StoatyDive: Evaluation and Classification of Peak Profiles for Sequencing Data

- Manuscript Draft -

Manuscript Number: GIGA-D-20-00218R3
Full Title: StoatyDive: Evaluation and Classification of Peak Profiles for Sequencing Data
Article Type: Technical Note
Funding Information:
- Deutsche Forschungsgemeinschaft (322977937/GRK2344 MeInBio) Mr. Florian Heyl
- Deutsche Forschungsgemeinschaft (390939984 EXC-2189) Prof. Dr. Rolf Backofen

Abstract:

Background

The prediction of binding sites (peak calling) is a common task in the data analysis of methods such as crosslinking immunoprecipitation in combination with high-throughput sequencing (CLIP-Seq). The predicted binding sites are often further analyzed to predict sequence motifs or structure patterns. When looking at a typical result of such a high-throughput experiments, the obtained peak profiles differ largely on a genomic level. Thus, a tool is missing that evaluates and classifies the predicted peaks based on their shapes. We hereby present StoatyDive, a tool that can be used to filter for specific peak profile shapes of sequencing data such as CLIP.

Findings

With StoatyDive we are able to classify peak profile shapes from CLIP-seq data of the histone stem-loop-binding protein (SLBP). We compare the results to existing tools and show that StoatyDive finds more distinct peak shape clusters for CLIP data. Furthermore, we present StoatyDive’s capabilities as a quality control tool and as a filter to pick different shapes based on biological or technical questions for other CLIP data from different RNA binding proteins with different biological functions and number of RNA recognition motifs. We finally show that proteins involved in splicing, such as RBM22 and U2AF1, have potentially more sharper shaped peaks than other RNA binding proteins.

Conclusion

StoatyDive finally fills the demand for a peak shape clustering tool for CLIP-Seq data that fine tunes downstream analysis steps such as structure or sequence motif predictions and that acts as a quality control.
authors leave threads either as superficial data summaries or hanging as "it is worth to investigate..."; I don't find these helpful, either they are important for proving the utility of StoatyDive (and should be clarified and included) or are not (and maybe are better left out or mentioned in the discussion section).

\begin{answer}

We thank the reviewer for the suggestion. It was a bit difficult to know, which sections and comments the reviewer referred to. We further revised the text for better clarity in reflecting such kind of statement.

We removed instances of "it is worth to investigate" from the paper. All other "worth of investigate" are either explained in the paper or used as a phrase for an outlook in the conclusion section. We also removed or rephrased a lot of sentences with "might" or phrases that were still vague.

\end{answer}

While many of these I think can be corrected with simple text editing, for me in the first section that issue remains - there's 3 figures and significant text devoted to this section, and at the end of that section I still don't have a good sense of what the authors think is going on here with respect to the CV distribution between replicates. Is it within range of what's expected? Surprisingly low enough to suggest a data problem with replicate 1?

\begin{answer}

We state now in the paper: "The CV distribution of the input control was expected because an ideal control experiment should contain no real or not enriched binding events, that is to say, the value of all CVs is expected to be very small and close to $0$. But, the CV distribution of replicate 1 was more similar to the control experiment than to replicate 2. A CLIP experiment should result into a peak set with enriched regions and thus more specific peaks. The distribution of replicate 1 indicates some degree of variability in the binding events. However, we have to stress out that this does not necessarily depicts a poor quality of replicate 1 and 2."

And also in the methods section: "The user receives a first impression of the binding specificity of the protein of interest from the CV distribution. An unspecific binder has a CV distribution $\approx 0$. A more specific binder has a CV distribution $> 0$. The CV distribution can also be used as a quality control to compare control and signal experiments. A quality breach might have occurred if the distributions of the control and signal experiment almost look identical. A control experiment should normally have a CV distribution $\approx 0$, with only a very few binding sites showing higher CVs. A CLIP experiment, on the other hand, should contain more peaks with higher CVs and thus have a CV distribution that significantly differs in comparison to the control (Wilcoxon P-value $< 0.05$, see Supplementary Figure 1)."

For both paragraphs we generated a new plot that we included in the Supplementary Figure 1, stating in the paper: "Furthermore, we report a plot for the mean CV for the CLIP data in comparison to the size-matched input control of each protein. The control data tends to have a CV close to 0. The CV distributions between CLIP and control data always have a two-sided Wilcoxon test P-value $< 0.05$."

Consequently, we show empirically the difference of the CV distribution between CLIP and control data.

\end{answer}

I appreciate that they edited the language to weaken the claim that the data itself is irreproducible, but I think they may have misinterpreted my prior review comment in that sense - my concern wasn't really inclusion of that claim, it was having analyses to validate it. I don't think it makes sense to simply write around this question in this way - it's a critical part of convincing that the method is of broad utility.

To phrase it a different way - I think the rest of the manuscript is now solidly presented once you have confidence that the CV metric is robust. But I think it's still missing evidence for whether the difference between replicates is biological / experimental
noise or simply reflects extreme sensitivity / variability of the CV metric, and if it's the latter then it's problematic for believing the downstream analyses are robust.

Maybe as a more straightforward question - if one dataset (say the SLBP Replicate 2 one) is simply downsampled by splitting it in half into two pseudo-replicates, how consistent are the CV values for each peak? I think that would be convincing that the results seen in Fig 2 and 3 are real rather than noise.

\begin{answer}
Thank you very much for this question so that we can clarify. We tested your suggestion and split the dataset for replicate two into two pseudo-replicates using samtools (v 1.12). we state in the paper:

"We checked the robustness of the CV distribution and split the second replicate into two pseudo-replicates. We took randomly 50%, with and without replacement, of the reads for pseudo-replicate 1 and the other half for pseudo-replicate 2. Both scenarios gave us similar results. Without replacement, replicate 1 had a mean CV of $\approx 1.05$ and replicate 2 a value of $\approx 1.03$, with a P-value of two-sided Wilcoxon test of $0.768$. With replacement, replicate 1 had a mean CV of $\approx 1.05$ and replicate 2 a value of $\approx 1.02$, with a P-value of the Wilcoxon test of $0.7546$. For example, the sharp peak Figure\textsuperscript{a} had a CV in both scenarios of $\approx 5.2$ and $\approx 5.3$ for pseudo-replicate 1, and $\approx 4.8$ and $\approx 5.4$ for pseudo-replicate 2. The peak Figure\textsuperscript{a} was in the peak list on position $47$ and $42$ for pseudo-replicate 1, and $52$ and $45$ for pseudo-replicate 2. The broad peak Figure\textsuperscript{b} had a CV $\approx 0.001$ (position $662$) and $\approx 0.001$ (position $658$) for pseudo-replicate 1, and $\approx 0.001$ (position $667$) and $\approx 0.001$ (position $655$) for replicate 2. A peak stayed as sharp or broad despite the random sampling. Thus, the overall trend of the CV distribution and the peak position in the list were robust to some random noise."
\end{answer}

\section{Final remark:}
We thank the reviewer for his suggestion. All changes are marked in red. We included a new figure in the Supplementary Figure 1.

