UNDERSTANDING RNA REGULATION THROUGH ANALYSIS OF CLIP-SEQ DATA

APPROVED BY SUPERVISORY COMMITTEE

Yang, Xie, Ph. D. (Mentor)

Guanghua, Xiao, Ph. D. (Mentor)

Joshua Mendell, MD, Ph.D. (Chair)

David Mangelsdorf, Ph.D.

Michael Q. Zhang, Ph. D

## DEDICATION

I would like to thank both of my mentors, Drs. Yang Xie and Guanghua Xiao. Their mentoring and help for the past few years is the key element of every one of my accomplishments during my PhD study in UT Southwestern. My appreciation also goes to my thesis committee members, Drs. Joshua Mendell, David Mangelsdorf and Michael Q. Zhang, who provided me with very insightful suggestions on my current research and future directions. Further, I am very lucky to have the opportunities to collaborate with excellent scientists both on-campus and in other research institutes on a series of projects. Last but not least, I am really thankful of my wife and my friends who always support and encourage me during my research in academia and life in Dallas.

# UNDERSTANDING RNA REGULATION THROUGH ANALYSIS OF CLIP-SEQ DATA

by

TAO WANG

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center

Dallas, Texas

December, 2015

## Copyright

by

UNDERSTANDING RNA REGULATION THROUGH ANALYSIS OF CLIP-SEQ DATA

Tao Wang, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2015

Supervising Professor: Yang Xie, Ph.D. & Guanghua Xiao, Ph.D.

## Abstract

The past decades have witnessed a surge of discoveries revealing RNA regulation as a central player in cellular processes. The advent of cross-linking immunoprecipitation coupled with high-throughput sequencing (CLIP-Seq) technology has recently enabled the investigation of genome-wide RNA binding protein-RNA interactions, which is a very important component of RNA-regulation. However, proper and systematic bioinformatics analysis of CLIP-Seq data is still lacking and challenging. For the past few years, I have been devoting my research to methodological developments of CLIP-Seq data analysis, and developed MiClip and dCLIP for peak calling and differential analysis of CLLIP-Seq data, respectively. I have also applied my CLIP-Seq analysis pipelines in on-campus collaborating projects, in which I

identified ORF57 and nuclear AGO2 binding sites. Finally, I conducted analysis of public CLIP-Seq datasets to systematically characterize RNA binding protein targeting sites on circular RNAs.

# TABLE OF CONTENTS

DEDICATION ..... ii
Copyright ..... iv
Abstract ..... v
TABLE OF CONTENTS ..... vii
PRIOR PUBLICATION ..... x
LIST OF FIGURES ..... xiii
LIST OF TABLES ..... xiv
LIST OF ABBREVIATIONS ..... xv
CHAPTER ONE - INTRODUCTION ..... 1
1.1 RNA regulation and RNA binding proteins ..... 1
1.2 The development of CLIP-Seq technologies to profile RNA regulation ..... 2
1.3 Experimental steps of CLIP-Seq technologies ..... 3
1.3.1 Cross-linking ..... 3
1.3.2 Immunoprecipitation and enzymatic digestion ..... 4
1.3.3 Reverse transcription ..... 4
1.3.4 High-throughput sequencing ..... 5
1.4 Existing analysis methods and database servers ..... 6
1.4.1 CLIPZ ..... 6
1.4.2 StarBase v2 ..... 7
1.4.3 PARalyzer ..... 7
1.4.4 Piranha ..... 7
1.4.5 PIPE-CLIP ..... 8
1.4.6 wavClusteR ..... 8
1.4.7 PARma ..... 8
1.5 Experimental design and bioinformatics analysis considerations for CLIP-Seq technology ..... 9
1.5.1 Choosing a CLIP method ..... 9
1.5.2 Replicates ..... 10
1.5.3 Control experiments ..... 11
1.5.4 Sequencing depth ..... 11

1.5.5 Mapping ..... 12
1.5.6 PCR duplicates ..... 13
1.5.7 Intron-locating clusters and spliced-mapping reads ..... 14
1.5.8 Characteristic mutations in calling RBP binding sites ..... 14
CHAPTER TWO - METHODOLOGICAL DEVELOPMENTS ..... 16
2.1 Identifying RNA-RBP interactions through analysis of CLIP-Seq data ..... 16
2.1.1 Background and rationale ..... 16
2.1.2 Materials and methods ..... 18
2.1.2.1 CLIP-Seq datasets and mapping ..... 18
2.1.2.2 Finding CLIP clusters by overlapping CLIP-Seq tags ..... 18
2.1.2.3 Identify enriched regions (first round HMM) ..... 18
2.1.2.4 Identify reliable binding sites (second round HMM) ..... 20
2.1.2.5 Motif Analysis ..... 21
2.1.3 Analysis results ..... 22
2.1.4 Discussion ..... 29
2.2 Differential analysis of RNA-RBP binding strength in two conditions ..... 31
2.2.1 Background and rationale ..... 31
2.2.2 Materials and methods ..... 34
2.2.2.1 Data preprocessing ..... 34
2.2.2.2 Data normalization ..... 35
2.2.2.3 Hidden Markov Model (HMM) ..... 36
2.2.2.5 Implementation ..... 39
2.2.3 Analysis results ..... 39
2.2.3 Discussion ..... 42
CHAPTER THREE – REAL DATA ANALYSIS ..... 45
3.1 Identify nuclear AGO2 binding sites using PAR-CLIP ..... 45
3.1.1 Background ..... 45
3.1.2 Results ..... 46
3.1.3 Discussion ..... 48
3.2 Identify ORF57 binding sites using HITS-CLIP ..... 49
3.2.1 Background ..... 49
3.2.2 Results ..... 50

3.2.3 Discussion ..... 54
CHAPTER FOUR - IDENTIFY RBP BINDING SITES ON CIRCRNAS ..... 55
4.1 Background ..... 55
4.1.1 Circular RNA (circRNA) and its importance ..... 55
4.1.2 Profile RBP-circRNA interaction using CLIP-Seq data ..... 55
4.1.3 Challenges of mining RBP-circRNA interactions from CLIP-Seq data ..... 56
4.2 Bioinformatics pipeline development ..... 56
4.2.1 Download and curation of public CLIP-Seq data ..... 56
4.2.2 Linearization of circRNA library ..... 57
4.2.3 Competitive alignment of CLIP-Seq reads to circRNA library and reference genome ..... 58
4.3 Pipeline evaluation and downstream analysis ..... 59
4.3.1 Evaluation of the circRNA-identification pipeline ..... 60
4.3.2 Some RBP-bound circRNAs are predominantly located anti-sense to parental genes ..... 61
4.3.3 Some RBP-bound circRNAs show enriched sequence motifs ..... 63
4.3.4 Gene ontology enrichment of parental genes for circRNAs bound by RBPs ..... 63
4.4 Discussion ..... 65
CHAPTER FIVE - DISCUSSION ..... 67
BIBLIOGRAPHY ..... 69

# PRIOR PUBLICATION

**First or co-first author publication**

1. **Wang, T. ***, Zhan, X. *****, Bu, C. H. *****, Lyon, S. *****, Pratt, D., Hildebrand, S., Choi, J. H., Zhang, Z., Zeng, M., Wang, K. W., Turer, E., Chen, Z., Zhang, D., Yue, T., Wang, Y., Shi, H., Wang, J., Sun, L., SoRelle, J., McAlpine, W., Hutchins, N., Zhan, X., Fina, M., Gobert, R., Quan, J., Kreutzer, M., Arnett, S., Hawkins, K., Leach, A., Tate, C., Daniel, C., Reyna, C., Prince, L., Davis, S., Purrington, J., Bearden, R., Weatherly, J., White, D., Russell, J., Sun, Q., Tang, M., Li, X., Scott, L., Moresco, E. M., McInerney, G. M., Karlsson Hedestam, G. B., Xie, Y. and Beutler, B. Real-time resolution of point mutations that cause phenovariance in mice (2015) ***Proc Natl Acad Sci U S A.*** pii: 201423216. **(co-first author)**
2. **Wang, T.**, Xie, Y. and Xiao, G. dCLIP: a computational approach for comparative CLIP-seq analyses. (2014) ***Genome Biology***, 15, R11.
3. **Wang, T.**, Chen, B., Kim, M., Xie, Y. and Xiao, G. A Model-Based Approach to Identify Binding Sites in CLIP-Seq Data. (2014) ***PloS ONE***, 9, e93248.
4. Chen, X. *****, Zhao, C.*****, Li, X. *****, **Wang, T. ***, Li, Y., Cao, C., Ding, Y., Dong, M., Finci, L., Wang, J. H., Li, X. and Liu, L. Terazosin activates Pgk1 and Hsp90 to promote stress resistance. (2014) ***Nature Chemical Biology***, 11, 19-25. **(co-first author)**
5. Sei, E. *****, **Wang, T.***, Hunter, O. V., Xie, Y. and Conrad, N. K. HITS-CLIP analysis uncovers a link between the Kaposi's sarcoma-associated herpesvirus ORF57 protein and host pre-mRNA metabolism. (2015) ***PLOS Pathogens.*** 24;11(2):e1004652. **(co-first author)**
6. Chu, Y. *****, **Wang, T. ***, Dodd, D., Xie, Y., Janowski, B. A. and Corey, D. R. Intramolecular Circularization Increases Efficiency of RNA Sequencing. (2015) ***Nucleic Acid Research.*** **(co-first author)**

7. Eduati F*, Mangravite L*, **Wang T***, Tang H*, Bare C, Huang R, Norman T, Kellen M, Menden M, Yang J, Zhan X, Zhong R, Xiao G, Xia M, the NIEHS-NATS-UNC DREAM Toxicogenetics Collaboration, Friend S, Dearry A, Simeonov A, Tice R, Rusyn I, Wright F, Stolovitzky G, Xie Y[#], Saez-Rodriguez J[#] . Opportunities and limitations in the prediction of population responses to toxic compounds assessed through a collaborative competition. (2015) ***Nature Biotechnology.*** (In press) **(co-first author)**
8. Tao Wang, Guanghua Xiao, Yongjun Chu, Michael Q. Zhang, David R. Corey, and Yang Xie. Design and bioinformatics analysis of genome-wide CLIP experiments. (2015) ***Nucleic Acid Research.***

**Co-author publications**

9. Kwon, I., Xiang, S., Kato, M., Wu, L., Theodoropoulos, P., **Wang, T.**, Kim, J., Yun, J., Xie, Y. and McKnight, S.L. Poly-dipeptides encoded by the C9orf72 repeats bind nucleoli, impede RNA biogenesis, and kill cells. (2014) ***Science***, 345, 1139-1145.
10. Augustyn, A., Borromeo, M., **Wang, T.**, Fujimoto, J., Shao, C., Dospoy, P. D., Lee, V., Tan, C., Sullivan, J. P., Larsen, J. E., Girard, L., Behrens, C., Wistuba, II, Xie, Y., Cobb, M. H., Gazdar, A. F., Johnson, J. E. and Minna, J. D. ASCL1 is a lineage oncogene providing therapeutic targets for high-grade neuroendocrine lung cancers. (2014) ***Proc Natl Acad Sci U S A.***, 111, 14788-14793.
11. Bansal, M., Yang, J., Karan, C., Menden, M. P., Costello, J. C., Tang, H., Xiao, G., Li, Y., Allen, J., Zhong, R., Chen, B., Kim, M., **Wang, T.**, Heiser, L. M., Realubit, R., Mattioli, M., Alvarez, M. J., Shen, Y., Community, Nci-Dream, Gallahan, D., Singer, D., Saez-Rodriguez, J., Xie, Y., Stolovitzky, G., Califano, A. and NCI-Dream Community. A community computational challenge to predict the activity of pairs of compounds. (2014) ***Nature Biotechnology***, 32, 1213-1222.
12. Yun, J., **Wang, T.** and Xiao, G. Bayesian hidden Markov models to identify RNA-protein interaction sites in PAR-CLIP. (2014) ***Biometrics***.

13. Yun J., **Wang T.**, Wang X., Xiao G. Identification of RNA-protein binding sites in HITS-CLIP using heterogeneous logit models via semi-supervised learning. (2015) ***Annals of Applied Statistics.*** (Under revision)

# LIST OF FIGURES

FIGURE 1 The number of related articles found by Google Scholar.

FIGURE 2 Cartoon representation of MiClip algorithm.

FIGURE 3 MiClip analysis results on exemplary datasets.

FIGURE 4 miRNA seed motifs detected by MiClip on exemplary datasets.

FIGURE 5 Schematic representation of the dCLIP pipeline.

FIGURE 6 The analysis of the FMR1 dataset by dCLIP.

FIGURE 7 Preparation of RC-Seq libraries and analysis of PAR-CLIP derived RNAs.

FIGURE 8 Identification of enriched clusters mapping to the KSHV genome.

FIGURE 9 Identification and characterization of enriched clusters mapping to the human genome.

FIGURE 10 Cartoon of the pipeline for identifying RBP-circRNA interactions using CLIP-Seq data.

FIGURE 11 Decision tree to determine whether each alignment of a CLIP-Seq read is in a linear transcript or a circRNA.

FIGURE 12 Some RBP-bound circRNAs are predominantly located anti-sense to parental genes.

FIGURE 13 Some RBP-bound circRNAs show enriched sequence motifs

FIGURE 14 Gene ontology enrichment of parental genes for circRNAs bound by RBPs

# LIST OF TABLES

TABLE 1 Summary of CLIP-Seq analysis software programs and databases

TABLE 2 The enrichment of the top 10 miRNA seed sequences within the 5,795 clusters.

TABLE 3 Previous publications that reported large scale existence and locations of circRNAs

TABLE 4 Number of CLIP-Seq reads identified to support each of the 15 most enriched PolII-associated circRNA that have been experimentally validated before.

# LIST OF ABBREVIATIONS

| | |
|---|---|
| RNA | Ribonucleic acid |
| RBP | RNA binding protein |
| DNA | Deoxyribonucleic acid |
| miRNA | microRNA |
| lncRNA | long noncoding RNA |
| mRNA | messenger RNA |
| circRNA | circular RNA |
| nt | Nucleotide |
| Bp | Base pair |
| ROC | Receiver operating characteristic |
| AUC | Area under curve |
| HMM | Hidden Markov Model |
| HITS-CLIP | High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation |
| PAR-CLIP | Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation |
| iCLIP | individual-nucleotide resolution Cross-Linking and Immunoprecipitation |
| NGS | Next Generation Sequencing technology |

| | |
|---|---|
| RNase | Ribonuclease |
| SDS-PAGE | Sodium dodecyl sulfate polyacrylamide gel electrophoresis |
| ChIP-Seq | ChIP-sequencing |
| RNA-Seq | RNA sequencing |
| KSHV | Kaposi's sarcoma-associated herpesvirus |
| KS | Kaposi's sarcoma |
| PEL | Primary effusion lymphoma |
| KICS | KSHV-inflammatory cytokine syndrome |
| MCD | Multicentric Castleman's disease |
| CLASH | Cross-linking, ligation, and sequencing of hybrids |
| RIP-Seq | RNA-immunoprecipitation sequencing |

# CHAPTER ONE - INTRODUCTION

## 1.1 RNA regulation and RNA binding proteins

The diversity of RNA in sequence and structure underpins much of cell heterogeneity and complexity. RNA-binding proteins (RBPs) are proteins that bind to double- or single-stranded RNAs in cells and form ribonucleoprotein complexes with the bound RNAs. Located in either the nucleus or cytoplasm, or both, they engage in every step of the post-transcriptional modification process, including alternative splicing, regulation of mRNA levels, transport between cellular compartments, alternative polyadenylation, transcript stability, *etc.* (14,15). For example, the TIAR protein has been shown to be transported from the nucleus to the cytoplasm during Fas-mediated apoptotic cell death (16). One example of an intra-nuclear RBP is Yra1p, which has been found to be involved in mRNA export (17). Cytoplasmic RBPs, on the other hand, include Unr, which has been shown to be required for internally initiating the translation of human rhinovirus RNA (18).

RBPs bind target RNAs by recognizing their sequences or/and RNA secondary structures through RNA-binding motifs. For example, the AUF1 protein recognizes RNAs through a signature motif composed of 29–39 nucleotides with high A and U contents and a secondary structure specific to the RNAs (20). Binding of RBPs with RNA targets can also be regulated through competition with other RBPs and non-coding RNAs (21,22). RBPs may influence the global coordination of gene expression by organizing nascent groups of RNAs into downstream chains of the post-transcriptional modification process, through what is known as the “RNA-operon” theory (24). RBPs have been implicated in various types of human diseases (14,25-28). For instance, the RBP Musashi1 was found to be related to many cancer types, including those of the breast, colon, medulloblastoma and glioblastoma, as well as to neurogenesis and neurodegenerative diseases (28). In addition, lack of FMRP results in a deficiency in human cognition and premature ovarian insufficiency (29), and the FET protein family is responsible for RNA editing and plays important roles in many diseases (30,31). In summary, studying RNA-protein

interactions is necessary to achieve a systematic understanding of transcription, translation and other biological processes.

### 1.2 The development of CLIP-Seq technologies to profile RNA regulation

CLIP (cross-linking immunoprecipitation) is a molecular biology technology that employs UV cross-linking and immunoprecipitation in order to identify RBP-RNA interactions (32,33). The advantage of CLIP lies in allowing identification of interactions within cells (where the crosslinking occurs) versus interactions that might occur after cells are lysed. CLIP increases the confidence that observed interactions are physiologically relevant and can better justify identification of candidates for experimental validation. In the early reports, CLIPed cDNAs were sequenced in a low-throughput manner that yielded a few hundred sequence reads. Recently, next-generation sequencing (NGS) techniques have been applied to globally analyzing transcriptional and post-transcriptional regulation, including mRNA sequencing (34), alternative splicing (35), and miRNA profiling (36). The combination of CLIP with NGS technology has greatly improved our ability to study RBP-RNA interactions on the genome scale (37). While earlier CLIP-Seq studies focused more on the binding of RBP to mRNAs, recent studies have implicated a wide range of regulatory functions of RBP binding sites in long noncoding RNA (lncRNA) (38), circular RNA (39) and mitochondrial RNA (40).

There are three major technologies for CLIP-Seq experiments: 1. HITS-CLIP (high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation) (37,41), which is the first version of CLIP-Seq-Seq technology; 2. Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) (42), which improved the signal-to-noise ratio of the characteristic mutations observed in sequencing data by use of nucleoside analog; and 3. Individual-nucleotide resolution CLIP (iCLIP) (43), which achieved a much higher efficiency in reverse-transcription compared with HITS-CLIP and PAR-CLIP. Throughout this text, I used CLIP-Seq as a generic name for HITS-CLIP, PAR-CLIP and iCLIP. The field of RNA-regulation has seen rapid growth for all versions of

CLIP-Seq technology (**Fig. 1**).

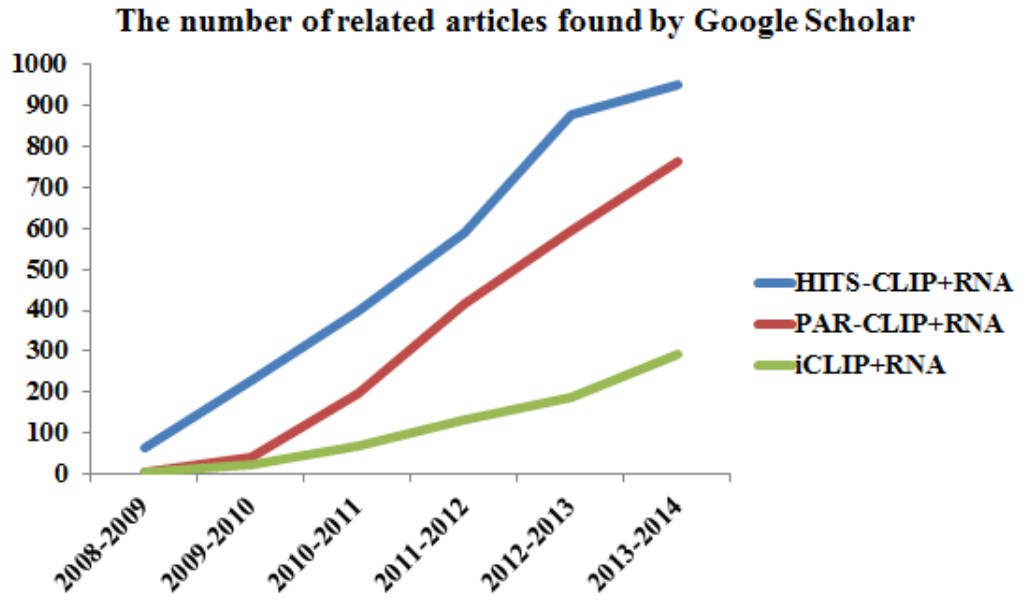



**Fig. 1.** Number of related scientific articles found by Google Scholar by searching for each of the key terms in the given interval of years. The key terms were searched together with "RNA" to yield more accurate citation counts.

### 1.3 Experimental steps of CLIP-Seq technologies

In general, CLIP-Seq technology involves cross-linking, partial RNA digestion, immunoprecipitation, reverse transcription and sequencing. The similarities and differences in the experimental procedures of these three CLIP methods are detailed below:

#### *1.3.1 Cross-linking*

The HITS-CLIP method treats the processed biomaterials from cells or tissues with UV light to cross-link RNAs with bound RBPs (32). It was the first CLIP-Seq platform developed for the genome-wide identification of RBP binding sites. Although successful, it is limited by its low efficiency of UV-induced crosslinking, which makes it difficult to locate high-confidence binding sites. The PAR-CLIP method resolves this efficiency problem by incorporating photoreactive ribonucleoside analogs, such as

4-thiouridine (4-SU) and 6-thioguanosine (6-SG), into living cells in the culture system before the UV light treatment (42). Although ribonucleoside analogs improve the signal-to-noise ratio in PAR-CLIP data, treatment of living animals with these chemicals could be toxic. iCLIP employs a UV cross-linking strategy similar to HITS-CLIP.

### *1.3.2 Immunoprecipitation and enzymatic digestion*

The immunoprecipitation step is similar for all HITS-CLIP, PAR-CLIP and iCLIP experiments. It generally involves bead preparation, cell lysis, partial RNA digestion immunoprecipitation, labeling, and SDS-PAGE. The purified protein-RNA complexes are then treated by proteinase K. In the RNA digestion step, substantial bias could be introduced due to sequence specificity and amount of RNase being used. Less bias is expected with a low sequence-specificity RNase, like RNase I, and mild digestion strength. Importantly, recombinant ligase and proteinase K enzymes contain bacterial RNAs, mostly rRNAs. If the 3' linker ligation is performed with free RNAs rather than with on-bead RNAs, these bacterial RNAs can also be cloned (44).

