CLIPSeqTools—a novel bioinformatics CLIP-seq analysis suite

MANOLIS MARAGKAKIS,^{1,2} PANAGIOTIS ALEXIOU,^{1,2} TADASHI NAKAYA,³ and ZISSIMOS MOURELATOS^{1,2}

¹Department of Pathology and Laboratory Medicine, Division of Neuropathology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

²PENN Genome Frontiers Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

³Laboratory of Neuroscience, Graduate School of Pharmaceutical Sciences, Hokkaido University, Sapporo 060-0812, Japan

ABSTRACT

Immunoprecipitation of RNA binding proteins (RBPs) after *in vivo* crosslinking, coupled with sequencing of associated RNA footprints (HITS-CLIP, CLIP-seq), is a method of choice for the identification of RNA targets and binding sites for RBPs. Compared with RNA-seq, CLIP-seq analysis is widely diverse and depending on the RBPs that are analyzed, the approaches vary significantly, necessitating the development of flexible and efficient informatics tools. In this study, we present CLIPSeqTools, a novel, highly flexible computational suite that can perform analysis from raw sequencing data with minimal user input. It contains a wide array of tools to provide an in-depth view of CLIP-seq data sets. It supports extensive customization and promotes improvisation, a critical virtue, since CLIP-seq analysis is rarely well defined *a priori*. To highlight CLIPSeqTools capabilities, we used the suite to analyze Ago-miRNA HITS-CLIP data sets that we prepared from human brains.

Keywords: Argonaute; CLIP-seq; HITS-CLIP; RNA binding protein; miRNA

INTRODUCTION

CLIP-seq or HITS-CLIP, which stands for high-throughput sequencing of RNA that is isolated after ultraviolet (UV) irradiation-induced crosslinking and immunoprecipitation (Chi et al. 2009; Moore et al. 2014), and related methods, such as individual-nucleotide resolution iCLIP (Konig et al. 2010), photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) (Hafner et al. 2010), and FAST-iCLIP (Flynn et al. 2015), are powerful techniques used to identify *in vivo* targets of RNA binding proteins (RBPs) that bind directly to RNAs or via an RNA guide, as is typically the case for Argonaute family proteins comprised of Ago and Piwi subclades (Liu et al. 2008; Kim et al. 2009). Agos are ubiquitously expressed and bind microRNAs or small interfering RNAs (siRNAs) to form RNA-induced silencing complexes (RISC)/miRNA Ribonucleoproteins (miRNPs) (Hammond et al. 2001; Mourelatos et al. 2002; Sontheimer 2005; Paroo et al. 2007) and use the bound miRNA or siRNA as a guide to identify and silence target RNAs (Lee et al. 1993; Elbashir et al. 2001; Hammond et al. 2001; Lewis et al. 2003; Kiriakidou et al. 2004; Rajewsky and Succi 2004). CLIP-seq has shown that miRNAs bind preferentially to the 3' untranslated region (UTR) and coding sequence (CDS) of protein-coding genes (Chi et al. 2009; Hafner et al. 2010; Zisoulis et al. 2010). Piwi proteins are expressed primarily in the germline, bind piRNAs and use them

to target and silence transposons (Seto et al. 2007; Klattenhoff and Theurkauf 2008; Ghildiyal and Zamore 2009; Senti and Brennecke 2010; Castaneda et al. 2011; Siomi et al. 2011; Ishizu et al. 2012; Pillai and Chuma 2012; Guzzardo et al. 2013; Weick and Miska 2014). CLIP-seq for Piwi proteins revealed that they may also bind directly to long RNAs without using piRNAs as guides, either in the process of piRNA biogenesis or to bind to mRNAs (Vourekas et al. 2012, 2015).

Although CLIP-seq involves laborious experimental work, the bottleneck and the associated cost are primarily related to managing and analyzing the acquired data. In contrast to RNA-seq, for which there are well-defined analysis outputs and several established tools (Katz et al. 2010; Anders et al. 2013), CLIP-seq usually involves analyses that are inherently nuanced and, more often than not, widely diverse (Chi et al. 2009; Cho et al. 2012; Huelga et al. 2012; Vourekas et al. 2012, 2015; Ibrahim et al. 2013; Nakaya et al. 2013; Chu et al. 2015). CLIP-seq analyses usually start with a series of well-defined processing steps such as adaptor trimming and alignment of the sequenced reads on a reference genome. However, the steps following the alignment usually need to be tailored and customized to specifically address the biological questions at hand. Such questions can vary greatly depending on the studied protein and can involve analysis steps that have never been addressed before and complex data queries.

© 2015 Maragkakis et al. This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://rnajournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Corresponding author: mourelaz@uphs.upenn.edu

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.052167.115>.