### Additional Information:

| Question                                      | Response |
|-----------------------------------------------|----------|
| Are you submitting this manuscript to a special series or article collection? | No       |
| **Experimental design and statistics**        | Yes      |
| Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. |          |
| Have you included all the information requested in your manuscript? | Yes      |
| **Resources**                                 | Yes      |

Powered by Editorial Manager\textsuperscript{®} and ProduXion Manager\textsuperscript{®} from Aries Systems Corporation
A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.

Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?

| Availability of data and materials | Yes |

All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.

Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?
TECHNICAL NOTE

StoatyDive: Evaluation and Classification of Peak Profiles for Sequencing Data

Florian Heyl¹,* and Rolf Backofen¹,2,*

¹Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany and ²Signalling Research Centres BIOSS and CIBSS, University of Freiburg, Schaenzlestr. 18, 79104 Freiburg, Germany

*heylf@informatik.uni-freiburg.de, backofen@informatik.uni-freiburg.de

Abstract

Background The prediction of binding sites (peak-calling) is a common task in the data analysis of methods such as crosslinking immunoprecipitation in combination with high-throughput sequencing (CLIP-Seq). The predicted binding sites are often further analyzed to predict sequence motifs or structure patterns. When looking at a typical result of such high-throughput experiments, the obtained peak profiles differ largely on a genomic level. Thus, a tool is missing that evaluates and classifies the predicted peaks based on their shapes. We hereby present StoatyDive, a tool that can be used to filter for specific peak profile shapes of sequencing data such as CLIP.

Findings With StoatyDive we are able to classify peak profile shapes from CLIP-seq data of the histone stem-loop-binding protein (SLBP). We compare the results to existing tools and show that StoatyDive finds more distinct peak shape clusters for CLIP data. Furthermore, we present StoatyDive’s capabilities as a quality control tool and as a filter to pick different shapes based on biological or technical questions for other CLIP data from different RNA binding proteins with different biological functions and numbers of RNA recognition motifs. We finally show that proteins involved in splicing, such as RBM22 and U2AF1, have potentially sharper-shaped peaks than other RNA binding proteins.

Conclusion StoatyDive finally fills the demand for a peak shape clustering tool for CLIP-Seq data that fine-tunes downstream analysis steps such as structure or sequence motif predictions and that acts as a quality control.

Key words: CLIP-Seq; data analysis; peak shape clustering; RNA; protein

Findings

Background

The biological function of a protein is determined by its interaction partners and the mode of interaction. Studying these interactions broadens our horizon about the cellular mechanisms such as alternative splicing and post-transcriptional regulation. Crosslinking immunoprecipitation in combination with high-throughput sequencing (CLIP-Seq) fathoms these interactions. CLIP-Seq investigates all interactions between an RNA binding protein (RBP) and its target RNAs [1]. CLIP-Seq thus scrutinizes the post-transcriptional regulation by RBPs. Prediction of binding regions (peak-calling) is a crucial step in the data analysis of methods such as CLIP-Seq. Before the peak analysis there is typically no evaluation and classification of the peak characteristics. Yet, the obtained peak set might have different peak profiles that are worth to filter to refine a downstream analysis. The different peak shapes are the result of several biological and technical problems.

Many RBPs have several binding domains with different binding affinities, and are often part of protein complexes, leading to an intricate binding pattern. As described in a review by Jankowsky, Eckhard and Harris, Michael E [2], there are specific and unspecific binders. Examples for unspecific binders are often RBPs that need to bind many RNAs such as mRNA export factors [3]. Another example of common unspecific binders are RNA helicases. However, even more specific
RBP's bind RNAs in large range of affinities, indicating that different binding sites vary in their binding specificity. While many factors, such as the affinity of an RBP for the binding site and the concentration of the protein and RNA, influence the binding specificity, it is likely that these factors are manifested in the CLIP binding profile landscape. At this point, however, no tool exists that can be used to study this possibility in more detail.

In addition, technical biases might change the peak profile landscape. Binding artifacts might be introduced during read library preparation. Protocol biases, for example, PAR-C~LIP biases that are introduced by endonuclease and photoactivatable nucleosides [4], might also affect the binding site predictions. On top, the peak caller itself might generate specific peak profiles and false positives, which the user might not want to have in their data.

This leads to many questions in the data analysis of binding sites that can currently not be answered adequately. Examples are: Does my protein of interest bind generally specific (Figure 1a) or unspecific (Figure 1b)? Does my RBP of interest have more than one binding motif? Does my experiment have any quality issues, meaning, do my reads come from unspecific bindings because of library preparation artifacts? Does my protocol generate biases? Do I have false positives in the set of predicted peaks from my peak caller of choice?

We hereby present StoatyDive, a tool to evaluate and classify peak profiles to help to answer the aforementioned questions. StoatyDive uses the whole peak profiles as well as predefined features to do a peak shape clustering for sequencing data. In this paper, we will test StoatyDive on CLIP data of the eCLIP protocol from the histone stem–loop–binding protein (SLBP) from the study by Van Nostrand et al. [5]. SLBP has been reported to be a histone mRNA export and translation factor [6]. StoatyDive delivers several plots and a table to assess the different binding profiles of a protein. The tool assists to select specific and unspecific binding sites and to find similar shaped peak profiles. Thus, we try to refine the obtained peaks of the SLBP data to find more specific sites of SLBP. It also helps as a quality assessment to validate a CLIP–Seq or any other binding experiment. Later in the paper, we use StoatyDive to investigate the peak profile landscape of different RBPs with different biological functions and different number of RNA recognition motifs (RRM). StoatyDive comes with some test data and a quick installation guide.

