

miR-eCLIP Data Review Guide

Introduction

Once miR-eCLIP samples have been prepped and sequenced, the resulting data is processed by our expert bioinformatics scientists to generate a dataset rich in information for miRNA-mRNA target sites. The miR-eCLIP data processing pipeline begins with UMI (unique molecular identifiers) trimming and adapter trimming of raw sequencing reads, then reads are filtered of repetitive genome elements such as rRNA and aligned to the reference genome (ie. human, mouse, etc.). For the non-chimeric reads that align to the genome, PCR duplicates are removed and clusters of reads are identified to define peaks in the data representing AGO2 binding sites. Peaks and genes containing peaks are analyzed in detail to reveal AGO2 binding features. Any reads not aligned initially are further processed to identify chimeric miRNA-mRNA reads. Efficiency of the chimeric miRNA to mRNA ligation is low, so only a small fraction of the reads will remain for chimeric read analysis. First, miRNAs are mapped to the set of putative chimeric reads and the best miRNA alignment is assigned to a read. The miRNA portion of reads is removed, and the remaining mRNA portion of the chimeric reads is mapped to the genome. Finally, PCR duplicates are removed, and peaks are called and annotated with the gene and feature information, as well as the number of chimeric reads for each miRNA that map to that mRNA peak (**Figure 1**). Following data analysis, a user will receive a login and key to download several data files from our secure SFTP server, including intermediate data and detailed reports summarizing the results for each sample in the miR-eCLIP experiment. The miR-eCLIP data deliverables can be complex; to assist in understanding the rich dataset delivered, this data review guide provides a step-by-step explanation of the results, describing the different components of the figures and tables in the final HTML reports and each data file type delivered. For additional guidance in understanding Eclipse Bio's data deliverables, please contact techsupport@eclipsebio.com.

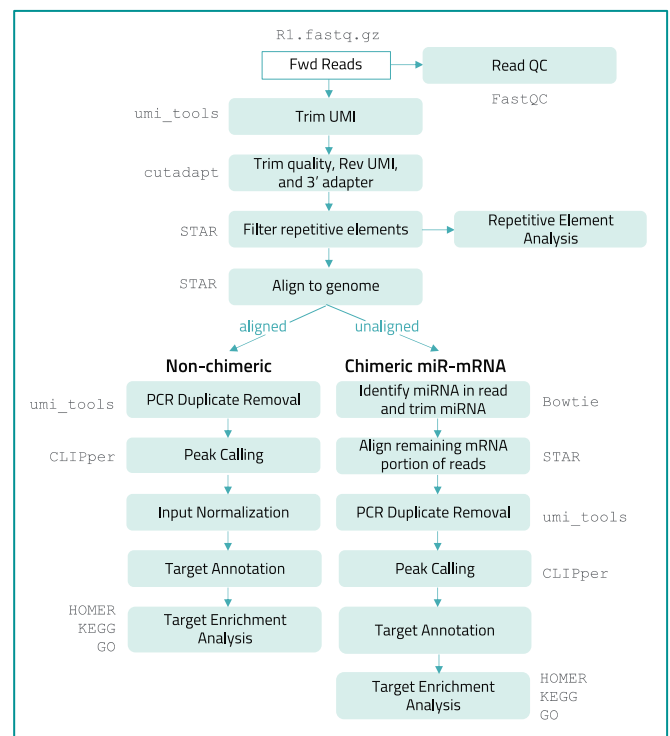


Figure 1. miR-eCLIP data analysis pipeline. Each analysis step is ordered from top to bottom with the publicly available tool used listed to the side of the step where available.

miR-eCLIP HTML Report

Two HTML reports are generated for each sample in the miR-eCLIP experiment; one provides an overview of the identified peaks from non-chimeric AGO2 reads with enrichment in the immunoprecipitation ("IP") versus a size-matched input control ("input"), and the other provides an overview of the identified peaks from miRNA-mRNA chimeric reads in the IP sample. The non-chimeric AGO2 HTML report summarizes only the highest confidence set of

enriched peaks within the dataset, with \log_2 fold change ≥ 3 and $-\log_{10}(\text{p-value}) \geq 3$. Additional true peaks may exist below these cutoffs, however using only the highest confidence set minimizes false positive signals in the results.

For miR-eCLIP experiments containing 2 or 3 replicates, an additional HTML report will be created summarizing the reproducible non-chimeric AGO2 peaks identified across sample replicates using Eclipse Bio IDR (Irreproducible Discovery Rate) analysis. The IDR analysis does not include chimeric reads. For IDR peaks \log_2 fold change values in the report correspond to the geometric mean of the \log_2 fold change across replicates, and the p-value corresponds to the minimum p-value across replicates. IDR HTML reports contain most of the plots and tables listed below but exclude the “Information scores for peak features pie chart”, the “Swarm plot of peak \log_2 fold changes for each feature”, and the “IP Repetitive Element Mapping Information Table”. For more information on IDR analysis refer to the **Data File Descriptions** section below.

1. Annotation and Feature Types

Peaks are annotated using transcript information from GENCODE or Ensembl. Each annotated transcript region is labeled with specific annotation feature types, first split by coding and non-coding transcripts, then by transcript regions, and then by intron/exon proximity regions. For overlapping transcript regions, the following hierarchy is used to label the region: coding sequence (CDS), 5' or 3' untranslated region (UTR), intron, non-coding exon, then non-coding intron. For example, if a region has a UTR of one transcript that overlaps an intron of another transcript, the region will be labeled UTR. For intron/exon proximity feature types, the definitions are as follows (**Figure 2**):

miRNA proximal: within 500 bp of an annotated miRNA

5' splice site (5' SS): within the first 100 bp of an intron (5' to 3' direction)

3' splice site (3' SS): within the last 100 bp of an intron (5' to 3' direction)

proximal intron: within 100 bp to 500 bp of the nearest exon

distal intron: greater than 500 bp away from the nearest exon

For non-coding transcripts there are two additional feature categories: **miRNA** and **tRNA**. Other annotations corresponding to non-coding transcripts will be labeled as **noncoding exon**, or with an intronic feature type from above followed by **(ncRNA)**.

