

Chapter 17

mCarts: Genome-Wide Prediction of Clustered Sequence Motifs as Binding Sites for RNA-Binding Proteins

Sebastien M. Weyn-Vanhentenryck and Chaolin Zhang

Abstract

RNA-binding proteins (RBPs) are critical components of post-transcriptional gene expression regulation. However, their binding sites have until recently been difficult to determine due to the apparent low specificity of RBPs for their target transcripts and the lack of high-throughput assays for analyzing binding sites genome wide. Here we present a bioinformatics method for predicting RBP binding motif sites on a genome-wide scale that leverages motif conservation, RNA secondary structure, and the tendency of RBP binding sites to cluster together. A probabilistic model is learned from bona fide binding sites determined by CLIP and applied genome wide to generate high specificity binding site predictions.

Key words RNA-binding protein (RBP), CLIP tag clusters, motif, Binding site prediction, mCarts

1 Introduction

RNA-binding proteins (RBPs) bind to short, degenerate sequences in RNA to regulate a number of post-transcriptional processes including alternative splicing, alternative polyadenylation, and RNA stability [1–3]. These short sequences have traditionally made it difficult to computationally predict RBP binding sites on a genome-wide scale with high specificity. As a result, previous approaches using consensus sequence [4] or position weight matrices (PWMs) [5] to search for functional binding sites have low discriminative power.

Recently, the development of HITS-CLIP (crosslinking and immunoprecipitation combined with high-throughput sequencing) and its variants has made it possible to map direct, in vivo RBP binding sites genome wide in a given condition [6–8]. In brief, UV light is used to crosslink protein-RNA complexes in direct contact in cells or tissue. The protein of interest is isolated along with its bound RNA fragments, which are then purified in very stringent conditions and sequenced in depth.

HITS-CLIP has resulted in a wealth of information about RBP function and provided new insights into RNA regulation. However, a HITS-CLIP experiment only provides a snapshot reflecting the experimental conditions rather than the complete landscape of protein-RNA interactions. Some transcripts may not be expressed under the conditions in which the HITS-CLIP experiment was performed, and thus will not be detected as bound by the RBP [9, 10]. In addition, some binding sites might escape the detection by HITS-CLIP due to technical issues that limit the complexity and depth of CLIP libraries [9]. Finally, while CLIP data provide evidence of protein-RNA interactions, they do not directly provide a mechanism for the specific recognition, which is coded in the RNA sequences and structures.

Several algorithms have been developed to take advantage of the HITS-CLIP data to generate genome-wide RBP binding profiles [11–13]. We previously developed mCarts which predicts clusters of RBP motif sites by integrating several intrinsic and extrinsic features of functional protein-RNA interactions, including the number and clustering of individual motif sites, their accessibility as determined by RNA secondary structures, and cross-species conservation [14]. A hidden Markov model (HMM) framework learns quantitative and subtle rules of these features from *in vivo* RBP binding sites derived from HITS-CLIP data and generates genome-wide predictions of new sites with high specificity and sensitivity. These predictions extend the RBP interaction map observed from HITS-CLIP data, providing a broader picture of RBP binding and regulation. This protocol describes the process of installing the mCarts software, training the mCarts model, and predicting RBP binding sites in the mouse genome for the RBP Nova, a neuron-specific RBP that binds to YCAY ($Y = C/U$) motifs to regulate alternative splicing [6].

Although downstream analysis after obtaining predicted binding sites is beyond the scope of this protocol and varies depending on specific RBPs, typical steps include cross-validation using CLIP data or other known list of binding sites and correlation of RBP binding with altered RNA splicing or gene expression upon RBP perturbation [14]. For example, we previously showed that a high validation rate was achieved when a subset of top candidate alternative exons with strong YCAY clusters as predicted by mCarts were tested for Nova-dependent splicing by RT-PCR. In addition, on a genome-wide scale, alternative exons with evidence of Nova binding from both CLIP data and mCarts predicted motif sites are more likely to have Nova-dependent splicing than those exons that are supported by CLIP or bioinformatics predictions alone. These analyses suggest that CLIP and mCarts data are complementary to each other and the combination of the two can help identify direct Nova targets more accurately.