In recent years, several tools have been published that perform parts of CLIP-seq analyses. dCLIP (Wang et al. 2014b) is a command line tool that uses a two-stage approach comprised of a normalization method and a hidden Markov model to identify differences in binding site intensity between CLIP-seq libraries. It focuses on identifying differential binding levels of common binding sites between two conditions. miCLIP (Wang et al. 2014a) consists of an R package together with a companion Galaxy interface and attempts to identify high-confidence binding sites from CLIP-seq data. It assigns a probability score for each potential binding site to prioritize subsequent validation experiments. PIPE-CLIP (Chen et al. 2014) is also a Galaxy tool that has a preprocessing step in which duplicate sequence reads are discarded. It uses the number of reads mapping in a region and characteristic mutations of the reads (Hafner et al. 2010; Zhang and Darnell 2011; Moore et al. 2014) to identify binding sites. Though this is a valid approach, there are caveats since the use of mutations might introduce biases such as the preference for TTT motifs in HITS-CLIP data (Sugimoto et al. 2012). Evidently, identifying binding sites is an important task but it also has a quite limited analysis scope. In the general case, a CLIP-seq experiment can reveal positional information for RBP sites but it can also reveal structural and mechanistic insight for the bound protein/RNA complexes (Huelga et al. 2012; Vourekas et al. 2012, 2015; Weyn-Vanhentenryck et al. 2014). Moreover, RBP function might differ depending on the genomic context, and a generic CLIP-seq analysis tool has to provide a straightforward mechanism to analyze data sets under different criteria and parameters. For example, reads and correspondingly binding sites that map within repeat regions are usually analyzed separately from reads that map within genic areas (Vourekas et al. 2012, 2015). None

of the published tools has clear mechanisms to support this type of parameterization and filtering.

Similarly there are other tools, like pyCRAC (Webb et al. 2014), HTSeq (Anders et al. 2015), and DESeq (Anders and Huber 2010) with additional functionalities, which can be used in the CLIP-seq context. Here, we present ClipSeqTools, a computational suite whose aim is to address certain shortcomings of CLIP-seq analyses. It is an easy to use, flexible tool that accepts raw reads as input and with minimal user parameterization, can provide a wide overview of the analyzed CLIP-seq data sets. To highlight the suite's capabilities, we used CLIPSeqTools on four novel Ago HITS-CLIP data sets to identify the *in vivo* miRNA-mRNA interactions in human brain. We exploited the flexibility of CLIPSeqTools to analyze these libraries using several different parameters in order to extract biological insight for varying genomic annotations and to evaluate the effect of using functional metrics such as evolutionary conservation on CLIP-seq analysis. To underline the flexibility of CLIPSeqTools, most of the figures in the manuscript are either direct output of the suite or have been minimally processed with common tools such as Excel.

RESULTS

The output of a CLIP-seq experiment (Fig. 1A) is typically one or more data files, most often in the FastQ format that contain sequenced reads. CLIPSeqTools is a suite of tools for the analysis of such data sets. The suite has a modular structure and comprises three toolboxes (Fig. 1B). The first toolbox, *clipseqtools-preprocess*, is used to process the raw sequencing data and prepare the files required by the subsequent tools. The second, *clipseqtools*, contains the analysis tools that will extract biological insight from the data sets. The third,

FIGURE 1. CLIPSeqTools workflow. (A) Schematic of a typical CLIP-seq experiment. After crosslinking, immunoprecipitation, and RNA isolation, cDNA libraries are prepared and sequenced and a file (usually FastQ) is created. (B) Schematic of CLIPSeqTools. The suite uses a FastQ file as input. The *clipseqtools-preprocess* toolbox can process the input file, do the alignment of the reads on a reference genome, and prepare all the data for the subsequent toolboxes. *clipseqtools* and *clipseqtools-compare* toolboxes can perform multiple analyses that cover a wide spectrum and export tables with results and figures for visualization.

clipseqtools-compare, is used to compare data sets with each other. To support users that are not comfortable running individual commands, the suite offers three easy-to-use pipelines, one for each toolbox that will run a preconfigured set of processing and analysis on the input data sets. These pipelines will use the raw CLIP-seq data as input and output a wide array of analyses. More experienced users who feel comfortable parameterizing command line applications, can run each tool independently and tailor analyses to specific needs or ask complex questions on specific subsets of the data sets.