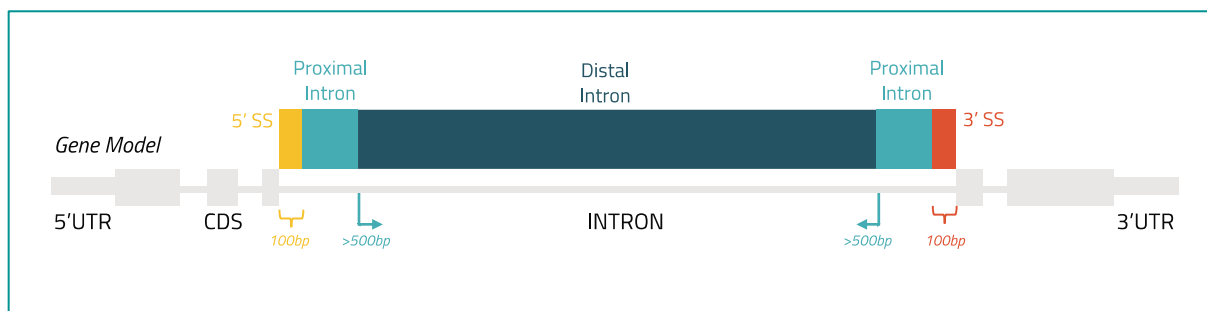
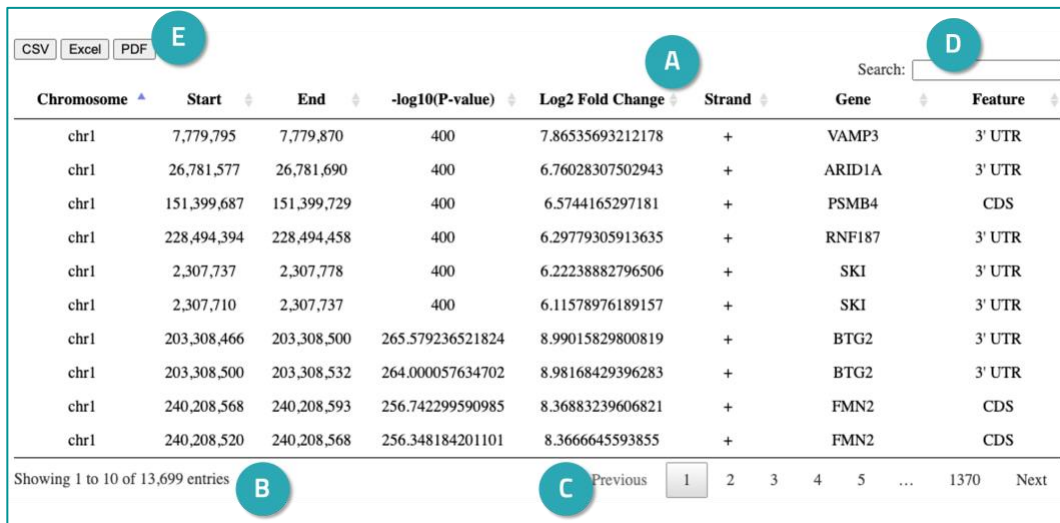


Figure 2. Feature type examples labeled on a gene model.

2. Peak information table

The peak information table contains the genome location of each non-chimeric AGO2 peak, the $-\log_{10}(\text{p-value})$, \log_2 fold change of IP vs. input, and the annotated gene and feature for that region (see **Annotation and Feature Type** details above).



The screenshot shows a web interface for the Peak Information Table. At the top left, there are buttons for 'CSV', 'Excel', and 'PDF' (labeled E). At the top right, there is a search bar (labeled D). The table has columns: 'Chromosome', 'Start', 'End', '-log10(P-value)', 'Log2 Fold Change' (labeled A), 'Strand', 'Gene', and 'Feature'. The table displays 10 rows of data. At the bottom left, it says 'Showing 1 to 10 of 13,699 entries' (labeled B). At the bottom right, there is a page navigator with 'Previous', '1', '2', '3', '4', '5', '...', '1370', and 'Next' (labeled C).

Chromosome	Start	End	$-\log_{10}(\text{P-value})$	Log2 Fold Change	Strand	Gene	Feature
chr1	7,779,795	7,779,870	400	7.86535693212178	+	VAMP3	3' UTR
chr1	26,781,577	26,781,690	400	6.76028307502943	+	ARID1A	3' UTR
chr1	151,399,687	151,399,729	400	6.5744165297181	+	PSMB4	CDS
chr1	228,494,394	228,494,458	400	6.29779305913635	+	RNF187	3' UTR
chr1	2,307,737	2,307,778	400	6.22238882796506	+	SKI	3' UTR
chr1	2,307,710	2,307,737	400	6.11578976189157	+	SKI	3' UTR
chr1	203,308,466	203,308,500	265.579236521824	8.99015829800819	+	BTG2	3' UTR
chr1	203,308,500	203,308,532	264.000057634702	8.98168429396283	+	BTG2	3' UTR
chr1	240,208,568	240,208,593	256.742299590985	8.36883239606821	+	FMN2	CDS
chr1	240,208,520	240,208,568	256.348184201101	8.3666645593855	+	FMN2	CDS

- A. Table Sorting:** Peaks are best sorted by \log_2 fold change to focus on the most enriched peaks over the input sample. To do so, simply click the arrows in the header of the table next to this column and it will resort by this value. This example shows the default sorting, with peaks listed in Chromosome order.
- B. Number of Entries:** The table contains as many peak entries as listed at the bottom left corner, displaying 10 entries at a time. This example has 13,699 total entries.
- C. Page Navigator:** The entries in the table can be viewed using the *Previous* and *Next* buttons on the bottom right. The next or previous 10 entries will load.
- D. Search:** The table is also searchable with keywords in order to list peaks on specific genes or features of interest. Multiple keywords can be searched with a space to separate each; for example, entering "VAMP3 3'UTR" into the search bar will return all 3'UTR peaks on the gene VAMP3.
- E. File Export:** Selecting the buttons on the top left will export the table to a file in the selected format (CSV, Excel, or PDF). Exporting the table will preserve any sorting or filtering that has been performed. For example, searching for "3'UTR" and then exporting the table will only export the peaks on 3' UTRs.

3. Chimeric peak information table

The chimeric peak information table contains the genome location of each peak, the annotated gene name and Ensembl ID, the feature for that region (see **Annotation and Feature Type** details above), the specific miRNA targeting this location, and the number of chimeric reads for that miRNA in the peak. Each entry in the table describes the chimeric reads assigned to a single miRNA contained in a peak location. A peak will have multiple entries when more than one miRNA has chimeric reads within the peak. The number of chimeric reads are summed as a the fraction of the chimeric read overlapping the peak location, so if only 50% of a read overlaps, a count of 0.5 is added for that read.