2 Materials

2.1 Computer

1. This analysis requires a personal computer or cluster running a UNIX-based operating system (Linux or Mac OS X) with sufficient memory (8 GB RAM; 16 GB recommended) to run the software and enough storage for the data files (30 GB for mouse and 40 GB for human) and the CLIP data (up to a few GB, but varies).
2. The software package provides a set of command line tools implemented in perl and C++ and relies on several standard Unix tools such as awk and sort. Updates to the software, documentation, and protocol can be found at <http://zhanglab.c2b2.columbia.edu/index.php/MCarts>.
3. Commands that should be entered into the terminal (*see Note 1*) will be identifiable by a different typeface. The beginning of a command is indicated by a "\$" (which should not be entered on the command line). Commands often span multiple lines in the text, but they should be entered as a single line. For example:

```
$ perl ~/src/script.pl -option filename
```

2.2 The mCarts Software

1. Download the mCarts software and the required czplib perl libraries from the links provided at http://zhanglab.c2b2.columbia.edu/index.php/MCarts_Documentation#Download.
2. Install mCarts (v1.2.x or later) by running the following commands:

```
$ tar xzvf mCarts.v1.x.x.tgz
$ cd mCarts
$ make
```

The make command requires the GCC compiler as well as the Boost (<http://www.boost.org>) and popt libraries (<http://rpm5.org/files/popt/>), which are installable from your distribution's package manager.

3. Add mCarts to your path (optionally, add this command to your `.bash_profile`; *see Note 2*):

```
$ export PATH=~/.src/mCarts:$PATH
```

4. Add czplib to the perl libraries path by running the following command (this needs to be added to your `.bash_profile` if SGE/OGE is installed on your system; otherwise, optionally add this command to your `.bash_profile`; *see Note 2*):

```
$ tar xzvf czplib.v1.x.x.tgz
$ export PERL5LIB=~/.src/plib
```

5. Install the perl modules `Math::CDF` and `Bio::SeqIO` (e.g., using CPAN).

2.3 *The CIMS Software (for CLIP Data Processing)*

1. Download the CIMS package from the link provided at http://zhanglab.c2b2.columbia.edu/index.php/CIMS_Documentation#Download (suggested location is ~/src).
2. Decompress the CIMS package by running the following command:

```
$ tar xzvf CIMS.v1.x.x.tgz
```

2.4 *The Reference Library Files*

1. Download the reference library files for the organism corresponding to your CLIP data from the link provided at http://zhanglab.c2b2.columbia.edu/index.php/MCarts_Documentation#Download (suggested location is ~/data/):
mm10 library files: mCarts_lib_data_mm10.tgz.
hg19 library files: mCarts_lib_data_hg19.tgz.
For running the protocol with the sample data, download the mm10 database.

2. Decompress the library files:

```
$ tar xzvf mCarts_lib_data_mm10.tgz.
```

2.5 *The RepeatMasker Database*

A copy of the database as it stood at publication time is available at http://zhanglab.c2b2.columbia.edu/index.php/MCarts_Documentation#Download. Using this file as you follow the protocol will ensure that your results match ours, but we recommend using the latest version when performing your own analysis.

1. Go to the UCSC Genome Browser (<http://genome.ucsc.edu>).
2. Click “Tools > Table browser” at the top of the page.
3. Set “genome” to your organism of interest (for this protocol, select “mouse”).
4. Set “assembly” to the assembly matching your dataset (“mm10”).
5. Set “group” to “Variation and Repeats” (mouse) or “Repeats” (human).
6. Set “track” to “RepeatMasker.”
7. Set “output format” to “BED - browser extensible data.”
8. Set the name to something memorable, e.g., “mm10.rmsk.bed.”
9. Click “Get output.”
10. Click “get BED” to download the file (suggested location is ~/data/).

2.6 *The CLIP Data*

1. For following along with the protocol, download and decompress the sample Nova CLIP data, Nova_CLIP_uniq_mm10.bed, from the link provided at <http://zhanglab.c2b2.colum>

bia.edu/index.php/MCarts_Documentation#Download (*see Note 3* for details about this dataset):

```
$ gunzip Nova_CLIP_uniq_mm10.bed.gz
```

- Alternatively, provide a BED file for an RBP of your choosing. This file should be in BED format and should contain only unique CLIP tags that represent independent captures of protein-RNA interactions. If this is the case, the CLIP data must have been mapped and filtered properly with removal of PCR duplicate tags (*see Note 4*).