Processing of CLIP-seq data sets

We developed *clipseqtools-preprocess* (Fig. 1B) to process the raw sequencing files produced by sequencers and prepare the data required by the analysis tools. To make the toolbox more flexible each of the included tools can run independently but we have also bundled them in a preconfigured pipeline to be used by more inexperienced users. The pipeline will first remove adaptor sequences from the 3' end of reads (*cut_adaptor*) using the cutadapt program (Martin 2011). This step is required to improve mapping efficiency during the alignment step. The alignment step (*star_alignment*) is done using STAR, (Dobin et al. 2013) which is one of the most efficient aligners that support spliced alignments. Next, the alignment files are sanitized by removing unmapped reads and keeping one random mapping position for each read that maps equally well in more than one position (*cleanup_alignment*). In the case of Ago or Piwi proteins, these unmapped reads may be processed further to identify guide RNA–target RNA hybrid duplexes produced during the ligation steps, which may further aid identification of sites of RNA–RNA interactions (Helwak et al. 2013). The pipeline will then load the alignments on an SQLite database and will build the appropriate indexes based on the chromosome and alignment location for optimized range queries (*sam_to_sqlite*). Importantly, the aligned reads are then annotated with functional, genomic and positional information. Specifically, they are tagged based on whether they are part of genes, transcripts, coding transcripts, exons, introns, 5' UTRs, coding sequences, 3' UTRs, and repeat elements (*annotate_with_genic_elements*). They are also assigned a conservation score that is calculated as the average PhastCons (Siepel et al. 2005) conservation score in the genomic area where they are mapped (*annotate_with_conservation*). In addition, custom annotations can be applied if given in the SAM or BED file formats (*annotate_with_file*). All annotations can later be used to filter the CLIP-seq data and execute any analysis on custom subsets of the data. The toolbox also contains some helper tools that can be useful under certain conditions such as to process reads that contain barcodes or to collapse PCR duplicates (*trim_fastq*, *collapse_fastq*) (Fu et al. 2014).

Analysis on a single CLIP-seq data set

clipseqtools (Fig. 1B) offers a wide collection of tools and covers many aspects of CLIP-seq analysis. It uses the files pro-

duced by the preprocess toolbox and generates tables with results that can be imported into spreadsheets for additional processing by the user. Most of the tools also produce high quality graphs to visualize the results and facilitate the extraction of biological insight. Similar to the preprocessing toolbox, each of the included tools can run independently or through a preconfigured pipeline. The pipeline for this toolbox measures the read size distribution (*size_distribution*), the nucleotide composition along the reads (*nucleotide_composition*) and creates plots for visualization of the distributions. It also measures the read genomic distribution by counting the reads that map within genes, repeat regions, exons, introns, 5' UTRs, etc. (*genomic_distribution*) and measures the percent of genome that is covered by reads (*genome_coverage*). To evaluate the effect on the transcriptome, it measures binding levels by counting reads mapped on genes, gene transcripts, exons, and introns (*count_reads_on_genic_elements*) and also it identifies putative positional binding preferences within genic elements (5' UTR, coding sequence, 3' UTR, exons, and introns) by measuring how reads are distributed along their length (*distribution_on_genic_elements*, *distribution_on_introns_exons*). In addition, the toolbox assembles read-clusters, measures statistics for their size and number of contained reads (*cluster_size_and_score_distribution*), and creates histograms summarizing the results. It also measures motif enrichment within the reads and compares them to background reads (*nmer_enrichment_over_shuffled*). To address potential binding along exon–exon junctions it collects and reports statistics for long alignment gaps that span introns (*reads_long_gaps_size_distribution*). Importantly, all of the tools can be selectively applied on custom subsets of the data to enable the extraction of biological insight for varying classes of data such as conserved versus nonconserved reads. Additionally, the toolbox contains helper tools to extract reads in a BED format to visualize them in a genome browser (*export_bed*) or to evaluate evolutionary conservation (*conservation_distribution*).

Comparison of two CLIP-seq data sets

clipseqtools-compare supports studies in which two or more data sets need to be compared. Similar to *clipseqtools*, all the tools generate tables that can be imported in spreadsheets for further processing and most of them also produce figures that summarize the results. Each of the tools can run independently or through a preconfigured pipeline. The pipeline for this toolbox measures the overlap of two data sets by counting the number of overlapping reads (*libraries_overlap_stats*). This estimates the binding potential similarities between two libraries and can reveal cases where antagonistic or synergistic effects are in place. To identify positional relationships between the binding sites of two proteins such as in cases when one protein binds preferentially upstream of/ downstream from another one, the pipeline measures the read density of one data set around the reads of another one

(*libraries_relative_read_density*). Finally, to identify potential differential binding between two conditions, the pipeline performs upper quartile normalization (Bullard et al. 2010) on all conditions combined and compares the binding levels for genes, gene transcripts, introns, and exons (*compare_counts*). It creates graphs to visualize the results and make changes easier to identify. In addition the toolbox contains a tool to extract the counts for genes, gene transcripts, introns, and exons for all compared libraries into a file format that is compatible with DESeq (Anders and Huber 2010) for further processing and evaluation of binding-level differences (*join_tables*). All tools can be selectively applied on custom subsets of the data enabling the comparison of not only different data sets but even on different subsets of the same data set.