Data Preparation of SLBP and Analysis

We used eCLIP data of the histone stem–loop–binding protein (SLBP; ENCSR483NOP; GSEq1802; Van Nostrand et al. [5]). The data comprised 2 CLIP replicates and a size-matched input control from immortalised mylogenous leukemia cells (K562). We processed the data with the schnakemake pipeline SalamiSnake (https://github.com/BackofenLab/SalamiSnake, v0.0.1) for eCLIP data. SLBP has been reported to be cytoplasmic but to be present also in the nucleus [6]. Thus, we mapped the reads against the human genome (version hg38) with STAR [7] also taking the transcriptome into account. We predicted potential binding sites of SLBP with PureCLIP [8], which we ran for each CLIP replicate separately, taking the input control into account. We extended the predicted binding regions by 20 nucleotides left and right because PureCLIP often underestimates the binding region. We further fused the predicted peaks from each CLIP replicate with bedtools [9] to get a robust set of predicted binding sites, which resulted in 899 robust peaks. We executed StoatyDive (v1.1.0 with umap v0.2.5.0) with length normalization, a penalty for broader plateaus, and peak profile smoothing. The complete call was: StoatyDive.py -a peaks.bed -b reads.bam -c hg38.chrom.sizes.txt --peak_correction --scale_max 10 --border_penalty --sm.

Peak Profile Landscape Reveals Variability of Binding Sites

The user obtains from StoatyDive a distribution of the coefficient of variation (CV), calculated for each peak, to get a broad overview of the peak profile landscape of their experiment (see Methods). Broader peaks tend to have a $\text{CV} \approx 0$. Although the CV distributions of the input control and replicate 1 of the SLBP data differed significantly (one-sided Wilcoxon test P-value $\approx 0.03$), both contained a lot of regions with a CV $\approx 0$ (Figure 2, both with a mean of CV of 0.47). In contrast, the CV distribution of replicate 2 was distinct (P-value $< 0.05$ to input control and replicate 1) since it had more peaks with a higher CV (mean CV of 1.44) and thus more specific binding events (e.g., Figure 1a, CV $\approx 3$). Yet, some potential binding sites were more unspecific with a CV $\approx 0$ (e.g., Figure 1b, CV $\approx 0.0003$).

The CV distribution of the input control was expected because an ideal control experiment should contain no real or not enriched binding events, that is to say, the value of all CVs is expected to be very small and close to 0 (see Supplementary Figure 1). But, the CV distribution of replicate 1 was more similar to the control experiment than to replicate 2. A CLIP experiment should result into a peak set with enriched regions and thus more specific peaks. The distribution of replicate 1 indicates some degree of variability in the binding events. However, we have to stress out that this does not necessarily depict a poor quality of replicate 1 and 2. For a downstream analysis, for example the prediction of sequence motifs, it is worth investigating why the CV distribution of replicate 1 was very different to replicate 2. A sequence motif prediction depends on the selected binding sites, thus StoatyDive’s inspection allows the user to assess the binding sites using the CV distribution and a quick installation guide.

---

**Figure 1.** We show 2 significant peaks of a CLIP experiment for the protein SLBP (ENCSR483NOP, replicate 2). One can see peaks with drastically different peak profiles, pointing towards more specific (a) or unspecific (b) binding. Current analysis of CLIP binding sites is typically based on manual inspection of a few peaks. Thus a general tool is missing that allows to filter, cluster and quantify peak profiles and thereby refine downstream analysis tasks for data such as CLIP. StoatyDive assists to find and distinguish peaks like (a) and (b).

**Figure 2.** StoatyDive generates a CV distribution to evaluate the peak profile shapes, which can be used as a quality control. The CV distribution of the peak profiles of the input control and replicate 1 of the SLBP CLIP–Seq experiment are quite similar. In contrast, the CV distribution of replicate 2 is different. The number of uniquely mapped reads is listed below the sample name.
butions. This helps to appraise if SLBP has different binding mechanisms, which we further investigated and discussed in section "Information from Peak Profile Shapes". We checked the robustness of the CV distribution and split the second replicate into two pseudo-replicates. We took randomly 50%, with and without replacement, of the reads for pseudo-replicate 1 and the other half for pseudo-replicate 2. Both scenarios gave us similar results. Without replacement, replicate 1 had a mean CV of $\approx 1.05$ and replicate 2 a value of $\approx 1.02$, with a $P$-value of the Wilcoxon test of 0.768. With replacement, replicate 1 had a mean CV of $\approx 1.05$ and replicate 2 a value of $\approx 1.03$, with a $P$-value of two-sided Wilcoxon test of 0.768. For example, the sharp peak in Figure 1a had a CV in both scenarios of $\approx 5.2$ and $\approx 5.3$ for pseudo-replicate 1, and $\approx 4.8$ and $\approx 5.4$ for pseudo-replicate 2. The peak Figure 1a was in the peak list on position 47 and 42 for pseudo-replicate 1, and 52 and 45 for pseudo-replicate 2. The broad peak Figure 1b had a CV $\approx 0.001$ (position 662) and $\approx 0.001$ (position 658) for replicate 1, and $\approx 0.001$ (position 667) and $\approx 0.001$ (position 655) for replicate 2. A peak stayed as sharp or broad despite the random sampling. Thus, the overall trend of the CV distribution and the peak position in the list were robust by including random noise.

Checking the CV distributions of other CLIP-Seq datasets such as TAF15, TARDBP, and HNRNPA1 (Supplementary Figure 1), which we will analyze further in a later section, we saw that the CV distributions also had differences between the replicates (two-sided Wilcoxon test $P$-value < 0.05). However, this does not mean a low quality of the data and just highlights that it is important to do replicates in order to quantify biological and technical variance as noted in a previous CLIP study [10]. To investigate further differences between the two replicates, we split the peak set into peaks overlapping with exons and introns (see Figure 3). SLBP is a translation and transport factor, which is present in the cytoplasm as well as nucleus [6, 11]. Therefore, the replicates could have different binding events, where one replicate might have more events in cytoplasm and the other more in the nucleus. Replicate 2 had 136 exonic and 136 intronic peaks more than replicate 1. Notably, we can see a CV difference when comparing the intronic peaks of the two replicates, with a mean CV of 0.23 for replicate 1 and 1.26 for replicate 2 (one-sided Wilcoxon test $P$-value < 0.05). The exonic peaks on the other hand were more similar (mean CV = 0.48 and 0.90, respectively), but still the CV distributions were significantly different (one-sided Wilcoxon test $P$-value < 0.05). Therefore, StoatyDive showed a variability of binding events (intronic versus exonic) between the two replicates.

Seven Different Peak Shapes in the SLBP Data

For a more detailed analysis, we classified the peaks of replicate 1 and 2 with the help of StoatyDive (Figure 4). The procedure is mentioned in the methods. StoatyDive has found 7 distinguishable peak profiles for both replicate 1 and 2. We looked more closely at the profiles of replicate 2 (Figure 4b). Cluster 2 and 5, which are set apart clearly from clusters 1, 3, 4, and 6, are characterized by plateau–shaped profiles. The other groups had profiles with mountain–like shapes with peaks tending to become broader and fuzzier in the order of clusters 3, 1, 6, and 4. To return to our initial examples (Figure 1), peak profile Figure 1a was classified by StoatyDive as a small, centered mountain (Figure 4b1), whereas peak profile Figure 1b was classified as a very broad profile (Figure 4b2).