3 Methods

This protocol assumes that the mCarts software is located in the directory `~/src/`, that the mCarts library files are located in `~/data/`, and that the CLIP data file is in the current working directory. Adjust the paths accordingly. As you are progressing through the protocol, you can compare the number of lines in each file with those provided in Table 1.

Table 1
The number of lines expected in each file (obtained using `wc -l filename`)

Filename	Lines
Nova_CLIP_uniq_mm10.bed	4,401,394
Nova_CLIP_uniq_mm10.cluster.0.bed	1,766,040
Nova_CLIP_uniq_mm10.cluster.bed	249,393
Nova_CLIP_uniq_mm10.tag.exact.bedGraph	5,112,140
Nova_CLIP_uniq_mm10.cluster.PH.detail.txt	249,393
Nova_CLIP_uniq_mm10.cluster.PH.center.bed	249,393
Nova_CLIP_uniq_mm10.cluster.PH.center.ext50.bed	249,393
Nova_CLIP_uniq_mm10.cluster.PH.center.ext50.normsk.bed	142,786
mm10.exon.uniq.ext1k.bed	272,046
Nova_CLIP_uniq_mm10.cluster.PH.center.ext50.normsk.ext1k.bed	61,028
Nova_CLIP_uniq_mm10.cluster.PH15.center.ext50.normsk.ext1k.bed	7,700
↘CLIP.pos.bed	7,700
mm10.exon.uniq.ext1k.noCLIP.bed	112,798
↘CLIP.neg.bed	112,798

3.1 Generate the Positive Training File by Identifying Regions with Strong CLIP Tag Clusters

Since Nova is known to be an important splicing factor, we will limit the CLIP tag clusters to exons and flanking intronic sequences for training.

1. Identify CLIP tag clusters by grouping overlapping CLIP tags (this step is slightly different from our previous method to generate Nova clusters. *See Note 3* for discussion comparing this to previous Nova results):

```
$ perl ~/src/CIMS/tag2cluster.pl -v -s -maxgap "-1"
Nova_CLIP_uniq_mm10.bed Nova_CLIP_uniq_mm10.cluster.0.bed
```

2. Select the clusters containing > 2 tags:

```
$ awk '$5>2' Nova_CLIP_uniq_mm10.cluster.0.bed >
Nova_CLIP_uniq_mm10.cluster.bed
```

3. Create a bedGraph file, which is used to determine the CLIP tag coverage at each position in the genome:

```
$ perl ~/src/CIMS/tag2profile.pl -ss -exact -of bedgraph -n Nova -v
Nova_CLIP_uniq_mm10.bed Nova_CLIP_uniq_mm10.tag.exact.bedGraph
```

4. Determine the peak heights of the clusters:

```
$ perl ~/src/CIMS/extractPeak.pl -s --no-match-score 0 -of detail -v
Nova_CLIP_uniq_mm10.cluster.bed Nova_CLIP_uniq_mm10.tag.exact.bedGraph
Nova_CLIP_uniq_mm10.cluster.PH.detail.txt
```

5. Determine the center position of the clusters:

```
$ awk '{print
$1"\t"int(($8+$9)/2)}'\t"int(($8+$9)/2)+1"\t"$4"\t"$7"\t"$6}'
Nova_CLIP_uniq_mm10.cluster.PH.detail.txt >
Nova_CLIP_uniq_mm10.cluster.PH.center.bed
```

6. Extend the cluster centers 50 nt in each direction:

```
$ awk '{print $1"\t"$2-50"\t"$3+49"\t"$4"\t"$5"\t"$6}'
Nova_CLIP_uniq_mm10.cluster.PH.center.bed
> Nova_CLIP_uniq_mm10.cluster.PH.center.ext50.bed
```

7. Remove clusters that overlap with repetitive regions:

```
$ perl ~/src/CIMS/tagoverlap.pl -big -region mm10.rmssk.bed -r -v
Nova_CLIP_uniq_mm10.cluster.PH.center.ext50.bed
Nova_CLIP_uniq_mm10.cluster.PH.center.ext50.normssk.bed
```

8. Extend the known exons by 1000 nt in each direction:

```
$ perl ~/src/CIMS/bedExt.pl -l -1000 -r 1000 -chrLen
~/data/mCarts_lib_data_mm10/chrLen.txt -v
~/data/mCarts_lib_data_mm10/mm10.exon.uniq.bed mm10.exon.uniq.ext1k.bed
```