CLIPSeqTools captures Ago binding on miRNAs and their target sites

We performed Ago HITS-CLIP on three human brain samples, derived from surgical temporal lobectomies. We gener-

ated three libraries, one for each human brain, that were size selected for larger reads (H: high) and were expected to be enriched for miRNA targets. Additionally, we prepared one library from one brain specimen that was size selected for smaller reads (L: low) and was expected to be enriched for miRNAs. We applied CLIPSeqTools on the Ago HITS-CLIP data sets to identify the in vivo miRNA-mRNA interactions and miRNA expression in human brain and to highlight the functionalities of our computational suite. In total, we analyzed ~45 million mapped reads.

Reads from libraries 1H, 2H, and 3H map primarily in genic areas and particularly within coding transcripts (~70%) with a strong preference for the coding region and the 3' UTR (Fig. 2A). In contrast, library L contains far fewer reads mapping to coding transcripts (~30%, data not shown) and unlike the H libraries, consists mostly of shorter reads, which is a direct reflection of the fact that H libraries are preferentially enriched for RNA targets while the L library contains mostly miRNAs (Fig. 2B). The specificity of the data sets is highlighted further by results of motif analysis on

FIGURE 2. CLIPSeqTools analysis for the binding profile of Ago HITS-CLIP. (A) Ago HITS-CLIP read genomic distribution. Colors correspond to annotation filtering. For example, reads overlapping repeat regions are excluded from all categories within the orange area (No Repeat). Error bars correspond to 1 SD for the 3 H libraries. The plot is based on the output of the *genomic_distribution* tool from *clipseqtools* toolbox. (B) Read size distribution for libraries 1H (dark blue), 2H (lighter blue), and 3H (light blue) that are enriched for miRNA targets and library L (red) that is enriched for miRNAs. The plot is based on the output of the *size_distribution* tool from *clipseqtools* toolbox. (C) Motif enrichment in library L. Violin plots are calculated for the enrichment *P*-values of miRNA seeds and all other 6mers. The plot is based on the output of the *nmer_enrichment_over_shuffled* tool from *clipseqtools* toolbox. (D) Nucleotide composition at the 5' end of reads from library L (red) and human miRNAs from miRBase (blue). The plot is based on the output of the *nucleotide_composition* tool from *clipseqtools* toolbox. (E) Read distribution in bins along the 5' UTR (red), CDS (orange), and 3' UTR (blue) of coding transcripts. Error bars correspond to 1 SD for the 3 H libraries. The plot is based on the output of the *distribution_on_genic_elements* tool from *clipseqtools* toolbox. (F) Read distribution in bins along the introns (red) and exons (orange) of coding transcripts. Error bars correspond to 1 SD for the 3H libraries. The plot is based on the output of the *distribution_on_introns_exons* tool from *clipseqtools* toolbox.

the L library, which reveals a high enrichment for miRNA seed sequences (nucleotides 2–7 from the miRNA 5' end) (Fig. 2C) and a strong preference for a uridine at the 5' end of reads, consistent with the similar preference of miRNAs at that position (Fig. 2D). The read distribution of the H libraries along the coding transcripts (Fig. 2E) shows an increased occupancy at the 5' and 3' end of 3' UTRs consistent with the known, increased miRNA binding site functionality in these areas (Grimson et al. 2007; Boudreau et al. 2014). Interestingly, we observed reduced occupancy at the 5' and 3' ends of the coding region suggesting reduced miRNA accessibility in these areas. The read distribution along introns indicates increased binding activity at the 5' end of intronic regions (Fig. 2F) and may suggest binding of Ago in transcripts with retained introns (Nakaya et al. 2013). However, it is important to note that the level of Ago binding in intronic areas is rather small (~10% of total reads) suggesting that even if such a mechanism exists its effects would be mild on the transcriptome.

Limited Ago recruitment on exon–exon junctions

Alternative splicing affects nearly 95% of mammalian genes (Kornblihtt et al. 2013). During mRNA maturation introns are removed and some of the exons of a gene are included or excluded from the final, processed mRNA. Through this process, multiple different transcripts can be produced by a single gene and depending on the combination of exons that are linked together miRNA binding sites may be incorporated or omitted from the final mRNA. To assess potential roles of splicing in influencing Ago recruitment we interrogated Ago binding on exon–exon junctions. We found that ~8% of the reads that map in genes have alignment gaps that span introns longer than 100 nt (Fig. 3A). To evaluate the significance of junction binding we assigned a new random binding position for each read on each gene transcript using an in house tool separate from CLIPSeqTools. For all three of the H libraries the percentage of random reads that overlapped with exon–exon junctions was significantly lower (P -value < 0.01, 1000 permutations) than the original reads, indicating limited recruitment of Ago on exon–exon junctions.

