameter settings. Thus, depending on the data, it might be worth to apply and compare the results of different peak calling techniques.

In the example case (see Figure 3) Piranha did not include the actual binding motif, which forms a stem-loop structure recognized by SLBP. Instead, bins get called in the upstream vicinity where most read starts occur. This is expected and not a flaw of the algorithm, since Piranha only takes read starts into account. It is known that double-stranded regions are less efficiently crosslinked in CLIP, which would explain the upstream accumulation of crosslink events. On the other hand, transcriptome-wide RBP binding preferences, whether sequence- or structure-dependent or both, are usually not known in advance, and thus one has to rely on the called sites to extract these preferences. Clearly, one may extend the called sites to include more nucleotides, but this can increase the amount of noise and other potentially (non-)RBP specific motifs returned by the analysis. All tested tools can be a reasonable choice, depending on the CLIP-seq protocol, but one should keep in mind their assets and drawbacks (Table 2). For a comprehensive analysis, we recommend trying more than one peak caller, especially if control experiments and replicates are present, which should become standard in future CLIP-seq experiments. A compound strategy, where the steps of site definition and their statistical evaluation are split between programs, could further improve results. Newly developed peak callers should combine the aforementioned strengths of the described programs. In addition, a more thorough study with true positive sets for RBPs targeting structure and sequence features could help to answer the question of which peak caller is suitable in which scenario.

In the last step, which is more or less protocol-independent, motifs are determined and binding models are inferred from the regions identified by the peak caller. The importance of this step is currently largely underestimated. However, without training binding models, published CLIP-seq data can hardly be utilized as they are. In the worst case, the direct use of regions identified in CLIP-seq data on different cells / conditions can lead to wrong conclusions concerning the underlying regulatory mechanisms. The reason is simply that a CLIP-seq experiment is expression-dependent, and binding sites in lowly or not expressed genes are not discovered. If an RNA is expressed in the currently investigated cell type but not in the cell type used for the original CLIP-seq experiment, then binding models can be applied to determine potential missing binding sites. Utilizing these prediction approaches in combination with validation experiments can therefore largely extend the explanatory power of CLIP-seq datasets.

5 Acknowledgments

We thank Daniel Maticzka and Philip Uren for their valuable comments on the manuscript.

6 Funding

This work was funded by the Baden-Württemberg-Stiftung (BWST_NCRNA_008), the German Research Foundation (DFG grant BA2168/11-1 SPP 1738) and the BMBF Verbundprojekt Deutsches Netzwerk für Bioinformatik-Infrastruktur (de.NBI).

References

[1] Kevin V Morris and John S Mattick. The rise of regulatory RNA. *Nat. Rev. Genet.*, 15(6):423–437, 2014.

[2] Alexander G Baltz, Mathias Munschauer, Björn Schwanhäusser, Alexandra Vasile, Yasuhiro Murakawa, Markus Schueler, Noah Youngs, Duncan Penfold-Brown, Kevin Drew, Miha Milek, Emanuel Wyler, Richard Bonneau, Matthias Selbach, Christoph Dieterich, and Markus Landthaler. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell*, 46(5):674–690, 8 June 2012.

[3] Alfredo Castello, Bernd Fischer, Katrin Eichelbaum, Rastislav Horos, Benedikt M Beckmann, Claudia Strein, Norman E Davey, David T Humphreys, Thomas Preiss, Lars M Steinmetz, Jeroen Krijgsveld, and Matthias W Hentze. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, 149(6):1393–1406, 8 June 2012.

[4] Debashish Ray, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, Hong Na, Manuel Irimia, Leah H Matzat, Ryan K Dale, Sarah A Smith, Christopher A Yarosh, Seth M Kelly, Behnam Nabet, Desirea Mecenas, Weimin Li, Rakesh S Laishram, Mei Qiao, Howard D Lipshitz, Fabio Piano, Anita H Corbett, Russ P Carstens, Brendan J Frey, Richard A Anderson, Kristen W Lynch, Luiz O F Penalva, Elissa P Lei, Andrew G Fraser, Benjamin J Blencowe, Quaid D Morris, and Timothy R Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, 11 July 2013.