9. Determine which exonic regions (exons±1000 nt) contain CLIP clusters:

```
$ perl ~/src/CIMS/tagoverlap.pl -region mm10.exon.uniq.ext1k.bed -ss --
keep-score --keep-tag-name --complete-overlap --non-redundant -v
Nova_CLIP_uniq_mm10.cluster.PH.center.ext50.normssk.bed
Nova_CLIP_uniq_mm10.cluster.PH.center.ext50.normssk.ext1k.bed
```

10. Select the top clusters based on peak height (PH):

```
$ awk '$5>=15'
Nova_CLIP_uniq_mm10.cluster.PH.center.ext50.normsk.ext1k.bed >
Nova_CLIP_uniq_mm10.cluster.PH15.center.ext50.normsk.ext1k.bed
```

This results in 7700 regions spanning 770,000 nucleotides. *See Note 5* for information about picking cluster threshold.

11. Create a symbolic link to the positive region file, which makes future commands clearer and easily reusable with a different training file:

```
$ ln -s Nova_CLIP_uniq_mm10.cluster.PH15.center.ext50.normsk.ext1k.bed
CLIP.pos.bed
```

3.2 Generate the Negative Training File by Filtering Out Any Regions with CLIP Tags

1. Select exonic regions (exons \pm 1000 nt) that contain no CLIP tags (n.b. tags, not clusters):

```
$ perl ~/src/CIMS/tagoverlap.pl -big -region Nova_
CLIP_uniq_mm10.bed -
ss --keep-score -r -v mm10.exon.uniq.ext1k.bed
mm10.exon.uniq.ext1k.noCLIP.bed
```

This results in 112,798 regions spanning 252,523,292 nucleotides.

2. Create a symbolic link to the negative region file, which makes future commands clearer and easily reusable with a different training file

```
$ ln -s mm10.exon.uniq.ext1k.noCLIP.bed CLIP.neg.bed
```

3.3 Train the mCarts Model

1. Run the mCarts training:

```
$ mCarts -ref mm10 -f CLIP.pos.bed -b CLIP.neg.bed
-lib
~/data/mCarts_lib_data_mm10 -w YCAY --min-site 3
--max-dist 30 --train-
only -v Nova_HMM_D30_m3
```

The whole genome is divided into a number of smaller splits for parallelization. Individual jobs are submitted to the queuing system when it is detected (Oracle Grid Engine (OGE), formerly known as Sun Grid Engine or SGE, is currently supported); jobs are run locally otherwise, in which case 24–36 h of runtime should be expected. Additional details on mCarts are worth noting (*see Note 6*).

If the program finished without errors, the following files should be created in the *Nova_HMM_D30_m3* directory:

- BLS (directory)
- formatted (directory)
- model.txt

- params.txt
 - train_neg.txt
 - train_pos.txt
2. To visualize the model, open the *models.txt* file (located in the *Nova_HMM_D30_m3* output directory) in Microsoft Excel. For each of the following categories, create a line graph comparing the positive to the negative regions for distance (distance between neighboring motif sites). There is a long tail for the distance parameters, so visualizing the score for all 1000 nt is not necessary (try ~100 nt). For conservation_0 (intron), conservation_1 (CDS), conservation_2 (5' UTR), conservation_3 (3' UTR), and accessibility, create a scatterplot comparing the positive and negative regions, using the “#” row for the *x*-axis. The “#” row indicates Branch Length Score (BLS) for conservation and degree of single strandedness for accessibility. The results for Nova are shown in Fig. 1.

3.4 Run the Model on the Whole Genome

1. Run the Nova model on the mm10 genome:

```
$ mCarts -v --exist-model ./Nova_HMM_D30_m3
```

If the program finished without errors, the following additional files should be created in the *Nova_HMM_D30_m3* directory:

- cluster.bed
- out (directory)
- qsub (directory; only if SGE is available)
- scripts (directory; only if SGE is available)
- scripts.list (only if SGE is available)

2. Convert the motif cluster BED file into a bedGraph file:

```
$ perl ~/src/CIMS/tag2profile.pl -ss -exact -weight -of bedgraph -n
"Nova_motif" -v ./Nova_HMM_D30_m3/cluster.bed
./Nova_HMM_D30_m3/cluster.bedGraph
```

3.5 Visualizing and Interpreting the Results

1. From the plots generated by the model training, we observe the following:
 - YCAY motifs are clustered more closely in positive training regions.
 - Positive regions are more accessible (more single stranded).
 - Positive regions have higher conservation in the 5' UTR, CDS, intron, and 3' UTR.
2. The cluster.bedGraph file generated by mCarts can be loaded into a genome browser such as the UCSC Genome Browser.

