Mobile DNA manuscript – **User Manual** –

"*sRNAPipe*: a GALAXY-based pipeline for bioinformatic in-depth exploration of small RNAseq data"

Prerequisites to use *sRNAPipe*:

The initial input of the *sRNAPipe* is a collection of single-end sequencing data in a fastq phred+33 format (Galaxy fastqsanger format), following adapter removal, and a list of input multi-fasta references (genome, transcripts, TEs, rRNAs, tRNAs, snRNAs, miRNAs, fasta format). If your uploaded file has not been recognized with the required format, the format can be changed using the Pencil icon to access Datatypes.

The dataset used in the submitted manuscript (sequencing data from Dennis C. et al. 2016, SRR4428936.fastq), as well as small artificial datasets to test the pipeline very quickly, within a few minutes, are available in the History panel on the right.

To run *sRNAPipe*,

1) Open *sRNAPipe* (in-depth study of small RNA) software in the Tools panel on the left.

2) To run *sRNAPipe* with the sequencing data from Dennis C. et al. 2016, SRR4428936.fastq, enter all the parameters as described here:

sRNAPipe In-depth study of small RNA (Galaxy Version 0.0.1)   ▼ Options

fastqsanger (Q33)

[📄 | 🗐 | 📁] 154: SRR4428936.fastq   ▼

Additional Fastq Files

＋ Insert Additional Fastq Files

Will you select a reference genome from your history or use a built-in index?

Use a built-in index   ▼

   Select a reference genome

   Drosophila melanogaster flybase6.03   ▼

Will you select transcripts database from your history or use a built-in index?

Use a built-in index   ▼

   Select a transcripts reference

   Drosophila melanogaster transcripts flybase6.03   ▼

Will you select TE database from your history or use a built-in index?

Use a built-in index   ▼

   Select a TE reference

   Drosophila all TE RepBase 19.5   ▼

Will you select miRNA database from your history or use a built-in index?

Use a built-in index   ▼

   Select a miRNA reference

   Drosophila melanogaster hairpin mirBase 21   ▼

Will you select snRNA database from your history or use a built-in index?

Use a built-in index   ▼

   Select a snRNAs reference

   Drosophila melanogaster snRNA flybase6.03   ▼

Will you select rRNAs database from your history or use a built-in index?

Use a built-in index   ▼

   Select a rRNAs reference

   Drosophila melanogaster rRNA flybase6.03   ▼

Will you select tRNA database from your history or use a built-in index?

Use a built-in index   ▼

   Select a tRNA reference

   Drosophila melanogaster tRNA flybase6.03   ▼

At least the sequencing data, genome, transcripts, TEs, and miRNA files are required to run *sRNAPipe*. Files for snRNAs, rRNAs and tRNAs may be skipped by choosing "None" in answer to "Will you select snRNA/rRNAs/tRNA database from your history or use a built-in index?".

To run *sRNAPipe* with small test-data files (available via GitHub: https://github.com/brassetjensen/sRNAPipe) for a quick test, enter all the parameters as described here:

sRNAPipe In-depth study of small RNA (Galaxy Version 0.0.1)   ▾ Options

fastqsanger (Q33)
[📄] [🗐] [🗀]   1: reads-sample.fastq   ▾

Additional Fastq Files
+ Insert Additional Fastq Files

Will you select a reference genome from your history or use a built-in index?
Use one from the history   ▾

  Select a reference from history
  [📄] [🗐] [🗀]   2: genome-small.fa   ▾

Will you select transcripts database from your history or use a built-in index?
Use one from the history   ▾

  Select a reference from history
  [📄] [🗐] [🗀]   3: transcripts-file-small.fa   ▾

Will you select TE database from your history or use a built-in index?
Use one from the history   ▾

  Select a reference from history
  [📄] [🗐] [🗀]   4: TE-file-small.fa   ▾

Will you select miRNA database from your history or use a built-in index?
Use one from the history   ▾