CSV Excel PDF **E**

D Search:

Chromosome	Start	End	Strand	Gene	Ensembl ID	Feature	miRNA	Number of chimeric reads
chr1	52,086,842	52,087,008	+	BTF3L4	ENSG00000134717.18	3' UTR	hsa-miR-194-5p	2,017
chr1	52,086,842	52,087,008	+	BTF3L4	ENSG00000134717.18	3' UTR	hsa-miR-196b-5p	0.833
chr1	52,086,842	52,087,008	+	BTF3L4	ENSG00000134717.18	3' UTR	hsa-miR-19a-3p	22.785
chr1	52,086,842	52,087,008	+	BTF3L4	ENSG00000134717.18	3' UTR	hsa-miR-19b-3p	14.848
chr1	52,086,842	52,087,008	+	BTF3L4	ENSG00000134717.18	3' UTR	hsa-miR-1307-5p	1.000
chr1	52,086,842	52,087,008	+	BTF3L4	ENSG00000134717.18	3' UTR	hsa-miR-20a-5p	1.000
chr1	52,086,842	52,087,008	+	BTF3L4	ENSG00000134717.18	3' UTR	hsa-miR-26a-5p	2.000
chr1	52,086,842	52,087,008	+	BTF3L4	ENSG00000134717.18	3' UTR	hsa-miR-17-5p	0.957
chr1	75,787,516	75,787,574	+	RABGGTB	ENSG00000137955.16	CDS	hsa-miR-20a-5p	26.927
chr1	75,787,516	75,787,574	+	RABGGTB	ENSG00000137955.16	CDS	hsa-miR-423-3p	1.113

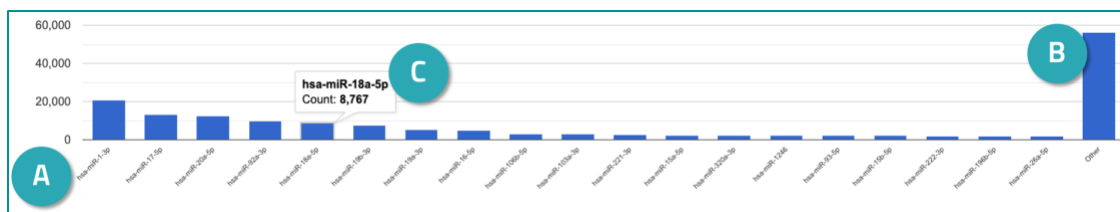
Showing 1 to 10 of 24,679 entries **B**

Previous 1 2 3 4 5 ... 2468 Next **C**

- A. Table Sorting:** Chimeric read peaks are best sorted by *Number of chimeric reads* to focus on the highest confidence miRNA targets. To do so, simply click the arrows in the header of the table next to this column and it will resort by this value. This example shows the default sorting, with peaks listed in Chromosome order.
- B. Number of Entries:** The table contains as many entries as listed at the bottom left corner, displaying 10 entries at a time. There may be multiple entries per chimeric read peak if multiple miRNAs target that peak location. This example has 24,679 total entries.
- C. Page Navigator:** The entries in the table can be viewed using the *Previous* and *Next* buttons on the bottom right. The next or previous 10 entries will load.
- D. Search:** The table is also searchable with keywords in order to list peaks on specific genes or features of interest. Multiple keywords can be searched with a space to separate each; for example, entering "VAMP3 3'UTR" into the search bar will return all 3'UTR peaks on the gene VAMP3.
- E. File Export:** Selecting the buttons on the top left will export the table to a file in the selected format (CSV, Excel, or PDF). Exporting the table will preserve any sorting or filtering that has been performed. For example, searching for "3'UTR" and then exporting the table will only export the peaks on 3' UTRs.

4. Distribution of miRNAs in the chimeric sample

The bar plot is only provided for chimeric peak HTML reports and displays the chimeric read distribution across the top 20 miRNAs assigned to chimeric reads. This plot is useful to highlight whether specific miRNAs are dominating the chimeric miRNA-mRNA reads in a sample and is particularly useful when certain miRNAs are enriched or overexpressed in a miR-eCLIP sample.



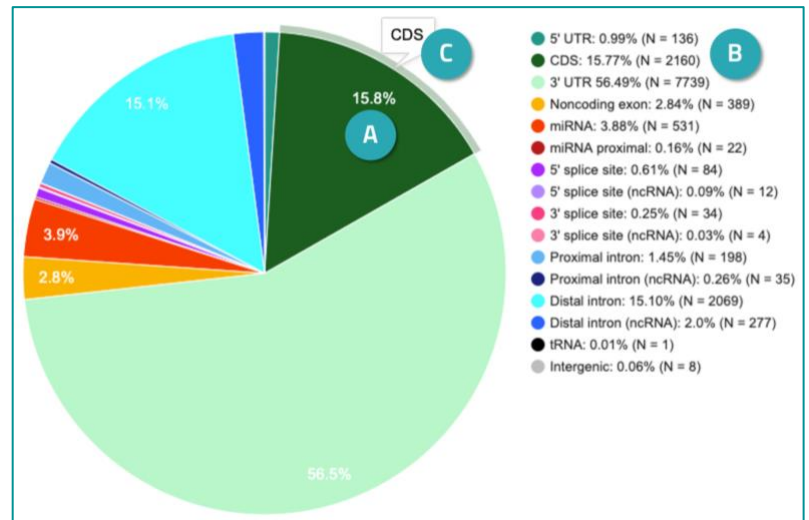
- A. X-axis miRNAs:** Each bar is the count of total chimeric reads assigned to the labeled miRNA.
- B. Other:** The miRNAs with the top 20 chimeric read counts are shown, with the remainder of chimeric reads summed in the *Other* bar on the right.

- C. Hover:** Hovering the cursor on each bar will display the chimeric read count for that miRNA. In this example, hsa-miR-18a-5p has 8,767 chimeric read counts.

5. Relative frequency of peaks that map to each feature pie chart

The pie chart depicts the relative frequency of peaks that map to each feature type (see **Annotation and Feature Type** details above).

- A. Pie Slice:** Each pie slice is colored according to the color legend on the right and the size of the slice is labeled with a percent rounded to the nearest tenth.
- B. Color Legend:** The legend lists each of the feature types with an *N* value representing the number of peaks associated with that feature type. The *N* value divided by the total number of peaks gives the percentage that is listed in the legend and on the pie chart.



- C. Hover:** Hovering the cursor on each slice will display the feature type for that slice.

