RNA-Protein Interaction (RPI) Identifier Final Report

Neel Mehtani, Sriram Kidambi, Ellie Lai, Suzi Kim

Introduction

RNA-binding proteins (RBPs) have important functions in the regulation of gene expression. They play a role in the post-transcriptional process in eukaryotes, such as splicing, mRNA transport, mRNA translation, and mRNA decay. RBPs can bind to double or single-stranded RNA and contain various structural motifs, allowing for unique RNA-binding activity and protein-protein interaction. Since there are over 2,000 RBPs that interact with transcripts in various ways, understanding more about their binding interactions is crucial for understanding more about the roles they play in different cellular processes.

Recent technologies have worked to characterize these RNA-binding protein interactions by approaching it in 2 ways:

1) characterizing proteins bound to an RNA of interest (RNA-centric)
2) characterizing RNAs bound to a protein of interest (protein-centric)

We are interested in looking at RBPs in a protein-centric approach, which is done experimentally via CLIP-seq methods, such as HITS-CLIP, PAR-CLIP, and eCLIP. CLIP methods like PAR-CLIP and HITS-CLIP are technically challenging, with high experimental failures. As such, eCLIP-seq has become a recent, enhanced protocol of CLIP-seq that improves specificity in the discovery of authentic binding sites at the nucleotide-level and works to address those challenges.

With the development of eCLIP-seq, data on various RBPs has been generated and organized as part of the ENCODE project, a public research consortium aimed at identifying all functional elements in human and mouse genomes. However, pipelines need to be developed to be able to analyze eCLIP data effectively. Currently, there is no universal standard for eCLIP-seq analysis and quantification is dependent on study aims and features of the eCLIP data itself. As such, navigating eCLIP analysis is challenging and confusing, particularly for those who are outside of the field.

With this in mind, we have created an eCLIP-analysis pipeline that works to integrate various software tools in a streamlined, user-friendly way, with the goal of investigating RNA-binding protein interactions. Our pipeline works with paired-end eCLIP datasets, retrievable from the ENCODE website. These datasets are then processed through our pipeline to receive functional information about a RNA-binding protein of interest. This entire workflow is executable through the command line on a high performance computing cluster, and all the necessary scripts are downloadable from GitHub. We recommend using a computing server, such as BRIDGES-2, to be able to achieve all computational requirements easily. Here we introduce an efficient pipeline that works to find potential conserved motifs of any RNA-binding protein.

Methods and Materials

Materials:

eCLIP against RBFOX2 in the K562 cell line:
https://www.encodeproject.org/experiments/ENCSR756CKJ/

eCLIP control against RBFOX2 in the k56s cell line:
https://www.encodeproject.org/experiments/ENCSR051IXX/

Implementation:

First, the eCLIP paired-end data was downloaded from the ENCODE website and stored in a separate file for ease of access. Both replicates were used for the experimental data, and replicates from the control experiment were also used. Pre-processing steps (cutadapt, FastQC) were done to ensure the reads were ready for genome alignment, which was performed using STAR. After genome alignment, the unique reads were sorted and merged to ensure there were no PCR duplicates. Then, peak calling was performed on the set of reads from both control and experimental data. From there, motif detection was carried out to look at the preferential binding motifs of the RNA-binding protein of interest. All of these steps are shown in the workflow (Figure 1).

All these steps were then encapsulated in batch scripts to streamline the analysis, and a stepwise procedure (detailed in ‘Launching the Pipeline’) helps the user initiate our pipeline efficiently, All aspects of the pipeline have been uploaded to GitHub, where users can find the directory and easily access the analysis pipeline on their own machines.