It is to mention that constant profiles (Figure 4) represent a constant read coverage throughout the whole peak. Because of the max–min normalization of the profile (see Methods), the value becomes 0, that is to say, the profile is not empty.

---

**Figure 4.** Results of the peak profile clustering with StoatyDive (procedure described in the methods). We applied StoatyDive to the SLBP data [5]. StoatyDive has found 7 different peak profile shapes in the data of replicate 1 (a1–γ) and replicate 2 (b1–γ) of SLBP. We present one example profile for each cluster with the number of peaks on top. For replicate 1 and 2 we could separate between very thin and specific mountains such as Figure 1a and very broad profiles like Figure 1b. We also found peaks shaped like plateaus, such as 3a, and constant peaks, for example, 7a. The profiles also vary slightly between groups. For example, 6b has more than one spiky mountain in contrast to 1b.
In contrast, peaks shaped like plateaus have not a constant value since their coverage changes at a few positions. Furthermore, the number of clusters depend on the optimization of StoatyDivE, but can also be defined by the user.

The high distance between cluster 2 and 5 is the result of the difference between the profile borders (compare Figure 4 b2 and b5). Where cluster 2 had profiles with lots of values in the left or right side of the peak profile, cluster 5 has occupied the center of the peak profile.

SLBP has been reported as an mRNA export and translation factor [6]. Thus, it is worth to investigate if peaks like Figure 1a are more informative for a translation factor than peaks like Figure 1b. That is to say, Figure 1a might be more suited for sequence and structure predictions than peak Figure 1b. Therefore, we will do a deeper inspection of group 1, 3, 4 and 6 later in the next section of the paper.

We also checked the uniqueness of the shapes by analyzing the peaks based on the reads from the size–matched input control (see Supplementary Figure 2). StoatyDivE had just identified 4 different clusters, encompassing mountain-shaped peaks, as well as plateaus, and constant peaks. It was to be expected to find similar shapes in the control because the biggest challenge for peak-calling is the identification of enriched sites with different shapes between control and CLIP data [12, 13]. In a future version of StoatyDivE, we will include a mode to check peak shapes between samples to see if we could improve peak-calling results with a peak shape comparison.

### Information from Peak Profile Shapes

We made the assumption that replicate 1 might have more unspecific and less distinguishable profiles than replicate 2 based on the different CV distributions (Figure 2). Thus, we counted the number of peaks in each cluster for replicate 1 and 2 (Table 1). From our robust 899 peaks, in replicate 1 we had ≈ 19% peaks being a sharp mountain shape (Figure 4 a4 and a6), ≈ 53% being a broader mountain (Figure 4 a1, a2 and a5), ≈ 6% peaks with plateaus (Figure 4 a3), and ≈ 22% constant shaped peaks (Figure 4 a7). Replicate 2, on the other hand, had ≈ 29% sharply mountain-shaped peak profiles (see Figure 4 b1 and b3), so 94% more than replicate 1. This corroborated the assumption that replicate 1 had broader and more unspecific sites. Thus, replicate 2 had only ≈ 49% broad peak profiles (Figure 4 b4 and b6), and only 2 constant peak profiles (Figure 4 b7). Yet, replicate 2 had ≈ 21% peaks with plateaus (Figure 4 b2 and b5).

We further investigated the biological function of different peak profiles of replicate 2. Since SLBP targets histone mRNAs [6], we intersected known annotated mRNAs of histones with the peaks of the different profile clusters (Table 1). From the 899 peaks, only ≈ 13% of replicate 2 overlapped with mRNAs of histones. Yet, of these 118 peaks almost all came from group 1, 3, 4, and 6. These groups were either spiky, or broader mountain–shaped peak profiles. For example, we found a sharper peak located on RNU7-1 RNA (U7 small nuclear 1) that contains a stem loop that might be potentially targeted by SLBP [14]. The peak got a CV of 3.9 and was classified into the peak profile cluster 3 of replicate 2. Only 5 peaks intersected with histone mRNAs that had a profile shaped like a plateau (Figure 4 b5). This endorsed the assumption that peak profiles shaped like plateaus were less informative. The observation also suggests that broader profiles were still informative because some of them overlapped with histone mRNA.

Next, we used MEME–ChIP [15] to search for sequence motifs associated with the different peak shape groups of the second replicate of SLBP. We have found 2 significantly enriched motifs associated with the plateau peaks and 3 motifs associated with the sharper peaks (Table 2). Yet, both the plateaus and the sharper peaks had two similar sequences motifs. Both motifs (G)GUCU(UU) and (CA)GACA(C) were higher enriched in the sharper–shaped peaks. On the other hand, we found > 10 enriched motifs for the broader–shaped peaks. The motifs of that peak set were very different to the motifs of the plateaus and sharper–shaped peaks. Even the first 3 significantly enriched motifs had more noise and consequently were less enriched than the motifs of the other 2 peak shape groups. The E–values of all motifs were also > 1000 times higher for broader peaks than for sharper peaks. Peaks shaped like plateaus were slightly more significant than broader peaks.

On further inspection of those peak motifs in histone mRNAs, we found that the 3 motifs of the sharper shape peaks covered 10 more histone mRNAs (49 in total) than the broader–shaped peaks (39 in total). This endorsed the observation that broader peaks encompassed more noise. For example, the second motif of the broad peaks CA(A/C)CAAG came close to the third sequence motif, with the sequence AC(A/C)ACAAAG, of the sharper shape peak group. The second motif had the highest occurrence in histone mRNAs for the broader peaks. Thus, the broader peak set includes true binding sites but with some higher additional noise. Furthermore, we already showed that the plateau group might also hold some peaks that are true binding sites (Table 1), which was confirmed by the similar sequence motifs to the sharper peak set. We could confirm that the 2 sequence motifs that are present in plateaus were also present in histone mRNAs (Table 1). All in all, just the sequence motifs analysis showed how different the outcome of subsequent tasks can be for different peak shape groups. Yet, we cannot confirm the biological truth behind those motifs as it requires further experiments to verify them.