6. Information scores for peak features pie chart

The pie chart is only provided for non-chimeric AGO2 HTML reports and depicts the distribution of information scores for each feature type. This chart is useful for seeing if one feature contains very high confidence peaks; for example, in AGO2 miR-eCLIP experiments, the information score for the miRNA feature tends to be very large, meaning that miRNA peaks contain many reads. The more reads found in a peak, the higher the information score for that peak will be, so peaks with higher p-values tend to have higher information scores. The actual values of the information score are not very important; what's more important is ranking the information scores between features. Information scores are calculated as follows:

c_i = number of eCLIP reads overlapping a peak

i_i = number of input reads overlapping a peak

$$p_i = c_i / \text{total eCLIP reads}$$

$$q_i = i_i / \text{total input reads}$$

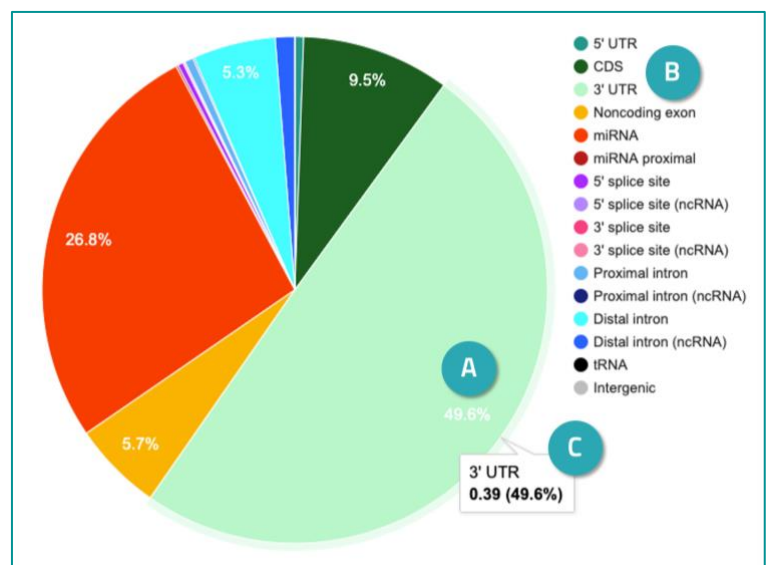
$$s_i = p_i \times \log_2 (p_i / q_i)$$

Information score =

$$\sum s_i \text{ for all peaks in a feature}$$

- A. Pie Slice:** Each pie slice is colored according to the color legend on the right and the size of the slice is labeled with a percent rounded to the nearest tenth.

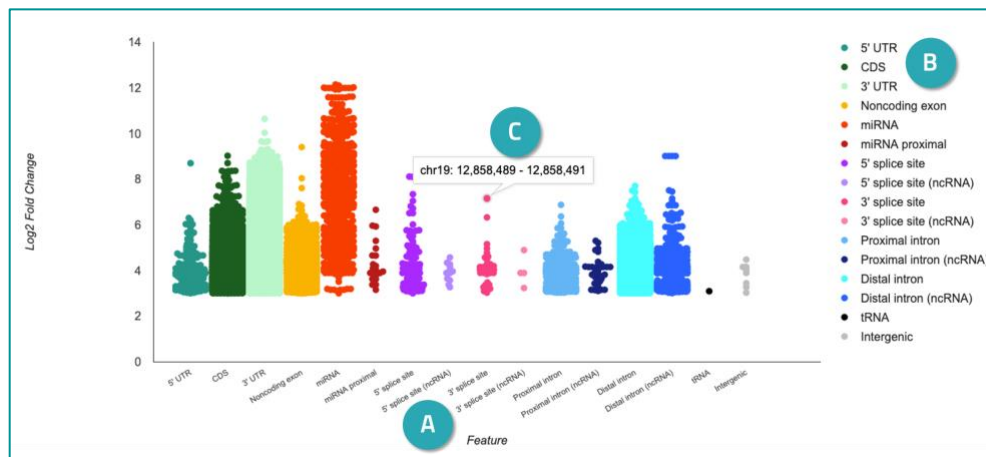
- B. Color Legend:** The legend lists each of the feature types and the color of the pie slice for that feature.



- C. Hover:** Hovering the cursor on each slice will display the feature type for that slice and the information score. In this example, the information score for 3' UTR is 0.39.

7. Swarm plot of peak log₂ fold changes for each feature

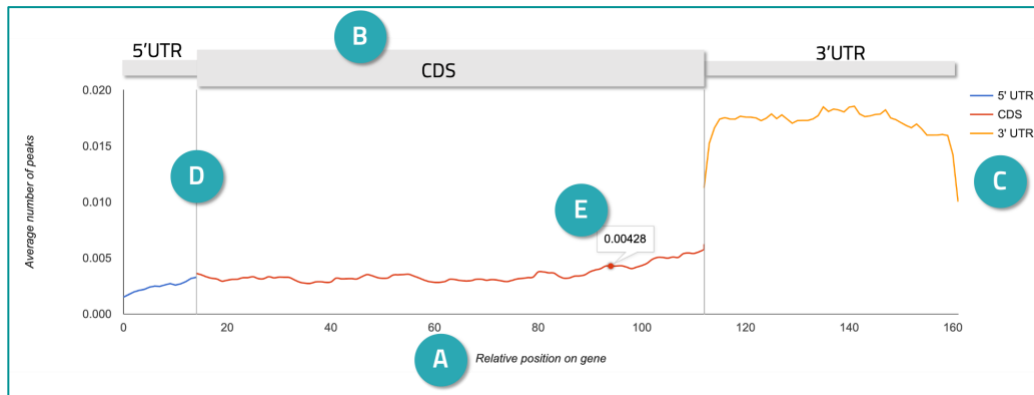
The swarm plot is only provided for non-chimeric AGO2 HTML reports and shows a data point for each peak, where peaks are sorted by feature type on the x-axis and the log₂ fold change is plotted on the y-axis (see **Annotation and Feature Type** details above). This plot is useful for seeing the general spread of log₂ fold enrichment values for each feature type and clearly displays the feature types of the most enriched peaks.



- A. X-axis Feature Types:** Each value along the x-axis displays the sets of peaks that are in that feature type and colored according to the legend on the right.
- B. Color Legend:** The legend lists each of the feature types and the color of the data points for that feature.
- C. Hover:** Hovering over a data point on the swarm plot will give the genomic location (chromosome, start, and end) of the corresponding peak. In this example, the peak hovered over is located on chromosome 19, starting at position 12,858,489 and ending at position 12,858,491.