Effect of evolutionary conservation on Ago read cluster definition

CLIPSeqTools assembles clusters by merging overlapping reads and assigns the number of clustered reads as a cluster score. We noticed that ~80% of the created clusters contain only a few reads and most often a single or just two reads (Fig. 3B). Interestingly though, these low score clusters contain <1% of the total read space (Fig. 3B). We therefore hypothesized that such clusters and the corresponding reads represent either background nonspecific binding or reads wrongly identified due to unknown artifacts or sequencing errors.

Technical and biological CLIP replicates can be used to estimate such background. However, complex tissues will inherently have high variance that eventually results in reduced sensitivity for higher specificity. Another approach, possibly complementary to using replicates, is to use functional metrics to filter background. We set out to identify whether evolutionary conservation can effectively filter background, nonspecific binding from data sets and improve specificity. For this, we reran *clipseqtools* introducing a filtering step to keep only reads that are highly conserved across species. To distinguish highly and lowly conserved reads, we used a conservation score threshold of 500 (Fig. 3D). Interestingly, the analysis indicates that highly conserved reads are clustered more densely together, and the number of clusters with single or only few reads reduces by >20% (Fig. 3C). This indicates that, although rarely accounted for, evolutionary conservation can help to extract more specific results from CLIP-seqs.

Target enrichment in genomic elements

miRNPs depend on sequence complementarity to recognize miRNA targets. Specifically, the miRNA seed has been shown to play the most critical role in target recognition (Lewis et al. 2003; Kiriakidou et al. 2004). Therefore, we reasoned that the genomic classes (3' UTR, CDS, Intron, Intergenic) that are under heavier miRNA targeting would contain on average more sequences complementary to the seed (seed targets) of the most highly expressed miRNAs. To test this, we first measured the miRNA expression levels (Supplemental Table 1) by counting the number of reads of library L that are contained within the miRNA coordinates provided by miRBase. To evaluate the targeting load for each genomic class we measured the seed target enrichment in the reads of each class for the top one hundred expressed miRNAs and compared them with the enrichment for all other 6mers. Seed targets are most highly enriched in 3' UTRs while the CDS, intronic, and intergenic classes have a targeting load very similar to each other (Fig. 3E). This indicates that despite the fact that the CDS contains more reads, miRNPs mostly target 3' UTRs (Reczko et al. 2012). Interestingly, all classes score much higher than a random predictor indicating that the reads associated with these classes represent bona fide miRNA targets.

Ago CLIP-seq library comparisons

To evaluate the biological variability with respect to miRNA targeting between the three human brain samples we compared the 3 H libraries with each other. Out of all 1H reads, 30% overlap with 2H and 45% with 3H. Similarly, 40% and 38% of 2H reads overlap with 1H and 3H, while library 3H had 37% and 30% overlap with 1H and 2H, respectively. To assess potential differences between the three brain samples we measured differential binding on protein-coding genes for all pairwise combinations of the libraries.

FIGURE 3. Functional analysis and library comparisons of Ago HITS-CLIP with CLIPSeqTools. (A) Length distribution of alignment gaps in reads. The plot shows the percent of reads that contain alignment gaps with length as grouped on the x-axis. The plot is based on the output of the *reads_long_gaps_size_distribution* tool from *clipseqtools* toolbox. (B) Cluster and read distribution for varying cluster score groups. The plot shows for library 1H the percent of clusters with given cluster scores (black line) and the percent of reads that are contained in clusters of given cluster scores (red line). The plot is based on the output of the *cluster_size_and_score_distribution* tool from *clipseqtools* toolbox. (C) Average percent of assembled clusters for each cluster score. Blue bars correspond to all the reads while the red bars correspond to the highly conserved ones (PhastCons > 500). Error bars correspond to 1 SD for the 3H libraries. The plot is based on the output of the *cluster_size_and_score_distribution* tool from *clipseqtools* toolbox. (D) Evolutionary conservation distribution for reads of library 1H. The plot is based on the output of the *conservation_distribution* tool from *clipseqtools* toolbox. (E) 6mer enrichment in reads of library 1H that are contained in 3' UTRs (red), CDS (orange), intergenic regions (light blue), and introns (blue). All 6mers are ranked by enrichment value and annotated as positive if they correspond to the reverse complement of a seed of the top 100 expressed miRNAs. The y-axis shows the number of positive 6mers found up to each rank. The gray dashed line shows the theoretical plot for random ranking. (F) Correlation plot for gene Argonaute binding levels for all three high libraries. On the lower left side of the plot the circle size and color correspond to the Pearson correlation coefficient. The actual correlation coefficient values are written within the circles. The corresponding scatter plots for the gene Argonaute binding levels are presented on the upper right part of the plot. The plot is based on the output of the *compare_counts* tool from *clipseqtools-preprocess* toolbox.

The binding levels were normalized across libraries using upper quartile normalization (Bullard et al. 2010). We observed high variability among libraries for lowly targeted genes but much less variability for highly targeted ones (Fig. 3F; Supplemental Tables 2, 3, and 4). This indicates that although binding positions may vary between samples, the overall effect, especially for highly targeted genes, is stable.