### *1.3.3 Reverse transcription*

In HITS-CLIP experiments, the remaining cross-linked amino acid(s) are attached to the RNAs, which then become an obstacle for reverse transcription. The reverse transcriptase can read through these obstacles on cDNAs with a certain probability, but errors, reflected as mutations after sequencing, may be introduced on the cross-linking sites. In PAR-CLIP, chemical property changes as a result of the nucleoside analog treatment and UV light stimulus will lead to mis-incorporation of dG rather than dA (4SU treatment), which will be reflected as mutations in the sequencing data. These cross-linking induced mutations could serve as markers for RBP binding sites and are sometimes referred to as "characteristic mutations". In HITS-CLIP experiments, the characteristic mutations could be substitutions, insertions, deletions or a combination of the above, depending on specific RBPs. For example, it has been shown

that deletions are preferably induced in Argonaut (AGO) HITS-CLIP experiments (45). On the other hand, PAR-CLIP experiments induce a specific type of substitution depending on the nucleotide analog used: applying 4SU or 6SG leads to T->C or G->A substitutions, respectively (10).

In reverse transcription, a significant number of cDNAs will be truncated at the attached residues since the reverse transcriptase fails to read through these obstacles. These truncated cDNAs are normally not sequenced in HITS-CLIP and PAR-CLIP. The iCLIP procedure is designed to capture these truncation sites of cDNA fragments with high efficiency through replacement of the intermolecular ligation procedure with intramolecular circularization. Therefore, the 5' ends of the sequencing reads, rather than characteristic mutations, are supposed to accurately mark the RBP targeting sites (43).

### *1.3.4 High-throughput sequencing*

cDNA libraries can be subject to deep sequencing. Since the RNAs are sheared into short fragments of 20-100bp, it was initially thought that single-end sequencing would usually be sufficient to cover whole cDNA fragments (46). However, some experiments require libraries of RNA fragments that are longer than those that could be covered by single-end sequencing, mainly due to dissimilar preferences in the library size selection step. Paired-end sequencing may be desirable in these cases so that whole cDNA fragments can be covered, because the lengths of RBP-RNA contact regions are comparable to the length of sequencing reads. Argonaut protein (AGO) is a key protein involved in RNAi that forms critical complexes with micro RNAs. AGO-RNA contact regions were estimated to be around 60bp long (41). Therefore, exact coverage is important since identification of RBP binding sites usually requires a much higher resolution compared to ChIP-Seq experiments for transcription factors, whose resolution requirements are at least a few hundred base pairs (47).

### 1.4 Existing analysis methods and database servers

The quick development of CLIP-Seq technologies has posed new and challenging analytical problems to the bioinformatics communities, ranging from peak-calling of RBP binding sites to downstream analysis to make biological discoveries. In this section, I will give an overview of existing bioinformatics analysis software and databases for CLIP-Seq experiments. **Table 1** summarizes the major software programs, pipelines and databases that have been developed so far. I will discuss some of these in more details in this section.

| Software/Database | Type | Comment | Citation |
|---|---|---|---|
| CLIPZ | Database | Can carry out simple bioinformatics analysis | (1) |
| StarBase v2 | Database | Contains CLASH datasets as well | (2,3) |
| doRiNA | Database | Focuses on miRNA biology | (4,5) |
| CLIPdb | Database | Contain uniformly identified binding sites of publicly available genome-wide CLIP datasets | (6) |
| PARalyzer | Software | Peak-finding algorithm for PAR-CLIP dataset only | (7) |
| Piranha | Software | Peak-finding and differential binding detection algorithm | (8) |
| PIPE-CLIP | Software | Peak-finding algorithm | (9) |
| wavClusteR | Software | Peak-finding algorithm for PAR-CLIP dataset only | (10) |
| PARma | Software | Differential binding detection algorithm for AGO PAR-CLIP dataset only | (11) |
| PAR-CLIP HMM | Software | Peak-finding algorithm employing Hidden Markov Model | (12) |
| GraphProt | Software | Peak-finding algorithm that can handle both RNAcompete and genome-wide CLIP data flexibly | (13) |
| Pyicos | Software | Peak-finding algorithm that can handle ChIP-Seq, genome-wide CLIP and RNA-Seq data flexibly | (19) |
| miRTarCLIP | Software | Peak-finding algorithm that employs a novel C to T reversion strategy in PAR-CLIP dataset analysis | (23) |

**Table 1** Summary of CLIP-Seq analysis software programs and databases

#### *1.4.1 CLIPZ*

CLIPZ is mainly a database for CLIP-Seq datasets. There were 94 publicly-visible samples stored on CLIPZ as of April 2015. CLIPZ also provides simple bioinformatics analysis for stored samples. It first aligns the sequencing reads to genomes and transcriptomes, allowing alignments with more than one error

(substitution, insertion or deletion). Then it generates clusters of sequencing reads and computes statistics like T->C substitutions for PAR-CLIP dataset. Finally, CLIPZ allows users to sort the clusters based on these computed features.

1.4.2 StarBase v2

StarBase v2 is a database designed for decoding pan-cancer and interaction networks of RBPs, mRNAs and various types of non-coding RNAs from CLIP-Seq datasets and CLASH datasets (48). As of April 2015, StarBase v2 contained 111 CLIP-Seq datasets from 37 studies. StarBase v2 processes all the stored datasets and presents the analysis results through disparate portals such as miRNA-lncRNA interactions, miRNA-target interactions, protein-mRNA interactions, and function predictions. The analysis conducted by StarBase v2 mostly relies on previously published software, such as PARalzyer for PAR-CLIP dataset analysis and TargetScan (49) and other similar pipelines for miRNA target site predictions.

### *1.4.3 PARalyzer*

PARalyzer is a popular peak-calling algorithm for PAR-CLIP datasets only. PARalyzer employs a non-parametric kernel-density estimation classifier to identify the RNA-RBP interaction sites using both total binding intensity information and T->C mutation information. It provides a dozen parameters, such as minimum number of reads and minimum number of conversions for a cluster, to help users filter the final results.

1.4.4 Piranha

Piranha is mainly a peak-calling algorithm, but it also provides a way to detect differential binding across a range of conditions. All reads are binned and each bin represents a genomic interval. Piranha allows the users to flexibly choose an underlying model, including Poisson distribution and Negative

Binomial distribution. It permits users to add additional covariates such as mutation data or transcript abundance data in a regression framework. This enables Piranha to incorporate mutation data in peak-finding or to conduct a differential binding analysis.

### 1.4.5 PIPE-CLIP

PIPE-CLIP is a Galaxy-based comprehensive online pipeline for CLIP-Seq data analysis. It processes BAM files by filtering out reads that do not meet mismatched numbers and/or aligned read-length criteria and by removing PCR duplicates according to reads locations or sequences. Then it applies zero-truncated negative binomial regression to identify the enriched clusters and fits a binomial distribution to assess the significance of featured mutations/truncations. After that, enriched clusters with significant mutations/truncations are reported as binding sites.

### 1.4.6 wavClusteR

wavClusteR is designed for identifying RBP peaks in a single PAR-CLIP experiment. It defines a mixture model where the first component indicates random substitutions, which are not induced by cross-linking, and the second component indicates cross-linking-induced substitutions that serve as markers of RBP-protein binding sites. wavClusteR relies on the assumption that all types of non-experimentally-induced substitutions have approximately the same distribution as the first component, while only PAR-CLIP-induced T->C mutations exist in the second component of the mixture model. However, this may not be the case for tumor cell lines where the background mutation profiles are distinct for each type of substitution (50).

### 1.4.7 PARma

PARma is a tool for differential AGO PAR-CLIP data analysis. In PARma, a statistical model and a novel pattern discovery tool are iteratively applied to estimate probabilities and to assign the most

probable miRNAs until convergence. The statistical model is composed of three independent parts that consider the T->C mutation frequencies as well as the properties of the nucleotide compositions at both ends of the sequencing reads. The PARma algorithm addresses several important issues in the data preprocessing step, such as the handling of spliced-mapping reads and consideration of experimental replicates. However, it can only be applied to differential AGO PAR-CLIP datasets.

## 1.5 Experimental design and bioinformatics analysis considerations for CLIP-Seq technology

### *1.5.1 Choosing a CLIP method*

The goal of a specific study is the primary consideration for choosing a CLIP method. Whether the experiment is to be done *in vivo* is one reason for favoring HITS-CLIP or iCLIP over PAR-CLIP, since the ribonucleoside analog treatment could be toxic. This is why HITS-CLIP and iCLIP have broad application in cultured cells, animal tissues and plants. On the other hand, if the study wishes to reach a higher resolution at determining binding sites, PAR-CLIP or iCLIP should be favored. This is because PAR-CLIP has a much higher proportion of reads with characteristic mutations on cross-linking sites compared with HITS-CLIP, and in iCLIP truncation sites can be directly used to accurately map interaction events. Thirdly, iCLIP is technically more challenging compared with HITS-CLIP and PAR-CLIP, which has probably limited its use. iCLIP requires the protein-bound RNA to be mildly digested by an endonuclease, which ensures the reads originated from truncated cDNA are long enough to be aligned. Therefore, a researcher needs to first experimentally determine the best condition to achieve an acceptable partial RNA digestion. In addition, iCLIP implements cDNA circularization and re-linearization steps. These steps require researchers to properly cut desired bands from polyacrylamide gels and carry out product elution and isolation. RNA obtained from CLIP techniques are generally in minute amounts. Extra manipulations on hardly-detectable cDNA will give an extra challenge to preparing an iCLIP sequencing library.

### *1.5.2 Replicates*

In RNA-Seq experiments, it has been shown that increasing the number of biological replications consistently improves expression level quantifications and increases the statistical power to detect differentially expressed genes (51). It has become a routine for most RNA-Seq experiments to have replicates to improve the data quality and reproducibility. For CLIP-Seq experiments, there is no rigorous study on how the replicates affect the experimental results. Many CLIP-Seq studies are based on a very limited number (1-5) of replicated experiments, and replicates are often pooled or only one of them is used in analysis (30,52). The number of replicates that should be obtained depends on many factors, including the goal of the experiments, the variations of experiments, the sequencing depth and also the binding patterns of specific RBPs. For example, if the goal of the study is to conduct a comparative analysis between CLIP-Seq conditions, then the quantification of within- *vs.* between-group variation is very important and replicates will be of great value. The number of replicates to conduct can also take into consideration previously published studies for the experimental variations and binding patterns.

With respect to bioinformatics analysis, it is undesirable to pool the replicates. As each replicate could have a different sequencing depth, pooling will tend to down-weigh the replicates with less-sequenced reads. Moreover, the variation information of binding intensity at each binding site is lost after pooling. A measurement called biologic complexity (BC) has been applied to identifying RBP binding sites using replicates (41). Other than BC, PARma is the only algorithm that can consider replicates in its statiscal algorithm (11). The DESeq package implements a statistical model that can incorporate replicate information to call differentially expressed regions (53). It was originally proposed for ChIP-Seq and RNA-Seq data, but could be adapted to CLIP-Seq studies where replicates are available (54). However, more advanced statistical approaches are also needed to address specific data features from CLIP-Seq experiments to better analyze such data with replicates more efficiently. In summary, no rigorous and comprehensive study has been conducted to investigate the effects of the number of replicates on statistical power and the accuracy of binding site detection for CLIP-Seq experiments. Future studies and

the development of bioinformatics tools for analyzing such experiments with replicates would improve the experimental design and data analysis.

### *1.5.3 Control experiments*

Most recently published CLIP-Seq studies did not use background control experiments for identification of binding sites. Accordingly, few analysis approaches could process the sequencing data with both CLIP-Seq and control conditions. Since CLIP-Seq experiments involve stringent washes, such experiments without controls still identified high-confident RBP binding sites. However, generating control experiments for CLIP studies would improve the analysis and interpretation of the results. First, the ranking of identified binding sites from analyzing CLIP-Seq data is usually biased towards abundantly expressed genes. If the CLIP cluster binding intensities are not normalized by control experiments and some clusters with high apparent binding strength could simply be intermediate-level-binding-strength sites on highly expressed RNA transcripts. Therefore, having background control experiments could help reduce such bias. Secondly, background RNA-Seq experiments could also help to identify SNPs in cell lines or tissue samples, as we have previously mentioned. In addition, if the study's goal is to understand RBP functions such as splicing, conducting an RNA-Seq experiment will help to discern which sites are functionally relevant. Konig et al. suggested a few ways to conduct background experiments (43) for iCLIP experiments, such as no-antibody sample, non-crosslinked cells, or immunoprecipitation from a knockout condition. Liu *et al* experimentally showed that input RNA or RNA-Seq experiment is also a good control (55). Again, which type of control experiment to conduct also depends on the specific goal of the study.

### *1.5.4 Sequencing depth*

There is no consensus on the required sequencing depth for CLIP-Seq experiments, which can range from less than 10 million reads to more than 300 million reads for one experiment. The early studies

generated low numbers of reads, while more recent studies generated much deeper reads for an RBP under one treatment. Due to the generally limited complexities of the cDNA libraries, the very deep sequencing may not necessarily capture more unique events of RBP-RNA interactions for HITS-CLIP and PAR-CLIP experiments. The library complexities vary greatly for different CLIP experiments depending on many factors (6). One factor is how many binding sites the RBP under investigation truly binds. If the RBP has very specific binding sites, the expected library complexity would be small. Overall, the type of CLIP-Seq experiment, cost of sequencing, and the number of true binding sites of the RBP should all be considered in determining the proper sequencing depth for the CLIP-Seq experiments. Readers may refer to another review that thoroughly discusses the matter of sequencing depth in genomics studies (56).

### *1.5.5 Mapping*

Aligning the reads to a genome or transcriptome is the first step in CLIP-Seq analysis. Mapping to a genome is usually chosen since there are sometimes many CLIP clusters that locate within-reference gene introns. Mapping to a transcriptome or to both genome and transcriptome would be a good choice when the focus of the study is to detect RBP binding sites on mature RNAs that have already been spliced. In general, an aligner such as Gsnap that can handle short deletions and spliced-mapping will be a good choice. Gsnap is preferred by the CLIPper software and it scored high in a systematic comparison of RNA aligners (57).

Another issue to consider is whether rRNAs, tRNAs and other types of repetitive sequences are of interest or should be removed by screening them from the pool. If not, mapping to a pre-masked genome or removing rRNAs at the experimental stage using kits like Ribo-Zero may be more efficient. But this may not be the case with experiments that are conducted to make a comparative analysis, where 18S rRNAs can be used as a control invariant gene (58) Also, it is common practice for CLIP-Seq data mapping to discard reads that can be mapped to multiple locations (30,59,60). However, some RBPs may

have real binding sites in genes that have multiple copies in the genome. In such cases, discarding non-uniquely mapped reads will result in the loss of some true binding sites.

### *1.5.6 PCR duplicates*

Since CLIP-Seq experiments involve PCR amplification from cDNA libraries with limited complexities, removal of PCR duplicates amplified from common unique cDNA fragments is an important step. After duplicate removal, the size of the sequencing data usually drops dramatically. There are a few ways to define PCR duplicates in CLIP-Seq data. (1) Introducing random barcodes into the cDNA adaptor. This approach has been primarily applied to iCLIP experiments and made it relatively easy to define PCR artifacts from the iCLIP data. Barcoding can give the clearest answer to whether a sequencing read is a PCR duplicate, and in fact it can also be applied to HITS-CLIP and PAR-CLIP, though this is not commonly done yet. PIPE-CLIP (9) has a bioinformatics procedure that can remove PCR duplicates according to barcodes for genome-wide CLIP data of all three sorts. (2) For HITS-CLIP and PAR-CLIP, earlier studies defined PCR duplicates as sequencing reads having the same aligned genomic starting sites, and duplicates were collapsed to a single sequencing tag (45). This may be too conservative, which usually leads to a collapsed sequencing read dataset that is less than 1/10 of its original size. (3) Another popular approach adopted in many studies (61-63) is to define reads that have exactly the same mapping coordinates as PCR duplicates. (4) Alternatively, it is also possible to define PCR duplicates as those having the same nucleotide sequence. Unfortunately, to our knowledge there hasn't been any strict comparison reported in the literature to help select the best approach from (2)-(4) for HITS-CLIP and PAR-CLIP, and the scenario is even more complicated for paired-end sequencing reads. One consideration to choose among approaches (2)-(4) is the number of reads left after duplicate removal. If this step is too stringent, too few reads may be left for downstream analysis.

### *1.5.7 Intron-locating clusters and spliced-mapping reads*

Most CLIP-Seq experiments do not distinguish nucleic RNAs from cytoplasmic RNAs because the RNA is obtained from whole cells. Since libraries could contain cDNAs converted from nascent pre-mRNAs, it is possible that a significant portion of CLIP reads will be mapped to reference gene introns. For example, in a few published studies, the proportions of intron-locating reads could be as low as 15% but also as high as 90% (64-67). This proportion depends on both the compartment of the cell that is being investigated and the property of the RBP under investigation. For example, Chu *et al* found through PAR-CLIP that nucleic AGO2 preferentially binds intron regions while cytoplasmic AGO2 mainly binds 3' UTR regions (68).

On the other hand, there are varying amounts of cDNAs generated from mature mRNAs in the libraries. Therefore, some of the sequencing reads could be mapped across splicing junctions. As a result, it is sometimes important to use an aligner that can handle splice junction mapping, or alternatively, to map the sequencing reads to the transcriptome in addition to the reference genome. However, usually fewer than 5% of all CLIP reads are mapped across splice junctions, due to two possible reasons: (1) only a small fraction of RBP binding sites are close to or on top of splice junctions, or (2) current aligners are not very efficient in mapping reads across splice junctions.

### *1.5.8 Characteristic mutations in calling RBP binding sites*

The read counts are usually the primary measure for peak-calling from most algorithms, but the characteristic mutations induced by cross-linking procedures have also been proved to be useful for peak calling algorithms. In HITS-CLIP and PAR-CLIP, the cross-linking procedure induces mismatches in the final sequencing data, which could be used to pinpoint the location of RBP target sites at single-base-pair resolution and have been used to improve the binding target identifications. However, the proportion of sequencing reads with characteristic mutations vary greatly from 20%-80% for PAR-CLIP data

(30,42,69,70). For HITS-CLIP data, the proportion is only round 10% (52) and even as small as <1% in one case (71). A recent study by Bahrami-Samani *et al* that investigated CLIP-Seq data from 20 public studies yielded similar percentages (72). Moreover, mutant bases are usually sparsely spread within CLIP clusters, leading to usually small ratios of mutant tags out of total tags on the exact cross-linking sites. Low mutant tag ratios in some experiments could be problematic for bioinformatics pipelines for analyzing HITS-CLIP and PAR-CLIP data that rely one mutation ratios, such as wavClusteR (10). On the other hand, there may be a small number of bases covered by CLIP clusters that show close to 100% mutant rates, which are likely SNPs in the cell lines or tissue samples instead of true RBP binding sites. To address these issues involving mutations, wavClusteR (10) introduces a parameter that effectively discards bases with mutation rates higher than a user-defined cutoff. Other ways to solve this problem include conducting control RNA-Seq experiments to detect SNPs or comparing results to databases of known SNPs. These observations, in addition to the obscurity of true characteristic mutations for some HITS-CLIP data, suggest that although characteristic mutation can help pinpoint the binding site and increase peak calling accuracy in most cases, careful examination of mutations from CLIP-Seq experiments, especially from HITS-CLIP data, are necessary.

# CHAPTER TWO - METHODOLOGICAL DEVELOPMENTS

CLIP-Seq data offers the opportunities to understand RBP-RNA binding on the genome-wide scale. However, the analysis of CLIP-Seq data is non-trivial. The first step of bioinformatics analysis on CLIP-Seq data is generally identification of RBP binding sites on RNAs in one condition or identification of differential RBP binding sites on RNAs across different conditions. There are certain special properties of CLIP-Seq data that need to be taken care of and also utilized during these analytical processes. I developed user-friendly algorithms based on HMM to tackle these problems, respectively.

## 2.1 Identifying RNA-RBP interactions through analysis of CLIP-Seq data

### *2.1.1 Background and rationale*

Most studies that conducted CLIP-Seq experiments are interested in finding RBP binding sites in their culture or tissue system. In most studies, *ad hoc* methods are employed to process CLIP-Seq data. Generally, CLIP clusters are formed by overlapping the short sequencing tags and simple cut-offs are applied to produce lists of reliable CLIP clusters (30,37,41). However, these methods are sensitive to the choice of cut-off values, and there is no confidence values associated with the identified binding sites. To better analyze CLIP-Seq datasets, a few bioinformatics tools have been developed recently. CLIPZ is an online server for analyzing CLIP-Seq datasets (1). However, it still works under an *ad hoc* framework --- no significant levels are given for the resulting binding sites. *wavClusteR* (73) is designed to analyze PAR-CLIP experiments. It assumes a two-component mixture model based on relative substitution frequencies (RSFs) for identifying reliable binding sites, and employs wavelet transformation for resolving peak boundaries. *PARalyzer* (7) is also designed to analyze PAR-CLIP experiments. *PARalyzer* identifies reliable binding sites as nucleotides with a minimum read depth and having a higher likelihood of T->C conversion than non-conversion. *PARalyzer* and *wavClusteR* could not be easily extended to

other types of CLIP-Seq datasets due to the underlying model assumptions. Besides, StarBase v2.0 (74) is a comprehensive database with more than 100 CLIP-Seq datasets.

None of the above mentioned algorithms considered the spatial dependency structure in the CLIP-Seq data for increasing signal-to-noise ratio of called results. The spatial dependency structure refers to the fact that neighboring regions are more likely to have the same properties, and vice versa. RNA-RBP interactions usually occupy a certain length of RNA fragments, which implicates the spatial dependency structure. Hidden Markov Model (HMM) is a stochastic statistical model that can utilize this feature for inferring true status of each unit in a chain, which is specialized in solving the spatial dependency problem.

Here, I present a model-based approach, MiClip, for analyzing both HITS-CLIP and PAR-CLIP data. This approach has been implemented in the R statistical environment (75). It first removes duplicates and finds CLIP clusters, then divides the task of identifying reliable binding sites into two rounds of Hidden Markov Model (HMM). The first HMM infers enriched *vs.* non-enriched regions in CLIP clusters, and the second HMM infers binding sites of RBPs *vs.* non-binding sites within the enriched regions. Finally, the reliable binding sites and the CLIP clusters containing these sites are reported in a user-friendly format. I have tested this algorithm on two datasets and shown that MiClip provides a general and efficient framework for identifying high-confidence RBP binding sites at high resolution. In the AGO HITS-CLIP dataset that was shown here, the signal/noise ratios of miRNA seed motif enrichment produced by the MiClip approach are between 17% and 301% higher than those by the *ad hoc* method for the top 10 most enriched miRNAs. To facilitate the application of the algorithm, I have released an R package, *MiClip (http://cran.r-project.org/web/packages/MiClip/index.html),* and public web-based graphical user interface software (*http://galaxy.qbrc.org/tool_runner?tool_id=mi_clip*) for customized analysis.