8. Peak Metagene Plot

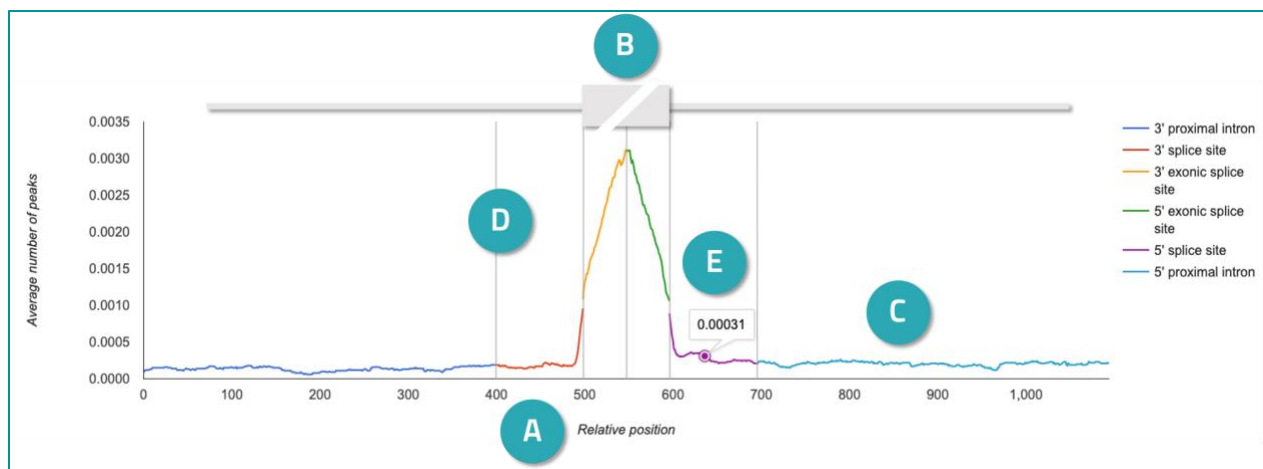
Metagene plots show the average number of peaks mapped to the 5'UTR, CDS, and 3'UTR. To create this plot, the number of peaks is calculated for each region of every annotated gene, the lengths of the regions are normalized, and the average number of peaks for a set number of positions along each region is calculated. This plot is useful for seeing the pattern of enrichment over the 5' UTR, CDS, and 3'UTR regions of genes. In addition, this plot easily highlights when peaks are concentrated in one part of the feature; for example, peaks may be found more often at the beginning or the end of the CDS region. For this particular example, the AGO2 RNA binding protein has peaks mostly found in 3'UTR regions of genes.



- A. X-axis relative position on gene:** Each value along the x-axis displays the sets of peaks that are in that feature type and colored according to the legend on the right. Normalized 5'UTR, CDS and 3'UTR regions are displayed (0-160 bp on the x-axis).
- B. Gene position:** This plot shows three gene regions normalized for length. A gene model graphic in gray has been added here to help with this visual.
- C. Colored lines:** The lines are each colored per region based on the region color legend on the right.
- D. Vertical gray lines:** Vertical lines delineate each region based on the region color legend on the right.
- E. Hover:** Hovering over a data point on the line will give the y-axis value (average number of peaks) for that region position. In this example, the CDS position near x=95 has an average of 0.00428 peaks.

9. Peak Metaintron Plot

Metaintron plots show the average number of peaks mapped to intronic features flanking exons. To create this plot, the number of peaks is calculated for each region of every gene, the lengths of the regions are normalized, and the average number of peaks for a set number of positions along each region is calculated.

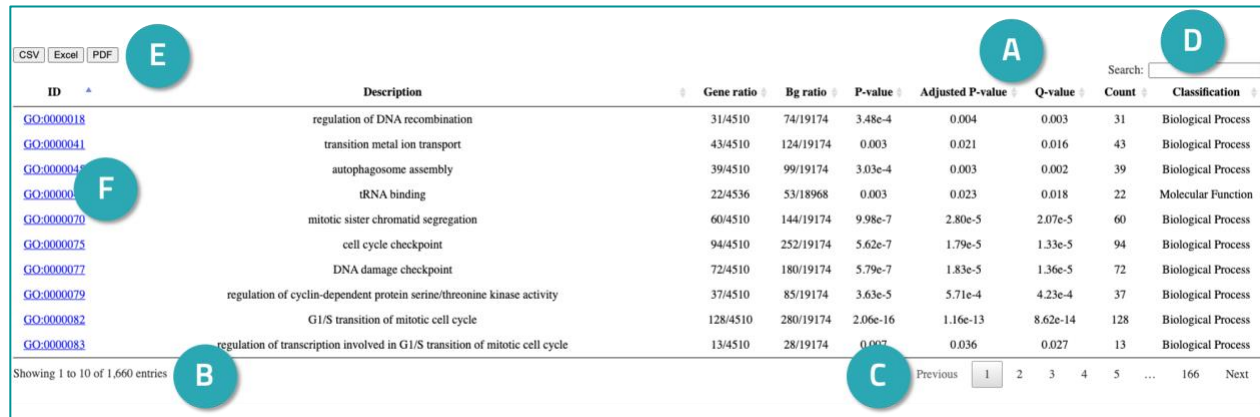


- A. X-axis relative position around exons:** Each value along the x-axis displays the sets of peaks that are in that feature type and colored according to the legend on the right. 500 bp of intronic space upstream of an exon, 100 bp of exon, and 500 bp downstream of an exon are displayed (0-1,100 bp on the x-axis). Any peaks labeled as "distal intron" are excluded from this plot since they are located more than 500 bp from an exon.

- B. Exon position:** This plot is centered on exons (from 500-600 bp on the x-axis) displaying the binding on just the most 50 bp of the exon edges and then the remainder of the plot extends into the flanking intron regions. Note that peaks in the exonic splice site regions are otherwise labeled with the feature type "CDS" in all other parts of this document. Binding on exons >50 bp into an exon is hidden from this plot as the focus here is on flanking intronic regions. A gene model graphic in gray has been added here to help with this visual.
- C. Colored lines:** The lines are each colored per region based on the region color legend on the right.
- D. Vertical gray lines:** Vertical lines delineate each region based on the region color legend on the right.
- E. Hover:** Hovering over a data point on the line will give the y-axis value (average number of peaks) for that region position. In this example, the 5' splice site position near x=640 has an average of 0.00031 peaks.

10. GO Term Enrichment Analysis

The Gene Ontology Consortium (GO Consortium) has defined sets of terms to label all genes with different function types. Additionally, each GO term is sorted into three different classifications: Biological Process, Cell Component, and Molecular Function. For genes with at least one peak, the tool *clusterProfiler* (<https://guangchuangyu.github.io/>) is applied to identify GO terms that are enriched within the set of targets.



The screenshot shows a table of GO term enrichment results. Annotations A-F highlight specific features: A points to the 'Adjusted P-value' header, B points to the 'Showing 1 to 10 of 1,660 entries' text, C points to the 'Previous' and 'Next' navigation buttons, D points to the search bar, E points to the file export buttons (CSV, Excel, PDF), and F points to a GO term ID link.