[5] Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. A census of human RNA-binding proteins. *Nat. Rev. Genet.*, 15(12):829–845, December 2014.

[6] Stefanie Gerstberger, Markus Hafner, Manuel Ascano, and Thomas Tuschl. Evolutionary conservation and expression of human RNA-binding proteins

and their role in human genetic disease. *Adv. Exp. Med. Biol.*, 825:1–55, 2014.

[7] Silvia Carolina Lenzken, Tilmann Achsel, Maria Teresa Carrì, and Silvia M L Barabino. Neuronal RNA-binding proteins in health and disease. *Wiley Interdiscip. Rev. RNA*, 5(4):565–576, July 2014.

[8] Aditi Gupta, Gupta Aditi, and Gribskov Michael. The role of RNA sequence and structure in RNA–Protein interactions. *J. Mol. Biol.*, 409(4):574–587, 2011.

[9] Eckhard Jankowsky and Michael E Harris. Specificity and nonspecificity in RNA-protein interactions. *Nat. Rev. Mol. Cell Biol.*, 16(9):533–544, September 2015.

[10] Yoichiro Sugimoto, Alessandra Vigilante, Elodie Darbo, Alexandra Zirra, Cristina Militti, Andrea D'Ambrogio, Nicholas M Luscombe, and Jernej Ule. hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by staufen 1. *Nature*, 519(7544):491–494, 26 March 2015.

[11] Yasuhiro Murakawa, Murakawa Yasuhiro, Hinz Michael, Mothes Janina, Schuetz Anja, Uhl Michael, Wyler Emanuel, Yasuda Tomoharu, Mastrobuoni Guido, Caroline C Friedel, Dölken Lars, Kempa Stefan, Schmidt-Supprian Marc, Blüthgen Nils, Backofen Rolf, Heinemann Udo, Wolf Jana, Scheidereit Claus, and Landthaler Markus. RC3H1 post-transcriptionally regulates A20 mRNA and modulates the activity of the IKK/NF-κB pathway. *Nat. Commun.*, 6:7367, 2015.

[12] Ibrahim Avsar Ilik, Jeffrey J Quinn, Plamen Georgiev, Filipe Tavares-Cadete, Daniel Maticzka, Sarah Toscano, Yue Wan, Robert C Spitale, Nicholas Luscombe, Rolf Backofen, Howard Y Chang, and Asifa Akhtar. Tandem stem-loops in rox RNAs act together to mediate X chromosome dosage compensation in drosophila. *Mol. Cell*, 51(2):156–173, 25 July 2013.

[13] Xiao Li, Hilal Kazan, Howard D Lipshitz, and Quaid D Morris. Finding the target sites of RNA-binding proteins. *Wiley Interdiscip. Rev. RNA*, 5(1):111–130, January 2014.

[14] Donny D Licatalosi, Aldo Mele, John J Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A Clark, Anthony C Schweitzer, John E Blume, Xuning Wang, Jennifer C Darnell, and Robert B Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469, 27 November 2008.

[15] Valentine Murigneux, Jérôme Saulière, Hugues Roest Crollius, and Hervé Le Hir. Transcriptome-wide identification of RNA binding sites by CLIP-seq. *Methods*, 63(1):32–40, 1 September 2013.

[16] Tao Wang, Guanghua Xiao, Yongjun Chu, Michael Q Zhang, David R Corey, and Yang Xie. Design and bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Res.*, 43(11):5263–5274, 23 June 2015.

[17] Eric L Van Nostrand, Stephanie C Huelga, and Gene W Yeo. Experimental and computational considerations in the study of RNA-Binding Protein-RNA interactions. *Adv. Exp. Med. Biol.*, 907:1–28, 2016.

[18] Jernej Ule, Kirk B Jensen, Matteo Ruggiu, Aldo Mele, Aljaž Ule, and Robert B Darnell. Clip identifies nova-regulated rna networks in the brain. *Science*, 302(5648):1212–1215, 2003.

[19] Jernej Ule, Kirk Jensen, Aldo Mele, and Robert B Darnell. Clip: a method for identifying protein–rna interaction sites in living cells. *Methods*, 37(4):376–386, 2005.

[20] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Jr, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 2 April 2010.