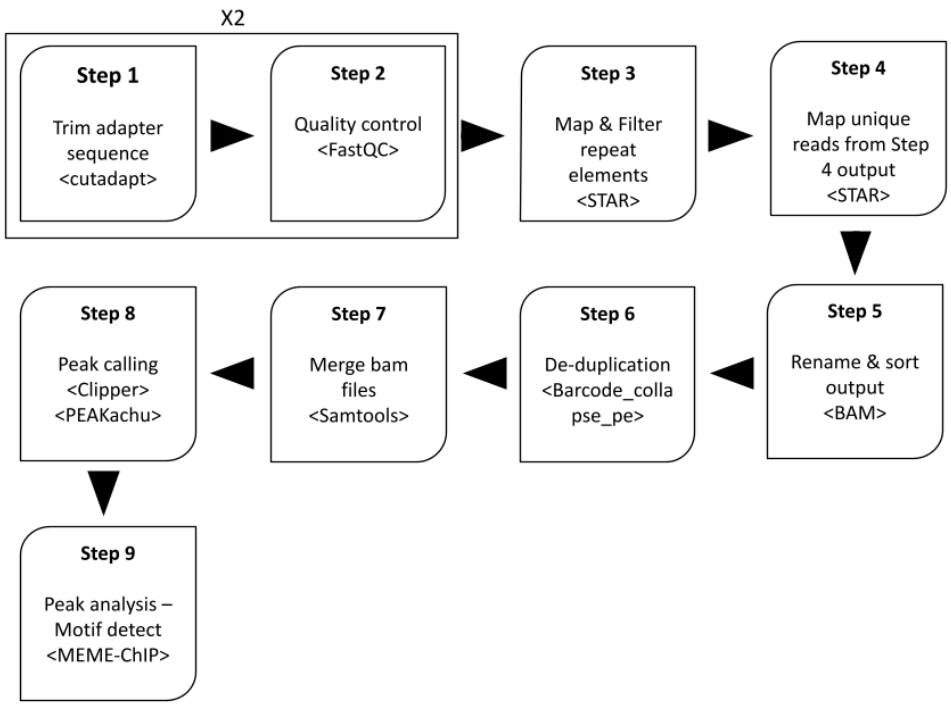

Figure 1. Workflow chart for RPI pipeline

Fastq Data

Usable input data can be found on the ENCODE website by looking for the ENCORE eCLIP data repository and navigating to 'File Details for a specific dataset', which will show replicates of raw sequencing data in a fastq file type for controls and experimental samples. All data inputted into the pipeline must be paired-end library type and are labeled as PE. Many proteins will have multiple biological replicates that should be used for analysis. Additionally, under 'Summary' details, information about control eCLIP experiments are linked, which will be utilized for ensuring experimental conditions are met.

The specific samples of interest will be stored in a separate file called "sample-links.txt" with their corresponding url links and name of the files retrieved from that link when called using a wget command in the setup (as shown below).

```
Sample,Read1,Read2,filename1,filename2
control,www.link1,www.link2,file-at-link1,file-at-link2
sample1,www.link1,www.link2,file-at-link1,file-at-link2
sample2,www.link1,www.link2,file-at-link1,file-at-link2
```

During the automated setup processing, the files for reads 1 and 2 are renamed as read1.fastq and read2.fastq for all control/experimental sample reads. This helps to generalize the pipeline workflow.

Cutadapt

Adapters are necessary for PCR amplification and sequencing. However, adapters may be sequenced if the machine goes past the read. As such, Cutadapt is a tool to trim adapters and adapter-dimers and will remove adapters from both the 3' and 5' end of sequenced reads.

For more information about cutadapt:
https://cutadapt.readthedocs.io/en/stable/guide.html#basic-usage

Fastqc

Fastqc is a tool to assess the quality of the raw data for high throughput sequence data. This ensures that the data is pre-processed correctly such that alignment to a reference genome can occur smoothly.

For more information about Fastqc and installation:
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

STAR

STAR is an alignment tool that does 2 main methods:

- generate genome index files from user-supplied reference genome sequences (FASTA file type) and annotation files (gtf file type)
- Map reads to the genome file

For more information about STAR:
https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf
To create the STAR environment:
https://github.com/alexdobin/STAR

UMI-tools

UMI-tools is a package designed to deal with Unique Molecular Identifiers (UMIs) that are present in high sequencing experiments. Through a UMI, identical copies arising from distinct molecules can be distinguished from those arising through PCR amplification of the same molecule. As such UMI-tools helps to ensure we have unique reads for peak calling.

For more information about UMI-tools:
https://genome.cshlp.org/content/27/3/491
For information on downloading UMI-tools:
https://github.com/CGATOxford/UMI-tools

Samtools

Samtools is a set of operations that can manipulate alignments in SAM and BAM-formatted files. Its tools include format conversions, sorting, merging, indexing, and retrieving reads in any regions swiftly.

For more information on Samtools:
http://www.htslib.org/doc/samtools.html

PEAKachu

PEAKachu is a peak-calling tool specifically for CLIP and eCLIP data, which will find possible binding motifs. Unlike other peak calling tools, PEAKachu is able to incorporate control data, which allows us to identify significant binding regions.