### Table 1. Number of peaks of SLBP for different shape groups.

| Replicate | Total | Sharp | Broad | Plateau | Constant |
|-----------|-------|-------|-------|---------|----------|
| 1         | 899   | 171   | 481   | 51      | 196      |
| 2         | 899   | 265   | 444   | 188     | 2        |

| Peak Summits in Histone mRNAs |
|-------------------------------|
| 1 | 116 | 22 | 86 | 6 | 2 |
| 2 | 118 | 42 | 71 | 5 | 0 |

### Table 2. First 3 MEME–ChIP motifs for the different peak shape groups of SLBP with the E-value, the portion of sequences that have the motif, and the number of peaks that have the motif and which also intersect with histone mRNAs.

| Shape | Motif 1 | Motif 2 | Motif 3 |
|-------|---------|---------|---------|
| Broad | 3.5e–4; 5.85% | 1.5e–3; 5.40% | 1.5e–3; 5.40% |
|       | 13      | 18      | 8       |
| Sharp | 5.7e–12; 20.00%| 8.6e–9; 18.87%| 1.0e–3; 9.06%|
|       | 23      | 17      | 9       |
| Plateau | 1.6e–5; 15.42%| 1.6e–3; 12.23%| |
|        | 3       | 2       |         |
Optimizing StoatyDive with the data of SLBP

We investigated the peaks of the second replicate of SLBP further and took the CLIPper peaks (ENCFF127WAK) from the study by Van Nostrand et al. [5]. We wanted to check the specificity and sensitivity of StoatyDive for different CV cutoffs and peak sizes (see Supplementary Table 3). PureCLIP does not give a FC or P-value, so it was only possible to calculate those features with the CLIPper peaks. Since SLBP binds mainly histone mRNAs [6], we defined peaks in histones with a log2 fold change (LFC) > 1 and a P-value < 0.05 as true positives. A true negative was a non-significant peak in a region that did not overlap with a histone. We investigated only peaks where we could calculate a CV and used StoatyDive with a peak size of 30 (median), 40 (Q3), 70 (Q3 + 1.5 * IQR) and a maximum peak length of 201 nucleotides. The peak sizes were chosen based on the length of all peaks (see Supplementary Figure 4). The Matthews correlation coefficient (MCC) is more informative in case of imbalanced datasets, which was the case of the SLBP data. The true negative (TNR) was always at 1.0 because the peak set had no peaks in histones that were not significant (no false positives). We investigated the cluster with the highest (Main Cluster) and second highest (Second Cluster) number of peaks in histones. Looking solely at the clustering, we achieved the highest true positive rate (TPR) (0.69) and MCC (0.66) with a peak length of 70 than does the set of all CLIPper peaks with a TPR of 0.48 and MCC of 0.57. This was achieved by the second cluster, which pointed out that the cluster had to be carefully chosen in order to remove some noise in the data (artifacts). The cluster had also the highest enrichment with a mean LFC of 3.4 (median P-value < 0.05) than does all CLIPper peaks with a value of 1.9 (median P-value = 0.038). Furthermore, the same peak length of 70 nucleotides achieved the highest TPR (0.61) and MCC (0.68) when we filtered the peaks based on a CV threshold of 0.2. However, the enrichment was not as good as with the clustering by StoatyDive, where we achieved a mean LFC of 3.4 in the second cluster that was higher than the resulting mean LFC of 1.6 (median P-value = 0.108) with the CV cutoff of 0.2. Based on these different sets, we observed that the CV was sometimes lower in the main or second cluster. Together with the previous results, we concluded that broad and sharp peaks equally play an important role for SLBP.

Comparison to Existing Tools

To further validate StoatyDive, we applied 2 other peak shape clustering tools, namely FunChIP (Parodi et al. 16; version 0.99.4) and SIC–ChIP (Cremona et al. 17; current release), to the second replicate of SLBP. To have a better ground truth, we took 10 peaks of three different peak shape groups (broad, sharp, and plateau) to define a test set with three distinct peak shapes from real CLIP-Seq data (in total 30 peaks). We defined peaks as sharp (CV > 1.0) and broad (CV < 0.05) based on the calculated CV of StoatyDive. We selected peaks shaped like plateaus by inspecting them in a genome browser. We strictly used the output of the tested tools. Both tools were designed and tested for ChIP data. A peak shape clustering was so far not done for CLIP data and a specific tool for that data type did not exist to the best of our knowledge. We applied SIC–ChIP with N = 10 and toll = 10 (the default parameter set resulted in errors) and ran FunChIP according to the manual in Bioconductor with the smoothing parameter lambda = 103. StoatyDive classified most peaks correctly into the 3 peak shape groups (Figure 5a, accuracy (ACC) = 0.87). It sorted 4 peaks incorrectly that came from the broad and sharp peak group. Sharp and broad peaks are harder to cluster and a second factor such as the CV helps to give a final assessment over the shape of the peak. SIC–ChIP identified up to 6 different peak shapes (Figure 5b), whereas FunChIP found 3 (Figure 5c and d, ACC = 0.6). Furthermore, SIC–ChIP as well as FunChIP had clusters that are mixed and not as well separated as with StoatyDive. SIC–ChIP’s predefined shape indices were not enough to separate the peak shape profiles properly, as shown for one scatter plot (Figure 5b) with the highest explained variance (cluster separation). In turn, FunChIP performed slightly better, finding profiles with different summit intensities. However, for the smoothed (Figure 5c) as well as the smoothed and scaled profiles (Figure 5d) the clusters included a lot of profiles with different shapes. For example, cluster 2 and cluster 3 of the smoothed profiles seemed very similar. Thus, FunChIP’s approach to use the whole profile without any predefined features or dimensional reduction was also not enough to separate the peak shapes in the same way as with StoatyDive.

Investigation of eCLIP Protein Profiles

We further investigated the peak shapes of several proteins from the study of Van Nostrand et al. [5], namely: CPSF6, CSTF2T, EWSR1, LARP7, RBM22, SAFB2, SLBP, SLTM, TAF15, TRA2A, U2AF1, HNRNPA1, IGF2BP1, IGF2BP2, NONO, SRSF1, TARDBP, HNRNPM, U2AF2, and PTBP1. We took the robust peaks (peak-calling, IDR, signal normalization) and the bam files, which were used for the peak-calling, from each protein from the ENCODE database. We chose the data from the eCLIP experiment on R562 and focused on proteins where the biological and molecular function and the number of RRMs are clearly
Figure 6. (a) Log$_2$ fold change of the number of sharp peaks versus the number of broad peaks; (b) fraction of sharp peaks. UniProt lists the proteins CPSF6, CSTF2T, EWSR1, LARP7, RBM22, SAFB2, SLBP, SLTM, TAF15, TRA2A, and U2AF1 with one RRM and the proteins HNRNPA1, IGF2BP1, IGF2BP2, NONO, SRSF1, TARDBP, HNRNPM, U2AF2, and PTBP1 with at least two RRMs. Proteins with more than one RRM tend to have sharper–shaped peaks, but also equally have broader peaks. The proteins RBM22, U2AF1, TARDBP, HNRNPM, U2AF2, and PTBP1 had a higher number of sharper peaks and all of them are involved in RNA splicing (colored bars). Other proteins with different functions (stripes) often have less sharper peaks.