[21] Julian König, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.*, 17(7):909–915, July 2010.

[22] Eric L Van Nostrand, Gabriel A Pratt, Alexander A Shishkin, Chelsea Gelboin-Burkhart, Mark Y Fang, Balaji Sundararaman, Steven M Blue, Thai B Nguyen, Christine Surka, Keri Elkins, Rebecca Stanton, Frank Rigo, Mitchell Guttman, and Gene W Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, 13(6):508–514, June 2016.

[23] Brian J Zarnegar, Ryan A Flynn, Ying Shen, Brian T Do, Howard Y Chang, and Paul A Khavari. irCLIP platform for efficient characterization of protein-RNA interactions. *Nat. Methods*, 13(6):489–492, June 2016.

[24] Georges Martin and Mihaela Zavolan. Redesigning CLIP for efficiency, accuracy and speed. *Nat. Methods*, 13(6):482–483, 31 May 2016.

[25] Nazmul Haque and J Robert Hogg. Easier, better, faster, stronger: Improved methods for RNA-Protein interaction studies. *Mol. Cell*, 62(5):650–651, 2 June 2016.

[26] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 153(3):654–665, 25 April 2013.

[27] Michael J Moore, Troels K H Scheel, Joseph M Luna, Christopher Y Park, John J Fak, Nishiuchi Eiko, Charles M Rice, and Robert B Darnell. miRNA–target chimeras reveal miRNA 3'-end pairing as a major determinant of argonaute target specificity. *Nat. Commun.*, 6:8864, 2015.

[28] Stephen M Testa, Matthew D Disney, Douglas H Turner, and Kierzek Ryszard. Thermodynamics of RNA-RNA duplexes with 2- or 4-thiouridines: Implications for antisense design and targeting a group I intron †. *Biochemistry*, 38(50):16655–16662, 1999.

[29] Shivendra Kishore, Lukasz Jaskiewicz, Lukas Burger, Jean Hausser, Mohsen Khorshid, and Mihaela Zavolan. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods*, 8(7):559–564, July 2011.

[30] Anna-Carina Jungkamp, Marlon Stoeckius, Desirea Mecenas, Dominic Grün, Guido Mastrobuoni, Stefan Kempa, and Nikolaus Rajewsky. In vivo and transcriptome-wide identification of RNA binding protein target sites. *Mol. Cell*, 44(5):828–840, 9 December 2011.

[31] Yoichiro Sugimoto, Julian König, Shobbir Hussain, Blaž Zupan, Tomaž Curk, Michaela Frye, and Jernej Ule. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol.*, 13(8):R67, 3 August 2012.

[32] Michael T Lovci, Dana Ghanem, Henry Marr, Justin Arnold, Sherry Gee, Marilyn Parra, Tiffany Y Liang, Thomas J Stark, Lauren T Gehman, Shawn Hoon, Katlin B Massirer, Gabriel A Pratt, Douglas L Black, Joe W Gray, John G Conboy, and Gene W Yeo. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.*, 20(12):1434–1442, December 2013.

[33] Yu-Cheng T Yang, Chao Di, Boqin Hu, Meifeng Zhou, Yifang Liu, Nanxi Song, Yang Li, Jumpei Umetsu, and Zhi Lu. CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics*, 16(1):51, 2015.

[34] Matthias Dodt, Johannes Roehr, Rina Ahmed, and Christoph Dieterich. FLEXBAR—flexible barcode and adapter processing for next-generation sequencing platforms. *Biology*, 1(3):895–905, dec 2012.

[35] Haibin Xu, Xiang Luo, Jun Qian, Xiaohui Pang, Jingyuan Song, Guangrui Qian, Jinhui Chen, and Shilin Chen. Fastuniq: A fast de novo duplicates removal tool for paired short reads. *PLOS ONE*, 7(12):1–6, 12 2012.

[36] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), 2011.

[37] Felix Krueger. Trim galore. `https://github.com/FelixKrueger/TrimGalore`, 2016.

[38] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, page btu170, 2014.

[39] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[40] Thomas D Wu and Serban Nacu. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, 2010.

[41] Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M. Sharma, Philipp Khaitovich, Jörg Vogel, Peter F. Stadler, and Jörg Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9):e1000502, sep 2009.