DISCUSSION

CLIP-seq is emerging as a key experimental approach for the study of RNA binding proteins. In many cases, investigators

start analyzing CLIP-seq data sets without a priori knowledge of the inherent difficulties of analysis or without a clear path of how they should approach such analysis. In this work, we presented CLIPSeqTools, a novel, flexible, and easy to use suite that can perform a wide collection of analysis to provide a deep understanding of the analyzed protein's function. CLIPSeqTools provides an automated way to extract a very diverse collection of results starting from the raw sequencing files with minimal user customization. Additionally, we have highlighted the importance of functional metrics such as evolutionary conservation in CLIP-seq analysis. CLIPSeqTools makes the integration of such metrics in the analysis

straightforward, and this helps in the extraction of robust and possibly more biologically relevant results. CLIPSeqTools is suited to experienced and less experienced users and aims at minimizing the inertia of CLIP-seq adaptation as a valuable experimental approach.

MATERIALS AND METHODS

CLIPSeqTools implementation

CLIPSeqTools is an open source project written in the Perl and R programming languages and has been packaged for easy installation. The source code has been deposited in GitHub and complete documentation is available at <http://mourelatos.med.upenn.edu/clipseqtools/>.

CLIPSeqTools has been designed to be flexible and efficient. It can manage the amount of data produced by sequencing runs and also the diversity of the required analyses. To achieve this, we designed CLIPSeqTools to work with databases instead of flat files. Databases are designed to handle big data and are broadly supported and integrated within almost all major programming languages. In addition, they support efficient range queries that are particularly important in CLIP-seq analysis. Last but not least, they allow for the straightforward addition of extra annotation fields in an asynchronous way, which is not possible using flat files. CLIPSeqTools uses SQLite3 by default but supports other database engines as well.

CLIPSeqTools has been designed on top of the GenOO sequencing programming framework (Maragkakis et al. 2015). GenOO is an open-source, object-oriented Perl framework we developed specifically for the design and implementation of high-throughput sequencing analysis. GenOO models biological entities such as genes and transcripts as Perl objects, and includes relevant toolboxes, attributes and methods that allow for the manipulation of high-throughput sequencing data. Most importantly, GenOO integrates these elements in a simple and transparent way, which allows for the creation of complex analysis pipelines minimizing the overhead for the user.

The preprocessing toolbox of CLIPSeqTools uses internally the cutadapt and STAR aligner software to perform the adaptor trimming and the alignment of the reads, respectively. The default parameters and flags for these programs are *cutadapt* `<fastq_file> -n 3 -m 15 -e 0.25 -a <adaptor> -o <output_file>` and *STAR* `-genomeDir <star_genome> --readFilesIn <fastq_file> --runThreadN <threads> --outSAMattributes All --outFilterMultiMapScoreRange 0 --alignIntronMax 50000 --outFilterMatchNmin 15 --outFilterMatchNminOverLread 0.9 --outFileNamePrefix <output_prefix>`.

CLIPSeqTools runs

We run the CLIPSeqTools pipelines using several different filters on each data set. Initially we run the pipelines with the default parameters on the full data sets. To analyze reads in the 3' UTRs and CDS we filtered the reads to keep those that are contained in coding transcripts and not contained in repeat regions. Similarly, to analyze intergenic areas we filtered reads to keep those that map outside the boundaries of any transcript, repeat element, or miRNA genes.

Ago CLIP-seq experiment

HITS-CLIP was performed on three normal, live, human brain samples as previously described in Nakaya et al. (2013) with minor modifications: The 2A8 anti-Ago monoclonal antibody was tethered to Protein A Dynabeads (Invitrogen) that were precoated with rabbit anti-mouse IgG (Jackson ImmunoResearch). After washes, the 2A8-containing Dynabeads were used for CLIP. Following on-bead ligation of radiolabeled 3'-adaptor, samples were resolved by NuPAGE (Invitrogen), transferred to nitrocellulose and the cross-linked RNA-protein complexes were visualized by autoradiography. Two areas of the nitrocellulose filter were excised: the main radioactive signal, enriched in miRNAs, and a region just above it, enriched in larger RNAs (miRNA targets). After ligation of 5'-adaptor and RT-PCR, the cDNA library enriched in miRNAs (designated L: low) and the cDNAs libraries enriched in longer RNAs (designated H: high) were resolved in 8% PAGE gels, gel-purified, PCR reamplified, and sequenced by Illumina as described previously (Nakaya et al. 2013).

Ago CLIP-seq processing

Ago reads were processed using CLIPSeqTools. The adaptor was trimmed and the reads were aligned to the mm9 genome that was downloaded from UCSC. Then the reads were annotated with gene and RepeatMasker information also downloaded from the UCSC genome browser. The PhastCons data were downloaded from UCSC (<http://hgdownload.cse.ucsc.edu/goldenPath/mm9/phastCons30way/placental/>).