listed on UniProt. Thus, we wanted to investigate if the number of RRMs or the function of the protein by any means affected the shape of the peak profiles and consequently led to more or less broader peaks. We therefore took the files from ENCODE and merged both bam files (replicates) for the coverage. We then used StoatyDive with --peak_correction --scale_max 10 --border_penalty --sm --peak_length 77 -k 3. All peaks were therefore extended or shrunk to a length of 77 nucleotides. This was based on the observation that the third quartile of all peaks from all proteins was 77 nucleotides long (see Supplementary Figure 4). In addition, StoatyDive achieved for the SLBP data a better TPR and MCC with a peak size of 70 and a CV threshold of 0.2 (see Supplementary Figure 3). The results of SLBP showed that it may be wise to combine the clustering and the CV threshold to assess the profile landscape of other proteins. We therefore defined a peak as sharp if it had a CV > 0.2 and it fell into a cluster that was generally sharper. A cluster was declared as sharp if the median CV of the cluster was bigger than the median CV of the whole peak set. All other peaks were classified as broad.

The proteins LARP7, RBM22, SLBP, U2AF1, IGF2BP2, NONO, TARDBP, HNRNPM, U2AF2, and PTBP1 had a higher number of sharper–shaped peaks, whereas the rest of the proteins had a higher number of broader–shaped peaks relative to the other shape (Figure 6a). Protein IGF2BP1 was almost half sharp and half broad peaks. UniProt lists the proteins CPSF6, CSTF2T, EWSR1, LARP7, RBM22, SAFB2, SLBP, SLTM, TAF15, TRA2A, U2AF1 with one RRM and the proteins HNRNPA1, IGF2BP1, IGF2BP2, NONO, SRSF1, TARDBP, HNRNPM, U2AF2, PTBP1 with at least two RRMs.

We could observe a trend between the number of RRMs and the number of sharper–shaped peaks (Figure 6a). From 11 proteins with 1 RRM just 4 had sharper peaks than broader-shaped peaks and 8 out of 9 proteins with at least 2 RRMs had sharper-shaped peaks. Furthermore, the proteins HNRNPM, U2AF2, and PTBP1 have all more than 2 RRMs (3, 3, and 4, respectively), thus it might be possible that an increasing number of RRMs results in sharper peaks. Figure 6b shows more clearly that proteins with more than one RRM tend to have sharper peaks (two-sided Wilcoxon test P-value $\approx 0.046$). It is possible that RNA–protein interactions become more specific with > 1 RRM and so the separation between more specific and unspecific binding sites is stricter.

Another observation was that shapes of splicing factors (RBM22, U2AF1, TARDBP, HNRNPM, U2AF2, and PTBP1) tend to be sharper than broader (Figure 6a). Yet, proteins, such as SRSF1, had broader peaks and are also involved in splicing. On the other hand, proteins not involved in splicing such as EWSR1, SAFB2, SLTM, and TAF15 clearly showed more broader–shaped peaks. The proteins SLBP, HNRNPA1, IGF2BP1, IGF2BP2, and NONO almost had an equal number of sharper or broader–shaped peaks, and all these proteins have multiple functions, contributing to at least two biological processes, such as transport and translation in the case of SLBP [6]. SRSF1 is also a multi–functional protein. Perhaps, that is the reason why it has broader peaks even though it is a splicing factor.

We also checked whether our result that RBPs involved in splicing have sharper peaks can have technical reasons. In this case, a splicing related protein could have more peaks that are split over two exons, which are detected by the peak caller as two separate but sharp peaks (split peak). So we investigated the number of peaks that fall into introns (90% overlap) for the proteins that are involved in the splicing process. We used bedtools set to a strict overlap (intersect –u -s -f 0.9) to in-
vestigate potential split peaks. The proteins PTBP1 (≈ 85%), RBM22 (≈ 62%), TARDBP (≈ 79%), and HNRNPM (≈ 87%) had more than 50% of peaks in introns, which reflected the assumption of split peaks. However, the proteins U2AF1 (≈ 17%), and U2AF2 (≈ 15%) had more peaks in exon regions, where the possibility of split peaks might still occur. It is important to note that this does not mean that the aforementioned proteins bind generally more introns or exons. As there exist no tool that can correct for split peaks, a further analysis for these proteins was required. We checked for potential split peaks by extending the peaks that fall completely into exons by 5 nucleotides to each side. Next, we intersected those extended peaks with introns to see if they are close to the exon boundaries. We found for U2AF1 only 27 peaks (0.72%) and for U2AF2 5 peaks (0.40%) that are potential split peaks, again deflecting the assumption of a technical artifact in the peak set of splicing factors. Thus, the sharpness of the peaks of U2AF1 and U2AF2 was potentially not the result of the peak-calling.

Potential Implications

StoatyDive is a powerful tool that can evaluate and classify peak profiles. It can be used in any sequencing data analysis that involves the prediction of binding sites such as CLIP-Seq or ChIP-Seq. Within this work, we provided an example for SLBP to show the usability of StoatyDive. First, it is possible to assess the quality of an experiment such as CLIP. The CV is just one quality factor and we recommend to test other features as well, such as the read coverage correlation. Second, StoatyDive assists to evaluate the binding specificity of the protein. The normalized CV distribution produced by StoatyDive provides valuable information for the user. A protein that binds very specific will have a distribution concentrated around a normalized CV of one. A protein with a lot of unspecific bindings will have a normalized CV distribution ≈ 0. Third, StoatyDive helps to filter for specific and unspecific binding sites to investigate if the protein has multiple protein domains that have different binding mechanisms. A finer distinction can be made with the classification mode of StoatyDive. This helps to identify peak profiles with a specific shape and filter them based on the corresponding biological question and function of the protein. Fourth, the results of StoatyDive can be used to validate a peak caller (e.g., PureCLIP), that is to say, one can assess how many false positives are in the peak sets based on the shape. Different peak caller might result in disparate peak sets and consequently different peak profile shapes.

We could show with StoatyDive that proteins with a higher number of RRMs tend to have sharper binding profiles. However, we had not taken any other RNA binding domains into account apart from RRMs and we would need to investigate more proteins to be certain about this trend. But, we could demonstrate that splicing factors tend to have more sharper peaks in comparison to proteins with other functions.