### *2.1.2 Materials and methods*

#### 2.1.2.1 CLIP-Seq datasets and mapping

The AGO HITS-CLIP dataset described in Chi, *et al.* (41) was downloaded from http://AGO.rockefeller.edu. The alignment of the HITS-CLIP data was done by Novoalign (*Novocraft Technologies*) to the mm10 reference genome. After alignment, the resulting SAM files from Brain A-E were pooled in the HITS-CLIP dataset.

#### 2.1.2.2 Finding CLIP clusters by overlapping CLIP-Seq tags

The MiClip package took the alignment SAM format files as input. In each dataset, duplicate reads that have the same mapping coordinates (including strand) were identified and collapsed to a single "tag". Tags overlapping by at least one nucleotide were grouped together to form CLIP clusters, and those not overlapping with any other tags were discarded. The deletions on each base were counted as mutation events for the AGO dataset.

#### 2.1.2.3 Identify enriched regions (first round HMM)

To identify enriched regions, CLIP clusters were divided into bins of 5bp. Let $x_t^{(k)}$ be the total tag count in the *t*-th bin of *k*-th cluster, so cluster *k* could be represented as a series of tag count numbers: $\vec{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, ..., x_{T_k}^{(k)})$. Here I used HMM to determine the enriched regions from observed tag counts.

The HMM has two states: $\begin{cases} I_t^{(k)} = 0 & \text{if bin t is non-enriched} \\ I_t^{(k)} = 1 & \text{if bin t is enriched} \end{cases}$. Poisson model is a popular model to fit the count data (76,77). Given state $I_t^{(k)}$, the observed tag counts were modeled by a two-component

Poisson mixture model: $\begin{cases} X_t^{(k)} \sim Poisson(\lambda_0) \mid I_t^{(k)} = 0 \\ X_t^{(k)} \sim Poisson(\lambda_1) \mid I_t^{(k)} = 1 \end{cases}$, so the emission probability could be written as $\Pr(X_t^{(k)} = x \mid \lambda_0, \lambda_1, \omega) = (1-\omega)\frac{\lambda_0^x e^{-\lambda_0}}{x!} + \omega\frac{\lambda_1^x e^{-\lambda_1}}{x!}, \quad (\lambda_0 < \lambda_1)$, where the ω is the proportion of enriched bins in the CLIP clusters. The transition matrix Π is a 2×2 matrix, where element $\pi_{r,s}$ is the transition probability $\Pr(I_t^{(k)} = s \mid I_{t-1}^{(k)} = r)$. I estimated the $\lambda_0$, $\lambda_1$ and ω parameters first from the observed data using method of moments (78), and then used the standard Viterbi algorithm (79) to infer the hidden states $I_t^{(k)}$, namely the enriched *vs.* non-enriched bins. The Viterbi algorithm determines the hidden state of each bin according to the criterion that posterior probability of each bin in the inferred state (enriched or non-enriched) should be larger than the posterior probability of this bin in the other state, given the model and observation. Finally, the adjacent enriched bins were concatenated into enriched regions. A cartoon illustration of how first round HMM works is shown in **Fig. 2a.**

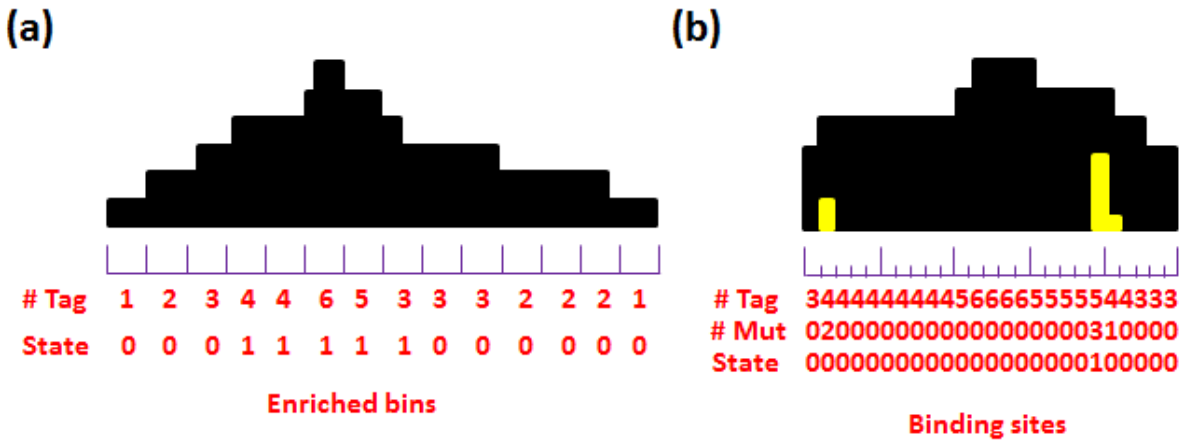



**Fig. 2.** Cartoon representation of MiClip algorithm (a) A cartoon of the first HMM. At the bottom, "# Tag" refers to the rounded number of tags covering each bin and "State" shows whether a bin was inferred to be "enriched". (b) A cartoon of the second HMM. At the bottom, "# Tag" refers to the number of total tags covering each base, "# Mut" refers to the number of mutant tags covering each base and "State" shows whether a base was inferred to be a "binding site".

### 2.1.2.4 Identify reliable binding sites (second round HMM)

To identify reliable binding sites, the enriched regions were further divided into bins of 1bp. Let $(m_b^{(n)}, x_b^{(n)})$ be the number of mutations and total tag count in the *b*-th base pair of the *n*-th enriched region. This HMM was designed to have two states: $\begin{cases} D_b^{(n)} = 0 & \text{if base pair b is not a binding site} \\ D_b^{(n)} = 1 & \text{if base pair b is a binding site} \end{cases}$

The observed number of mutations $M_b^{(n)}$ given the tag count $X_b^{(n)}$ given $D_b^{(n)}$ was modeled by $\begin{cases} M_b^{(n)} \mid X_b^{(n)} \sim ZIB(p_0, X_b^{(n)}, \varphi) \mid D_b^{(n)} = 0 \\ M_b^{(n)} \mid X_b^{(n)} \sim Bin(p_1, X_b^{(n)}) \mid D_b^{(n)} = 1 \end{cases}$, here a zero inflated binomial distribution (ZIB) (80) with probability $p_0$, size $X_b^{(n)}$ and inflation parameter φ was used to model the background mutations,

such as random sequencing errors at non-binding sites ( $D_b^{(n)} = 0$ ), and a binomial distribution with probability $p_1$ and size $X_b^{(n)}$ was used to model the cross-linking induced mutations at RBP binding sites ( $D_b^{(n)} = 1$ ). So, the emission probability could be written as:

$$\Pr(\mathrm{M}_b^{(n)} = m \mid X_b^{(n)} = x, p_0, p_1, \theta, \varphi) = (1-\theta)\left[\varphi I(m=0) + (1-\varphi)\binom{x}{m} p_0^m (1-p_0)^{x-m}\right] + \theta\left[\binom{x}{m} p_1^m (1-p_1)^{x-m}\right]$$

where θ is the proportion of binding sites in enriched regions. I estimated the parameters as follows: First, from the density plot of mutation rates *(m/x)*, assume that I could observe two modes $\hat{f}_1$ and $\hat{f}_2$, where $\hat{f}_1$ corresponds to the probability for success of the background ZIB component and $\hat{f}_2$ corresponds to the probability of success for the binomial component. Then, I chose a parameter *c* specified by the user, so that $\hat{f}_1 < c < \hat{f}_2$. The bins with mutation ratio $\frac{m}{x} < c$ were used to estimate $p_0$ and φ for ZIB distribution using method of moments, and the remaining bins were used to estimate $p_1$ for the binomial distribution. According to our simulation studies (data not shown), the estimation procedure is robust and the choice of parameter c will not greatly change the estimated parameters. Again, the standard Viterbi algorithm (79) was used to infer the hidden states, and the probability of being a reliable binding site $\Pr(\mathrm{D}_b^{(n)} = 1 \mid \vec{\mathrm{X}}, \vec{\mathrm{M}})$ was calculated for each base pair *b*. The enriched part of the peak in **Fig. 2a** is shown as a cartoon in **Fig. 2b** to illustrate how second round HMM works.

### 2.1.2.5 Motif Analysis

Exact matches to the 7-mer seed motifs of the top 10 most enriched miRNAs were scanned by an in-house Perl script (this script was included in the exec folder of the *MiClip* package for users to replicate our results). The Perl script scanned the 7-mers through the genomic sequences covered by all significant clusters in the HITS-CLIP dataset and reported the locations of matches. To estimate the signal to noise ratio, 40,000 background sequences (with the same length as the mean length of the identified clusters)

that have no overlapping regions with the target sequences were randomly chosen from the mouse genome. The same scanning procedure was done in these background sequences to calculate the signal/noise ratio for miRNA seed motif enrichment. The relative distances from the binding site to the centers of the matches within each cluster were calculated. If a cluster has more than one possible binding site, the shortest distance was kept for analysis.

### *2.1.3 Analysis results*

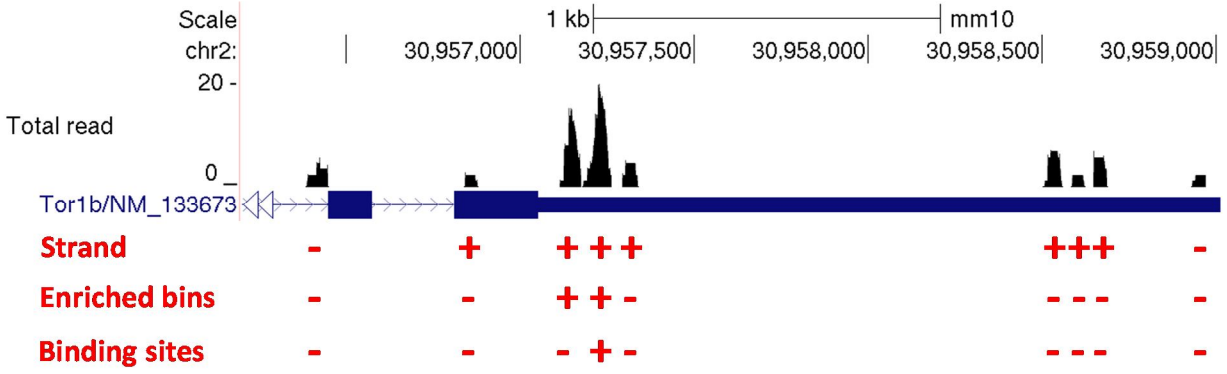
In the AGO HITS-CLIP study, AGO protein bound to mouse brain RNAs was purified by UV-irradiating P13 neocortex and immunoprecipitation. After purification, complexes of two different modal sizes were observed: 110 kD complexes harboring miRNAs (miRNA library) and 130kD complexes harboring mainly mRNAs (mRNA library). The MiClip analysis was carried out only on the mRNA library. All five replicates of the AGO mRNA datasets were pooled, resulting in a total of around 26 million reads. Around 22 million reads were aligned successfully to the mm10 reference genome by Novoalign. Removing duplicates has been shown to be important in CLIP-Seq analysis, and many recent studies adopt slightly different methods of collapsing duplicate reads in specific datasets (23,52,65). To be general, MiClip reads with exactly the same chromosome, strand, start site and end site were defined as duplicates and collapsed to one tag. About 1.6 million unique tags were kept after removing duplicates. Then, tags overlapping by at least one nucleotide were grouped together to form CLIP clusters. Around 380,000 clusters with two or more overlapping tags were formed. According to the study by Zhang, *et al.* (45), deletion is the characteristic marker mutation for protein-mRNA interaction sites in AGO and Nova HITS-CLIP experiments. Thus, I only counted the occurrences of deletions on each genomic site in our analysis for this dataset.

To identify the enriched regions, all clusters were divided into bins of 5bp, resulting in a total of 3,525,678 bins. On average, each cluster was divided into ~9 bins. All the bins derived from the same cluster were defined as one observation sequence. A two-component-Poisson mixture model was fitted

for all the tag counts on the bin-level, and the status of being enriched or non-enriched for each bin was inferred from the first round HMM. After the first round HMM, 291,160 enriched bins were identified, which correspond to 39,471 clusters. Then, adjacent enriched bins were concatenated together and formed 41,078 enriched regions. This number is larger than 39,471, because some clusters have multi-modal peaks, which results in multiple enriched regions inside one cluster.

To identify reliable binding sites, total tag and mutant tag information was collected on a single nucleotide basis within the enriched regions, resulting in 1,441,030 bases. Then, a mixture model of a zero-inflated binomial distribution and a binomial distribution was fitted for the total tag numbers and mutant tag numbers of all 1,441,030 bases. The zero-inflated binomial distribution was used to encompass background mismatches on non-binding sites, such as random PCR and sequencing errors, and the binomial distribution was used to encompass mismatches induced by cross-linking on binding sites. The status of binding and non-binding at each base pair was then inferred by using the second round HMM, and 6,867 single-nucleotide binding sites were identified. There are 5,795 out of all 39,471 enriched CLIP clusters containing binding sites, and most of them contain only one binding site per cluster. I also randomly permuted the total tag counts and mutant tag counts data, and the *MiClip* algorithm found a total of 665 CLIP clusters with at least one reliable binding site in the permuted data. Therefore, the False Discovery Rate (FDR) is 0.11.

**(a)**




**(b)**

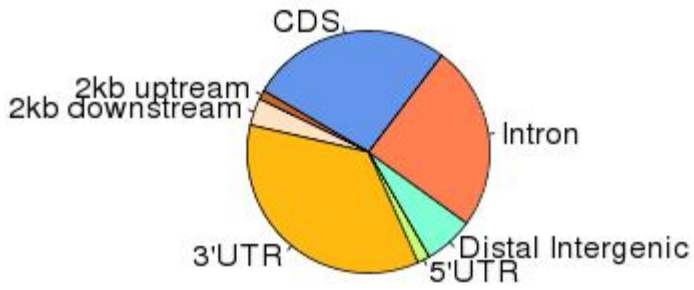



**Fig. 3.** MiClip analysis results on exemplary datasets (a) The inference results of MiClip on AGO CLIP clusters. At the bottom, "Strand" refers to the strands of clusters, "Enriched bins" shows whether a cluster contains enriched bins and "Binding sites" shows whether a cluster contains binding sites. This region covers part of the Tor1b gene. (b) The percentages of all 5,795 reliable clusters identified by *MiClip* and all reads as background control mapped to each annotation category. CDS represents coding sequence, upstream 2k represents 2kb regions upstream of the 5'UTRs, downstream 2k means 2kb regions downstream of 3'UTRs and Intergenic represents distal genomic regions not included in all the other categories.

All CLIP clusters were marked as to whether they contain enriched regions, and for clusters with enriched regions they were marked as to whether they contain at least one binding site (**Fig. 3a**). A CLIP cluster containing at least one enriched bin with binding site(s) was reported as a reliable cluster. The

orange bars show non-enriched bins, the red bars show enriched bins and the blue arrow points to the binding site. For each identified binding site, MiClip produced a probability score, which could be used to prioritize subsequent validation experiments. A site with a higher probability score means this site is a more reliable binding site. When I aligned the reliable clusters to the mouse genome and summarized the genomic locations of these clusters (**Fig. 3b**), I discovered that the largest portion of the clusters (35%) falls into 3'UTR. This is followed by coding sequences (27%) and intronic regions (25%), while 5'UTR, 5'UTR extended regions and 3'UTR extended regions each contain 1.6%, 1.2% and 3.6%, respectively, of the clusters. Within mRNAs, the clusters are highly enriched in 3'UTR (~55%). This observation is consistent with previous knowledge of miRNA regulation (49). Also the percentage of each annotation type is distinctly different from the background distribution of the annotation types of all the reads as control (**Fig. 3b**), supporting the algorithm's ability of filtering for true binding sites in the 3'UTR. Therefore our results suggest that the MiClip method is able to find reliable CLIP clusters with functional significance.

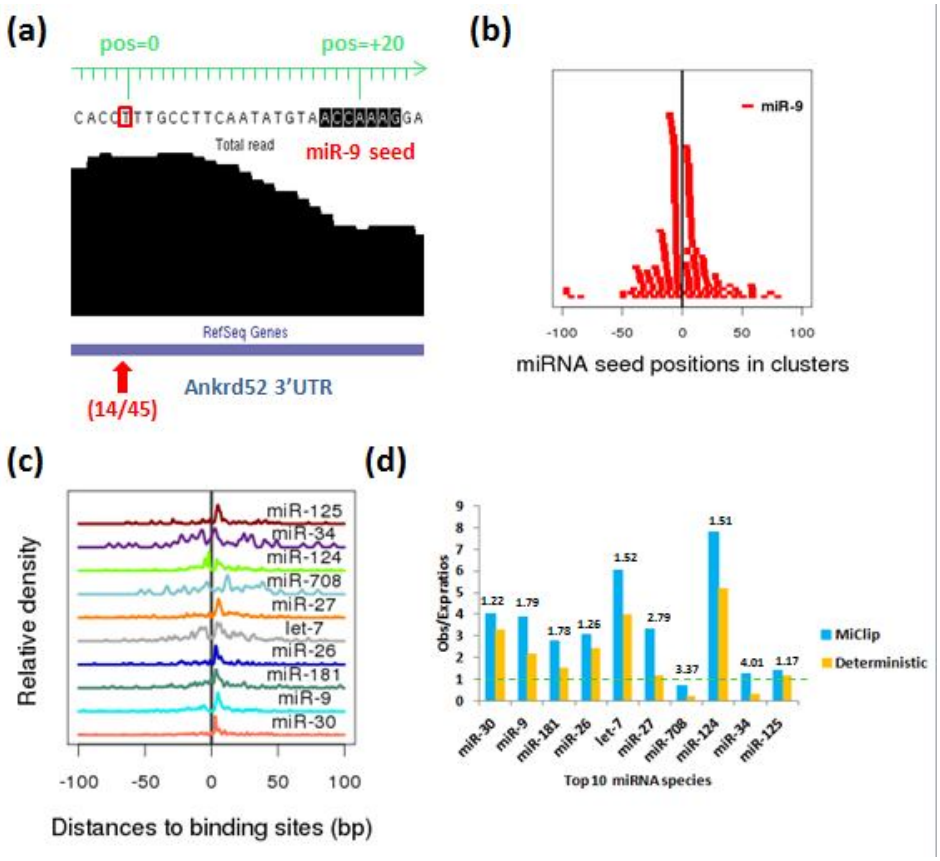



**Fig. 4.** miRNA seed motifs detected by MiClip on exemplary datasets (a) An exemplary cluster with one reliable binding site and a seed match for miR-9. The red rectangle is the binding site, which has 14 mutant tags out of 45 total tags. The miR-9 seed motif is ACCAAAG, which locates 20 bp downstream of the binding site. (b) The positions of seed matches of miR-9 (positions 2-8) within the 5,795 significant clusters. 4 out of 199 seed motifs are more than 100 bp upstream or downstream away from the binding sites, which are ignored. X-axis is the relative position with respect to the binding site within each cluster. (c) The pileup of seed matches of the top 10 miRNAs (positions 2-8) within the 5,795 significant clusters drawn as density curves. 49 out of 1557 seed motifs (all top 10 miRNAs) are more than 100 bp upstream or downstream away from the binding sites, which are ignored. (d) The signal/noise ratio of miRNA seed motif enrichment. The blue bars show enrichment ratios for MiClip and the orange bars show enrichment ratios for the *ad hoc* method. The numbers on top of these bars show the signal/noise ratio for clusters found by MiClip divided by the signal/noise ratio for clusters found by the *ad hoc* method.

To further validate the MiClip approach, I tried to correlate identified CLIP mRNA clusters with miRNA seed matches. A recent X-ray study suggests that the AGO proteins function by forming ternary structures with miRNA and mRNA (81). By mapping short reads from the miRNA library, Chi, *et al.* were able to identify the most enriched miRNA species and rank the miRNA species by their abundance, with the most abundant miRNA being miR-30. I scanned for the 7-mer (position 2-8) seed motif matches for the top ten most enriched miRNAs within the 5,795 clusters. **Fig. 4a** shows an exemplary significant CLIP cluster with a motif match for miR-9. The red arrow points to the identified binding site, which has 14 deletions, and the probability of this site being a true binding site is > 0.999. A miR-9 seed motif match occurs at 20bp away downstream of the binding site. I calculated the relative distances to the centers of the miR-9 seed motifs from the binding sites within all reliable clusters containing miR-9 motifs. For the example shown in **Fig. 4a**, this distance is +20. Then, I plotted the positions of conserved miR-9 seed matches relative to binding sites according to the calculated distances (**Fig. 4b**). I found that most of the miR-9 seed matches are within -50 to +50 bp of binding sites and form a sharp peak around position 0, which are the binding sites. Also, I plotted the distances relative to the binding sites for all the top 10 miRNAs in **Fig. 4c**, and in this figure I plotted the pileup of seed motifs as density curves, for the sake of clarity. The seed matches for all top 10 miRNA motifs also tend to form very sharp peaks towards position 0, confirming the validity of the identified clusters and binding sites. Interestingly, the vertical line at pos=0 does not cross many motif matches in **Fig. 4b** and **Fig. 4c**, indicating the binding sites are not located directly within seed motifs. This is because in the AGO HITS-CLIP experiment, miRNA and mRNA were paired at the seed motif, and thus partially protected from cross-linking. Similar phenomena were observed in previous work by another group (7). Another interesting point to note is that seed motifs of some miRNAs, like miR-9, are predominantly downstream of binding sites, while motifs of other miRNAs, like let-7, locate both upstream and downstream of binding sites in large numbers. Overall, these results confirm the validity of the identified binding sites by MiClip, and show that using binding sites rather than peak summits, as in the original study, is more informative for finding protein binding locations.

| Rank | miRNA | Seed motif | # clusters with seed match | Percentage of clusters with seed match | Percentage of background sequences with seed match | Signal/Noise |
|---|---|---|---|---|---|---|
| 1 | miR-30 | TGTTTAC | 162 | 2.79% | 0.69% | 4.04 |
| 2 | miR-9 | ACCAAAG | 199 | 3.43% | 0.88% | 3.89 |
| 3 | miR-181 | TGAATGT | 165 | 2.84% | 1.03% | 2.75 |
| 4 | miR-26 | TACTTGA | 102 | 1.76% | 0.57% | 3.08 |
| 5 | let-7 | CTACCTC | 225 | 3.88% | 0.64% | 6.06 |
| 6 | miR-27 | ACTGTGA | 189 | 3.26% | 0.98% | 3.32 |
| 7 | miR-708 | AGCTCCT | 46 | 0.79% | 1.16% | 0.68 |
| 8 | miR-124 | GTGCCTT | 318 | 5.48% | 0.7% | 7.82 |
| 9 | miR-34 | CACTGCC | 66 | 1.13% | 0.90% | 1.25 |
| 10 | miR-125 | CTCAGGG | 85 | 1.46% | 1.05% | 1.39 |

Table. 2 The enrichment of the top 10 miRNA seed sequences within the 5,795 clusters.