ID	Description	Gene ratio	Bg ratio	P-value	Adjusted P-value	Q-value	Count	Classification
GO:0000018	regulation of DNA recombination	31/4510	74/19174	3.48e-4	0.004	0.003	31	Biological Process
GO:0000041	transition metal ion transport	43/4510	124/19174	0.003	0.021	0.016	43	Biological Process
GO:0000044	autophagosome assembly	39/4510	99/19174	3.03e-4	0.003	0.002	39	Biological Process
GO:0000000	tRNA binding	22/4536	53/18968	0.003	0.023	0.018	22	Molecular Function
GO:0000070	mitotic sister chromatid segregation	60/4510	144/19174	9.98e-7	2.80e-5	2.07e-5	60	Biological Process
GO:0000075	cell cycle checkpoint	94/4510	252/19174	5.62e-7	1.79e-5	1.33e-5	94	Biological Process
GO:0000077	DNA damage checkpoint	72/4510	180/19174	5.79e-7	1.83e-5	1.36e-5	72	Biological Process
GO:0000079	regulation of cyclin-dependent protein serine/threonine kinase activity	37/4510	85/19174	3.63e-5	5.71e-4	4.23e-4	37	Biological Process
GO:0000082	G1/S transition of mitotic cell cycle	128/4510	280/19174	2.06e-16	1.16e-13	8.62e-14	128	Biological Process
GO:0000083	regulation of transcription involved in G1/S transition of mitotic cell cycle	13/4510	28/19174	0.007	0.036	0.027	13	Biological Process

- A. Table Sorting:** Terms are best sorted by adjusted p-value in order to see the most significantly enriched function types. To do so, simply click the arrows in the header of the table next to this column and it will resort by this value.
- B. Number of Entries:** The table contains as many GO term entries as listed at the bottom left corner, displaying 10 entries at a time.
- C. Page Navigator:** The entries in the table can be viewed using the *Previous* and *Next* buttons on the bottom right. The next or previous 10 entries will load.
- D. Search:** The table is also searchable with keywords in order to list ontologies or classifications of interest.
- E. File Export:** Selecting the buttons on the top left will export the table to a file in the selected format (CSV, Excel, or PDF). Exporting the table will preserve any sorting or filtering that has been performed.
- F. Term Links:** Clicking the term ID will go to the AmiGO 2 webpage with details for the selected GO term.

11. KEGG Enrichment Analysis

The KEGG (Kyoto Encyclopedia of Genes and Genomes) database contains annotations for genes that are involved in certain pathways. For all genes with at least one peak, the tool *clusterProfiler* is applied to identify KEGG annotations that are enriched within the set of targets.

CSV | Excel | PDF

Search:

ID	Description	Gene ratio	Bg ratio	P-value	Adjusted P-value	Q-value	Count
hsa00310	Lysine degradation	35/2007	59/7431	1.67e-7	4.28e-6	2.92e-6	35
hsa01521	EGFR tyrosine kinase inhibitor resistance	31/2007	79/7431	0.012	0.043	0.029	31
hsa01522	Endocrine resistance	46/2007	98/7431	1.70e-5	2.01e-4	1.37e-4	46
hsa01523	Platinum drug resistance	31/2007	73/7431	0.003	0.014	0.010	31
hsa03040	RNA transport	63/2007	171/7431	0.003	0.014	0.009	63
hsa03045	mRNA surveillance pathway	39/2007	91/7431	7.47e-4	0.005	0.003	39
hsa03018	RNA degradation	33/2007	79/7431	0.003	0.014	0.010	33
hsa03060	Protein export	13/2007	23/7431	0.003	0.013	0.009	13
hsa04010	MAPK signaling pathway	109/2007	295/7431	8.87e-5	8.26e-4	5.63e-4	109
hsa04012	ErbB signaling pathway	35/2007	85/7431	0.003	0.014	0.010	35

Showing 1 to 10 of 85 entries




Previous 1 2 3 4 5 ... 9 Next

- A. Table Sorting:** Terms are best sorted by adjusted p-value in order to see the most significantly enriched function types. To do so, simply click the arrows in the header of the table next to this column and it will resort by this value.
- B. Number of Entries:** The table contains as many KEGG pathway entries as listed at the bottom left corner, displaying 10 entries at a time.
- C. Page Navigator:** The entries in the table can be viewed using the *Previous* and *Next* buttons on the bottom right. The next or previous 10 entries will load.
- G. Search:** The table is also searchable with keywords in order to list pathway types of interest.
- D. File Export:** Selecting the buttons on the top left will export the table to a file in the selected format (CSV, Excel, or PDF). Exporting the table will preserve any sorting or filtering that has been performed.
- E. Term Links:** Clicking the term ID will go to the KEGG database webpage with details for the selected pathway.

12. HOMER Motif Results

The HOMER algorithm (<http://homer.ucsd.edu/>) identifies enriched sequence motifs within peaks *de novo*. The reverse complement seed region of highly enriched miRNAs are expected to be reported enriched motifs on mRNA targets.

Total target sequences = 13695
Total background sequences = 32644
* - possible false positive

Rank	Motif	P-value	log P-value	% of Targets	% of Background	STD(Bg STD)
1		1e-1227	-2.825e+03	13.72%	1.32%	32.2bp (71.8bp)
2		1e-662	-1.526e+03	10.75%	1.72%	37.3bp (75.8bp)
3		1e-496	-1.144e+03	6.38%	0.73%	35.5bp (62.1bp)

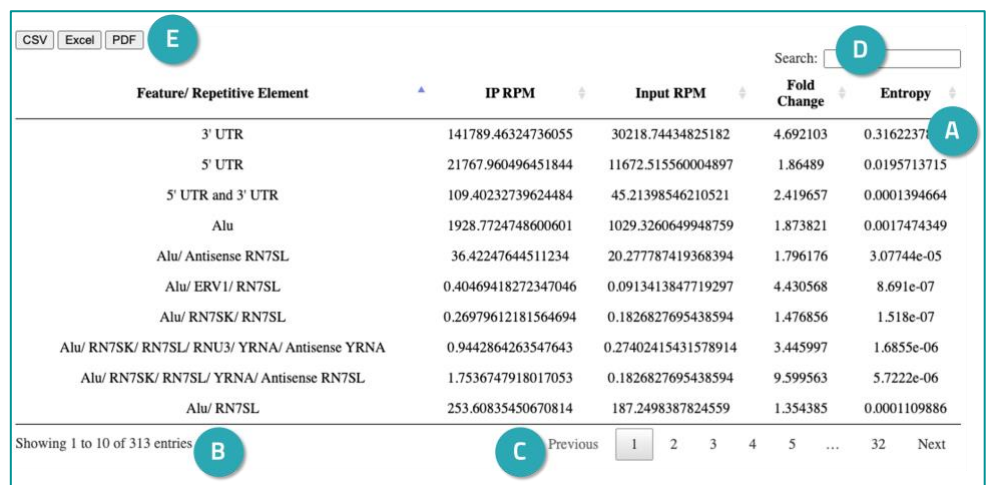
- A. Total sequences:** The number of peak sequences used for the motif enrichment calculation is listed above the table. Additionally, the number of background sequences are also listed, which are a randomly distributed set of sequences according to the set of peaks provided to the algorithm.
- B. Ranked Entries:** The table contains motifs that were found to be significantly enriched in the set of peak sequences. The table is ranked by *P*-value, with the most enriched motif at the top.
- C. Motif image:** The motif is displayed with colored bases (blue C, red U, gold G, and green A). The height of the letter is proportional to the relative frequency of that base at that motif position.