StoatyDive is a very powerful, well documented, and easy to apply tool that refines the binding site detection in the data analysis such as CLIP-Seq. Nevertheless, StoatyDive is a very general tool. In the future it is worth to investigate, if StoatyDive can be used with different types of peak-calling outputs and data types of sequencing data (e.g., ChIP-Seq, ATAC-Seq, Ribo-Seq, and others). It serves as a quality control and filtering step to select specific binding profiles, which therefore allows to improve other binding site prediction tools such as DeepBind [18], or any other subsequent analysis tasks, to increase the accuracy for the prediction.

Methods

Peak Correction, Extension and Coverage Calculation

StoatyDive was implemented in python (>= 3.6) and R (>= 3.4.4). The tool needs three files: the predicted binding regions of a peak-calling algorithm in bed6 format, a bam or bed file that was used for the peak-calling (experiment or control), and a tabular file of the chromosome size of the reference genome (Figure 7).

First, StoatyDive checks if a peak profile needs to be centered (peak correction). In the default mode, the profiles are centered by a convolution with a standard normal distribution. The maximum value of the convolution gives the nucleotide shift of the peak profile to center the peaks. So the window with the peak length is shifted to the center of the peak (Figure 7 step 1). With this approach we retain the context and take care of two problems. First, peak callers often produce peaks that are not correctly centered. Second, dimensionality reduction methods, such as uniform manifold approximation and projection for dimension reduction (uMAP; McInnes et al. [19]), are not translation invariant. Thus, two profiles with the same shape but in a different relative genomic position might end up in different locations in the new dimensional space.

After the peak correction, StoatyDive extends the peaks by default to the maximal peak length of the given peak set (Figure 7 step 2). This removes the peak length as a potential feature for the evaluation and classification. StoatyDive then calculates the read coverage (Figure 7 step 3) for each position inside a peak with the help of bedtools [9].

Evaluation of Peak Profiles

With the results of bedtools, StoatyDive evaluates every peak \( i \) from the total set of \( k \) peaks. StoatyDive will estimate the read count for every peak as a negative binomial \( X_i ∼ NB(r_i, p_i) \) with the hyperparameters \( r_i \) (number of hits) and \( p_i \) (probability of a hit). It then calculates the coefficient of variation (CV) for every peak. A simple estimation of the variance is not enough because the profile depends on the read coverage. Thus, to be able to compare each peak profile we have to normalize for the expected number of reads to adjust the variance. So the CV for each peak,

\[
CV_i = \sqrt{\frac{1 - p_i}{r_i}},
\]

is calculated with the estimated hyperparameters. In the last step, StoatyDive normalizes the CV score by the max and min of all scores,

\[
CV_i' = \frac{CV_i - \min(\nu CV)}{\max(\nu CV) - \min(\nu CV)}.
\]

At the end, our defined CV score will range from \( CV_i' \in [0, \infty] \) and the normalized score from \( CV_i' \in [0, 1] \), with a \( CV_i' = 0 \) for a more unspecific binding and \( CV_i' = 1 \) for a more specific one.

Classification of Peak Profiles

StoatyDive classifies the peak profiles in an unsupervised manner using uMAP [19] and k-means clustering [20]. Yet before clustering, StoatyDive processes the peak profiles. First, the profiles are normalized based on the individual maximum and minimum read count, since we are only interested in the shape of the profiles and not in the absolute read counts (Figure 7 step 4). So assuming each peak \( X_i \) has \( x_1, x_2, x_3, \ldots, x_n \) nucleotides, we
The normalized CV score of peaks that are covering a small appendage of a specific binder has a CV distribution \( \approx 0 \). A more specific binder has a CV distribution \( > 0 \). The CV distribution can also be used as a quality control to compare control and signal experiments. A quality breach might have occurred if the CV distributions of the control and signal experiments almost look identical. A control experiment should normally have a CV distribution that significantly differs in comparison to the control (Wilcoxon P-value \(< 0.05\), see Supplementary Figure 1).

The normalized CV distribution helps to evaluate the peaks based on the individual experiments. An empirical threshold is set at a CV of 0.2 (Equation 1), below which binding sites are deemed unspecific. The user can change the threshold. Keep in mind, the threshold for the normalized CV is relative in accordance to the individual experiment.

For the peak classification, StoatyDive generates a plot of the \( k \)-means optimization and a plot of the dimensional reduction with uMAP, which can be used to readjust the number of \( k \) clusters if this is necessary. The user also receives a set of example peak profiles and smoothed peak profiles of each cluster, which can be used to investigate the identified shapes. For a general trend, StoatyDive delivers average profiles for each cluster.

The final output of StoatyDive is a CV sorted table of the whole peak set, from the highest to the lowest CV. Each peak is labeled with 0, for more specific binding sites, and 1, for more unspecific sites. The table also lists for each peak the cluster number (group number) of the peak profile shape.

### Important Options of StoatyDive

The peak correction (Figure 7 step 1) can be turned off. The user can also change the translocation scheme of the peak profiles to shift them based on the maximal value (summit). The maximum translocation scheme is useful for nucleotide specific events such as truncation events in the case of iCLIP data [23]. StoatyDive has also the option for a different CV score that penalizes peaks within broad plateaus. StoatyDive then adjusts the CV score of peaks that are covering a small appendage of a read stack. Furthermore, the user can provide a maximal score to StoatyDive to normalize the CV distribution (Equation 2). This option helps to compare the CV distribution between experiments in accordance to their disparate peak sizes and total amount of reads. StoatyDive also has a threshold for the normalized CV score to divide the peaks into more specific and more unspecific binding sites, which the user can change.

StoatyDive has two major parameters for the peak profile classification (Figure 7 step 6). First, the user can adjust the maximal amount of potential peak clusters identified by the k-
means clustering. Yet, the final number of peak clusters will be optimized by StoatyDive. The parameter is an upper bound. However, the user has the option to force StoatyDive to use k specific clusters. The smoothing (Figure 7 step 4) of the peak profiles can also be adjusted by the user. The default was optimized with different test sets. Increasing the parameter (> default) might underfit the smoothing and thus lead to fewer peak clusters. A lower value (< default) might overfit and so lead to more clusters. The smoothing can also be turned off, but it is recommended to turn it on.

Availability of Supporting Source Code and Requirements

Project name: StoatyDive
Project homepage: https://github.com/BackofenLab/StoatyDive
Conda: https://anaconda.org/bioconda/stoatydive
Operating system(s): Unix
biotools:StoatyDive
RRID:SCR_018796

Availability of Supporting Data and Materials

StoatyDive provides a small dataset for a test run, which can be found in the github repository. The whole eCLIP data used in this paper, such as SLBP or RBPFOX2, is listed in the supplementary of the study by Van Nostrand et al. [5].