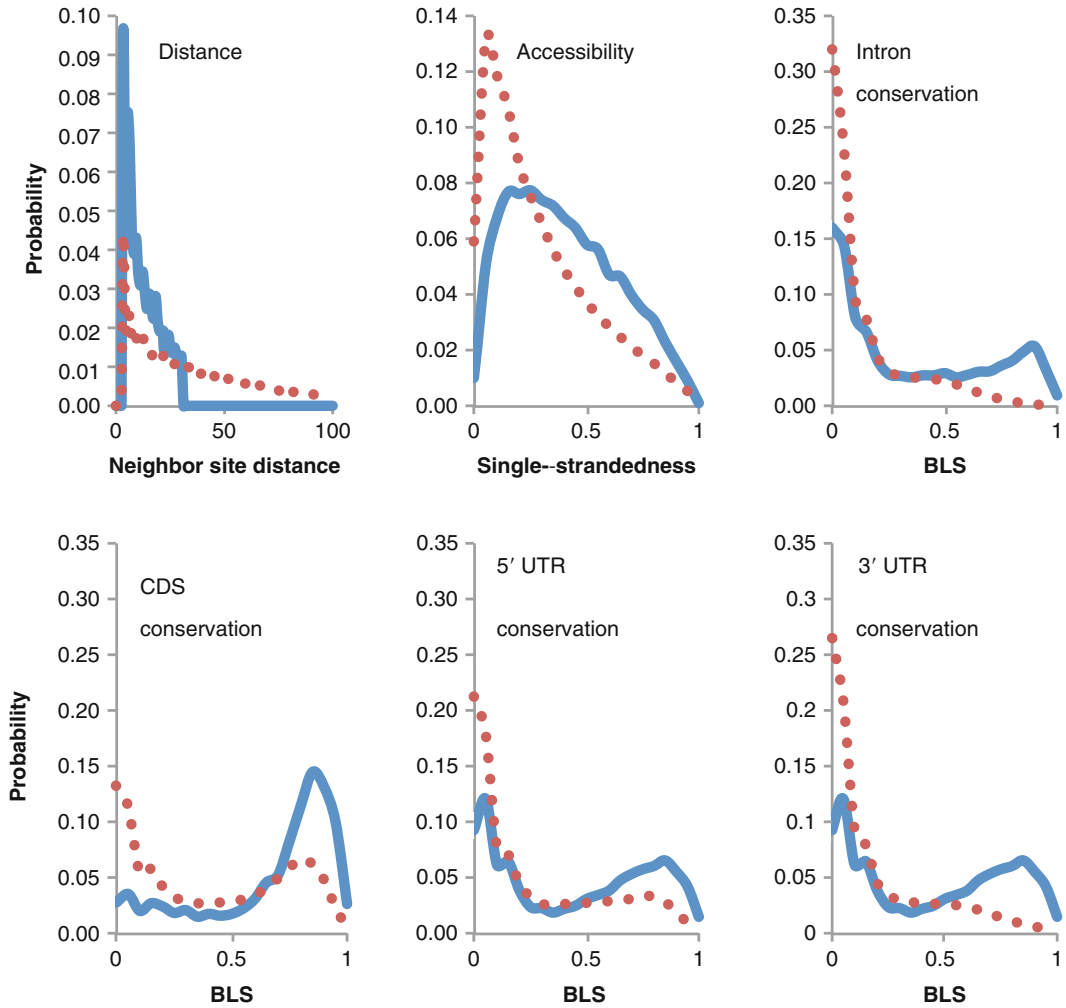


Fig. 1 Features of positive (*solid line*) and negative (*dashed line*) Nova YCAY clusters as determined by the mCarts model