Moreover, I calculated the percentages of the 5,795 clusters with seed motif matches for the top 10 miRNAs, as well as the percentages in 40,000 background sequences, for computing the signal/noise ratios for the top 10 miRNAs, as in the original publication (41) and others (82) in the field of nucleotide motif discovery. I found that 9 out of 10 miRNA seed matches are enriched (**Table 2**). In the original publication, the authors also provided enrichment ratios for the top 10 miRNAs from their list of significant clusters, but the enrichment ratios are not as high as calculated from the results of MiClip (**Fig. 4d**). For each miRNA, signal/noise ratio by the MiClip method over the *ad hoc* method ranges from 1.17 for miR-125 to 4.01 for miR-34 (**Fig. 4d**). In conclusion, MiClip is able to find AGO binding sites and reliable CLIP clusters with better biological significance than the *ad hoc* method used in the original work.

### *2.1.4 Discussion*

In this study, I presented the MiClip approach for identifying reliable protein-RNA binding sites and clusters in CLIP-Seq datasets. MiClip is a model-based approach that can identify high-confidence binding sites using probability scores. Different from crosslinking-induced mutation sites (CIMS) analysis (45) that only looks at mutation rates, MiClip approach analyzes both tag counts and numbers of mutations simultaneously to improve detection power. It employs two HMMs, one for searching enriched regions at 5bp resolution, and the other for searching binding sites at single base-pair resolution. The two-stage approach handles large sequencing data efficiently, while identifying binding sites at high-resolution. One potential of MiClip is that it requires only 2 parameters to control the model fitting. The first one is the cutoff for truncating the counts of bins with extremely large count data (for the first HMM) and the second one is the parameter c in parameter estimation (for the second HMM). In comparison, *PARalyzer* requires more than 8 parameters, and *wavClusteR* requires 3 parameters to control model fitting. Some of these parameters are not intuitive and informative, so it could be difficult to decide the best values for these parameters. Usually the more parameters there are, the more difficult and confusing it will be. In the *MiClip* software, I provided default values for the 2 parameters, based on our experience with several CLIP-Seq datasets.

Choosing the characteristic marker mutation is important when analyzing CLIP-Seq datasets. The marker mutation type for PAR-CLIP dataset is easy to determine, because it depends solely upon the type of analog used in the experiment. However, for HITS-CLIP experiments, choosing the right type of mutation as the marker for binding events could be difficult. According to Zhang, *et al*. (45), deletion is the marker mutation for AGO and Nova HITS-CLIP experiments, but this may not hold true for other proteins. In fact, our unpublished data shows that for a certain human protein, I might need to include all types of mutations as marker mutations, because a well-defined target transcript bound by this protein contains large and comparable amounts of deletions, insertions and substitutions. In such cases, it is a

good idea to run *MiClip* multiple times, each time specifying a different marker mutation or mutation combination, in order to compare the results and see which setting leads to the most reasonable results.

The MiClip algorithm takes full advantage of the unique properties of HMM, which is the core of the MiClip algorithm. First, HMM is a powerful method to identify hidden states with spatial dependencies between neighboring observations. CLIP clusters formed by overlapping short tags should have inherent spatial dependency features. For example, bins with tag intensity of 5 can be inferred either as enriched or non-enriched with similar probabilities. But bins with tag intensity of 5 should have higher probability of being inferred as enriched when their neighboring bins have bigger tag counts, which is how spatial dependency plays a role in statistical inference. For inference of enrichment, I implemented the Poisson model. For future studies, it would be interesting to investigate whether a Negative Binomial mixture model could help improve the inference accuracy. Secondly, protein binding events will lead to sequencing tag pile up, as well as sequencing mismatches in a random process, which can be naturally reflected as an emission function. In the AGO HITS-CLIP data, genomic sites with mutation ratio around 0.18 have similar probabilities of being inferred as binding sites or non-binding sites. Here, using binomial distributions to model the number of mutations, while considering the total tag counts, is better than looking at the mutation rate alone. As a simplified example, let us assume the mutation probabilities are 0.2 and 0.1 for a binding site and a non-binding site, respectively. The probability of observing 3 mutations from 10 tags is 0.201 and 0.057 at a binding and non-binding site, respectively, while the probability of observing 30 mutations from 100 tags is 0.0052 and 1.84e-8 at a binding and non-binding site, respectively. As a result, although the mutation rates are the same (equals 0.3), the probability of being a binding site for a site with 10 tag counts is different from a site with 100 tag counts. Overall, MiClip's analytic power is derived from the appropriate use of HMM. Besides, this model-based approach is able to provide a probability value for each identified binding site, which helps researchers plan subsequent experiments.

The *MiClip* R package is designed to be flexible in analyzing both HITS-CLIP and PAR-CLIP datasets. In MiClip, the marker mutations introduced in HITS-CLIP and PAR-CLIP experiments could be defined as deletion, insertion, substitution, or some combination of these mutation types. For example, if the characteristic mutations induced for a specific protein are both deletions and insertions, the MiClip method can take both types of mutations into consideration at the same time. I also added an option in the *MiClip* package to incorporate background sequencing data for normalization purpose, if such data are available. Conducting background profiling of gene expression as a control for CLIP-Seq experiments has not become a standard procedure yet, but could be very helpful for improving the accuracy of the identification of RBP targeting sties. The package can handle both single-end and paired-end CLIP-Seq data. Users could run this package on UNIX, Mac OS or PC machines. In addition, the *MiClip* package is highly efficient. It takes 45 minutes for *MiClip* to analyze the FUS PAR-CLIP dataset with sequencing depth ~4.2 million reads, compared with *wavClusteR* which takes 2 hours to analyze the same dataset. For the AGO HITS-CLIP dataset with ~26 million reads, *MiClip* only takes 60 minutes to process. Although it takes *PARalyzer* about 30 minutes to process the FUS dataset, it can only accept Bowtie-format output files from the Bowtie aligner, which severely limits its application.

## 2.2 Differential analysis of RNA-RBP binding strength in two conditions

### *2.2.1 Background and rationale*

Although a few bioinformatics tools like PARalyzer, CLIPZ, wavClusteR, and miRTarCLIP (1,7,10,23) have been developed to analyze single CLIP-Seq dataset, the quantitative comparison of multiple CLIP-Seq datasets has only recently gained interest in the field (83-85). Piranha (8) has been developed for CLIP-Seq and RIP-Seq (86) data analysis, and also provide a procedure for comparative analysis. However, the comparative analysis procedure in the Piranha is relatively *ad hoc*, and does not utilize the spatial dependency among neighboring genomic locations, which is an important characteristic in creating differential binding profiles. A straightforward way to compare RNA-RBP interaction profiles

across conditions is to analyze individual CLIP-Seq data separately to identify the peaks (or binding sites) for each condition and then use coordinate overlapping or similar approaches to obtain common and differential binding sites. However, this *ad hoc* approach compares the results qualitatively but not quantitatively. For example, if a region is bound by an RBP under two distinct conditions (*e.g.* wild type *vs.* knockout) with both significant but different binding intensities, the *ad hoc* approach will not be able to detect this region as a differential binding site. In addition, this *ad hoc* approach is over-sensitive to the cutoffs used for analyzing individual data, and has been shown to underestimate the similarity of two samples when applied to the analysis of multiple ChIP-Seq experiments (87,88). Therefore, a computational approach that can compare different CLIP-Seq datasets simultaneously and quantitatively is needed.

The main challenge to comparing genome-level sequencing profiles across conditions quantitatively is that the next-generation sequencing data usually contains relatively low signal-to-noise ratios (89,90). Differences in background levels further complicated the analysis. To address these problems, several computational approaches were developed for comparative ChIP-Seq analysis, including ChIPDiff (91), ChIPnorm (92), MAnorm(93), and dPCA(94). These computational approaches have greatly facilitated the understanding of dynamic changes of protein-DNA interactions across conditions. However, these computational approaches cannot be directly applied to CLIP-Seq data in order to identify differential RNA-protein interactions, due to some inherent differences between ChIP-Seq and CLIP-Seq data. First of all, CLIP-Seq data is strand-specific, while the tools designed for ChIP-Seq experiments do not consider strands of peaks. Second, CLIP-Seq experiments usually induce additional characteristic mutations in high-throughput sequencing reads, but the mutation information in the raw sequencing data is simply discarded in the bioinformatics software designed for ChIP-Seq data analysis. Third, CLIP-Seq reads are usually short, and the reads are not shifted or extended when counting tag intensities, but shifting or extension of reads is a necessary step in ChIP-Seq analysis (77). Fourth, CLIP-Seq requires a much higher resolution (close to single nucleotide) in detection of RBP-binding sites, but ChIP-Seq

software usually work on a much lower level of resolution. For example, ChIPDiff works is limited to 1kb and ChIPnorm typically on a resolution of a few hundred base pairs. In addition, the method proposed by Bardet *et al.* (87) is not bundled as a portable software and takes about two days to finish. Therefore, I have developed the dCLIP software for detecting differential binding regions in comparing two CLIP-Seq experiments.

dCLIP is a two-stage computational approach for comparative CLIP-Seq analysis. As the first stage, a modified MA-plot approach was designed specifically to normalize CLIP-Seq data across datasets in order to obtain high resolution results. As the second stage, a hidden Markov model (HMM) was developed to detect common/different RBP-binding regions across conditions. The HMM has a great advantage in modeling the dependency among adjacent genomic locations, which leads to improved performance in identifying differential binding sites. Here, I show that dCLIP can accurately identify RBP differential binding sites through the comparative analysis of four differential CLIP-Seq datasets.

### *2.2.2 Materials and methods*

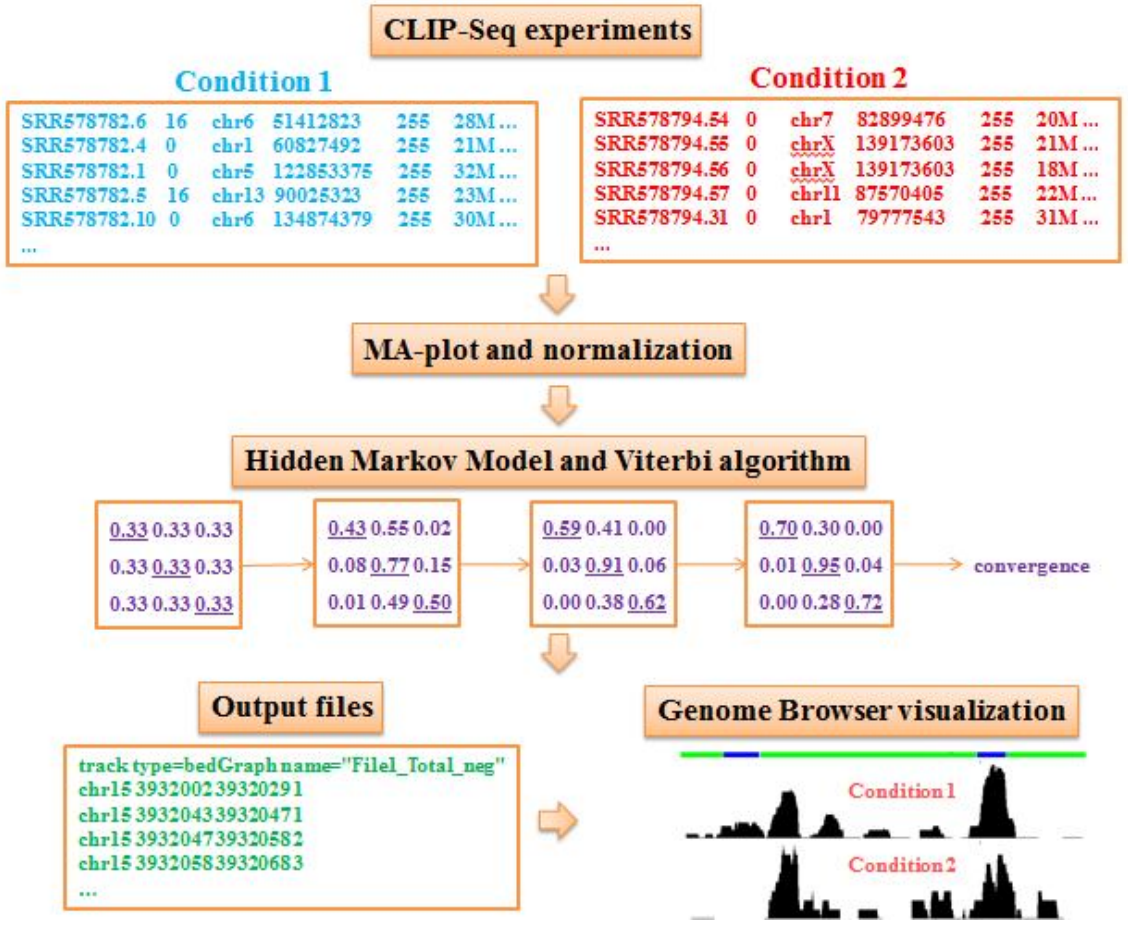



**Fig. 5** Schematic representation of the dCLIP pipeline. A summary of the major steps of dCLIP is provided as a flow chart. The format of the input and output files is also provided in the flow chart.

2.2.2.1 Data preprocessing

An overview of the dCLIP pipeline is shown in **Fig. 5**. Data preprocessing is conducted in a strand-specific manner. For HITS-CLIP and PAR-CLIP, duplicate reads with the same mapping coordinates and the same strand are first collapsed to unique tags. The characteristic mutations are collected on all tags and written to separate output files. CLIP clusters are defined as contiguous regions of nonzero coverage in either condition and are identified by overlapping CLIP tags from both conditions. The tags that

comprise each cluster retain their original condition identity. As a high resolution is needed for CLIP-Seq analysis, dCLIP divides the clusters into bins of small length (the default is 5 bp) and calculates tag counts in each bin for both conditions. More specifically, the number of tags covering each base is calculated and the counts on all bases in each bin are summed to be the tag intensity count for that location. Therefore, the i-th bin in the j-th cluster has a pair of data points $x_i^{(j)} = (x_{i,1}^{(j)}, x_{i,2}^{(j)})$, where $x_{i,1}^{(j)}$ is the tag intensity count for the first condition, $x_{i,2}^{(j)}$ was the tag intensity count for the second condition.

iCLIP dataset preprocessing mainly follows that of Konig *et al.* (95), with minor modifications. Sequencing reads with the same random barcode represent PCR duplicates. Duplicates were removed and barcodes were trimmed from the unique tags before mapping to the reference genome. A helper script, remove_barcode.pl, is provided in the dCLIP software to help users remove barcodes from Fastq sequencing files. After mapping, the first nucleotide upstream of each mapped cDNA, defined as the "cross-link nucleotide," is expanded by a few nucleotides (specified by the users) in both downstream and upstream directions from its location, namely adding 1 to the tag counts on all bases in this short window. Therefore, the total tag count on each base is calculated as the sum of expanded cDNA counts covering that base and the mutant tag count will always be zero. Similarly, cDNA counts in both experimental conditions are summarized on the bin-level in regions of non-zero coverage.

2.2.2.2 Data normalization

Due to the different sequencing depths of the two CLIP-Seq samples, a normalization step is essential for an unbiased comparison. However, the common method of normalizing by total number of tags in high-throughput sequencing studies could be problematic, because of different signal-to-noise ratios for different samples. I implemented the MA-plot normalization method, which was originally designed for normalizing microarray data (96), and later applied to ChIP-Seq analysis (93). When applying MA-plot method to normalizing microarray data, generally the expression value for each gene body is used as a

unit in normalization. When applying MA-plot method to normalizing multiple ChIP-Seq data as in (93), read counts in the 1000bp windows centered on the summits of peaks are used as a data unit in normalization. However, in dCLIP, I modified the MA-plot method to normalize count data on the bin level, as high resolution is required in CLIP-Seq data analysis. The $(M_i^{(j)}, A_i^{(j)})$ value of each bin is then defined as

$$M_i^{(j)} = \ln(x_{i,1}^{(j)} + c) - \ln(x_{i,2}^{(j)} + c)$$
$$A_i^{(j)} = \ln(x_{i,1}^{(j)} + c) + \ln(x_{i,2}^{(j)} + c)$$

A small number c is added to each count value to avoid logarithm of 0 count. I assumed that both conditions share a large number of common binding regions with similar binding strength. Therefore, a linear regression line $M = a + b \times A$ is fitted to bins whose $x_{i,1}^{(j)}$ and $x_{i,2}^{(j)}$ values are both larger than a user-defined cutoff. As common binding sites should have similar binding strengths, the parameters derived from the regression model should capture the true scaling relationship between the two samples. This scaling relationship is extrapolated to the whole dataset, by subtracting a fitted M value from the linear regression model from the raw M value of every bin in all clusters. The adjusted M value is used in the following data analysis.

2.2.2.3 Hidden Markov Model (HMM)

HMM is a statistical Markov model in which the system being modeled is assumed to have spatial dependency between neighboring data units. RBP-RNA interactions involve a short stretch of RNA, which could span up to a few bins (41). This ensures the strong auto-correlation of tag counts in neighboring bins, which can be modeled by HMM. Therefore, I applied HMM to identify common/differential binding regions from the adjusted M values. As these adjusted M values come from many individual CLIP clusters, the HMM model has multiple observation sequences. During the

statistical inference, all observation sequences share the same transition matrix and the same emission function.

The HMM has three possible states for each i-th bin in the j-th cluster:

$$\begin{cases} I_i^{(\mathrm{j})} = 0 & \text{stronger binding in condition 1} \\ I_i^{(\mathrm{j})} = 1 & \text{non-differential binding site} \\ I_i^{(\mathrm{j})} = 2 & \text{stronger binding in condition 2} \end{cases}$$

Accordingly, the transition matrix Π is a 3×3 matrix, whose element $\pi_{r,s}$ ias the transition probability $\Pr(I_i^{(\mathrm{j})} = s \mid I_{i-1}^{(\mathrm{j})} = r)$. Given state $I_i^{(j)}$, the adjusted M values are fitted by a three-component normal mixture model. Because the common peaks that are determined by similar mechanism in both conditions are normalized towards the same binding strength, the middle normal component is assigned a mean of 0. To avoid unreasonable assignment of bins to hidden states when the adjusted M values are extremely large or small, the three normal components are all assumed to have the same variance. Also in order to simplify the problem, the means of first and third normal components are assumed to have the same absolute value but different sign.

To estimate the parameters for the HMM, I adopted an empirical-based method by fitting the adjusted M values to a three-component Gaussian mixture model.

$$f(M_i^{(j)} \mid \sigma, \mu, \mathrm{p}) = \mathrm{p} \times \frac{1}{\sqrt{2\pi}\sigma} \times e^{\frac{(\mathrm{M}_i^{(j)}+\mu)^2}{2\sigma^2}} + (1-2\,\mathrm{p}) \times \frac{1}{\sqrt{2\pi}\sigma} \times e^{\frac{(\mathrm{M}_i^{(j)})^2}{2\sigma^2}} + p \times \frac{1}{\sqrt{2\pi}\sigma} \times e^{\frac{(\mathrm{M}_i^{(j)}-\mu)^2}{2\sigma^2}}.$$

Since I assume that most sites would not show changes in their binding between conditions, the second component should dominate the mixture distribution. Then the first and the third components could be treated as outliers if I solely focus on the second component. Then I apply a MAD (median absolute

deviation) method (97) to robustly estimate the standard deviation to estimate $\sigma$ , by equating $\hat{\sigma}$=median(|M-median(M)|)$\times$1.4826 .

The other parameters $p$ and $\mu$ are estimated by a recombinant method that combines method of moments estimator and maximum likelihood estimator (98). Simply speaking, the second moment and sample second moment of the mixture distribution are given by

$$\mu_2 = p\times(\mu^2+\hat{\sigma}^2)+(1-2\,\mathrm{p})\times\hat{\sigma}^2+p\times(\mu^2+\hat{\sigma}^2)$$
$$\hat{\mu}_2 = \frac{\sum(M_i^{(j)})^2}{n}$$

By equating the above two formulas, I could get a constraining relationship between $p$ and $\mu$ . The likelihood function was written as

$$L(\mathrm{p},\mu\,|\,M_i^{(j)},\hat{\sigma})$$
$$=\prod_{i,j} f(M_i^{(j)}\,|\,\hat{\sigma},\mu,\mathrm{p})$$
$$=\prod_{i,j}(\mathrm{p}\times\frac{1}{\sqrt{2\pi}\hat{\sigma}}\times e^{\frac{(\mathrm{M}_i^{(j)}+\mu)^2}{2\hat{\sigma}^2}}+(1-2\,\mathrm{p})\times\frac{1}{\sqrt{2\pi}\hat{\sigma}}\times e^{\frac{(\mathrm{M}_i^{(j)})^2}{2\hat{\sigma}^2}}+p\times\frac{1}{\sqrt{2\pi}\hat{\sigma}}\times e^{\frac{(\mathrm{M}_i^{(j)}-\mu)^2}{2\hat{\sigma}^2}})$$

So, using grid approximation, I obtain a pair of $\hat{p}$ and $\hat{\mu}$ which maximizes the likelihood function and also maintains the constraint at the same time.

The emission probabilities are calculated from the fitted model and fixed for each bin in different states before the iterations of HMM start. To find the chain of hidden states most likely given the observations and the model, a Viterbi dynamic-programming algorithm is employed to infer the hidden state $I_i^{(j)}$.

2.2.2.4 Data visualization

Finally, adjacent bins inferred to be in the same state are concatenated into continuous regions. A BED file is then generated to be uploaded to the UCSC Genome Browser, each entry of which is one continuous region in the same state. In addition, a TXT file is also generated that describe in more detail the inference results of each bin. Eight bedGraph files are generated that store the total or mutant tag counts for both conditions and both strands. These files can also be directly uploaded to the UCSC Genome Browser for visualization.