- D. Percentages:** The percentage of total peaks containing each motif is useful for understanding how much of the data this motif represents compared to the background enrichment.

13. IP Repetitive Element Mapping Information Table

As part of the miR-eCLIP analysis pipeline, reads are mapped to a database of repetitive elements and removed from the set of reads aligned to the genome. In order to understand what has been filtered, a separate repetitive read analysis is done using all reads including those mapped to repetitive elements and the genome. Enriched feature types or repetitive element types are displayed in this table with the IP RPM, Input RPM, Fold Change, and Entropy. Entropy values are a measure of confidence, where the higher entropy values give higher confidence in the feature type enrichment in the read data. This table is only provided in the non-chimeric AGO2 peak HTML reports.

- A. Table Sorting:** Features are best sorted by entropy, in order to look at the most confident fold change results. To do so, simply click the arrows in the header of the table next to this column and it will resort by this value. Sorting by IP RPM or Input RPM is another good sorting option to see the



The screenshot shows a table with the following columns: Feature/ Repetitive Element, IP RPM, Input RPM, Fold Change, and Entropy. The table is sorted by Entropy in descending order. Annotations A-E highlight specific features: A points to the Entropy column header, B points to the 'Showing 1 to 10 of 313 entries' text, C points to the 'Previous' and 'Next' navigation buttons, D points to the search bar, and E points to the export buttons (CSV, Excel, PDF).

Feature/ Repetitive Element	IP RPM	Input RPM	Fold Change	Entropy
3' UTR	141789.46324736055	30218.74434825182	4.692103	0.3162237
5' UTR	21767.960496451844	11672.515560004897	1.86489	0.0195713715
5' UTR and 3' UTR	109.40232739624484	45.21398546210521	2.419657	0.0001394664
Alu	1928.7724748600601	1029.3260649948759	1.873821	0.0017474349
Alu/ Antisense RN7SL	36.42247644511234	20.277787419368394	1.796176	3.07744e-05
Alu/ ERV1/ RN7SL	0.40469418272347046	0.0913413847719297	4.430568	8.691e-07
Alu/ RN7SK/ RN7SL	0.26979612181564694	0.1826827695438594	1.476856	1.518e-07
Alu/ RN7SK/ RN7SL/ RNU3/ YRNA/ Antisense YRNA	0.9442864263547643	0.27402415431578914	3.445997	1.6855e-06
Alu/ RN7SK/ RN7SL/ YRNA/ Antisense RN7SL	1.7536747918017053	0.1826827695438594	9.599563	5.7222e-06
Alu/ RN7SL	253.60835450670814	187.2498387824559	1.354385	0.0001109886

- most abundant features or repetitive elements found in the sample. Note that high fold changes are not necessarily indicative of AGO2 binding; a feature should have both a high fold change and a high entropy value. High fold changes for some repetitive elements may be present due to low read numbers rather than real binding.
- B. Number of Entries:** The table contains as many feature/repetitive element entries as listed at the bottom left corner, displaying 10 entries at a time.
- C. Page Navigator:** The entries in the table can be viewed using the *Previous* and *Next* buttons on the bottom right. The next or previous 10 entries will load.
- D. Search:** The table is also searchable with keywords in order to list repetitive elements or features of interest.
- E. File Export:** Selecting the buttons on the top left will export the table to a file in the selected format (CSV, Excel, or PDF). Exporting the table will preserve any sorting or filtering that has been performed.

Data File Descriptions

There are several types of data files available for download for each sample. Files will be labeled with the sample name, followed by the suffix detailed below.

***.fastq.gz**

Raw sequencing reads in FASTQ format and gzipped. This file is meant for users that want to run a data analysis pipeline beginning with raw reads. The general format of a FASTQ file contains four lines per read with the following information:

Line 1: unique sequence identifier, and may be followed by an optional description

Line 2: raw sequence

Line 3: "+", and may be followed by the unique sequence identifier and description

Line 4: sequence quality values

***.adapterTrim.fastq.gz / *.adapterTrim.round2.fastq.gz**

Reads with UMI relocated to the read name and trimmed of adapter at the 3' end, also in FASTQ format and gzipped. This file is meant for those who want to visualize cleaned reads and run a data analysis pipeline beginning with trimmed reads. The general format of a FASTQ file is detailed above (for **.fastq.gz**). The ***adapterTrim.fastq.gz** files are for IP samples, and the ***adapterTrim.round2.fastq.gz** files are for input samples.

***.adapterTrim.round2.rmRep.sorted.rmDup.sorted.bam**

Non-chimeric reads that have been filtered of repetitive elements, aligned to the reference genome, and removed of PCR duplicates. BAM files are compressed binary versions of Sequence Alignment/Map (SAM) files. BAM files are used as input to downstream data analysis and for visualizing read alignments in a genome browser (Note: the corresponding ***bam.bai** index file is also required for loading BAM files into a genome browser). To explore the contents of a BAM file, it will need to be uncompressed back to SAM format using the command line tool `samtools` (available through <http://samtools.sourceforge.net/>). A SAM file begins with a header detailing the reference genome and is followed by one read alignment per line. The general format for each alignment in a SAM file is as follows:

Column 1: query template name (ie: read name)

Column 2: bitwise flag (see <https://broadinstitute.github.io/picard/explain-flags.html> for a guide to SAM flags)

Column 3: reference sequence name (ie: chromosome for alignment)

Column 4: 1-based leftmost mapping position

Column 5: mapping quality

Column 6: Cigar string

Column 7: reference name of the mate/ next read

Column 8: position of the mate/ next read

Column 9: observed template length

Column 10: segment sequence

Column 11: ASCII of Phred-scaled base quality+33

***.adapterTrim.round2.rmRep.sorted.rmDup.sorted.bam.bai**

The index file associated for the above non-chimeric read BAM file. BAM index files are typically required when interacting with BAM files either via command line or in a genome browser. BAM index files should be stored in the same folder as BAM files so that tools can locate them.

***CombinedID.merged.r2.norm.[neg/pos].bw**

RPM-normalized genomic non-chimeric read coverage Bigwig files. Bigwig files are binary files used to display continuous data in a genome browser. Files ending in ***.pos.bw** are only the reads mapped to the "+" strand, while

files ending in ***.neg.bw** are only the reads mapped to the “-” strand. Coverage is normalized by RPM = (reads aligned to region / total aligned reads) x 1,000,000 so that coverage can visually be compared across samples.