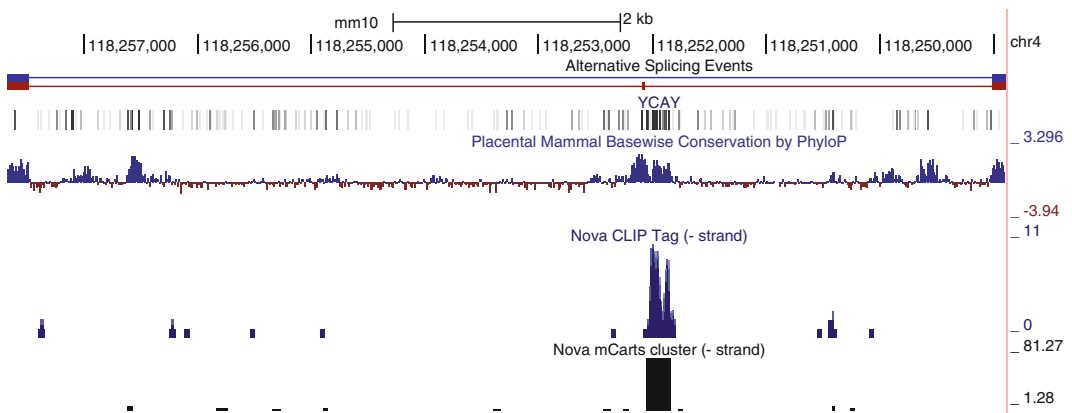


Fig. 2 Exon 6 of Ptpfr contains a cluster of highly conserved YCAYs. The motif cluster predicted by mCarts matches these and the binding profile determined by HITS-CLIP

This allows for the visualization of RBP binding clusters and their associated scores (Fig. 2). Figure 2 shows exon 6 of *Ptprf*, which contains 22 highly conserved YCAY elements and whose inclusion has been previously shown to be activated by Nova [15].

4 Notes

1. This protocol assumes familiarity with the UNIX command line. There are many great introductory resources available (e.g., ref. [16, 17]), but instruction in its use is beyond the scope of this protocol.
2. Unix-based operating systems contain a special file, `~/.bash_profile`, which is automatically executed upon starting the bash shell. To avoid having to add software to your path manually each time you open a new terminal window, you can add the commands directly to `~/.bash_profile`. Simply edit the file and add the commands of interest, then reload the profile manually using:

```
$ . ~/.bash_profile
```

Note the "." at the beginning.
3. The sample CLIP data we provide for this protocol is from ref. [18]. It consists of 4,401,528 unique tags originally mapped to mm9. We used the LiftOver utility (*see Note 4*) to translate the coordinates to mm10, resulting in 4,401,394 unique tags. Another important detail to note is that the results presented in this protocol will differ slightly from those presented in previous work [14, 18] due to the use of a different clustering algorithm. The method described here is more straightforward and has been successfully used in subsequent work [19].
4. Regarding data pre-processing, stringent mapping and filtering of CLIP data are critical for defining robust RBP binding sites. Detailed discussion of CLIP data processing is beyond the scope of this protocol, but readers are referred to the CIMS software package we developed [8]. It is often the case that the raw CLIP data for your RBP of interest was aligned to an earlier version of the reference genome. For example, the Nova data in this protocol was previously mapped to mm9. To convert the mm9 coordinates to mm10 coordinates, we use the LiftOver utility developed by the UCSC Genome Browser group (<https://genome-store.ucsc.edu>) [20]. The required chain files can be downloaded from UCSC as well (<http://hgdownload.cse.ucsc.edu/downloads.html>). For converting mm9 to mm10, download and unzip <http://hgdownload.cse.ucsc.edu/goldenPath/mm9/liftOver/mm9ToMm10.over.chain.gz>, then execute the following command:

```
$ liftOver Nova_CLIP_unique_tag_mm9.bed mm9ToMm10.over.chain
Nova_CLIP_uniq_mm10.bed Nova_CLIP_unique_tag_mm9mm10.unmapped
```

In some cases, such as this one, the BED file contains track lines that LiftOver can't handle (you will get an error). To get rid of these lines:

```
$ grep -v "track" file.bed > file.noheader.bed
```

5. To focus the model training on the most robust clusters, we pick the set of clusters with the greatest peak height (PH). The cutoff value depends on specific datasets, but in our experience based on cross-validation analyses, the exact value does not greatly affect the outcome. We generally pick a threshold where at least 5000–6000 confident clusters are obtained to reduce the variation in parameter estimation. The following command provides a summary of the peak heights, listing (1) the PH, (2) the number of clusters with that PH, and (3) the cumulative number of clusters at that peak height:

```
$ cut -f5 Nova_CLIP_uniq_mm10.cluster.PH.center.ext50.normsk.ext1k.bed
| sort -nr | uniq -c | awk 'BEGIN{cumul=0} {print $2"\t"$1"\t"$1+cumul;
cumul=$1+cumul}'
```

For this dataset, we choose to set the cutoff at 15, which corresponds to 7700 clusters.