2.2.2.5 Implementation

The dCLIP software was implemented in the Perl programming language and Perl (version >= 5.16) together with two Perl modules PDL and PDL::Stats are needed to run the program. The implementation is supported on all major operation platforms.

The dCLIP software inputs SAM format alignment files of the two conditions to be compared. The SAM format files can be in single-end mode or paired-end mode. The users can specify parameters such as bin size, minimal number of tags in a cluster, the number of nucleotides to expand for cDNA counts (iCLIP), the type of characteristic mutations to be profiled and the stop conditions for the HMM.

### *2.2.3 Analysis results*

To show the application of dCLIP in a real dataset, I applied the dCLIP software to a PAR-CLIP dataset where the RBP under investigation is FMR1 (29). The FMR1 RBP family comprises three members, FMR1, FXR1, and FXR2. FMR1 encodes for many isoforms, of which isoform 7 is predominantly expressed (99). The authors identified two major binding motifs of FMR1, ACTT/ACTG and AGGA/TGGA. The authors generated a recombinant FMR1 isoform 7 protein with a point mutation I304N in the KH2 domain. Through electromobility shift assays (EMSAs) and PAR-CLIP experiments conducted with the wild-type and I304N proteins, the authors found the KH2 domain to be specific for

binding to the ACTT/ACTG motif. Therefore, diminished binding to the ACTT/ACTG motif, rather than the AGGA/TGGA motif, should be the primary effect of the point mutation.

I downloaded the raw sequencing files from GSE39686. Adapters were trimmed and the sequencing reads were aligned to the hg19 genome using Bowtie (100). Then I analyzed the mapping files with the dCLIP software. dCLIP found a total of 9,859 FMR1 isoform 7 binding sites that have stronger binding strength in the wild-type than in the I304N mutant condition and have at least an average tag intensity of 3 in the wild-type condition. I showed one such binding site in **Fig. 6a**. This binding site locates in the 3'UTR of the Smad4 gene. The blue bar marks the binding region that has reduced binding upon mutation. Both the total tag counts and T->C mutation counts are shown.

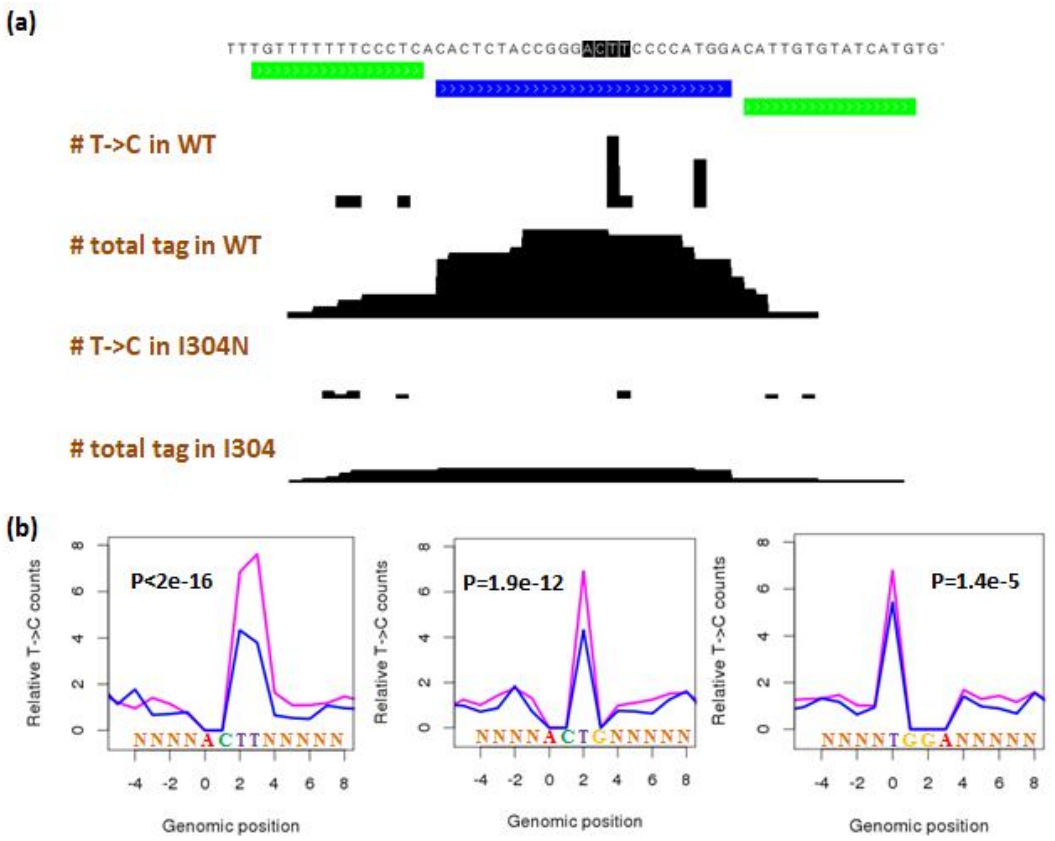



**Fig. 6** The analysis of the FMR1 dataset by dCLIP. (a) An example of FMR1 binding site with stronger binding in the wild-type condition than the I304N condition. The total tag counts and T->C mutant tag counts are shown. Green bars indicate common binding regions, and blue bars indicate regions with stronger binding in the wild-type than the I304N condition. The peak heights are scaled proportionally to the total sequencing depths of the two samples. (b) The relative counts of T->C mutations on top of all ACTT/ACTG and TGGA motifs found within the 9,859 binding sites. The T->C mutation counts on the T bases in these motifs are divided by the total T->C counts in a 30 bp window as background distribution. As the sequences surrounding these motifs vary and for each base outside these motifs only a fraction of the 9,859 binding sites have T base, they are all marked as N. The P values shown are for testing the differences in the proportions of T->C counts on top of each motif out of the total T->C counts in the 30-bp window between the wild-type condition (pink line) and I304N condition (blue line).

I further calculated the number of T->C mutations that occur on top of all ACTT, ACTG and TGGA motifs found within those 9,859 binding sites in both the wild-type and I304N condition (**Fig. 6b**). The T->C mutation counts on the T bases in these motifs are divided by the total T->C counts in a 30 bp window as the background distribution. As the AGGA motif does not have a T base, there will not be any T->C mutation on top of this motif and this motif was not included in this analysis. The normalized number of T->C mutations in the I304N condition was smaller than the number of T->C mutations in the wild type condition for the ACTT/ACTG motif as well as the TGGA motif, consistent with these sites having weaker binding in the I304N condition. The extent by which the relative T->C mutation counts decreased in the I304N condition was much more significant for the ACTT/ACTG motif (Pval<2e-16 for ACTT, Pval=1.9e-12 for ACTG) than the TGGA motif (Pval=1.4e-5). This is expected because the I304N point mutation locates in the KH2 domain responsible for binding to the ACTT/ACTG motif. As ACTT/ACTG and TGGA/AGGA motif always occur in adjacent or nearby regions on the genomic sequence, a loss of binding affinity to the ACTT/ACTG motifs by the I304N mutation should lead to a secondary, weaker effect on the binding of the protein to neighboring TGGA/AGGA motifs. Overall, the analysis of this FMR1 PAR-CLIP dataset shows that dCLIP also performs well on PAR-CLIP datasets.

### *2.2.3 Discussion*

The two-stage procedure implemented in dCLIP includes an MA normalization step and a hidden Markov model (HMM) to identify differential/common binding sites. The MA normalization is a critical step to make the CLIP-Seq data comparable across conditions. The straightforward rescaling by the total number of reads across samples is not appropriate for comparative CLIP-Seq analysis because the signal/noise ratio usually varies across different conditions. The modified MA plot normalization method in dCLIP not only addresses the issue of different signal/noise levels effectively, but also works on much smaller units than those used for microarray and ChIP-Seq data analysis, allowing dCLIP to detect binding sites of higher resolution required for CLIP-Seq data analysis. To reduce potential bias and conduct rigorous comparison across different conditions, I recommend adopting the same experimental

and bioinformatics procedures, such as RNase digestion, high-throughput sequencing and alignment, for both conditions.

The HMM plays a key role in identifying differential/common binding sites of two CLIP-Seq samples in the dCLIP software. HMM can increase signal/noise ratios for sequencing data analysis, because it takes into account the correlation between consecutive bins. This is particularly important for CLIP-Seq data, because of small bin size and high correlations between consecutive bins. The HMM in dCLIP defined a common binding state and two differential binding states. One thing to note for the three-state HMM is that the identified differential binding sites, for example (enriched, non-enriched), may actually only have a small tag enrichment in condition 1, and an even smaller tag enrichment in condition 2. Therefore, the differential binding sites need to be ranked and screened as such sites may not be of real interest to biologists. The analysis of the miR-155/AGO HITS-CLIP dataset, for example, sets a cutoff of average tag intensity of 30 in the wild-type condition.

The dCLIP software was benchmarked against the Piranha software (data not shown). Piranha incorporates covariates which could represent transcript abundance, count data in the second condition or positional mutation information. However, the covariate is incorporated in the statistical model in the exactly same way no matter which type of data it actually represents. This design enables Piranha to be easily applied to a wide variety of CLIP-Seq data analysis scenarios. However, this one-for-all method also harms the detection power of RBP binding regions of interest in each specific scenario, as different data types have their unique properties and should be treated differently. The dCLIP method, on the other hand, is specialized in comparing two CLIP-Seq experiments and was shown to perform better than Piranha in identifying differential binding sites. Therefore dCLIP should be a better choice when the users are interested in identifying differential or common RBP-binding sites.

In summary, I present a new computational approach, dCLIP, for the comparative analysis of CLIP-Seq data. dCLIP was implemented as an easy-to-use command line tool in the Perl programming

language. The dCLIP software is able to handle HITS-CLIP, PAR-CLIP and iCLIP datasets, and can take single-end or paired-end sequencing files as input. The dCLIP software is strand-sensitive and is able to detect differential binding sites at almost single-base resolution. It also correctly keeps all of the characteristic mutation information for later analysis. Real data analysis shows that dCLIP can accurately identify differential binding regions of RBPs and outperforms another CLIP analysis program, Piranha (8). I anticipate that the dCLIP software will become a helpful tool for biologists and bioinformaticians for comparative CLIP-Seq data analysis.

# CHAPTER THREE – REAL DATA ANALYSIS

In two collaborating projects, I applied the aforementioned developed methods to analyze the HITS-CLIP and PAR-CLIP data that my collaborators generated. In the first project, I identified the AGO2 binding sites in both nuclear and cytoplasmic compartments, which revealed interesting differences in AGO2 binding preferences in these two parts of cells. In the second project, I identified ORF57 binding sites in both human and virus genomes. This work greatly expanded the array of genes targeted by ORF57, while previously only a few targets are known to be bound by ORF57. This chapter is devoted to a description of the results and conclusions I generated from these efforts.

## 3.1 Identify nuclear AGO2 binding sites using PAR-CLIP

### *3.1.1 Background*

RNA sequencing (RNA-Seq) has become a widely used tool for investigating gene expression (101). Millions of sequence "reads" in combination with bioinformatics analysis and experimental validation can provide new insights into fundamental cellular processes. The usefulness of RNA-Seq, however, is often limited by the amount of input RNA needed to yield meaningful data. Another limitation comes from the lengths of RNA species in certain studies such as sequencing of small RNAs and RNA fragments (102). Our collaborators exploited the principle that *intra*molecular reactions are more favorable than *inter*molecular reactions by developing a sequencing methodology that uses RNA self-circularization (RC-Seq). Using this method, I obtained nuclear RNA samples after UV-crosslinking to protein (CLIP-Seq), while the use of standard methods did not yield data. Therefore, I were able to compare AGO2 binding site differences in cellular fraction and nuclear fraction using PAR-CLIP.

### *3.1.2 Results*

I sequenced the nuclear Ago2 PAR-CLIP library derived from our RC-Seq protocols and analyzed the results. The aligned sequencing data showed a dominant T to C mutation, consistent with incorporation of 4-SU and successful adaptation of the PAR-CLIP protocol to RC-Seq (**Fig. 7a**) (103). I found that 10-12 % of uniquely aligned reads had T to C mutations, a typical rate observed during PAR-CLIP. I obtained an average of 40 million raw reads and 9 million uniquely aligned reads from duplicate determinations. Using MiClip, a program weighing the T to C mutations and optimized to search for significant RNA clusters from CLIP-Seq datasets, I identified 7839 clusters from two biological replicates. The sequencing data were reproducible between duplicate experiments with a strong concordance between grouped aligned reads (**Fig. 7b**). Of 7839 total clusters, 7187 appeared in both replicates and the coefficient of determination was $R^2$ = 0.94. Subsequent genomic annotation shows that more than 50% of clusters are localized within intronic regions (**Fig. 7c**). This data is in a sharp contrast to data from three cytoplasmic samples that showed most clusters within the 3'-untranslated region of mRNAs (**Fig. 7d**), likely reflecting differing roles for cytoplasmic and nuclear RNAi (11,12).

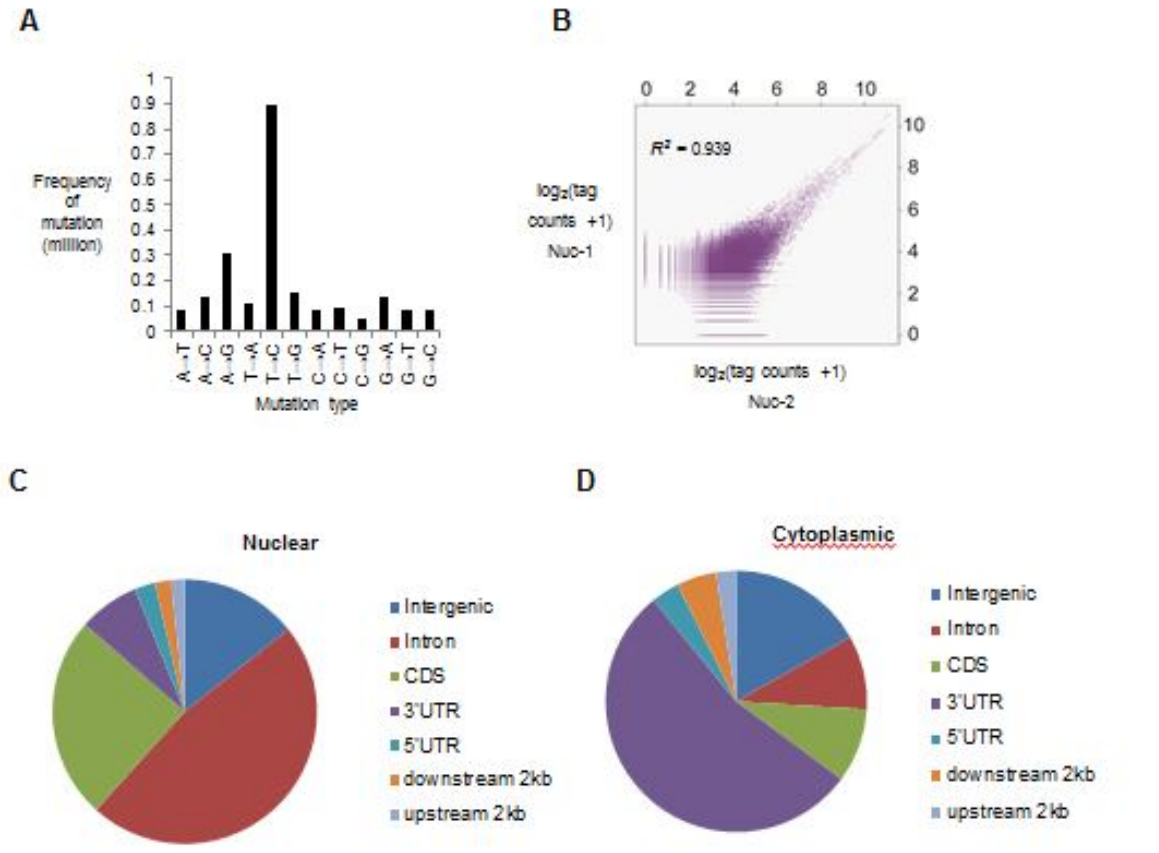



**Fig. 7.** Preparation of RC-Seq libraries and analysis of PAR-CLIP derived RNAs. (**a**) Detection of mutations in RC-Seq library derived from Ago2-PAR-CLIP, showing the expected T to C transition mutation from the incorporation of 4-SU. (**b**) Scatter plot showing the reproducibility of RC-Seq data from 2 biological nuclear replicates (Nuc-1 and Nuc-2). Coefficient of determination, $R^2$ = 0.939. The unique tags from the each replicate were clustered and the clusters were binned by a length of 10bp. The natural logarithm transformed count data (plus 1) in each bin of a cluster are plotted to show the correlation between the replicates. (**c**) Genetic annotation of Ago2 significant binding clusters identified in two nuclear samples (Nuc-1 and Nuc-2). (**d**) Genetic annotation of Ago2 significant binding clusters identified in three cytoplasmic samples. MiCLIP program was used in searching the significant clusters in aligned datasets.

To further evaluate the quality of our data, it would be useful to examine the overlap between the nuclear Ago2 binding sites detected by RC-Seq and those determined by other methods using Ago2 PAR-CLIP. I found six accessible Ago2 PAR-CLIP datasets, RNA ligation and whole cells-based (39,103-107). The hit lists vary, with the number of significant binding sites ranging from 2,000 to 44,000. All of these

six datasets were generated from using whole cell lysate for immunoprecipitation, making an exact comparison to our nuclear data difficult. In addition, the studies used different anti-Ago2 antibodies and cell lines. Notwithstanding these differences, I carried out a comparison to determine overlap in potential binding sites between these previous studies and our own, and the previous studies with one another. I found that ≤10% of the clusters overlapped between the individual whole cell datasets and our nuclear data. The overlapping percentage among any two of the six datasets ranged from 15% to 45%. When comparing our cytoplasmic Ago2 datasets to those six published datasets from whole cell lysate, similar overlapping cluster percentage (15%-50%) was obtained. These comparisons suggest that our RC-Seq method identifies many of the same cytoplasmic significant clusters as had been observed previously, but that the identities of cytoplasmic and nuclear clusters are substantially different.

### *3.1.3 Discussion*

RNA sequencing (RNA-Seq) is a powerful tool for analyzing the identity of cellular RNAs but is often limited by the amount of material available for analysis. Here our collaborators report a method for obtaining strand-specific small RNA libraries for RNA sequencing that requires picograms of RNA. Using the new method, our collaborators generated CLIP-Seq libraries from nuclear RNA that had been UV-crosslinked and immunoprecipitated with anti-Argonaute 2 (Ago2) antibody. Computational protocols were developed to enable analysis of data derived from circular RNA and I observe substantial differences between recognition by Ago2 of RNA species in the nucleus relative to the cytoplasm. This RNA self-circularization approach to RNA sequencing (RC-Seq) allows data to be obtained using small amounts of input RNA that cannot be sequenced by standard methods.

## 3.2 Identify ORF57 binding sites using HITS-CLIP

### *3.2.1 Background*

Kaposi's sarcoma-associated herpesvirus (KSHV; HHV-8) is a human gamma herpesvirus and the etiological agent for Kaposi's sarcoma (KS), primary effusion lymphoma (PEL), KSHV-inflammatory cytokine syndrome (KICS), and some cases of multicentric Castleman's disease (MCD) (108,109). One KSHV factor critical for viral gene expression is the ORF57 protein (Mta) (110,111). While no host homologs are known, every herpesvirus encodes a homolog of ORF57 and each is essential for virus replication (112,113). ORF57 is multifunctional, but most of its known activities are associated with posttranscriptional regulation of gene expression. ORF57 stabilizes viral RNAs in the cell nucleus, independent of its reported ability to export intronless RNAs. This function was first suggested by the observation that the levels of the polyadenylated nuclear (PAN) RNA is up-regulated by co-expression of ORF57 in transient transfections (114,115). Direct determination of PAN RNA half-lives further showed increases in PAN RNA levels upon co-expression of ORF57. In addition, ORF57 promotes regulation of host gene expression (116), genome instability (117), translation (118), and may be involved in transcription (115).

Because of the importance of ORF57 RNA-binding for function, a comprehensive understanding of the RNAs bound by ORF57 during lytic infection is required to understand ORF57's essential activities in viral replication. In this study, I have adapted high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) for identification of ORF57 targets during lytic reactivation. As predicted, I identified CLIP tags mapping to the 5´ end of PAN RNA and additionally observed ORF57 interactions with RNAs generated at the KSHV origins of lytic replication (oriLyt). Examination of host targets revealed ORF57 binding sites near the 5´ end of a subset of the transcripts and these often mapped close to the first exon-intron junction. Our collaborator then monitored the RNA levels of four potential ORF57 targets (BTG1, EGR1, TNFSF9, and ZFP36) at various times following

lytic induction. Interestingly, the levels of these ORF57-bound pre-mRNAs persisted longer than controls, suggesting that ORF57 may be stabilizing these cellular pre-mRNAs.

### *3.2.2 Results*

I expanded the dCLIP pipeline for CLIP-Seq analysis to process the replicates simultaneously to identify continuous regions of the genome where CLIP tags were enriched when compared to the equivalent region in the input samples. I dubbed these regions enriched clusters and they represent RNA fragments bound by ORF57. The general workflow for enriched cluster identification is given in Fig. 8a. This analysis led to 2,229 and 219 enriched clusters mapping to the human and viral genomes, respectively. Our analysis identified 219 enriched clusters in the viral genome. The enriched clusters mapped broadly across the KSHV genome and were observed on plus and minus strand RNAs, consistent with a general role for ORF57 in KSHV RNA biogenesis (Fig. 8b). As expected, I identified enriched clusters in PAN RNA and the identified clusters were located near the 5´ end of the RNA, which demonstrates that our HITS-CLIP analysis successfully identified ORF57-bound RNA fragments.

