**What is miRspring:**

The miRspring document is a html document that visualizes miRNA sequencing from a mapped small RNA data set. In addition there are a range of analysis tools that will identify various miRNA processing features. The miRspring document is compatible with most internet browsers and does not need any internet connectivity to function.

**What miRspring is not:**

miRspring will only visualize and analyse data from one sample and therefore cannot compare different data sets. You can open a data set per tab within a browser and then navigate between tabs. The analysis output tables do provide a means to be copy data to an external program that is capable of comparing different data sets.

**Creating a miRspring document**

The first step in creating a miRspring document is to map the raw sequence reads to reference, typically either a complete genome or a list of precursor sequences (i.e. miRBase), and creating a Binary aligned mapped (BAM) file. The BAM file must be sorted and indexed, which if not created in this way can be formatted using SAMTOOLS.

The two scripts that are required to create a miRspring document are dependent on the files and programs listed in table 1. The pipeline was implemented with two scripts to provide a method to combine outputs of multiple BAM files of the same data set (Figure 1), for example if a barcoded sample was spread across multiple lanes and then each lane mapped independently.

**Table 1**: File and software dependencies required to make a miRspring document.

|  |  |  |
| --- | --- | --- |
| Program/File | miRspring Scripts | URL |
| Samtools | BAM\_to\_Intermediate.pl |  |
| mature.fa | Intermediate\_to\_miRspring.pl |  |
| mirbase genome gff file | BAM\_to\_Intermediate.pl  Intermediate\_to\_miRspring.pl |  |
| hairpin.fa | BAM\_to\_Intermediate.pl  Intermediate\_to\_miRspring.pl |  |

The first script, called “BAM\_to\_intermediate.pl” creates a intermediate text file of starting positions and associated counts for expressed miRNAs. Multiple files can be concatenated to generate one large file that the Intermediate\_to\_miRspring.pl script can generate a miRspring document.

**Script BAM\_to\_intermediate.pl**

Usage:BAM\_to\_mirspring.pl <options>

REQUIRED

-bam <input bam file>

-ml <minimum length>

-s <three letter species code>

-out <output file name with full path>

OPTIONAL

-mm <mismatches: '0' or '1', default is 1>

-gff <Gene features file (gff), default is defined in script>

-mat <Mature sequence file, default is defined in script>

-pre <Precursor file, default is defined in script>

-flank <length of flanking sequence>

-ref <format of reference: '0' (genome) or '1' (custom), 0 (genome) is default

**Script Intermediate\_to\_miRspring.pl**

Usage: Intermediate\_to\_miRspring.pl <options>

REQUIRED

-in <input file>

-s <three letter species code>

OPTIONAL

-out <output file name with full path>

-mm <mismatches: '0' or '1', default is 1>

-gff <Gene features file (gff), default is defined in script>

-mat <Mature sequence file, default is defined in script>

-pre <Precursor file, default is defined in script>

-flank <length of flanking sequence>

-ref <format of reference: '0' (genome) or '1' (custom), 0 (genome) is default

-mintag <only include precursors with at least this number of tags, default = 0>

-comment <message for top of miRspring document>

**Shell script setup examples**

**The miRspring document**

**Introduction:**

The miRspring document is comprised of a “Global” and “Focused” view. The “Global” view summarises the data set with a XY scatter plot and a tabulated list of miRNA counts. Surrounding the XY scatter plot are a number of selectable buttons. The buttons on the right hand side Y-axis select the “processing feature” to be plotted on the graph, while the buttons above the graph create tabulated lists of data that the user can export for downstream analysis. Selecting a data point on the XY scatter plot or the “details” button within the table will navigate the miRspring document to the focused view where individual sequence data of that particular miRNA is displayed.

**Global view**

**X-Y scatter plot**

The “global” visualization XY scatter plot tool can selectively prepare one of six “processing feature” graphs. In each case the X-axis represents the rank or cumulative abundance of processed miRNA (log scale) and the Y-axis the miRNA processing feature (length, 5’ isomiRs, 3’ isomiRs, arm processing, non-canonical processing, seed frequency). Two other XY scatter plot tools, “cumulative distribution” and “threshold”, summarize the distribution of miRNA sequences within the data set. It should be noted that each data point represents the accumulation of all miRNA derived from that precursor. Moving the mouse pointer over the graph will highlight the name of the miRNA precursor, and selecting the data point will navigate to the focused view of that miRNA

ALL versus miRBase defined miRNA.

***Length:*** The average length of ALL processed miRNAs derived from the precursor

***5’ IsomiR:*** The percentage of 5’ isomiRs of all processed miRNAs derived from the precursor. A 5’ isomiR is defined as any miRNA whose 5’ end does not start as defined by miRBase but is within the defined window (see options) to be accepted as a mirBase processed miRNA.

***3’ isomiR:*** The percentage of 3’ isomiRs of all processed miRNAs derived from the precursor. A 3’ isomiR is defined as any miRNA whose 3’ end that is not finish as defined by miRBase. Note a miRNA can only be a 3’ isomiR if the 5’ end starts within the defined window (see options).

***Arm processing:*** This is calculated from the following formula, where 5p and 3p refers to the number of miRBase defined miRNA processed from the 5p or 3p arm.