For more information on PEAKachu/downloading PEAKachu:
https://github.com/tbischler/PEAKachu

Launching the Pipeline:

Bridges-2 is a supercomputer that provides powerful general-purpose computing, computation intensive analysis, and pre- and post-processing. Considering this pipeline is designed for eCLIP data, running this on a supercomputer is highly recommended for computational and memory allocation resources.

Our pipeline can be installed in the Bridges environment by cloning its git repository (https://github.com/sriskid/group1-bdip.git) after logging into Bridges-2 with your XSEDE credentials on port 2222. Launching the pipeline is based on the assumption that the user has XSEDE credentials and also has a /group_num directory in the $PROJECT/../shared/ workspace, corresponding to your group number. Further, it requires two prerequisites after which the pipeline will be ready to launch.
Prerequisite 1: Cloning git repository and automating workspace setup
Prerequisite 2: Creating Conda environments

Running the Pipeline:

Here we lay out the detailed outline of how this pipeline is run. However, the code snippets shown here are put in separate scripts that are then called on by a batch script the user runs. This ensures a more efficient, user-friendly approach to using our pipeline.

Batch scripts with the commands for the pipeline are found in the manual

The core of our pipeline implementation can be executed in a modular-step wise fashion from top-to-bottom based on the workflow, using the commands from the manual to accomplish the following tasks.

Step 1: Trim Adapter Sequence

For trimming the adapter sequence, cutadapt is used to ensure that the eCLIP data is properly cleaned prior to alignment. Because cutadapt can only process one site of the read pair, cutadapt must be done twice for both the forward and reverse pair. As such, the end result are sequences with two adapters removed from both the 3′ and 5′ end.

Step 2: Quality Control

Fastqc is run after each cutadapt round to examine whether the libraries look right and the data is usable. It is also run after alignment of the sequence to a reference genome to check if duplication has been accounted for. Below is an example of what a FastQC quality plot would look like, with the blue line showing duplication level distribution and the red line showing an ideal curve after deduplication. This example shows that this sequence data set has a high level of duplication.

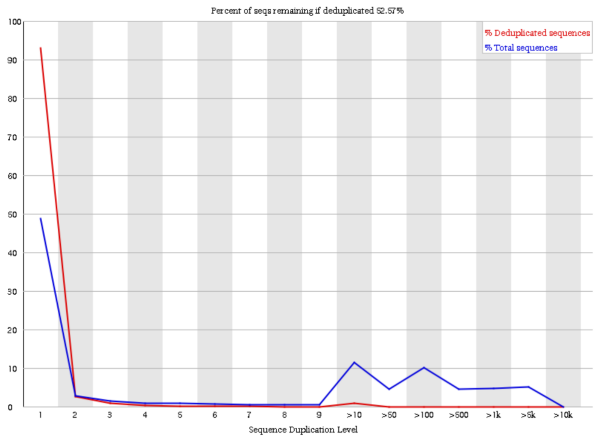

Figure 2. FastQC quality plot

It is important to note that in this pipeline, cutadapt and FastQC are ran twice for two runs of adapter finding and removal.

Step 3: Map and Filter Repeat Elements
Step 4: Map Unique Reads
Step 5: Rename and Sort Output

Although steps 3-5 are divided into separate steps, these steps can be tied as one big step. Using STAR rmRep, the output from cutadapt and FastQC round2 are mapped to a human specific version of RepBase to filter out repeating elements (**step 3**), which controls artifacts. **Step 4** takes the output from step 3 and uses STAR genome mapping to map unique reads to the human genome. STAR creates the genome index from the FASTA and gtf files supplied via the user to then map the specified eCLIP reads. For **step 5**, the output of the STAR alignment is renamed and sorted to a BAM file that will be used for further analysis.

Step 6: Deduplication

UMI-tools is a necessary processing step after alignment. UMI-tools is used to identify any duplicated reads that may be mapped redundantly to the genome. Identification of duplicated PCR reads that have the same UMI in the same location will be removed to have only one read for each UMI. This removal is done via a Python script called barcode_collapse_pe.py. Then, deduplication quality control can be checked again using FastQC.

Step 7: Sort and Merge Files

Prior to barcode collapsing, the bam file output is sorted using Samtools to ensure that read pairs are adjacent. After barcode collapsing, the bam file output is sorted again and merged so that the technical replicates become one file so that peak calling can be done on this final bam file.