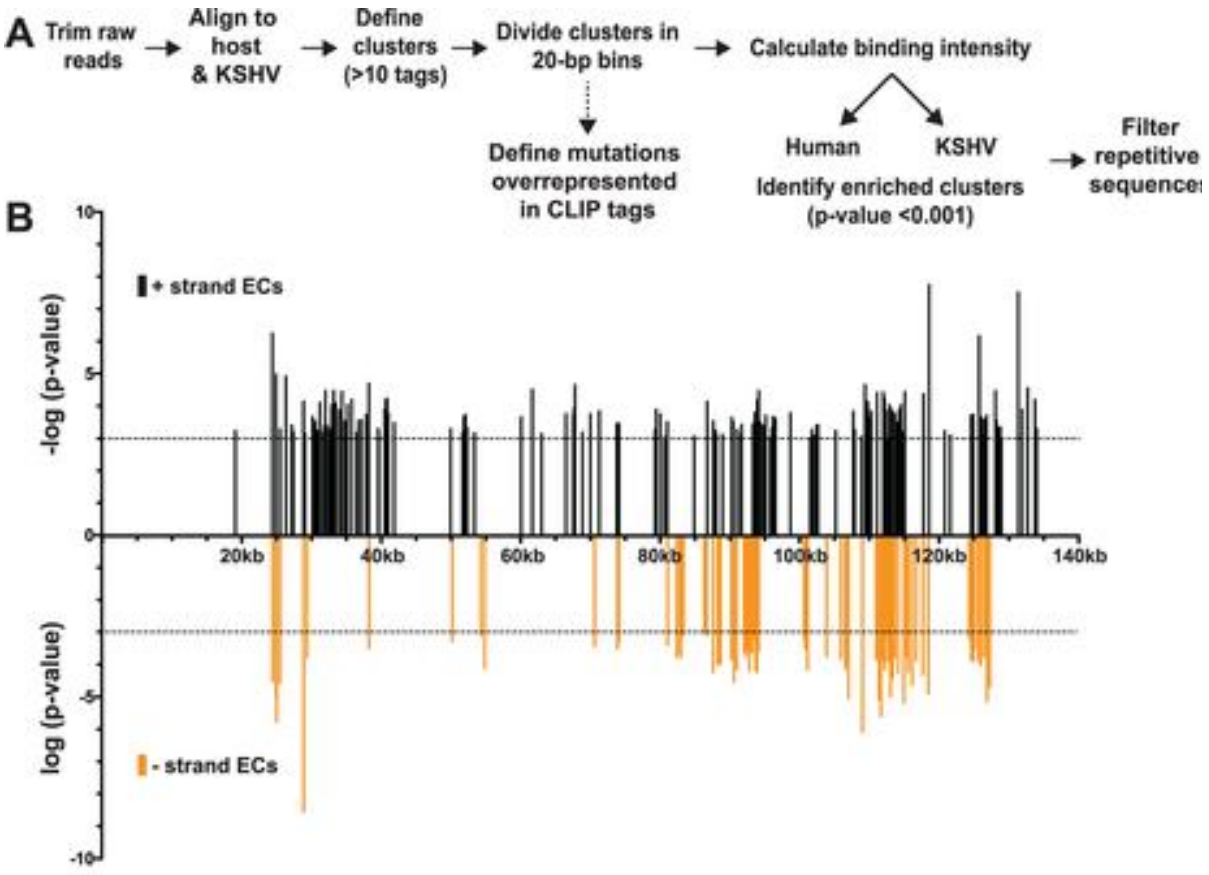



**Fig. 8** Identification of enriched clusters mapping to the KSHV genome. (a) Outline of the bioinformatics pipeline used to identify enriched clusters. (b) Genomic location of KSHV enriched clusters. The x-axis represents position on the KSHV genome (U75698) and the midpoint of each cluster was used as the x-coordinate. KSHV enriched clusters ranged from 19–3679 nt; the mean enriched cluster length was 176 nt and the median was 79 nt. A statistical cutoff of 0.001 was used to define enriched clusters (dashed lines). For display, the RNA fragments mapping to the KSHV plus strand were assigned -log10(p-values) (black) while the minus strand clusters are displayed as log10(p-values) (orange).

A total of 2,229 enriched clusters mapped to the human genome and these ORF57-bound RNA fragments correspond to ~700 unique host genes. The human enriched clusters were derived from clusters with tag counts spanning several orders of magnitude, so I can be confident that our bioinformatics pipeline has a wide dynamic range (Fig. 9a). I determined where the enriched clusters mapped in relationship with specific gene features. Over one third of the clusters mapped to introns, while 27% was found in coding sequences (Fig. 9b). To look for biases in the location of enriched clusters across genes, I

calculated where each enriched cluster midpoint maps as a fraction of the length of that specific gene. The results were subsequently compiled to a single model gene (Fig. 9c). While enriched clusters were found throughout the length of target genes, I observed a clear overrepresentation of enriched clusters near the 5´ ends of genes (Fig. 9c). The enriched clusters concentrated at 5´ ends (Fig. 8c, bracket) were examined based on where they mapped to gene features (Fig. 9d). As expected for 5´-enriched fragments, I detected increases in the percentage of enriched clusters mapping to the 5´ UTR and upstream 2 kb and decreases in the intergenic regions, downstream 2 kb and 3´ UTR annotations. I further observed a small increase in the percent mapping to intronic regions (36% in the total and 43% in the 5´-most clusters), but it is unclear whether this increase is significant. Next, I determined the distances between the transcription start sites (TSS) and the 5´ enriched clusters and observed that the 5´ enriched clusters do not peak directly at the TSS, but rather ~300–500 bp downstream of the TSS (Fig. 9e). In contrast, when I examined the distances between the 5´ enriched clusters and the first exon-intron boundary, I observed a peak coincident with this boundary (Fig. 9f). Consistent with the observed 43% intronic reads, the peak is not solely on the exonic sequence but spans the exon-intron junction. Taken together, these data show that a subset of the ORF57-bound RNA fragments map to the 5´ end of the transcript and are particularly concentrated near the 5´-most exon-intron boundary.

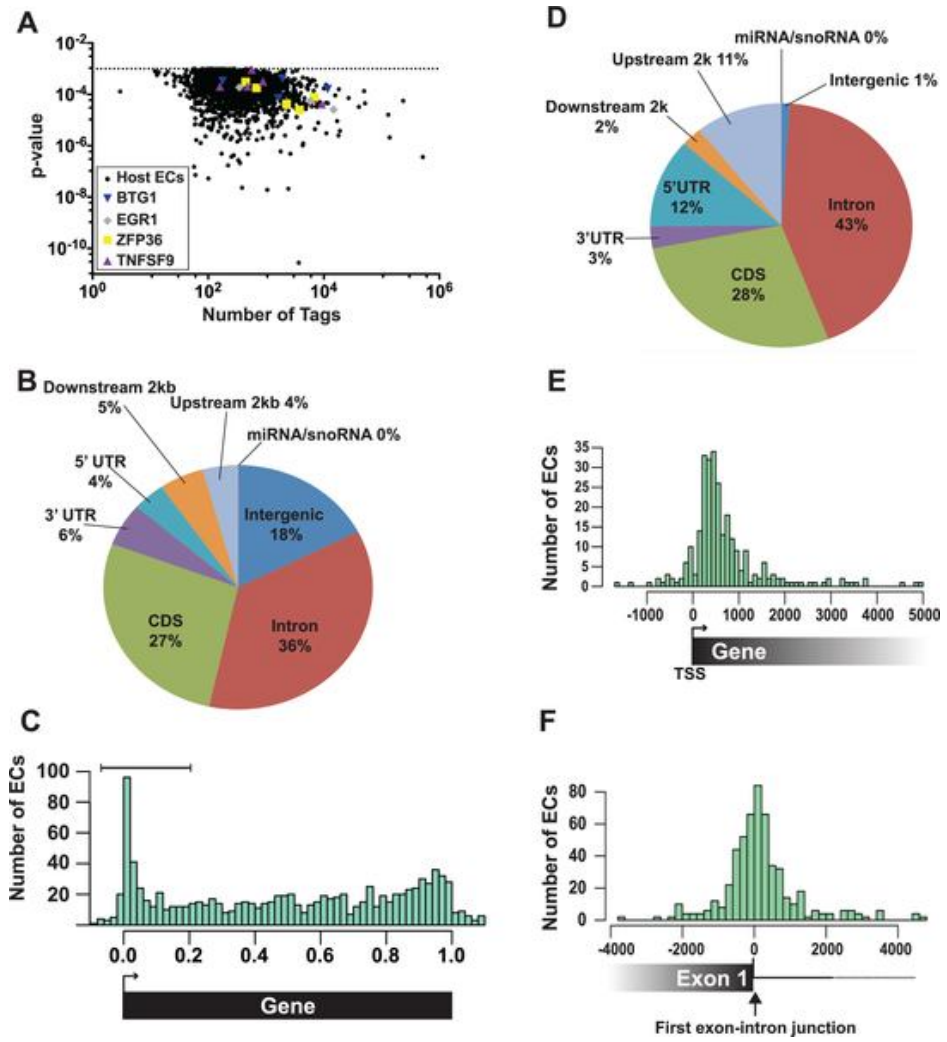



**Fig. 9** Identification and characterization of enriched clusters mapping to the human genome. (A) Graph of the 2,229 host enriched clusters with their p-values plotted on the y-axis and the total tag count for one of the three biological replicates plotted on the x-axis. Enriched clusters from four selected candidate genes are shown in color. (B) Pie chart of the gene feature annotations of human enriched clusters. If a cluster spanned multiple different annotations, the cluster annotation was split proportionally among the annotations. Upstream and downstream 2 kb refer to clusters mapping within 2 kb immediately flanking the nearest annotated gene. (C) The bar graph compares the number human enriched clusters (ECs) relative to their position on the closest annotated transcript. 0.0 and 1.0 correspond to the transcription start site (TSS) and the poly(A) site, respectively. Importantly, the length in this graph is relative to the annotated gene and is not a measure of the distance in base pairs. The brackets denote the 5´-enriched clusters examined in D-E. (D) Pie graph classifies the 5´-enriched clusters based on their gene feature annotations (n = 288). (E) Bar graph comparing the number (y-axis) and position (x-axis) of 5´ enriched clusters relative to the transcription start site (TSS). The distance is measured in base pairs from the TSS (TSS = 0). (F) Bar graph comparing the number (y-axis) and position (x-axis) of 5´ enriched clusters relative to the 5´-most exon-intron junction. The distance is measured in base pairs from the first exon-intron boundary, which was set at zero.

### *3.2.3 Discussion*

Inspection of the sequence traces confirmed the presence of enriched clusters mapping toward the 5´ ends of these RNAs but not in the GAPDH or β-actin controls. Multiple enriched clusters were found in each of these RNAs: three, five, four, and four enriched clusters were identified for EGR1, ZFP36, BTG1, and TNFSF9, respectively. Using these genes as examples, our collaborator used experimental strategies to show that candidate ORF57 target pre-mRNAs persist longer than control pre-mRNAs over the course of KSHV lytic reactivation (data not shown).

Because of its critical role in the viral life cycle, a mechanistic understanding of ORF57 functions is essential to the understanding of KSHV replication and pathogenesis. Our high throughput screening of ORF57-bound RNAs begins to address interactions of ORF57 with viral and host RNAs. This work extends existing data supporting a general role for ORF57 in the stabilization of a wide variety of viral RNAs. In addition, these data suggest that ORF57 nuclear RNA stabilization function is not restricted to viral RNAs, but further modulates the processing and decay of host transcripts during lytic reactivation. Ongoing studies seek to identify the precise molecular mechanisms of ORF57 interactions with the host cell RNA decay machinery that promote the stabilization of viral and host transcripts in the nucleus. In addition, it is of great interest to determine how changes in gene expression induced by ORF57 binding of host RNAs affect viral replication and/or pathogenesis.

# CHAPTER FOUR - IDENTIFY RBP BINDING SITES ON CIRCRNAS

## 4.1 Background

### *4.1.1 Circular RNA (circRNA) and its importance*

circRNA is formed when the 3' and 5' ends of part of the linear transcript are joined. The joining points are characterized by GU/AG splicing signal. circRNAs are mainly found in cytoplasmic fractions (119) and do not seem to have polyA tail, although one circRNA was recently found to be translatable (120). Reverse complementary sequences in flanking introns are shown to be necessary for some circRNAs, while not for others (121). There is evidence showing that their expression is regulated during EMT (122) or neuronal development (123). One study finds that circRNAs compete with pre-mRNA splicing (124). And there are two circRNAs cDR1as and Sry that are known to act as miRNA sponges (125). Li et al found that a special class of circRNAs could regulate transcription in the nucleus (126). In human samples, Bachmayr-Heyda et al discovered a negative correlation of global circRNA abundance with proliferation (127).

### *4.1.2 Profile RBP-circRNA interaction using CLIP-Seq data*

Splicing factors are shown to regulate formation of circRNAs (120,121). Conn et al used PAR-CLIP to show that Quaking regulates formation of circRNAs via binding sites in introns (122). Li et al conducted Pol II CLIP-Seq and revealed a subclass of nucleus-locating circRNAs that are associated with Pol II (126). All of these argue for the need of utilizing CLIP-Seq data to reveal the role of RBPs when bound to circRNAs. HITS-CLIP, PAR-CLIP and iCLIP differ in the cross-linking strategies and library preparation procedures, but all three techniques generate sequencing reads whose genomic mapping positions overlap the location of RBP binding sites. If these sequencing reads can be mapped across the splicing joining sites of circRNAs, it would provide direct evidence of RBPs binding to circRNAs. To our knowledge, a systematic analysis of public CLIP-Seq datasets for this purpose is lacking so far.

#### *4.1.3 Challenges of mining RBP-circRNA interactions from CLIP-Seq data*

Guo et al developed a pipeline and applied it to ENCODE data (119). CIRI is another software to detect circRNAs from transcriptome data (128). A database called circBase has been built that merged datasets of circRNAs in different organisms (129). The rationales behind these different approaches are similar: an RNA-Seq read whose 3' end and 5'end map to an upstream and a downstream transcript part respectively in reverse configuration is evidence of circRNAs. It can be expected that non-polyA selected paired-end RNA-Seq data is most suitable for circRNA discovery. But for CLIP-Seq data, it is difficult to apply previously-developed pipelines directly due to two major challenges: (1) CLIP-Seq technology involves enzymatic digestion, leading to generally very short read length (2) CLIP-Seq data are almost exclusively single-end. (3) CLIP-Seq data usually have limited library complexity, yielding a high PCR duplicate rate. CIRI, for example, is known to have high false discovery rate for single-ended data, refuses to process data whose alignment length is smaller than 40nt, and cannot distinguish between PCR duplicates.

### 4.2 Bioinformatics pipeline development

#### *4.2.1 Download and curation of public CLIP-Seq data*

I have and will continue to download all public CLIP-Seq data from Human, Mouse and Drosophila Melanogaster mainly from GEO and also other data depositories. There are >140 independent studies containing >300 sequencing experiments and still growing. Most datasets are from Hela and HEK293 cells. For the current stage, I only downloaded the CLIP-Seq datasets with wild-type genetic background and no treatment or control treatment. The adaptor sequences are found by reading experimental protocols, FastQC detection or manual comparison. After trimming adaptors, CLIP-Seq reads that are too short were discarded from analysis. To tackle the high PCR duplicate rate problem caused by limited library complexity, the remaining reads with exactly the same nucleotide sequence were collapsed to unique tags.

### *4.2.2 Linearization of circRNA library*

It is disadvantageous to examine whether CLIP-Seq reads can be split and mapped on both ends of a splice junction as each segment of the single-ended read will be too short to be aligned unambiguously. I can circumvent this problem by linearize nucleotide sequences of previously identified circRNAs around their splice junctions and add them to the reference genome as additional chromosomes (**Fig. 10a**). In this way, the CLIP-Seq reads can be aligned as a whole, increasing the possibility of successful alignment. I have collected a series of published literature that identified circRNAs by mining RNA-Seq data in Human, Mouse and Drosophila (**Table 3**). I have also run CIRI on Encode (130-132) non-polyA selected paired-end RNA-Seq data to discover more circRNAs. There are currently 90 Human, 16 Mouse and 108 Drosophila samples suitable for this analysis that are available in Encode.

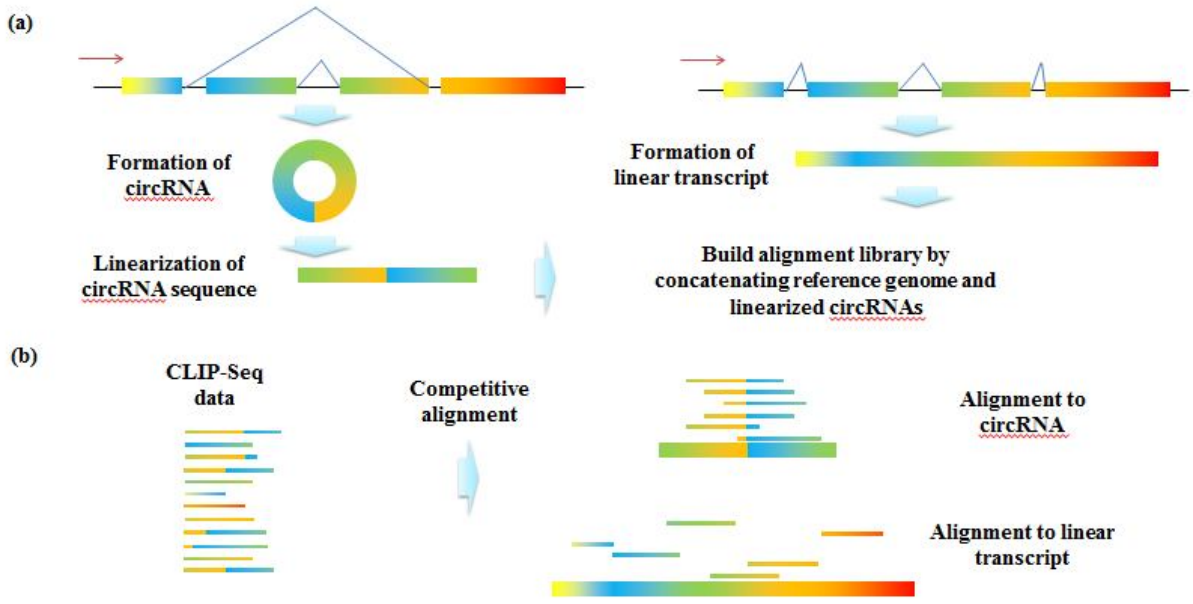



**Fig. 10** Cartoon of the pipeline for identifying RBP-circRNA interactions using CLIP-Seq data

**Table 3** Previous publications that reported large scale existence and locations of circRNAs

| Publication | Species | # circRNAs |
|---|---|---|

| | | |
|---|---|---|
| **(124)** | Drosophila | 4053 |
| **(133)** | Drosophila | 38115 |
| **(134)** | Human | 65731 |
| **(129)** | Human | 92375 |
| **(134)** | Mouse | 15849 |
| **(129)** | Mouse | 1903 |

### *4.2.3 Competitive alignment of CLIP-Seq reads to circRNA library and reference genome*

CLIP-Seq reads were mapped simultaneously to the linearized circRNA library and the normal reference genome (**Fig. 10b**). One read can be aligned to 0, 1 or more places. A decision tree was applied to determine whether each alignment of a CLIP-Seq read is within a circRNA or a linear transcript (**Fig. 11**). Of all alignments for each CLIP-Seq read, if the circRNA alignment has a longer alignment length and higher alignment score than any linear transcript alignment(s), if they are any, then this CLIP-Seq read was assigned to a circRNA. All datasets were scanned by this rule to count the number of CLIP-Seq reads associated with each circRNA.

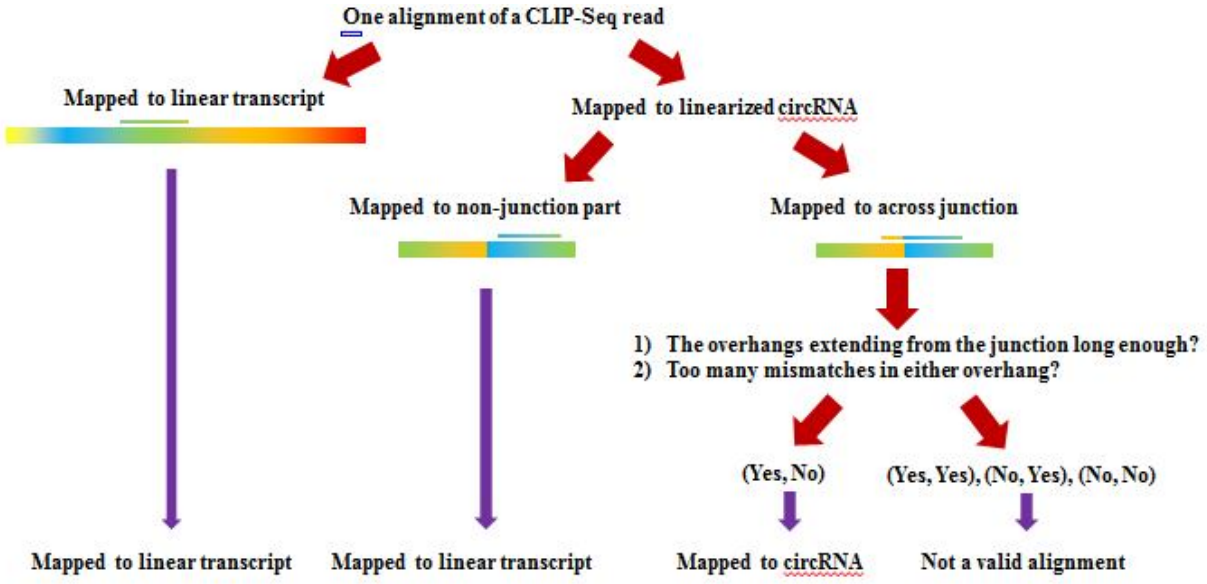



**Fig. 11** Decision tree to determine whether each alignment of a CLIP-Seq read is in a linear transcript or a circRNA.

### 4.3 Pipeline evaluation and downstream analysis

In this section, I first show some evidence that our novel algorithm is truly able to identify circRNAs bound by RBPs. Then I showed some preliminary downstream analysis results related to RBP binding bias, motif enriched on circRNAs, and gene ontology analysis.

### *4.3.1 Evaluation of the circRNA-identification pipeline*

| circRNA name | Chromosome | Start | End | Strand | # supporting reads |
|---|---|---|---|---|---|
| **circEIF3J** | chr15 | 44843074 | 44843720 | + | 2 |
| **circPAIP2** | chr5 | 138699448 | 138700432 | + | 2 |
| **circRSRC1** | chr3 | 157839892 | 157841780 | + | 3 |
| **circFUNDC1** | chrX | 44383248 | 44386611 | - | 3 |
| **circMIER1** | chr1 | 67423742 | 67428843 | + | 3 |
| **circSSR1** | chr6 | 7303783 | 7310262 | - | 2 |
| **circWDR60** | chr7 | 158662546 | 158669382 | + | 3 |
| **circRBM33** | chr7 | 155465561 | 155473602 | + | 12 |
| **circMAN1A2-1** | chr1 | 117944808 | 117957453 | + | 2 |
| **circMAN1A2-2** | chr1 | 117944808 | 117963271 | + | 9 |
| **circNAP1L4** | chr11 | 2972489 | 3000467 | - | 4 |
| **circBPTF** | chr17 | 65941525 | 65972074 | + | 2 |
| **circMAN1A2-3** | chr1 | 117944808 | 117984947 | + | 2 |
| **circCLTC** | chr17 | 57721637 | 57763169 | + | 0 |
| **circCDK11B** | chr1 | 1586823 | 1650894 | - | 42 |

**Table 4** Number of CLIP-Seq reads identified to support each of the 15 most enriched PolII-associated circRNA that have been experimentally validated before.