Additional files

Supplementary Figure 1. CV distributions of all other proteins analyzed for figure 6 with two-sided Wilcoxon test P-value, the number of uniquely mapped reads and the mean CV for each replicate. The two replicates quite often have different CV distributions. Furthermore, we report a plot for the mean CV for the CLIP data in comparison to the size-matched input control of each protein. The control data tends to have a CV close to 0. The CV distributions between CLIP and control data always have a two-sided Wilcoxon test P-value < 0.05.

Supplementary Figure 2. We applied StoatyDive to the size matched input control of the SLBP data [5]. StoatyDive has found 4 different peak profile shapes, broad (cluster 1), plateau (cluster 2), sharp (cluster 3), and constant (cluster 4). The supplements also include the average profiles for replicate 1 and 2 to show the overall trend of the clusters.

Supplementary Table 3. Mean CV, variance of the CV, mean log, fold change (LFC) enrichment between the control and CLIP experiment, median P-value, true positive rate (TPR), true negative rate (TNR), accuracy (ACC), and Matthews correlation coefficient (MCC) for the analyzed peaks (All Peaks) of replicate 2 (ENCFF127WAK) from the study by Van Nostrand et al. [5]. Features are listed for the peak shape cluster with the highest number of peaks in histones (Main Cluster) and second highest number (Second Cluster), and for the peaks with a CV smaller or bigger a threshold of 0.2, 0.5, and 0.8, using different peak lengths (30, 40, 70, and maximum peak length of 201 nucleotides). We achieved the best TPR, ACC and MCC with a peak length of 70 and with a CV cutoff of 0.2.

Supplementary Figure 4. Peak lengths of the peak set (ENCFF127WAK) for the second replicate of SLBP and peak lengths of all other proteins of the eCLIP data from the study by Van Nostrand et al. [5].

Supplementary Figure 5. All scatter plots from SIC-ChIP [17] for the artificial SLBP data.

Supplementary Figure 6. We tested different dimensional reduction methods such as PCA, SOM, and t-SNE on the CLIP data of SLBP. The PCA has no clear clusters for replicate 2, which is similar for t-SNE on replicate 1 and 2. Using an optimized SOM delivers a feature layer with a very high activated hidden unit for replicate 2. It is hard to see any distinct clusters from the counts (activation) of each hidden unit. uMAP can clearly separate the data into more defined clusters. Furthermore, it is much easier to interpret the results of uMAP, whereas an artificial neural network, such as a SOM, generates a feature layer (hidden layer) that is hard to explain.

Declarations

List of abbreviations

ACC: Accuracy; CLIP-Seq: Crosslinking immunoprecipitation in combination with high-throughput sequencing; CV: Coefficient of variation; LFC: log2 Fold Change; MCC: Matthews correlation coefficient; PCA: Principal component analysis; RBP: RNA-binding proteins; RRM: RNA recognition motifs; SOM: Self-organizing map; TPR: True positive rate; TNR: True negative rate; t-SNE: t-Distributed Stochastic Neighbor Embedding.

Ethical Approval

Not applicable

Consent for Publication

Not applicable

Competing Interests

The authors declare that they have no competing interests.

Funding

This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) grant 322977937/GRK2344 2017 MeInBio – BioInMe Research Training Group, and Germany’s Excellence Strategy (CIBSS – EXC-2189 – Project ID 390939984).

Author’s Contributions

F.H. performed the computational analysis and tool development. R.B. initialized the project, and supervised the research. F.H. and R.B. wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to Gokcen Eraslan for his support. We thank Milad Miladi, Florian Eggenhofer, and Wolfgang Maier for their
comments for the paper.

References

1. Lee FC, Ule J. Advances in CLIP technologies for studies of protein–RNA interactions. Molecular cell 2018;69(3):354–369.

2. Jankowsky E, Harris ME. Specificity and nonspecificity in RNA–protein interactions. Nature reviews Molecular cell biology 2015;16(9):533–544.

3. Müller-McNicoll M, Neugebauer KM. How cells get the message: dynamic assembly and function of mRNA–protein complexes. Nature Reviews Genetics 2013;14(4):275.

4. Corcoran DL, Georgiev S, Mukherjee N, Gottwein E, Skalsky RL, Keene JD, et al. PARalyzer: definition of RNA binding sites from PAR–CLIP short-read sequence data. Genome biology 2011;12(8):R79.

5. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, et al. Robust transcriptome-wide discovery of RNA–binding protein binding sites with enhanced CLIP (eCLIP). Nature methods 2016;13(6):508.

6. Sullivan KD, Mullen TE, Marzluff WF, Wagner EJ. Knockdown of SLBP results in nuclear retention of histone mRNA. Rna 2009;15(3):459–472.

7. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29(1):15–21.

8. Krakau S, Richard H, Marsico A. PureCLIP: capturing target–specific protein–RNA interaction footprints from single–nucleotide CLIP-seq data. Genome biology 2017;18(1):240.

9. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841–842.

10. Chen X, Chung D, Stefani G, Slack FJ, Zhao H. Statistical issues in binding site identification through CLIP–seq. Statistics and Its Interface 2015;8(4):419–436.

11. Erkmann JA, Wagner EJ, Dong J, Zhang Y, Kutay U, Marzluff WF. Nuclear import of the stem–loop binding protein and localization during the cell cycle. Molecular biology of the cell 2005;16(6):2960–2971.

12. Huppertz I, Attig J, D’Ambrogio A, Easton LE, Sibley CR, Sugimoto Y, et al. iCLIP: Protein–RNA interactions at nucleotide resolution. Methods 2014;65(3):274–287.

13. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA–and RNA-binding proteins by deep learning. Nature biotechnology 2015;33(8):831.

14. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426 2018.

15. Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society Series C (Applied Statistics) 1979;28(1):100–108.

16. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic acids research 2009;37(suppl_2):W202–W208.

17. Green PJ, Silverman BW. Nonparametric regression and generalized linear models: a roughness penalty approach. Chapman and Hall/CRC; 1993.

18. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Selected papers of hirotugu akaike Springer; 1998.p. 199–213.

19. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic acids research 2019;47(Suppl 2):W202–W209.
Click here to access/download
Supplementary Material
Supplements_1.pdf
Click here to access/download
Supplementary Material
Supplements_3.pdf
Click here to access/download
Supplementary Material
Supplements_4.pdf
Click here to access/download
**Supplementary Material**
Supplements_5.pdf
Click here to access/download
**Supplementary Material**
Supplements_6.pdf
Click here to access/download
Supplementary Material
main.tex