DATA DEPOSITION

Sequencing data were deposited at the Sequence Read Archive (SRA), project ID PRJNA299324.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank all members of the Mourelatos laboratory for helpful discussions. This work was supported by a Brody family fellowship to M.M. and by grants from the ALS Therapy Alliance and the National Institutes of Health (NIH) (GM072777) to Z.M.

Received April 13, 2015; accepted October 18, 2015.

REFERENCES

- Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.
- Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD. 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* **8**: 1765–1786.
- Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169.

- Boudreau RL, Jiang P, Gilmore BL, Spengler RM, Tirabassi R, Nelson JA, Ross CA, Xing Y, Davidson BL. 2014. Transcriptome-wide discovery of microRNA binding sites in human brain. *Neuron* **81**: 294–305.
- Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics* **11**: 94.
- Castaneda J, Genzor P, Bortvin A. 2011. piRNAs, transposon silencing, and germline genome integrity. *Mutat Res* **714**: 95–104.
- Chen B, Yun J, Kim MS, Mendell JT, Xie Y. 2014. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome Biol* **15**: R18.
- Chi SW, Zang JB, Mele A, Darnell RB. 2009. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460**: 479–486.
- Cho J, Chang H, Kwon SC, Kim B, Kim Y, Choe J, Ha M, Kim YK, Kim VN. 2012. LIN28A is a suppressor of ER-associated translation in embryonic stem cells. *Cell* **151**: 765–777.
- Chu Y, Wang T, Dodd D, Xie Y, Janowski BA, Corey DR. 2015. Intramolecular circularization increases efficiency of RNA sequencing and enables CLIP-Seq of nuclear RNA from human cells. *Nucleic Acids Res* **43**: e75.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.
- Elbashir SM, Lendeckel W, Tuschl T. 2001. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev* **15**: 188–200.
- Flynn RA, Martin L, Spitale RC, Do BT, Sagan SM, Zarnegar B, Qu K, Khavari PA, Quake SR, Sarnow P, et al. 2015. Dissecting noncoding and pathogen RNA-protein interactomes. *RNA* **21**: 135–143.
- Fu GK, Xu W, Wilhelmy J, Mindrinos MN, Davis RW, Xiao W, Fodor SP. 2014. Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. *Proc Natl Acad Sci* **111**: 1891–1896.
- Ghildiyal M, Zamore PD. 2009. Small silencing RNAs: an expanding universe. *Nat Rev Genet* **10**: 94–108.
- Grimson A, Farh KK, Johnston WK, Garrett-Engle P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27**: 91–105.
- Guzzardo PM, Muerdter F, Hannon GJ. 2013. The piRNA pathway in flies: highlights and future directions. *Curr Opin Genet Dev* **23**: 44–52.
- Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, et al. 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**: 129–141.
- Hammond SM, Boettcher S, Caudy AA, Kobayashi R, Hannon GJ. 2001. Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science* **293**: 1146–1150.
- Helwak A, Kudla G, Dudnakova T, Tollervey D. 2013. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153**: 654–665.
- Huelga SC, Vu AQ, Arnold JD, Liang TY, Liu PP, Yan BY, Donohue JP, Shiue L, Hoon S, Brenner S, et al. 2012. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep* **1**: 167–178.
- Ibrahim F, Maragkakis M, Alexiou P, Maronski MA, Dichter MA, Mourelatos Z. 2013. Identification of in vivo, conserved, TAF15 RNA binding sites reveals the impact of TAF15 on the neuronal transcriptome. *Cell Rep* **3**: 301–308.
- Ishizu H, Siomi H, Siomi MC. 2012. Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. *Genes Dev* **26**: 2361–2373.
- Katz Y, Wang ET, Airolidi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**: 1009–1015.
- Kim VN, Han J, Siomi MC. 2009. Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* **10**: 126–139.
- Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A. 2004. A combined computational-experimental approach predicts human microRNA targets. *Genes Dev* **18**: 1165–1178.
- Klattenhoff C, Theurkauf W. 2008. Biogenesis and germline functions of piRNAs. *Development* **135**: 3–9.
- Konig J, Zarnack K, Rot G, Curk T, Kayicki M, Zupan B, Turner DJ, Luscombe NM, Ule J. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**: 909–915.
- Kornblihtt AR, Schor IE, Alló M, Dujardin G, Petrillo E, Muñoz MJ. 2013. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol* **14**: 153–165.
- Lee RC, Feinbaum RL, Ambros V. 1993. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**: 843–854.
- Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. 2003. Prediction of mammalian microRNA targets. *Cell* **115**: 787–798.
- Liu X, Fortin K, Mourelatos Z. 2008. MicroRNAs: biogenesis and molecular functions. *Brain Pathol* **18**: 113–121.
- Maragkakis M, Alexiou P, Mourelatos Z. 2015. GenOO: a modern Perl framework for high throughput sequencing analysis. *bioRxiv* doi: 10.1101/019265.
- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17**: 10–12.
- Moore MJ, Zhang C, Gantman EC, Mele A, Darnell JC, Darnell RB. 2014. Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat Protoc* **9**: 263–293.
- Mourelatos Z, Dostie J, Paushkin S, Sharma A, Charroux B, Abel L, Rappsilber J, Mann M, Dreyfuss G. 2002. miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev* **16**: 720–728.
- Nakaya T, Alexiou P, Maragkakis M, Chang A, Mourelatos Z. 2013. FUS regulates genes coding for RNA-binding proteins in neurons by binding to their highly conserved introns. *RNA* **19**: 498–509.
- Paroo Z, Liu Q, Wang X. 2007. Biochemical mechanisms of the RNA-induced silencing complex. *Cell Res* **17**: 187–194.
- Pillai RS, Chuma S. 2012. piRNAs and their involvement in male germline development in mice. *Dev Growth Differ* **54**: 78–92.
- Rajewsky N, Socci ND. 2004. Computational identification of microRNA targets. *Dev Biol* **267**: 529–535.
- Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG. 2012. Functional microRNA targets in protein coding sequences. *Bioinformatics* **28**: 771–776.
- Senti KA, Brennecke J. 2010. The piRNA pathway: a fly's perspective on the guardian of the genome. *Trends Genet* **26**: 499–509.
- Seto AG, Kingston RE, Lau NC. 2007. The coming of age for Piwi proteins. *Mol Cell* **26**: 603–609.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.
- Siomi MC, Sato K, Pezic D, Aravin AA. 2011. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* **12**: 246–258.
- Sontheimer EJ. 2005. Assembly and function of RNA silencing complexes. *Nat Rev Mol Cell Biol* **6**: 127–138.
- Sugimoto Y, Konig J, Hussain S, Zupan B, Curk T, Frye M, Ule J. 2012. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol* **13**: R67.
- Vourekas A, Zheng Q, Alexiou P, Maragkakis M, Kirino Y, Gregory BD, Mourelatos Z. 2012. Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. *Nat Struct Mol Biol* **19**: 773–781.
- Vourekas A, Zheng K, Fu Q, Maragkakis M, Alexiou P, Ma J, Pillai RS, Mourelatos Z, Wang PJ. 2015. The RNA helicase MOV10L1 binds piRNA precursors to initiate piRNA processing. *Genes Dev* **29**: 617–629.