  Select a reference from history
  [📄] [🗐] [🗀]   5: mirbase-21-dme-hairpins-16jul2015.fa   ▾

Will you select snRNA database from your history or use a built-in index?
Use one from the history   ▾

  Select a reference from history
  [📄] [🗐] [🗀]   6: dmel-all-sn-snoRNA-r6.03.fasta   ▾

Will you select rRNAs database from your history or use a built-in index?
Use one from the history   ▾

  Select a reference from history
  [📄] [🗐] [🗀]   7: dmel-all-rRNA-r6.03.fasta   ▾

Will you select tRNA database from your history or use a built-in index?
Use one from the history   ▾

  Select a reference from history
  [📄] [🗐] [🗀]   8: dmel-all-tRNA-r6.03.fasta   ▾

3) Select the minimum and maximum size of the reads to analyze, the size of the siRNA and piRNA

minimum read size
18

maximum read size
29

lower bound of siRNA range
21

higher bound of siRNA range
21

lower bound of piRNA range
23

higher bound of piRNA range
29

4) Choose the maximum number of mismatches allowed for the mapping on the genome and on the TEs, and whether you wish to study eventual ping-pong partners ("PPPartners"). Click on Execute.

maximal genome mismatches
0

maximal TE mismatches
3

PPPartners
[Yes] [No]

✔ Execute

5) Running job items appear in yellow, and when achieved in green.

6) A first results item "sRNAPipe_fastq_output" is created containing the 26 fastq files generated when the job is completed. Each of these files can be visualized by clicking on the name "sRNAPipe_fastq_output", and downloaded for further analysis by clicking on its name and then on the Save icon. To come back to the History, click on "Back to <History name>"
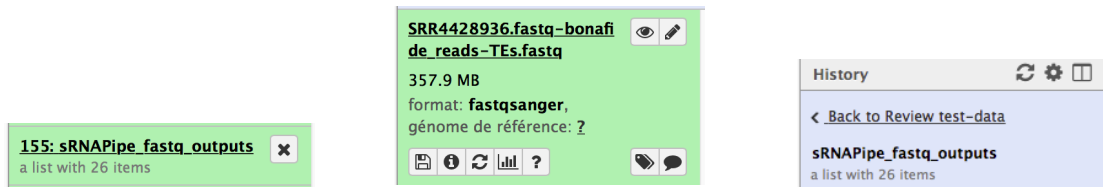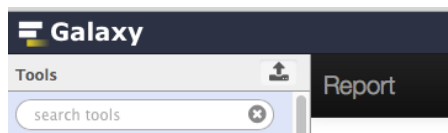

The "sRNAPipe_ fastq _output" files can be unhidden to be used for further analysis on the Galaxy platform by clicking on "hidden" in the History panel and then on "Rendre visible".

7) A second results item "sRNAPipe_<name input fastq file>" is created. The results can be viewed by clicking on the Eye icon.
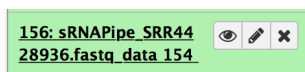
When you ran *sRNAPipe* with several fastq files, click on "Report" to view the respective results.

Results as obtained with the sequencing data from Dennis C. et al. 2016, SRR4428936.fastq:

a) On the first page:
- The size distribution of genome-mapped reads and the number of all genome-mappers and of genome-unique mappers (reads that map only once to the genome) can be visualized.

mappers #: 8605522
unique mappers #: 3157294

- Figures show mapping of genome-unique reads ("Uniquely mapped reads") and of all reads randomly assigned ("Reads randomly assigned") to the chromosomes. Small RNAs are mapped on plus and on minus strands, in blue and in red respectively.