5p/(3p + 5p)

Precursors that result in very high number of miRNA being processed from the 5p or 3p arm will display the data point closer to the top or bottom of the Y axis respectivey. Data points positioned in the middle of the Y-axis represent precursors that have a more equal 5p/3p processing.

***Non-canonical processing:*** The percentage of tags derived from the precursor having 5’ starting positions outside the defined window of sequences defined in miRbase.

***Seed Frequency:*** The number of sequence tags predicted to have a miRBase miRNA seed (X-axis) plotted against the actual seed distribution of the data set. Data points that sit on the diagonal line running from the bottom left hand corner to the upper right hand corner are miRNA seeds from the data set that are in agreement with the predicted distribution. Note that both axis are plotted on a log scale.

***Editing:*** The percentage of miRNAs derived from a precursor that contain mismatches to the reference.

***Cumulative distribution:*** The bottom left most data point represents the fraction of miRNA reads represented by the most abundant miRNA. Every subsequent data point tp the upper right is the sum of the previous data point and the next ranked miRNA. The 50th and 100th ranked miRNA are coloured green.

**DATA Analysis**

Export miR ID stats

Precursor ID, Total counts, Averagelength, Non canonical, 5p counts, 3p counts, 5p/(5p+3p), 5p 5’ isomiRs, 5p 3’ isomiRs, 3p 5’ isomiRs, 3p 3’ isomiRs, Editing counts

listmiRBase coordinates

precursor ID, 5p start, 5p end, 3p start, 3p end, Sequence

List seed abundance

Rank, Seed, Count, Seed details tab

Polycistronic miR counts

Genomic Coordinates, miR IDs, counts

Display sample info

Date/time, BAM header information (RG, PG, CO).

**Table**

The “global view” table is a list of all miRNAs identified to be expressed in the data set. The total counts, 5p and 3p processed miRNA counts for that miRNA are listed in the columns to the right. In addition there is a column that contains the 5p + 3p counts and non-canonical counts (ie anything that is not 5p or 3p). For each column of counts there is a DES and ASC button which will sort the column in descending or ascending order respectively. For miRNAs that do not have a miRBase entry for one of the stem arms a value of “undefined”. The sorting buttons can be used to identify abundant undefined processed miRNAs, eg selecting the ASCending button on the non-canonical processed column will rank all precursors that have abundant miRNAs that start at positions not defined in miRBase. The “view reads/details” button located in each row will navigate the miRspring document to the focused view.

**Focused view**

The “focused view” displays the precursor sequence, nucleotide frequency histogram, all sequence tags that align to a specific miRNA hairpin and a summary of processing statistics. If the miRNA is within the defined genomic distance of other miRNAs to be considered part of a cluster then a bar graph showing all members relative expression level is displayed. The bar graph of the miRNA on focus is drawn green while the others are brown. Navigation to the other miRNAs of the cluster is achieved by clicking the mouse pointer on the corresponding bar graph.

The miRBase defined positions of processed miRNA from within the hairpin is defined by a green text. The nucleotide frequency histogram is aligned to the precursor sequence and the regions that correspond to processed miRNA as defined by miRBase is also coloured green. The window used to accept sequence tags as miRbase processed miRNA is highlighted by the black thick line below the nucleotide histogram.

There are two modes of aligning sequence tags, “5’ collapsed” or “verbose”. The “5’ collapsed” displays all sequences that have a common 5’ end on one line. The 3’ variations are distinguished with a different font colour which correlates to a bar graph showing the relative abundance to other 3’ positions. The overall length of the most abundant 3’ sequence tag is displayed in the following column with the corresponding colour.

**Options**

A number of configurable options are available for the miRspring document. While these settings cannot be saved they can be permanently set when initially creating a document.

* miRBase window

The extended region surrounding the miRBase defined processed miRNA that will allow a sequencing tag to be counted as a processed 5p or 3p miRNA. The default setting is +-3nt.

* Sequence display format

The default setting is to display using the collapsed 5’ sequencing format where all sequences that have a common 5’ end to be displayed on one line. The 3’ variations are highlighted by differing font colours which correlate to the bar chart showing relative expression levels. The alternative is the more traditional verbose method where every unique sequence is listed independently along with the corresponding count.

* Mismatches

The miRspring document can display up to one mismatch per sequence and the mismatched nucleotide is displayed in lower case. Additionally sequences containing a mismatch are highlighted by the lack of flanking “.” or “-“ characters on the 3’ end. A number of options are available to the user to restrict what mismatches are shown, which include:

* 1. Only A to I
  2. Only C to T
  3. Only A to I AND C to T
  4. No mismatches

After resetting the display option the counts displayed in the “Global view” will be recalculated accordingly. Likewise the “Global view” data download buttons will only list sequence tags that meet the mismatch criteria.

* Normalisation

The counts displayed for each entry in the “global table” can be adjusted to display a normalised value. Counts per million (CPM) normalises the counts of each entry to the total count of either just miRNAs (CPMmi = counts per million miRNA) or the total counts of the complete data set (CPM = counts per million).

Alternatively the data can be normalised to a spike in.

* Cluster window

A miRNA cluster is any two or more miRNA located within the specified distance, irrespective of transcript orientation.

* Seed analysis options

The seed analysis tool is primarily designed to compare the seed sequence of miRNAs, however this can be extended to other small RNA classes by setting this option to “all”.