[42] Alexander Dobin. Star. `https://github.com/alexdobin/STAR`, 2016.

[43] Ayat Hatem, Doruk Bozdağ, Amanda E Toland, and Ümit V Çatalyürek. Benchmarking short sequence mapping tools. *BMC Bioinformatics*, 14(1):184, 2013.

[44] Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, Tyler Alioto, Jonas Behr, Paul Bertone, Regina Bohnert, Davide Campagna, Carrie A Davis, Alexander Dobin, Pär G Engström, Thomas R Gingeras, Nick Goldman, Gregory R Grant, Roderic Guigó, Jennifer Harrow, Tim J Hubbard, Géraldine Jean, André Kahles, Peter Kosarev, Sheng Li, Jinze Liu, Christopher E Mason, Vladimir Molodtsov, Zemin Ning, Hannes Ponstingl, Jan F Prins, Gunnar Rätsch, Paolo Ribeca, Igor Seledtsov, Botond Sipos, Victor Solovyev, Tamara Steijger, Giorgio Valle, Nicola Vitulo, Kai Wang, Thomas D Wu, Georg Zeller, Gunnar Rätsch, Nick Goldman, Tim J Hubbard, Jennifer Harrow, Roderic Guigó, and Paul Bertone. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods*, 10(12):1185–1191, nov 2013.

[45] Giacomo Baruzzo, Katharina E Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A FitzGerald, and Gregory R Grant. Simulation-based comprehensive benchmarking of rna-seq aligners. *Nature Methods*, 2016.

[46] Philip J Uren, Emad Bahrami-Samani, Suzanne C Burns, Mei Qiao, Fedor V Karginov, Emily Hodges, Gregory J Hannon, Jeremy R Sanford, Luiz O F Penalva, and Andrew D Smith. Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, 28(23):3013–3020, 1 December 2012.

[47] David L Corcoran, Stoyan Georgiev, Neelanjan Mukherjee, Eva Gottwein, Rebecca L Skalsky, Jack D Keene, and Uwe Ohler. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.*, 12(8):R79, 18 August 2011.

[48] Michael J Moore, Chaolin Zhang, Emily Conn Gantman, Aldo Mele, Jennifer C Darnell, and Robert B Darnell. Mapping argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat. Protoc.*, 9(2):263–293, February 2014.

[49] Sebastien M Weyn-Vanhentenryck, Aldo Mele, Qinghong Yan, Shuying Sun, Natalie Farny, Zuo Zhang, Chenghai Xue, Margaret Herre, Pamela A Silver, Michael Q Zhang, Adrian R Krainer, Robert B Darnell, and Chaolin Zhang. HITS-CLIP and integrative modeling define the rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep.*, 6(6):1139–1152, 27 March 2014.

[50] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[51] David L Corcoran, Stoyan Georgiev, Neelanjan Mukherjee, Eva Gottwein, Rebecca L Skalsky, Jack D Keene, and Uwe Ohler. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.*, 12(8):R79, 18 August 2011.

[52] Gabriel Pratt, Michael Lovci, Jill Moore, and ppliu. clipper: release to trigger doi, October 2014.

[53] Olga Botvinnik, Gabriel Pratt, Michael Lovci, ppliu, Leen, and Boyko Kakaradov. gscripts: release 0.1, October 2014.

[54] Erik Holmqvist, Patrick R Wright, Lei Li, Thorsten Bischler, Lars Barquist, Richard Reinhardt, Rolf Backofen, and Jörg Vogel. Global RNA recognition patterns of post-transcriptional regulators hfq and CsrA revealed by UV crosslinking in vivo. *EMBO J.*, 35(9):991–1011, 2 May 2016.

[55] David Langenberger, Clara Bermudez-Santana, Jana Hertel, Steve Hoffmann, Philipp Khaitovich, and Peter F Stadler. Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics*, 25(18):2298–2301, 15 September 2009.

[56] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550, 2014.

[57] Dazhi Tan, William F. Marzluff, Zbigniew Dominski, and Liang Tong. Structure of histone mrna stem-loop, human stem-loop binding protein, and 3'hexo ternary complex. *Science*, 339(6117):318–321, 2013.