- The corresponding bam and bedgraph files are accessible *via* the corresponding links on the upper left, as well as the counts of the mappers size distribution and lists of the mapped and unmapped reads.

alignment (bam)
bedgraph minus strand
bedgraph plus strand
bedgraph unique minus strand
bedgraph unique plus strand
mappers (txt)
mappers size distribution (txt)
unique alignment (bam)
unique mappers (txt)
unmapped (txt)

b) The second page is accessible by clicking on "View details" on the first page of results. Then the information for 4 subgroups, Bonafide reads, siRNAs, piRNAs and miRNAs, is accessible. For each subgroup, three distinct analyses are accessible in different folders: for genome-mapping reads, TE-mapping reads and transcript-mapping reads.


c) By clicking on Genome, TE or Transcripts folders, figures for genome-unique mappers and for all mappers and tables with read counts for the corresponding features (genome, TEs or transcripts) can be visualized, the corresponding sorted BAM files and bedgraphs for the plus and minus strand mappers can be downloaded.

Example: Click on Bonafide and then Genome folder:
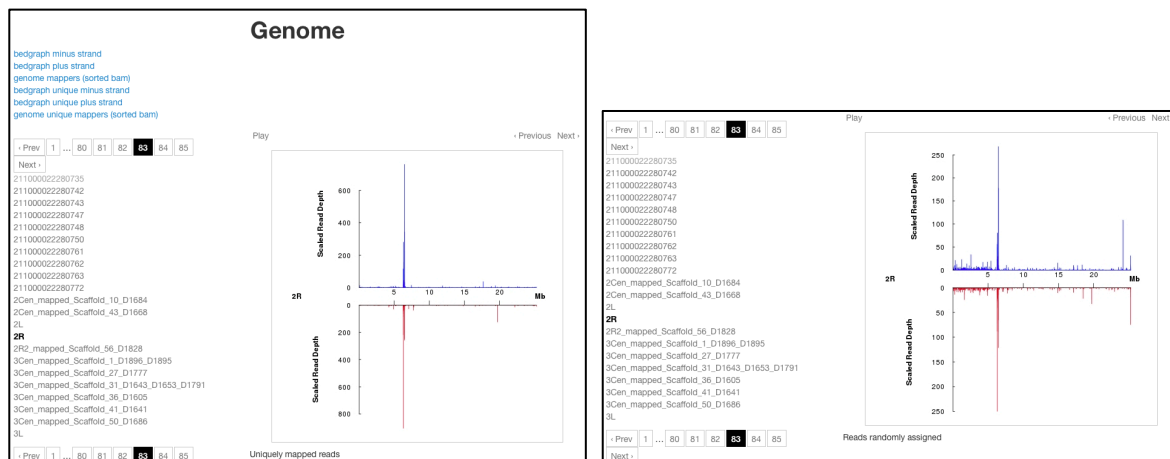
d) The results for Ping-pong partners are found only in the "piRNAs" subgroup "TE" folder, by clicking on "Ping Pong Partners".
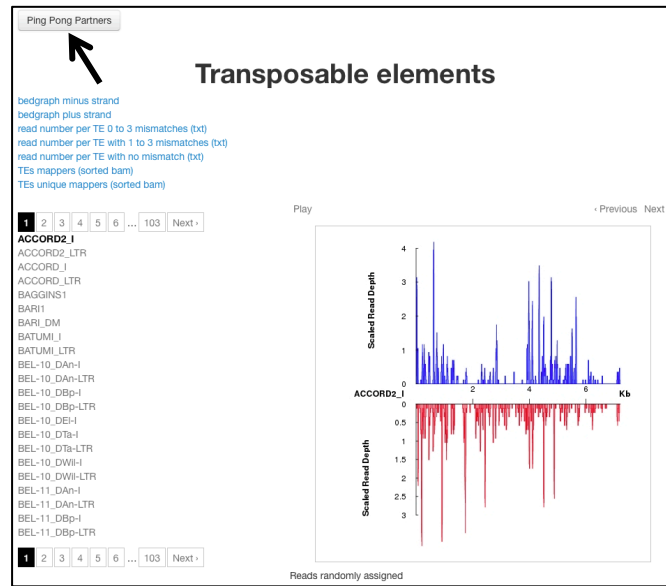
e) The ping-pong signature is analysed for each TE and a sum of all overlaps, the sum of 10-nt-overlaps, the mean, the standard deviation, the z-score and the p-value for each TE are summarized in a table.