Step 8: Peak Calling

PEAKachu is then called on the final bam file output processed by Samtools to find all possible binding motifs of the RNA-binding protein. Furthermore, PEAKachu incorporates the control data to ensure that all significant binding regions are found comparatively. The figure below is an example MA plot obtained from PEAKachu which shows the general trend of the log2 fold-change (y-axis) in dependence of the average mean of expression rate of the peaks (ref 5).

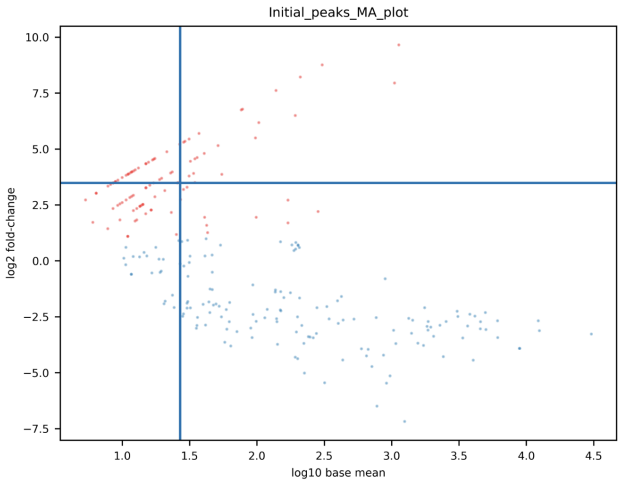

Figure 3. Example MA plot of PEAKachu.

Step 9: MEME-ChIP

MEME-ChIP is a comprehensive tool that is able to analyze motif detection for large DNA or RNA datasets. Outputs from peak analysis using PEAKachu are reformatted into bed files that are able to be analyzed for motifs with a user-defined width (i.e. 5 nucleotides). Additionally, the defined E-value will ensure that motifs found are significant. The figure below is an example plot of a sequence motif analyzed by MEME-ChIP. While the x-axis shows the nucleotide position of the motif, the y-axis shows the probability of the information regarding the nucleotide at the specific position. For example, the bigger letters G, C, A, T, G have higher probability at positions 2, 3, 4, 5, and 6, respectively, than letter T at position 1 (ref 5).

Figure 4. Plot of a sequence motif from MEME-ChIP.

Results

The end output of the peak calling (step 8) is a BED file, which contains all peaks of all unique reads from the eCLIP dataset, both experiment and control. While we were unable to get the functional analysis working, our expected output from the RBFOX2 protein would

be the binding motif: (U)GCAUG. Once we get the functional analysis working, we would compare our output to the known binding motif as validation that our analysis works as intended.

Discussion

In conclusion, our pipeline is able to successfully analyze eCLIP data for two experimental datasets and one control dataset, resulting in a BED file of analyzed peaks. From there, further functional analysis can be performed using MEME-ChIP and RCAS, which our pipeline is unable to work through currently. However, the re-created ENCODE pipeline is streamlined and user-friendly as the user can clone the Git repository, set up the appropriate environments, and work through the batch scripts efficiently without having difficulties working with each individual software.

Some limitations of our pipeline include the assumption that all computational work will be done on a supercomputer. This ensures that there is enough resources and memory to be able to carry out the computational tasks effectively. However, with a supercomputer, there may be additional wait times and the learning curve to understand a supercomputer environment can be tricky as the layout is very different from a local machine. Additionally, we recognize that improvements can be made on our pipeline. We would have liked to create a separate script that makes it easier for the user to set up the proper work environment without having to individually use different commands to do so. Furthermore, we would have liked to have figured out the functional analysis using MEME-ChIP fully.

References

1) Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, Onate KC, Graham K, Miyasato SR, Dreszer TR, Strattan JS, Jolanki O, Tanaka FY, Cherry JM. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018 Jan 4;46(D1):D794-D801. doi: 10.1093/nar/gkx1081. PMID: 29126249; PMCID: PMC5753278.

2) ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247. PMID: 22955616; PMCID: PMC3439153.

3) Ramanathan, M., Porter, D.F. & Khavari, P.A. Methods to study RNA–protein interactions. *Nat Methods* 16, 225–234 (2019). https://doi.org/10.1038/s41592-019-0330-1

4) Van Nostrand, E., Pratt, G., Shishkin, A. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* 13, 508–514 (2016). https://doi.org/10.1038/nmeth.3810

5) Florian Heyl, Daniel Maticzka, Bérénice Batut, 2022 CLIP-Seq data analysis from pre-processing to motif detection (Galaxy Training Materials).

https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/clipseq/tutorial.html Online.