6. In this mCarts protocol, we run the analysis with the following parameters:

- ref mm10: the reference genome being used.
- f CLIP.pos.bed: the foreground (positive) training set.
- b CLIP.neg.bed: the background (negative) training set.
- lib ~/data/mCarts_lib_data_mm10: the location of the mCarts library files for mouse.
- w YCAY: the motif we are searching for (IUPAC code is allowed).

mCarts currently does not accept “U” in the motif so be sure to provide a “T” instead (e.g., “TGCATG” instead of “UGCAUG”).

min-site 3: the minimum number of sites in a cluster.

max-dist 30: the maximum distance between neighboring sites in a cluster.

train-only: only train for now; we will test in the next step.

- v: verbose; print out what the software is doing.

The full mCarts documentation is available at http://zhan-glab.c2b2.columbia.edu/index.php/MCarts_Documentation and a full description of the methodology in ref. [14].

As of this writing, the direct software links are as follows:

mCarts: <http://sourceforge.net/p/mcarts/>.

czplib: <http://sourceforge.net/p/czplib/>.

CIMS: <http://sourceforge.net/p/ngs-cims/>.

Acknowledgements

The authors would like to thank Lauren E. Fairchild and Huijuan Feng for their assistance in testing the protocol and for providing feedback on the manuscript. This work was supported by grants from the National Institutes of Health (NIH) (R00GM95713) and the Simons Foundation Autism Research Initiative (297990 and 307711) to C.Z.

References

- Licalosi DD, Darnell RB (2010) RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet* 11:75–87
- Ray D, Kazan H, Cook KB et al (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499:172–177
- Cook KB, Kazan H, Zuberi K et al (2011) RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* 39:D301–D308
- Chasin LA (2007) Searching for splicing motifs. *Adv Exp Med Biol* 623:85–106
- Galarneau A, Richard S (2005) Target RNA motif and target mRNAs of the Quaking STAR protein. *Nat Struct Mol Biol* 12:691–698
- Licalosi DD, Mele A, Fak JJ et al (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456:464–469
- König J, Zarnack K, Rot G et al (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* 17:909–915
- Moore MJ, Zhang C, Gantman EC et al (2014) Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat Protoc* 9:263–293
- Darnell RB (2010) HITS-CLIP: panoramic views of protein–RNA regulation in living cells. *Wiley Interdiscipl Rev RNA* 1:266–286
- Blencowe BJ, Ahmad S, Lee LJ (2009) Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes Dev* 23:1379–1386
- Maticzka D, Lange SJ, Costa F et al (2014) GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol* 15:R17
- Cereda M, Pozzoli U, Rot G et al (2014) RNAmotifs: prediction of multivalent RNA motifs that control alternative splicing. *Genome Biol* 15:R20
- Han A, Stoilov P, Linares AJ et al (2014) De novo prediction of PTBP1 binding and splicing targets reveals unexpected features of its RNA recognition and function. *PLoS Comput Biol* 10:e1003442
- Zhang C, Lee K-Y, Swanson MS et al (2013) Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic Acids Res* 41:6793–6807
- Jelen N, Ule J, Živin M et al (2007) Evolution of nova-dependent splicing regulation in the brain. *PLoS Genet* 3:e173–e1847
- Stein LD (2002) *Unix survival guide*. John Wiley, Hoboken, NJ
- Buffalo V (2015) *Bioinformatics data skills*. O'Reilly Media, Sebastopol, CA
- Zhang C, Frias MA, Mele A et al (2010) Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science* 329:439–443
- Weyn-Vanhentenryck SM, Mele A, Yan Q et al (2014) HITS-CLIP and integrative modeling define the rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep* 6:1139–1152
- Hinrichs AS, Karolchik D, Baertsch R et al (2006) The UCSC genome browser database: update 2006. *Nucleic Acids Res* 34:D590–D598