ID	overlap sum	ten overlap sum	mean	standard deviation	z-score	p-value
ACCORD2_I	762	39	33.1304347826087	10.5965718321339	0.553911709406997	0.289819635540828
ACCORD2_LTR	1267	166	55.0869565217391	39.3887128430581	2.81585854100354	0.00243235406595044
ACCORD_I	6312	579	274.434782608696	80.8309164307798	3.76792978280929	8.23034928753019e-05
ACCORD_LTR	973	207	42.304347826087	41.7200137381785	3.94764136961916	3.94624570775326e-05
BAGGINS1	153161	15635	6659.17391304348	2082.26495937923	4.31060708510045	8.14034819840437e-06
BARI1	13	0	0.565217391304348	1.01407859040788	-0.557370401713154	0.711362808127372
BARI_DM	10053	1018	437.086956521739	189.232066646377	3.06984462926058	0.00107085077343727
BATUMI_I	383752	42525	16684.8695652174	5914.46249756989	4.36897358727014	6.24159308748595e-06
BATUMI_LTR	4710	627	204.782608695652	118.651098502304	3.55847856980565	0.000186504609273053
BEL-10_DAn-I	0	0	0	0	NA	NA
BEL-10_DAn-LTR	0	0	0	0	NA	NA
BEL-10_DBp-I	0	0	0	0	NA	NA

f) By clicking on a particular TE in the table, a histogram of the percentage of 5'-overlaps of reads in opposite orientation is accessed. Reads with or without ping-pong partners, in sense and in antisense orientation, can be downloaded for further analysis.
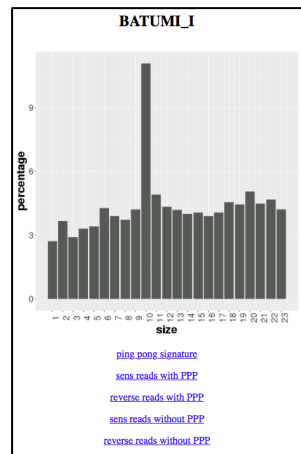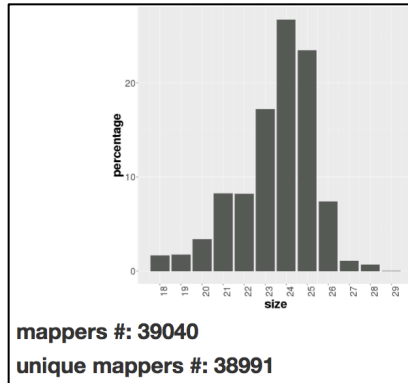

All results can be downloaded, either by a click on the corresponding link in the results folders, or as a whole by clicking on the name of the results item and then clicking the Save icon.
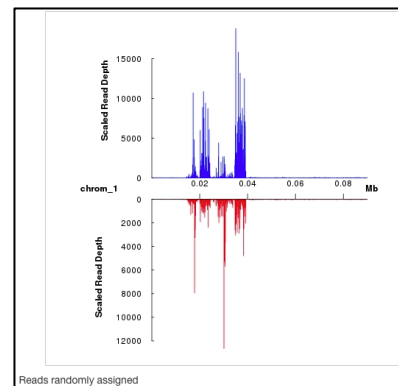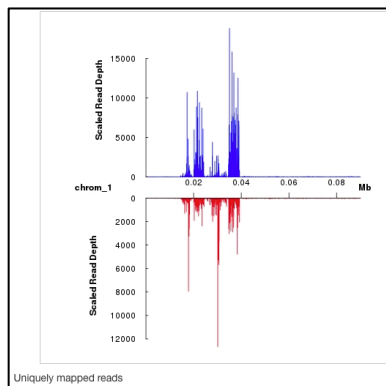
Results as obtained with the small data files for a quick test:

a) On the first page:

- The size distribution of genome-mapped reads and the number of all genome-mappers and of genome-unique mappers (reads that map only once to the genome) can be visualized.
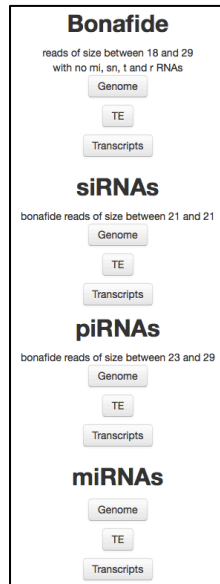

mappers #: 39040
unique mappers #: 38991

- Figures show mapping of genome-unique reads ("Uniquely mapped reads") and of all reads randomly assigned ("Reads randomly assigned") to the chromosomes. Small RNAs are mapped on plus and on minus strands, in blue and in red respectively.
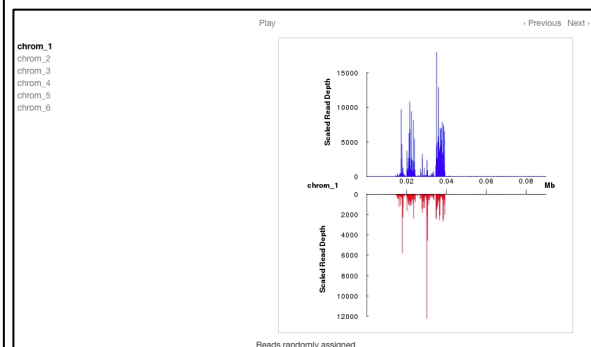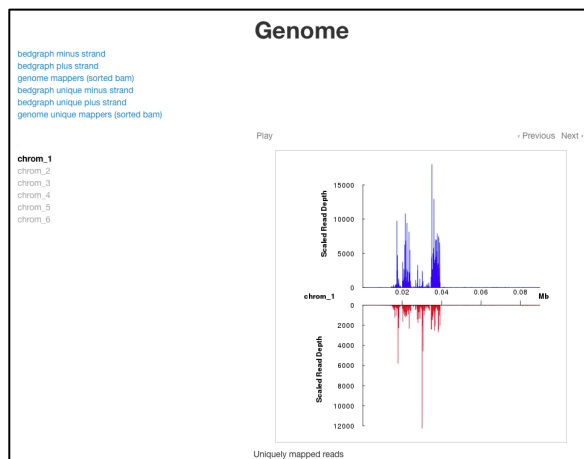

- The corresponding bam and bedgraph files are accessible *via* the corresponding links on the upper left, as well as the counts of the mappers size distribution and lists of the mapped and unmapped reads.

alignment (bam)
bedgraph minus strand
bedgraph plus strand
bedgraph unique minus strand
bedgraph unique plus strand
mappers (txt)
mappers size distribution (txt)
unique alignment (bam)
unique mappers (txt)
unmapped (txt)

b) The second page is accessible by clicking on "View details" on the first page of results. Then the information for 4 subgroups, Bonafide reads, siRNAs, piRNAs and miRNAs, is accessible. For each subgroup, three distinct analyses are accessible in different folders: for genome-mapping reads, TE-mapping reads and transcript-mapping reads.