- Wang T, Chen B, Kim M, Xie Y, Xiao G. 2014a. A model-based approach to identify binding sites in CLIP-Seq data. *PLoS One* **9**: e93248.
- Wang T, Xie Y, Xiao G. 2014b. dCLIP: a computational approach for comparative CLIP-seq analyses. *Genome Biol* **15**: R11.
- Webb S, Hector RD, Kudla G, Granneman S. 2014. PAR-CLIP data indicate that Nrd1-Nab3-dependent transcription termination regulates expression of hundreds of protein coding genes in yeast. *Genome Biol* **15**: R8.
- Weick EM, Miska EA. 2014. piRNAs: from biogenesis to function. *Development* **141**: 3458–3471.
- Weyn-Vanhentenryck SM, Mele A, Yan Q, Sun S, Farny N, Zhang Z, Xue C, Herre M, Silver PA, Zhang MQ, et al. 2014. HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep* **6**: 1139–1152.
- Zhang C, Darnell RB. 2011. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol* **29**: 607–614.
- Zisoulis DG, Lovci MT, Wilbert ML, Hutt KR, Liang TY, Pasquinelli AE, Yeo GW. 2010. Comprehensive discovery of endogenous Argonaute binding sites in *Caenorhabditis elegans*. *Nat Struct Mol Biol* **17**: 173–179.

RNA

A PUBLICATION OF THE RNA SOCIETY

CLIPSeqTools—a novel bioinformatics CLIP-seq analysis suite

Manolis Maragkakis, Panagiotis Alexiou, Tadashi Nakaya, et al.

RNA 2016 22: 1-9 originally published online November 17, 2015

Access the most recent version at doi:[10.1261/rna.052167.115](https://doi.org/10.1261/rna.052167.115)

Supplemental Material

<http://rnajournal.cshlp.org/content/suppl/2015/11/10/rna.052167.115.DC1>

References

This article cites 56 articles, 13 of which can be accessed free at:

<http://rnajournal.cshlp.org/content/22/1/1.full.html#ref-list-1>

Creative Commons License

This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://rnajournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to *RNA* go to:

<http://rnajournal.cshlp.org/subscriptions>