***basedon.peaks.l2inputnormnew.bed.compressed.bed**

Locations of input normalized non-chimeric read clusters (“peaks”) called using the CLIPper algorithm (<https://github.com/YeoLab/clipper>). The BED file contains the following tab-separated columns:

Column 1: Chromosome

Column 2: Start position

Column 3: Stop position

Column 4: $-\log_{10}(\text{p-value})$

Column 5: \log_2 fold enrichment in eCLIP vs. input

Column 6: Strand (+ or -)

Peaks provided are only called within annotated transcript regions. The p-value for each peak is calculated using a Yates’ Chi-Square test, or if the observed or expected read coverage is below 5 using a Fisher Exact Test. The peak p-values and fold enrichments in this file are based on eCLIP RPM coverage as compared to input RPM coverage. Other sample condition comparisons can also be evaluated in the same manner such as an eCLIP from a wild type sample vs. an eCLIP from a knock out (WT vs KO). This file can be loaded into a genome browser for visualization and manipulated with command line tools such as BEDTools (<https://bedtools.readthedocs.io/>).

***reverse_mir_alignments.filtered.tsv**

Results of the reverse mapping of mature miRNA sequences (obtained from miRbase) to putative chimeric reads with Bowtie. The file is filtered for only the best aligned miRNA for each putative chimeric read with tab-separated columns as follows:

Column 1: miRNA name

Column 2: Strand (+ or -)

Column 3: Read name (Reads with identical sequence are collapsed for uniqueness prior to reverse mapping of miRNA sequences, so only one read name from the set of identical reads is provided here.)

Column 4: Position of miRNA starting position within the read sequence (0-based).

Column 5: mature miRNA sequence aligned

Column 6: miRNA Phred base qualities (Set by default to 40 (“I”), since qualities are unavailable for the aligned miRNA FASTA file.)

Column 7: The number of other instances where the same miRNA sequence aligned against the same read reference characters as were aligned against in the reported alignment. This is *not* the number of other places the miRNA aligns with the same number of mismatches. The number in this column is generally not a good proxy for that number (e.g., the number in this column may be ‘0’ while the number of other alignments with the same number of mismatches might be large).

Column 8: Comma-separated list of mismatch descriptors (empty if no mismatches within miRNA alignment)

***reverse_mir_alignments.filtered.chimeric_candidates.STARAligned.outSo.rmDup.bam**

Chimeric reads that have the miRNA portion removed, the mRNA portion aligned to the reference genome, and removed of PCR duplicates. BAM files are used as input to downstream data analysis and for visualizing chimeric

read alignments in a genome browser (Note: the corresponding ***bam.bai** index file is also required for loading BAM files into a genome browser). To explore the contents of a BAM file, it will need to be uncompressed back to SAM format using the command line tool `samtools` (available through <http://samtools.sourceforge.net/>). See SAM file details above.

***reverse_mir_alignments.filtered.chimeric_candidates.STARAligned.outSo.rmDup.bam.bai**

The index file associated for the above chimeric read BAM file. BAM index files are typically required when interacting with BAM files either via command line or in a genome browser. BAM index files should be stored in the same folder as BAM files so that tools can locate them.

***reverse_mir_alignments.filtered.chimeric_candidates.STARAligned.outSo.rmDup.bed.annotated_w_mirnas.txt**

An annotated table of chimeric read alignments useful for seeing the genomic locations of chimeric reads, as well as the gene and miRNA that the read mapped to with the following tab-separated columns:

- Column 1: Chromosome
- Column 2: Start position
- Column 3: Stop position
- Column 4: Read name
- Column 5: Mapping quality
- Column 6: Strand
- Column 7: Feature annotation
- Column 8: Ensembl ID
- Column 9: Gene name
- Column 10: miRNA name
- Column 11: miRNA ID

***reverse_mir_alignments.filtered.chimeric_peaks.bed**

Locations of chimeric read clusters ("peaks") called using the CLIPper algorithm (<https://github.com/YeoLab/clipper>). The BED file contains the following tab-separated columns:

- Column 1: Chromosome
- Column 2: Start position
- Column 3: Stop position
- Column 4: miRNA name
- Column 5: RPM coverage of chimeric reads
- Column 6: Strand (+ or -)

Peaks provided are only called within annotated transcript regions. The p-value for each peak is calculated by CLIPper using a poisson distribution. Coverage is normalized by $RPM = (\text{reads aligned to region} / \text{total aligned reads}) \times 1,000,000$ so that coverage can be compared across samples. This file can be loaded into a genome browser for visualization and manipulated with command line tools such as BEDTools (<https://bedtools.readthedocs.io/>).

***reverse_mir_alignments.filtered.chimeric_candidates.STARAligned.outSo.rmDup.norm.[pos/neg].bw**

RPM-normalized coverage Bigwig files for the mRNA portion of chimeric reads. Bigwig files are binary files used to display continuous data in a genome browser. Files ending in ***.pos.bw** are only the chimeric reads mapped to the "+"

strand, while files ending in ***.neg.bw** are only the chimeric reads mapped to the “-” strand. Coverage is normalized by $RPM = (\text{reads aligned to region} / \text{total aligned reads}) \times 1,000,000$ so that coverage can visually be compared across samples.

***01v02.IDR.out.0102merged.bed** - *This file is only included if the experiment contained 2 or 3 replicate samples.*

Locations of input normalized non-chimeric AGO2 peaks found to be reproducible and significant across 2 or 3 replicates after Eclipse Bio IDR (Irreproducible Discovery Rate) analysis. To generate this set of filtered peaks, the IDR tool (<https://github.com/nboley/idr>) is run on the replicate input-normalized clusters, where clusters are ranked by \log_2 fold change and reproducible regions are identified. The regions are further filtered for those that fall within the coordinates of clusters called by CLIPper in each replicate. For each cluster region, the \log_2 fold change and p-value are calculated for IP vs. input (or IP vs. IP) for each replicate individually and then the geometric mean of all replicate \log_2 fold changes is calculated. Clusters that have a geometric mean of \log_2 fold changes > 3 and minimum $-\log_{10}(P\text{-value}) > 3$ are reported in the final IDR set of peaks. This BED file can be loaded into a genome browser for visualization and manipulated with command line tools such as BEDTools (<https://bedtools.readthedocs.io/>).

***report.html**

Report with tables and plots for enriched gene features, GO terms, KEGG pathways, motifs, and repetitive elements in enriched peaks. See the above section for a detailed review of each aspect of the HTML report.