To confirm the validity of the proposed bioinformatics pipeline, I investigated whether our novel pipeline is able to re-identify the top 15 experimentally-validated circRNAs bound by PolII in this study (126). In **Table 4**, I showed the number of CLIP-Seq reads identified by our algorithm that supports each circRNA. 14 of these 15 circRNAs are supported by at least one CLIP-Seq read(s). As for the other circRNA, no supporting reads are found because it has a lot of low complexity regions and was excluded from analysis. Overall, our pipeline is shown to be very accurate in identifying RBP-bound circRNAs. I may want to reduce the stringency of the low complexity filter for more comprehensive inclusion.

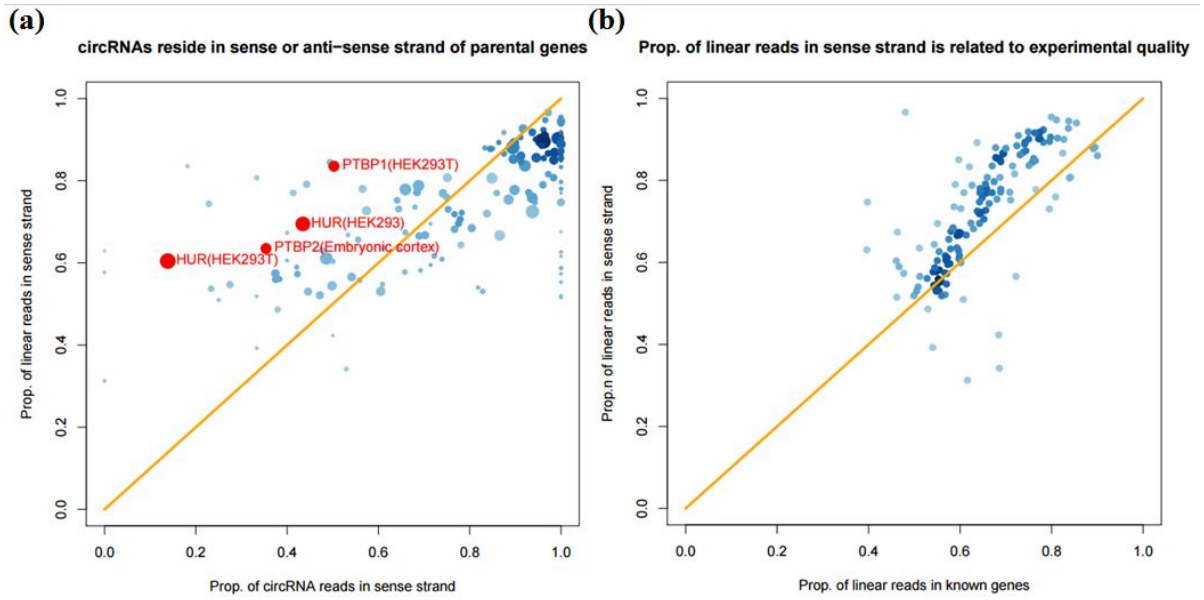



**Fig. 12** Some RBP-bound circRNAs are predominantly located anti-sense to parental genes. (a) Proportion of circRNA reads in sense strand vs. proportion of linear reads in sense strand. Dot size represents number of circRNA sequences in a study. (b) Proportion of linear reads in known genes vs. proportion of linear reads in sense strand.

### *4.3.2 Some RBP-bound circRNAs are predominantly located anti-sense to parental genes*

circRNAs are usually embodied within regular linear genes which are called parental genes of respective circRNAs. The circRNAs could be transcribed in the same direction (sense) or in the opposite direction (anti-sense) with respect to the transcription direction of parental genes, although the sense direction should be predominantly as circRNAs are spliced from transcripts of parental genes. I investigated whether RBP-bound circRNAs tend to transcribe in the sense direction or the anti-sense direction of parental genes. To do this, I used proportion of CLIP-Seq reads mapped to linear transcripts that are in the same or opposite direction of the linear genes as a control (**Fig. 12a**). The control is important since the strand-specific library preparation required for CLIP-Seq may not be perfect in some studies. To support the differing degree of library preparation in different studies, I found that the higher proportion of CLIP-Seq reads in anti-sense direction of linear genes for a certain study, the less likely the

sequencing reads from this CLIP-Seq study are mapped to known genes (**Fig. 12b**), which can be regarded as a rough measure of library quality. Overall, the proportion of bound circRNAs in sense direction is consistent with the proportion of CLIP-Seq reads in sense direction of linear genes for most studies. Interestingly, PTBP1, PTBP2 from two different studies showed much larger proportion of circRNA reads in the anti-sense direction of parental genes than the proportion of CLIP-Seq reads in the anti-sense direction of linear genes; HUR protein from two studies also consistently show that HUR tend to bind circRNA in anti-sense direction of parental genes. This suggests that strand bias of RBP-bound circRNA is a phenomenon of valid biological meaning, and suggest further investigation into anti-sense circRNAs which might play important regulatory roles.

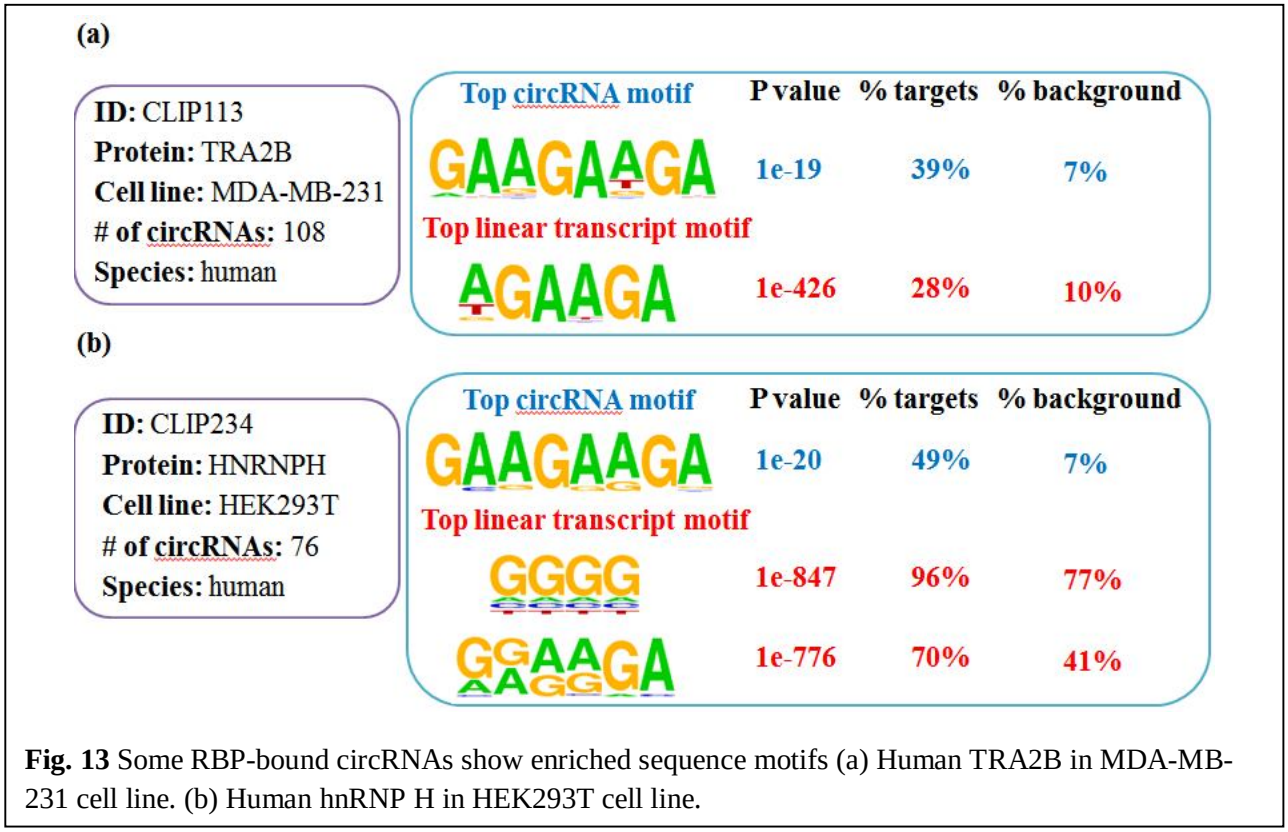



**Fig. 13** Some RBP-bound circRNAs show enriched sequence motifs (a) Human TRA2B in MDA-MB-231 cell line. (b) Human hnRNP H in HEK293T cell line.

### *4.3.3 Some RBP-bound circRNAs show enriched sequence motifs*

Next I asked whether RBPs bind to circRNAs through recognition of sequence motifs by carrying out HOMER motif search on regions of circRNAs bound by each RBP. To do this, I randomly sampled the whole genome (human, mouse or drosophila) to generate pseudo length-matched sequences to serve as background. I also searched sequence motif in CLIP-Seq reads mapped only to linear genes with the same background control. These linear transcript motifs can give hint to possible RBP recognition motifs on linear transcript, and may also help exclude the possibility that motifs found on circRNAs are reminiscent adaptor sequences that were not trimmed completely. Consistent with previous reports that relatively a minor fraction of RBPs recognize binding sites through motif matches, I failed to find enriched motifs on circRNAs in most studies. However, I found that human TRA2B seems to bind circRNAs through recognition of GAAGAAGA motifs (**Fig. 13a**). HOMER also identified a similar enriched motif for TRA2B with sequence logo AGAAGA on linear transcripts, which is consistent with previous reports (135). I also found that human hnRNP H seems to bind circRNAs through GAAGAAGA motifs (**Fig. 13b**). HOMER also finds that hnRNP H binds linear transcripts through similar motifs of GGGG or GGAAGA motifs, with GGGG being consistent with previous reports (136). The similarity of sequence motifs on linear transcripts identified by us with previous reports increased our confidence that sequence motifs on circRNAs are also real. These results suggest that some, but not all, RBPs recognize sequence motifs on target circRNAs. Besides, the same circRNA binding motif of GAAGAAGA by TRA2B and hnRNP H may hint some coordination between these 2 proteins.

### *4.3.4 Gene ontology enrichment of parental genes for circRNAs bound by RBPs*

We also investigated whether parental genes for circRNAs bound by each RBP shows enrichment in a certain biological functional category. Some RBP's bound circRNAs are too few in number (<30), and are excluded for this analysis. I downloaded the KEGG pathways from the GSEA database (137,138). I then calculated the hypergeometric test p value (log transformed) for whether the set of parental genes for

circRNAs bound by each RBP is enriched in the set of genes from each ontology. As a background control, I randomly sampled a set of genes with the same size as the set of parental genes for each RBP and calculated a background p value. I subtracted the averaged background p values from 100 randomizations from the true test p value for each RBP. **Fig. 14** shows the subtracted log transformed p values of the 186 KEGG pathways for each RBP. It seems that the parental genes do not show highly significant enrichment in any gene ontology for most RBPs except for human YBX1 protein in MDA cell line and human DDX21 in HEK293 cell line. The enriched ontology is SMALL CELL LUNG CANCER for YBX1 and RIBOSOME RNA PROCESSING for DDX21. Interestingly, the known function for YBX is involved in non-small cell lung cancer and one known function for DDX21 is coordination of ribosome RNA processing. This result further confirms that circRNAs are likely to be functionally important in biological regulation and raised the intriguing possibility that some RBPs may exert their functions through binding circRNAs.

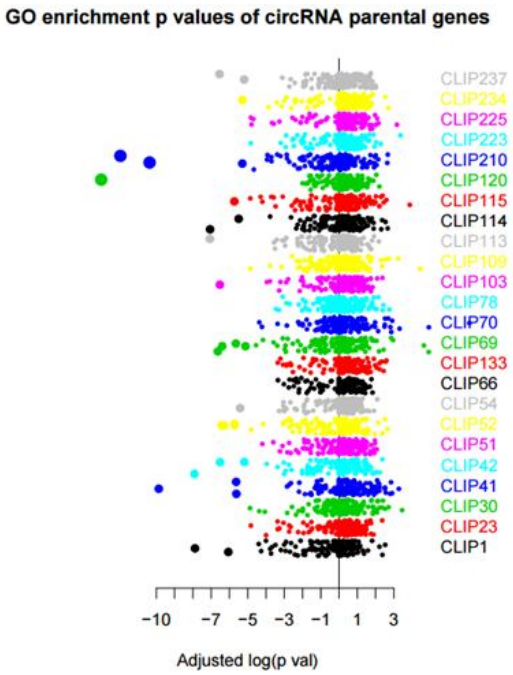



**Fig. 14** Gene ontology enrichment of parental genes for circRNAs bound by RBPs. The hypergeometric test p value (log transformed) for whether the set of parental genes for circRNAs bound by each RBP is enriched in the set of genes from each ontology is calculated. A set of genes with the same size as the set of parental genes for each RBP is randomly sampled 100 times and the resulting background p values were averaged and subtracted from the true test p value for each RBP.

## 4.4 Discussion

In previous chapters I have developd user-friendly algorithms for initial analysis of CLIP-Seq data and demonstrated the applications of these methods in our collaborator's datasets. Following peak-calling, downstream analysis will generally focus on characterization of RBP-RNA interaction sites, such as motif discovery and secondary structure prediction. These are the most generic bioinformatics analyses of CLIP-Seq data, which has greatly expanded our knowledge of RNA regulation. On the other hand, integrative analysis of CLIP-Seq data with other types of high-throughput data types can be a very interesting and productful research direction, but has so far not been intensively explored yet. One recent study (139) identified 22,735 RBP–lncRNA regulatory relationships from >100 public genome-wide CLIP datasets. This study serves as an example how integrative analysis could lead to meaningful

discoveries. In this chapter, I investigated the possibility of integrative analysis of CLIP-Seq datasets with circRNA data to characterize the function of RBP binding sites on circRNAs.

However, there are a few limitations and pitfalls of this study. (1) Only RBPs bound to junction sites of mature circRNAs are investigated. During the formation of circRNAs, the upstream and downstream introns surrounding the circRNAs will form a “stem”. In this process, RBPs, especially splicing factors, will play very important roles. This study didn’t investigate binding events on the stems. This study didn’t investigate RBP binding sites on non-junction part of circRNAs, either, because it is impossible to distinguish such binding events from those on the linear transcripts. (2) This study took a shortcut to identify RBP-circRNA interactions by aligning to a pseudo reference genome prepared from previously identified circRNAs. This convenience comes at the price of the inability to discovery novel circRNAs in the CLIP-Seq data. To partially solve this problem, I would collect as many as possible previous published circRNAs and to run CIRI exhaustively on all suitable paired-end non-polyA selected RNA-Seq data.

# CHAPTER FIVE - DISCUSSION

In this thesis, I have introduced and discussed the CLIP-Seq technology. The development of technology and bioinformatics in this field has greatly improved our capacity to study protein-RNA interactions and understand the functions of different RNA species in physiological and pathological process. I have then shown my work in methodological developments. The two software MiClip and dCLIP that I developed addressed two important questions in CLIP-Seq data analysis, respectively. Then I demonstrated the applications of these methods in real datasets and the interesting biological discoveries that have been made from them. Finally, I presented some preliminary results of integrative analysis of CLIP-Seq datasets with other types of data. For the moment, I focused on identify circRNA-RBP interactions from reanalyzing CLIP-Seq datasets. Overall, I contributed to the RNA regulation research community from the bioinformatics side with an emphasis on understanding RBP-RNA interactions.

There are several related technologies, such as CLASH and RIP-Seq, which may be complementary to genome-wide CLIP to study the function of RNAs. CLASH is short for cross-linking, ligation, and sequencing of hybrids. It was invented for characterizing intramolecular and intermolecular RNA-RNA interactions (48). Recently, this technology was adapted to straightforwardly detect miRNA-mRNA pairs as chimeric reads in high-throughput sequencing data (140). Integrative analysis can be carried out that combines CLASH data that can directly capture reliable miRNA-mRNA interactions and genome-wide CLIP data that focuses more on detecting RBP-RNA interactions. RNA immunoprecipitation sequencing (RIP-Seq) can also complement genome-wide CLIP for identifying RBP-RNA interactions (141). RIP-Seq bears some similarity to genome-wide CLIP, but lacks high-stringency washes and crosslinking of RBP to RNAs, which leads to high background noise and mis-interpretations in the data analysis. For example, RIP-Seq identifies both direct and indirect RBP-RNA interactions, while genome-wide CLIP can accurately identify direct RBP-RNA association events (142). However, genome-wide CLIP is more technically challenging and also requires high-quality antibodies to work properly. Therefore the data

from CLIP experiments and RIP-Seq experiments could be complementary in studying RBP-RNA bindings.

With the maturation of both the CLIP-Seq technology and its bioinformatics analysis pipelines, large amount of novel discoveries could be made by CLIP-Seq at a faster pace. For example, one direction for further study is to conduct CLIP-Seq experiments of different proteins under different treatments simultaneously in an experimental system to methodically understand and model transcriptional events. The MOV10 and UPF1 proteins have recently been shown to bind in close proximity and interact directly (143), pointing to the importance of studying the coordination pattern of RBPs and its functional impact. Another future direction could be to combine CLIP-Seq with other types of data, including ChIP-Seq, RNA-Seq and proteomics data for integrative analysis. EZH2 was reported to bind lncRNAs (70), despite its chromatin-binding capability and its role in epigenetic regulation. This intriguing phenomenon suggests that ChIP-Seq data and CLIP-Seq data can be analyzed together to reveal novel RNA-binding functions of well-characterized DNA-binding proteins. All of these efforts will help us better understand transcriptional regulation in biological systems.

# BIBLIOGRAPHY


1. Khorshid, M., Rodak, C. and Zavolan, M. (2011) CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic acids research*, **39**, D245-252.
2. Li, J.H., Liu, S., Zhou, H., Qu, L.H. and Yang, J.H. (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic acids research*, **42**, D92-97.
3. Yang, J.H., Li, J.H., Shao, P., Zhou, H., Chen, Y.Q. and Qu, L.H. (2011) starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic acids research*, **39**, D202-209.
4. Blin, K., Dieterich, C., Wurmus, R., Rajewsky, N., Landthaler, M. and Akalin, A. (2015) DoRiNA 2.0-upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic acids research*, **43**, D160-167.
5. Anders, G., Mackowiak, S.D., Jens, M., Maaskola, J., Kuntzagk, A., Rajewsky, N., Landthaler, M. and Dieterich, C. (2012) doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic acids research*, **40**, D180-186.
6. Yang, Y.C., Di, C., Hu, B., Zhou, M., Liu, Y., Song, N., Li, Y., Umetsu, J. and Lu, Z. (2015) CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC genomics*, **16**, 51.
7. Corcoran, D.L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R.L., Keene, J.D. and Ohler, U. (2011) PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome biology*, **12**, R79.
8. Uren, P.J., Bahrami-Samani, E., Burns, S.C., Qiao, M., Karginov, F.V., Hodges, E., Hannon, G.J., Sanford, J.R., Penalva, L.O. and Smith, A.D. (2012) Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, **28**, 3013-3020.
9. Chen, B., Yun, J., Kim, M.S., Mendell, J.T. and Xie, Y. (2014) PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome biology*, **15**, R18.
10. Sievers, C., Schlumpf, T., Sawarkar, R., Comoglio, F. and Paro, R. (2012) Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic acids research*, **40**, e160.
11. Erhard, F., Dolken, L., Jaskiewicz, L. and Zimmer, R. (2013) PARma: identification of microRNA target sites in AGO-PAR-CLIP data. *Genome biology*, **14**, R79.
12. Yun, J., Wang, T. and Xiao, G. (2014) Bayesian hidden Markov models to identify RNA-protein interaction sites in PAR-CLIP. *Biometrics*.
13. Maticzka, D., Lange, S.J., Costa, F. and Backofen, R. (2014) GraphProt: modeling binding preferences of RNA-binding proteins. *Genome biology*, **15**, R17.
14. Kechavarzi, B. and Janga, S.C. (2014) Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome biology*, **15**, R14.
15. Bahrami-Samani, E., Vo, D.T., de Araujo, P.R., Vogel, C., Smith, A.D., Penalva, L.O. and Uren, P.J. (2014) Computational challenges, tools, and resources for analyzing co- and post-transcriptional events in high throughput. *Wiley interdisciplinary reviews. RNA*.
16. Taupin, J.L., Tian, Q., Kedersha, N., Robertson, M. and Anderson, P. (1995) The RNA-binding protein TIAR is translocated from the nucleus to the cytoplasm during Fas-mediated apoptotic cell death. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 1629-1633.
17. Strasser, K. and Hurt, E. (2000) Yra1p, a conserved nuclear RNA-binding protein, interacts directly with Mex67p and is required for mRNA export. *The EMBO journal*, **19**, 410-420.