[58] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[59] Florian Hahne and Robert Ivanek. *Visualizing Genomic Data Using Gviz and Bioconductor*, pages 335–351. Springer New York, New York, NY, 2016.

[60] Daniel Maticzka, Sita J Lange, Fabrizio Costa, and Rolf Backofen. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.*, 15(1):R17, 22 January 2014.

[61] Daniela Schmitter, Jody Filkowski, Alain Sewer, Ramesh S Pillai, Edward J Oakeley, Mihaela Zavolan, Petr Svoboda, and Witold Filipowicz. Effects of dicer and argonaute down-regulation on mRNA levels in human HEK293 cells. *Nucleic Acids Res.*, 34(17):4801–4815, 13 September 2006.

[62] Roberto Ferrarese, Griffith R. 4th Harsh, Ajay K. Yadav, Eva Bug, Daniel Maticzka, Wilfried Reichardt, Stephen M. Dombrowski, Tyler E. Miller, Anie P. Masilamani, Fangping Dai, Hyunsoo Kim, Michael Hadler, Denise M. Scholtens, Irene L. Y. Yu, Jurgen Beck, Vinodh Srinivasasainagendra, Fabrizio Costa, Nicoleta Baxan, Dietmar Pfeifer, Dominik V. Elverfeldt, Rolf Backofen, Astrid Weyerbrock, Christine W. Duarte, Xiaolin He, Marco Prinz, James P. Chandler, Hannes Vogel, Arnab Chakravarti, Jeremy N. Rich, Maria S. Carro, and Markus Bredel. Lineage-specific splicing of a brain-enriched alternative exon promotes glioblastoma progression. *J Clin Invest*, 124(7):2861–2876, 2014.

[63] Yuanchao Xue, Yu Zhou, Tongbin Wu, Tuo Zhu, Xiong Ji, Young-Soo Kwon, Chao Zhang, Gene Yeo, Douglas L. Black, Hui Sun, Xiang-Dong Fu, and Yi Zhang. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol Cell*, 36(6):996–1006, 2009.

[64] Yuanchao Xue, Kunfu Ouyang, Jie Huang, Yu Zhou, Hong Ouyang, Hairi Li, Gang Wang, Qijia Wu, Chaoliang Wei, Yanzhen Bi, Li Jiang, Zhiqiang Cai, Hui Sun, Kang Zhang, Yi Zhang, Ju Chen, and Xiang-Dong Fu. Direct conversion of fibroblasts to neurons by reprogramming PTB-regulated microRNA circuits. *Cell*, 152(1-2):82–96, 2013.

[65] Gary D Stormo. Dna binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[66] Timothy L Bailey, Nadya Williams, Chris Misleh, and Wilfred W Li. Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic acids research*, 34(suppl 2):W369–W373, 2006.

[67] Barrett C Foat, S Sean Houshmandi, Wendy M Olivas, and Harmen J Bussemaker. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc. Natl. Acad. Sci. U. S. A.*, 102(49):17675–17680, 6 December 2005.

[68] Barrett C. Foat, Alexandre V. Morozov, and Harmen J. Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, 22(14):e141–9, 2006.

[69] Joshua A Granek and Neil D Clarke. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol.*, 6(10):R87, 30 September 2005.

[70] Hilal Kazan, Debashish Ray, Esther T Chan, Timothy R Hughes, and Quaid Morris. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput. Biol.*, 6:e1000832, 1 July 2010.

[71] Swetlana Nikolajewa, Rainer Pudimat, Michael Hiller, Matthias Platzer, and Rolf Backofen. BioBayesNet: a web server for feature extraction and bayesian network modeling of biological sequence data. *Nucleic Acids Res.*, 35(Web Server issue):W688–93, July 2007.

[72] Michael Hiller, Rainer Pudimat, Anke Busch, and Rolf Backofen. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res.*, 34(17):e117, 20 September 2006.

[73] Ye Ding and Charles E Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, 31(24):7280–7301, 15 December 2003.

[74] Martin Stražar, Marinka Žitnik, Blaž Zupan, Jernej Ule, and Tomaž Curk. Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics*, 32(10):1527–1535, 15 May 2016.