c) By clicking on Genome, TE or Transcripts folders, figures for genome-unique mappers and for all mappers and tables with read counts for the corresponding features (genome, TEs or transcripts) can be visualized, the corresponding sorted BAM files and bedgraphs for the plus and minus strand mappers can be downloaded.
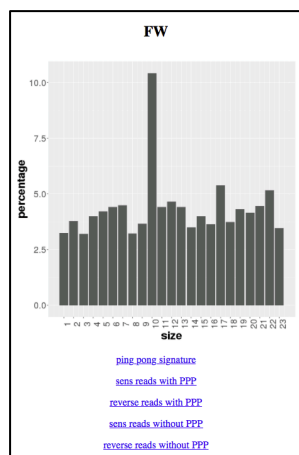
d) The results for Ping-pong signature ("PPPartners") are found only in the "piRNAs" subgroup "TE" folder, by clicking on "Ping Pong Partners".
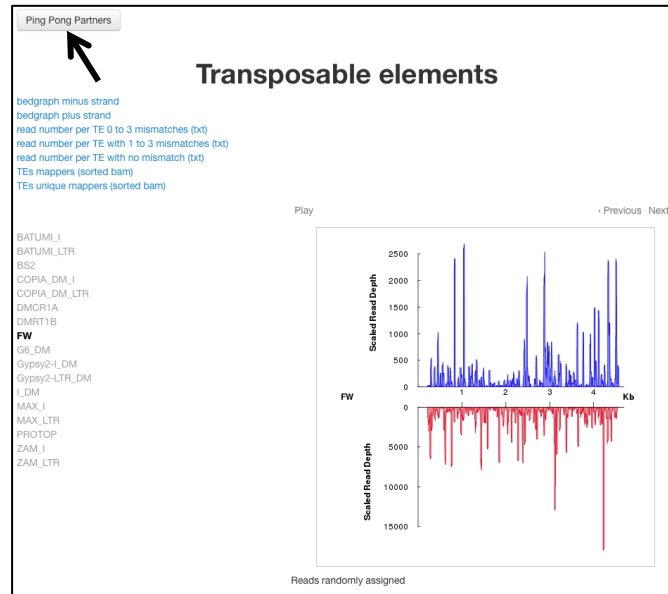
e) The ping-pong signature is analysed for each TE and a sum of all overlaps, the sum of 10-nt-overlaps, the mean, the standard deviation, the z-score and the p-value for each TE are summarized in a table.

ID	overlap sum	ten overlap sum	mean	standard deviation	z-score	p-value
BATUMI_I	92	22	4	6.26931797136554	2.87112570812537	0.00204506417965455
BATUMI_LTR	13	1	0.565217391304348	0.824941998304795	0.52704627669473	0.299080726341764
BS2	4418	610	192.086956521739	94.6508058457982	4.41531416181612	5.04316992888665e-06
COPIA_DM_I	0	0	0	0	NA	NA
COPIA_DM_LTR	0	0	0	0	NA	NA
DMCR1A	109	30	4.73913043478261	6.40799359023961	3.94208720865353	4.03877968415678e-05
DMRT1B	299	13	13	10.8266980794861	0	0.5
FW	9995	1042	434.565217391304	141.421610207381	4.29520482561295	8.7266059448643e-06
G6_DM	1084	129	47.1304347826087	28.7512636362392	2.84751189558847	0.0022031223713177
Gypsy2-I_DM	0	0	0	0	NA	NA
Gypsy2-LTR_DM	170	20	7.39130434782609	4.341299449171	2.90435981203315	0.00184002483108026
I_DM	2	1	0.0869565217391304	0.281771334713385	3.24037034920393	0.000596872722436026
MAX_I	54	11	2.34782608695652	2.44253424438055	3.5422938011818	0.00019833170283523
MAX_LTR	16	0	0.695652173913043	1.75535906419012	-0.396301923694456	0.654058845865745
PROTOP	49	11	2.1304347826087	2.30885531258291	3.84154224348902	6.11318444337927e-05
ZAM_I	0	0	0	0	NA	NA
ZAM_LTR	0	0	0	0	NA	NA

f) By clicking on a particular TE in the table, a histogram of the percentage of 5'-overlaps of reads in opposite orientation is accessed. Reads with or without ping-pong partners, in sense and in antisense orientation, can be downloaded for further analysis.

All results can be downloaded, either by a click on the corresponding link in the results folders, or as a whole by clicking on the name of the results item and then clicking the Save icon.