18. Hunt, S.L., Hsuan, J.J., Totty, N. and Jackson, R.J. (1999) unr, a cellular cytoplasmic RNA-binding protein with five cold-shock domains, is required for internal initiation of translation of human rhinovirus RNA. *Genes & development*, **13**, 437-448.
19. Althammer, S., Gonzalez-Vallinas, J., Ballare, C., Beato, M. and Eyras, E. (2011) Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. *Bioinformatics*, **27**, 3333-3340.
20. Mazan-Mamczarz, K., Kuwano, Y., Zhan, M., White, E.J., Martindale, J.L., Lal, A. and Gorospe, M. (2009) Identification of a signature motif in target mRNAs of RNA-binding protein AUF1. *Nucleic acids research*, **37**, 204-214.
21. Abdelmohsen, K., Panda, A.C., Kang, M.J., Guo, R., Kim, J., Grammatikakis, I., Yoon, J.H., Dudekula, D.B., Noh, J.H., Yang, X. *et al.* (2014) 7SL RNA represses p53 translation by competing with HuR. *Nucleic acids research*, **42**, 10099-10111.
22. Rossbach, O., Hung, L.H., Khrameeva, E., Schreiner, S., Konig, J., Curk, T., Zupan, B., Ule, J., Gelfand, M.S. and Bindereif, A. (2014) Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP L. *RNA biology*, **11**, 146-155.
23. Chou, C.H., Lin, F.M., Chou, M.T., Hsu, S.D., Chang, T.H., Weng, S.L., Shrestha, S., Hsiao, C.C., Hung, J.H. and Huang, H.D. (2013) A computational approach for identifying microRNA-target interactions using high-throughput CLIP and PAR-CLIP sequencing. *BMC genomics*, **14 Suppl 1**, S2.
24. Keene, J.D. (2007) RNA regulons: coordination of post-transcriptional events. *Nature reviews. Genetics*, **8**, 533-543.
25. Lukong, K.E., Chang, K.W., Khandjian, E.W. and Richard, S. (2008) RNA-binding proteins in human genetic disease. *Trends in genetics : TIG*, **24**, 416-425.
26. Castello, A., Fischer, B., Hentze, M.W. and Preiss, T. (2013) RNA-binding proteins in Mendelian disease. *Trends in genetics : TIG*, **29**, 318-327.
27. Wurth, L. (2012) Versatility of RNA-Binding Proteins in Cancer. *Comparative and functional genomics*, **2012**, 178525.
28. Glazer, R.I., Vo, D.T. and Penalva, L.O. (2012) Musashi1: an RBP with versatile functions in normal and cancer stem cells. *Frontiers in bioscience*, **17**, 54-64.
29. Ascano, M., Jr., Mukherjee, N., Bandaru, P., Miller, J.B., Nusbaum, J.D., Corcoran, D.L., Langlois, C., Munschauer, M., Dewell, S., Hafner, M. *et al.* (2012) FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature*, **492**, 382-386.
30. Hoell, J.I., Larsson, E., Runge, S., Nusbaum, J.D., Duggimpudi, S., Farazi, T.A., Hafner, M., Borkhardt, A., Sander, C. and Tuschl, T. (2011) RNA targets of wild-type and mutant FET family proteins. *Nature structural & molecular biology*, **18**, 1428-1431.
31. Neumann, M., Bentmann, E., Dormann, D., Jawaid, A., DeJesus-Hernandez, M., Ansorge, O., Roeber, S., Kretzschmar, H.A., Munoz, D.G., Kusaka, H. *et al.* (2011) FET proteins TAF15 and EWS are selective markers that distinguish FTLD with FUS pathology from amyotrophic lateral sclerosis with FUS mutations. *Brain : a journal of neurology*, **134**, 2595-2609.
32. Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A. and Darnell, R.B. (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Science*, **302**, 1212-1215.
33. Ule, J., Jensen, K., Mele, A. and Darnell, R.B. (2005) CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods*, **37**, 376-386.
34. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, **10**, 57-63.
35. Zhang, C. and Zhang, M.Q. (2012) Identification of Splicing Factor Target Genes by High-throughput Sequencing. Chap. 51 in Alternative pre-mRNA Splicing: Theory and Protocols. Eds.
36. Baker, M. (2010) MicroRNA profiling: separating signal from noise. *Nature methods*, **7**, 687-692.
37. Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464-469.

38. Kaneko, S., Bonasio, R., Saldana-Meyer, R., Yoshida, T., Son, J., Nishino, K., Umezawa, A. and Reinberg, D. (2014) Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. *Molecular cell*, **53**, 290-300.
39. Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M. *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, **495**, 333-338.
40. Zhang, X., Zuo, X., Yang, B., Li, Z., Xue, Y., Zhou, Y., Huang, J., Zhao, X., Zhou, J., Yan, Y. *et al.* (2014) MicroRNA directly enhances mitochondrial translation during muscle differentiation. *Cell*, **158**, 607-619.
41. Chi, S.W., Zang, J.B., Mele, A. and Darnell, R.B. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, **460**, 479-486.
42. Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.C., Munschauer, M. *et al.* (2010) PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. *Journal of visualized experiments : JoVE*.
43. Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M. and Ule, J. (2011) iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *Journal of visualized experiments : JoVE*.
44. Zhang, Y., Xie, S., Xu, H. and Qu, L. (2015) CLIP: viewing the RNA world from an RNA-protein interactome perspective. *Science China. Life sciences*, **58**, 75-88.
45. Zhang, C. and Darnell, R.B. (2011) Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nature biotechnology*, **29**, 607-614.
46. Zhang, M.Q. (2012) Dissecting Splicing Regulatory Network by Integrative Analysis of CLIP-Seq Data. Chap 12 in Bioinformatics for High Throughput Sequencing, eds., 209-218.
47. Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, **10**, 669-680.
48. Kudla, G., Granneman, S., Hahn, D., Beggs, J.D. and Tollervey, D. (2011) Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 10010-10015.
49. Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P. and Bartel, D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell*, **27**, 91-105.
50. Iengar, P. (2012) An analysis of substitution, deletion and insertion mutations in cancer genes. *Nucleic acids research*, **40**, 6401-6413.
51. Liu, Y., Zhou, J. and White, K.P. (2014) RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, **30**, 301-304.
52. Licatalosi, D.D., Yano, M., Fak, J.J., Mele, A., Grabinski, S.E., Zhang, C. and Darnell, R.B. (2012) Ptbp2 represses adult-specific splicing to regulate the generation of neuronal precursors in the embryonic brain. *Genes & development*, **26**, 1626-1642.
53. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome biology*, **11**, R106.
54. Sei, E., Wang, T., Hunter, O.V., Xie, Y. and Conrad, N.K. (2015) HITS-CLIP analysis uncovers a link between the Kaposi's sarcoma-associated herpesvirus ORF57 protein and host pre-mRNA metabolism. *PLoS pathogens*, **11**, e1004652.
55. Liu, Q.Z., X; Madison, B; Rustgi, A; Shyr, Y. (2015) Assessing Computational Steps for CLIP-Seq Data Analysis. *BioMed Research International*.
56. Sims, D., Sudbery, I., Ilott, N.E., Heger, A. and Ponting, C.P. (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews. Genetics*, **15**, 121-132.
57. Engstrom, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., Ratsch, G., Goldman, N., Hubbard, T.J., Harrow, J., Guigo, R. *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature methods*, **10**, 1185-1191.

58. Bas, A., Forsberg, G., Hammarstrom, S. and Hammarstrom, M.L. (2004) Utility of the housekeeping genes 18S rRNA, beta-actin and glyceraldehyde-3-phosphate-dehydrogenase for normalization in real-time quantitative reverse transcriptase-polymerase chain reaction analysis of gene expression in human T lymphocytes. *Scandinavian journal of immunology*, **59**, 566-573.
59. Sugimoto, Y., Konig, J., Hussain, S., Zupan, B., Curk, T., Frye, M. and Ule, J. (2012) Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome biology*, **13**, R67.
60. Webb, S., Hector, R.D., Kudla, G. and Granneman, S. (2014) PAR-CLIP data indicate that Nrd1-Nab3-dependent transcription termination regulates expression of hundreds of protein coding genes in yeast. *Genome biology*, **15**, R8.
61. Ince-Dunn, G., Okano, H.J., Jensen, K.B., Park, W.Y., Zhong, R., Ule, J., Mele, A., Fak, J.J., Yang, C., Zhang, C. *et al.* (2012) Neuronal Elav-like (Hu) proteins regulate RNA splicing and abundance to control glutamate levels and neuronal excitability. *Neuron*, **75**, 1067-1080.
62. Boudreau, R.L., Jiang, P., Gilmore, B.L., Spengler, R.M., Tirabassi, R., Nelson, J.A., Ross, C.A., Xing, Y. and Davidson, B.L. (2014) Transcriptome-wide discovery of microRNA binding sites in human brain. *Neuron*, **81**, 294-305.
63. Wang, T., Xie, Y. and Xiao, G. (2014) dCLIP: a computational approach for comparative CLIP-seq analyses. *Genome biology*, **15**, R11.
64. Haecker, I., Gay, L.A., Yang, Y., Hu, J., Morse, A.M., McIntyre, L.M. and Renne, R. (2012) Ago HITS-CLIP expands understanding of Kaposi's sarcoma-associated herpesvirus miRNA function in primary effusion lymphomas. *PLoS pathogens*, **8**, e1002884.
65. Macias, S., Plass, M., Stajuda, A., Michlewski, G., Eyras, E. and Caceres, J.F. (2012) DGCR8 HITS-CLIP reveals novel functions for the Microprocessor. *Nature structural & molecular biology*, **19**, 760-766.
66. Masuda, A., Andersen, H.S., Doktor, T.K., Okamoto, T., Ito, M., Andresen, B.S. and Ohno, K. (2012) CUGBP1 and MBNL1 preferentially bind to 3' UTRs and facilitate mRNA decay. *Scientific reports*, **2**, 209.
67. Ishigaki, S., Masuda, A., Fujioka, Y., Iguchi, Y., Katsuno, M., Shibata, A., Urano, F., Sobue, G. and Ohno, K. (2012) Position-dependent FUS-RNA interactions regulate alternative splicing events and transcriptions. *Scientific reports*, **2**, 529.
68. Chu, Y., Wang, T., Dodd, D., Xie, Y., Janowski, B.A. and Corey, D.R. (2015) Intramolecular circularization increases efficiency of RNA sequencing and enables CLIP-Seq of nuclear RNA from human cells. *Nucleic acids research*.
69. Friedersdorf, M.B. and Keene, J.D. (2014) Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome biology*, **15**, R2.
70. Kaneko, S., Son, J., Shen, S.S., Reinberg, D. and Bonasio, R. (2013) PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nature structural & molecular biology*, **20**, 1258-1264.
71. Sei, E., Wang, T., Hunter, O.V., Xie, Y. and Conrad, N.K. (2015) HITS-CLIP Analysis Uncovers a Link between the Kaposi's Sarcoma-Associated Herpesvirus ORF57 Protein and Host Pre-mRNA Metabolism. *PLoS pathogens*.
72. Bahrami-Samani, E., Penalva, L.O., Smith, A.D. and Uren, P.J. (2015) Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic acids research*, **43**, 95-103.
73. Sievers, C., Schlumpf, T., Sawarkar, R., Comoglio, F. and Paro, R. (2012) Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic acids research*.
74. Li, J.H., Liu, S., Zhou, H., Qu, L.H. and Yang, J.H. (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic acids research*, **42**, D92-97.
75. R Development Core Team. (2012). R Foundation for Statistical Computing.

76. Xing, H., Liao, W., Mo, Y. and Zhang, M.Q. (2012) A novel Bayesian change-point algorithm for genome-wide analysis of diverse ChIPseq data types. *Journal of visualized experiments : JoVE*, e4273.
77. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome biology*, **9**, R137.
78. Harter, S.P. (1975) Probabilistic Approach to Automatic Keyword Indexing .1. Distribution of Specialty Words in a Technical Literature. *J Am Soc Inform Sci*, **26**, 197-206.
79. Viterbi, A.J. (1967) Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *Ieee T Inform Theory*, **13**, 260-+.
80. Hall, D.B. (2000) Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics*, **56**, 1030-1039.
81. Wang, Y., Juranek, S., Li, H., Sheng, G., Tuschl, T. and Patel, D.J. (2008) Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature*, **456**, 921-926.
82. Hu, J.J., Li, B. and Kihara, D. (2005) Limitations and potentials of current motif discovery algorithms. *Nucleic acids research*, **33**, 4899-4913.
83. Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T. and Thun, M.J. (2008) Cancer statistics, 2008. *CA: a cancer journal for clinicians*, **58**, 71-96.
84. Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 13790-13795.
85. Loeb, G.B., Khan, A.A., Canner, D., Hiatt, J.B., Shendure, J., Darnell, R.B., Leslie, C.S. and Rudensky, A.Y. (2012) Transcriptome-wide miR-155 Binding Map Reveals Widespread Noncanonical MicroRNA Targeting. *Molecular cell*.
86. (!!! INVALID CITATION !!!).
87. Bardet, A.F., He, Q., Zeitlinger, J. and Stark, A. (2012) A computational pipeline for comparative ChIP-seq analyses. *Nature protocols*, **7**, 45-61.
88. Wang, X., Zang, M. and Xiao, G. (2012) Epigenetic change detection and pattern recognition via Bayesian hierarchical hidden Markov models. *Statistics in medicine*.
89. Soon, W.W., Hariharan, M. and Snyder, M.P. (2013) High-throughput sequencing for biology and medicine. *Molecular systems biology*, **9**, 640.
90. Hardcastle, T.J. (2013) High-throughput sequencing of cytosine methylation in plant DNA. *Plant methods*, **9**, 16.
91. Xu, H. and Sung, W.K. (2012) Identifying differential histone modification sites from ChIP-seq data. *Methods in molecular biology*, **802**, 293-303.
92. Nair, N.U., Sahu, A.D., Bucher, P. and Moret, B.M. (2012) ChIPnorm: a statistical method for normalizing and identifying differential regions in histone modification ChIP-seq libraries. *PloS one*, **7**, e39573.
93. Shao, Z., Zhang, Y., Yuan, G.C., Orkin, S.H. and Waxman, D.J. (2012) MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome biology*, **13**, R16.
94. Siegel, R., Naishadham, D. and Jemal, A. (2013) Cancer statistics, 2013. *CA: a cancer journal for clinicians*, **63**, 11-30.
95. Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M. and Ule, J. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, **17**, 909-915.
96. Smyth, G.K. and Speed, T. (2003) Normalization of cDNA microarray data. *Methods*, **31**, 265-273.
97. Rousseeuw, P.J. and Croux, C. (1993) Alternatives to the Median Absolute Deviation. *J Am Stat Assoc*, **88**, 1273-1283.

98. Ailliot, P., Thompson, C. and Thomson, P. (2011) Mixed methods for fitting the GEV distribution. *Water Resour Res*, **47**.
99. Ashley, C.T., Sutcliffe, J.S., Kunst, C.B., Leiner, H.A., Eichler, E.E., Nelson, D.L. and Warren, S.T. (1993) Human and murine FMR-1: alternative splicing and translational initiation downstream of the CGG-repeat. *Nature genetics*, **4**, 244-251.
100. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, **10**, R25.
101. Ozsolak, F. and Milos, P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics*, **12**, 87-98.
102. Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D.S., Busby, M.A., Berlin, A.M., Sivachenko, A., Thompson, D.A., Wysoker, A., Fennell, T. *et al.* (2013) Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nature methods*, **10**, 623-629.
103. Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129-141.
104. Skalsky, R.L., Corcoran, D.L., Gottwein, E., Frank, C.L., Kang, D., Hafner, M., Nusbaum, J.D., Feederle, R., Delecluse, H.J., Luftig, M.A. *et al.* (2012) The viral and cellular microRNA targetome in lymphoblastoid cell lines. *PLoS pathogens*, **8**, e1002484.
105. Gottwein, E., Corcoran, D.L., Mukherjee, N., Skalsky, R.L., Hafner, M., Nusbaum, J.D., Shamulailatpam, P., Love, C.L., Dave, S.S., Tuschl, T. *et al.* (2011) Viral microRNA targetome of KSHV-infected primary effusion lymphoma cell lines. *Cell host & microbe*, **10**, 515-526.
106. Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M. and Zavolan, M. (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature methods*, **8**, 559-564.
107. Lipchina, I., Elkabetz, Y., Hafner, M., Sheridan, R., Mihailovic, A., Tuschl, T., Sander, C., Studer, L. and Betel, D. (2011) Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. *Genes & development*, **25**, 2173-2186.
108. Ruocco, E., Ruocco, V., Tornesello, M.L., Gambardella, A., Wolf, R. and Buonaguro, F.M. (2013) Kaposi's sarcoma: etiology and pathogenesis, inducing factors, causal associations, and treatments: facts and controversies. *Clinics in dermatology*, **31**, 413-422.
109. Dittmer, D.P. and Damania, B. (2013) Kaposi sarcoma associated herpesvirus pathogenesis (KSHV)--an update. *Current opinion in virology*, **3**, 238-244.
110. Conrad, N.K. (2009) Posttranscriptional gene regulation in Kaposi's sarcoma-associated herpesvirus. *Advances in applied microbiology*, **68**, 241-261.
111. Majerciak, V. and Zheng, Z.M. (2009) Kaposi's sarcoma-associated herpesvirus ORF57 in viral RNA processing. *Frontiers in bioscience*, **14**, 1516-1528.
112. Swaminathan, S. (2005) Post-transcriptional gene regulation by gamma herpesviruses. *Journal of cellular biochemistry*, **95**, 698-711.
113. Sandri-Goldin, R.M. (2008) The many roles of the regulatory protein ICP27 during herpes simplex virus infection. *Frontiers in bioscience : a journal and virtual library*, **13**, 5241-5256.
114. Nekorchuk, M., Han, Z., Hsieh, T.T. and Swaminathan, S. (2007) Kaposi's sarcoma-associated herpesvirus ORF57 protein enhances mRNA accumulation independently of effects on nuclear RNA export. *Journal of virology*, **81**, 9990-9998.
115. Kirshner, J.R., Lukac, D.M., Chang, J. and Ganem, D. (2000) Kaposi's sarcoma-associated herpesvirus open reading frame 57 encodes a posttranscriptional regulator with multiple distinct activities. *Journal of virology*, **74**, 3586-3597.
116. Kang, J.G., Pripuzova, N., Majerciak, V., Kruhlak, M., Le, S.Y. and Zheng, Z.M. (2011) Kaposi's sarcoma-associated herpesvirus ORF57 promotes escape of viral and human interleukin-6 from microRNA-mediated suppression. *Journal of virology*, **85**, 2620-2630.

117. Jackson, B.R., Noerenberg, M. and Whitehouse, A. (2014) A novel mechanism inducing genome instability in Kaposi's sarcoma-associated herpesvirus infected cells. *PLoS pathogens*, **10**, e1004098.
118. Boyne, J.R., Jackson, B.R., Taylor, A., Macnab, S.A. and Whitehouse, A. (2010) Kaposi's sarcoma-associated herpesvirus ORF57 protein interacts with PYM to enhance translation of viral intronless mRNAs. *The EMBO journal*, **29**, 1851-1864.
119. Guo, J.U., Agarwal, V., Guo, H. and Bartel, D.P. (2014) Expanded identification and characterization of mammalian circular RNAs. *Genome biology*, **15**, 409.
120. Wang, Y. and Wang, Z. (2015) Efficient backsplicing produces translatable circular mRNAs. *Rna*, **21**, 172-179.
121. Starke, S., Jost, I., Rossbach, O., Schneider, T., Schreiner, S., Hung, L.H. and Bindereif, A. (2015) Exon circularization requires canonical splice signals. *Cell reports*, **10**, 103-111.
122. Conn, S.J., Pillman, K.A., Toubia, J., Conn, V.M., Salmanidis, M., Phillips, C.A., Roslan, S., Schreiber, A.W., Gregory, P.A. and Goodall, G.J. (2015) The RNA Binding Protein Quaking Regulates Formation of circRNAs. *Cell*, **160**, 1125-1134.
123. You, X., Vlatkovic, I., Babic, A., Will, T., Epstein, I., Tushev, G., Akbalik, G., Wang, M., Glock, C., Quedenau, C. *et al.* (2015) Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nature neuroscience*, **18**, 603-610.
124. Ashwal-Fluss, R., Meyer, M., Pamudurti, N.R., Ivanov, A., Bartok, O., Hanan, M., Evantal, N., Memczak, S., Rajewsky, N. and Kadener, S. (2014) circRNA biogenesis competes with pre-mRNA splicing. *Molecular cell*, **56**, 55-66.
125. Valdmanis, P.N. and Kay, M.A. (2013) The expanding repertoire of circular RNAs. *Molecular therapy : the journal of the American Society of Gene Therapy*, **21**, 1112-1114.
126. Li, Z., Huang, C., Bao, C., Chen, L., Lin, M., Wang, X., Zhong, G., Yu, B., Hu, W., Dai, L. *et al.* (2015) Exon-intron circular RNAs regulate transcription in the nucleus. *Nature structural & molecular biology*, **22**, 256-264.
127. Bachmayr-Heyda, A., Reiner, A.T., Auer, K., Sukhbaatar, N., Aust, S., Bachleitner-Hofmann, T., Mesteri, I., Grunt, T.W., Zeillinger, R. and Pils, D. (2015) Correlation of circular RNA abundance with proliferation--exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues. *Scientific reports*, **5**, 8057.
128. Gao, Y., Wang, J. and Zhao, F. (2015) CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome biology*, **16**, 4.
129. Glazar, P., Papavasileiou, P. and Rajewsky, N. (2014) circBase: a database for circular RNAs. *Rna*, **20**, 1666-1670.
130. Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.
131. Mouse, E.C., Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D.M., Groudine, M., Bender, M., Kaul, R. *et al.* (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome biology*, **13**, 418.
132. mod, E.C., Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L. *et al.* (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, **330**, 1787-1797.
133. Westholm, J.O., Miura, P., Olson, S., Shenker, S., Joseph, B., Sanfilippo, P., Celniker, S.E., Graveley, B.R. and Lai, E.C. (2014) Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell reports*, **9**, 1966-1980.
134. Rybak-Wolf, A., Stottmeister, C., Glazar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R. *et al.* (2015) Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Molecular cell*, **58**, 870-885.
135. Anko, M.L. and Neugebauer, K.M. (2012) RNA-protein interactions in vivo: global gets specific. *Trends in biochemical sciences*, **37**, 255-262.

136. Fisette, J.F., Toutant, J., Dugre-Brisson, S., Desgroseillers, L. and Chabot, B. (2010) hnRNP A1 and hnRNP H can collaborate to modulate 5' splice site selection. *Rna*, **16**, 228-238.
137. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545-15550.
138. Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, **34**, 267-273.
139. Li, J.-H., Liu, S., Zheng, L.-L., Wu, J., Sun, W.-J., Wang, Z.-L., Zhou, H., Qu, L.-H. and Yang, J.-H. (2015) Discovery of protein-lncRNA interactions by integrating large-scale CLIP-Seq and RNA-Seq datasets. *Frontiers in Bioengineering and Biotechnology*, **2**.
140. Helwak, A., Kudla, G., Dudnakova, T. and Tollervey, D. (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, **153**, 654-665.
141. Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M. and Lee, J.T. (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular cell*, **40**, 939-953.
142. Riley, K.J. and Steitz, J.A. (2013) The "Observer Effect" in genome-wide surveys of protein-RNA interactions. *Molecular cell*, **49**, 601-604.
143. Gregersen, L.H., Schueler, M., Munschauer, M., Mastrobuoni, G., Chen, W., Kempa, S., Dieterich, C. and Landthaler, M. (2014) MOV10 Is a 5' to 3' RNA helicase contributing to UPF1 mRNA target degradation by translocation along 3' UTRs. *Molecular cell*, **54**, 573-585.