proTRAC

version 2.4.4

- Documentation -

1. Scope and prerequisites

1.1 Introduction

proTRAC predicts and analyzes genomic piRNA clusters based on mapped piRNA sequence reads. proTRAC applies a sliding window approach to detect loci that exhibit high sequence read coverage. Subsequently, sequences mapped to these loci are analyzed with respect to typical piRNA and piRNA cluster characteristics to ensure high specificity. The latest proTRAC version can be found at https://sourceforge.net/projects/protrac/files and http://www.smallRNAgroup-mainz.de. Running proTRAC on your local machine requires the installation of a Perl interpreter. Perl is pre-installed on common Linux and Mac systems. For Windows you can download and install either StrawberryPerl (www.strawberryperl.com) or ActivePerl (www.activestate.com/activeperl/downloads). For piRNA cluster prediction proTRAC requires a genome file in FASTA format and a map file in so-called ELAND3 format or SAM format that contains information about the mapped sequence reads.

1.2 Changes compared to previous versions

2.0.5 → 2.1 proTRAC now produces output files in a very nice html format. Some default values changed to produce best results when using mammalian piRNA datasets.

2.1 → 2.1.2 Bug fixed: Pseudogenes, lincRNAs and antisense transcripts in piRNA clusters were displayed as protein coding (when implementing Ensembl gene sets information).

2.1.2 → 2.2.0 Headers in FASTA files that contain mapped sequence reads per piRNA cluster are now less confusing. Default minimum piRNA cluster size changed from 5kb to 1kb. Fixed minor bugs and revised some code to speed up computation. proTRAC finally accepts map files in SAM format as output by popular aligners such as Bowtie or STAR. But check out 3.1 for the required FASTA/FASTQ format for mapping.

2.2.0 → 2.3.0 Input files in SAM format must not be sorted any longer. Sequence data used to generate SAM files must not be collapsed any longer. GD image files are disabled, proTRAC no longer needs the GD module (installation of the GD module may fail on some computers for cryptic reasons). Bug fixed: Mapped reads and genes in clusters are displayed correctly in HTML output files (that bug was introduced with V.2.2.0).

2.3.0 → 2.3.1 Bug fixed: proTRAC failed to create an output folder when the input map file was not located in the same folder as proTRAC or was specified with the absolute path.

2.3.1 → 2.4.0 proTRAC now calculates the repeat content of every predicted cluster and total repeat content of all clusters compared to the whole genome (when providing a RepeatMasker output file). These values will occur in the summary file. We have also added further transcription factor binding sites to search within predicted clusters.

2.4.0 → 2.4.1 Bug fixed: proTRAC tried to calculate transposon content of clusters even though no RepeatMasker file was provided via command line options. Bug fixed: Wrong calculation of total gap size in genomes.

2.4.1 → 2.4.2 Bug fixed: Missing '>' character in 'clusters.fas' file.

2.4.2 → 2.4.3 New test procedure for piRNA cluster candidates rejects candidates if any sliding window with size n bp comprises more than i % of the total reads of the cluster. Reduces false-positive rate. Additional gtf output that contains piRNA cluster coordinates.

2.4.3 → 2.4.4 A bug prevented users from using the option –dens.

2. Quick Start

You need a genome file (e.g. `genome.fasta`) and a file that contains your small RNA sequence reads (`seq.fasta`). Download and unzip NGS TOOLBOX (a collection of simple Perl scripts) here: http://www.smallrnagroup.uni-mainz.de/software/TBr2.zip. Open a terminal/command prompt window and type the following commands:

```
perl TBr2_collapse.pl -i seq.fasta -o seq.collapsed
```

This will remove redundant sequences from `seq.fasta`

```
perl TBr2_duster.pl -i seq.collapsed
```
This will remove low-complexity reads from `seq.fasta.collapsed` and produce an output file named `seq.fasta.collapsed.no-dust`.
Download sRNAmapper and map your sequences to a genome using the following command (by default allows non-template 3' nucleotides which increases sensitivity):
```
perl sRNAmapper.pl -input seq.fasta.collapsed.no-dust -genome genome.fasta -alignments best
```
The map file will have the name `seq.fasta.collapsed.no-dust.map`.

Alternatively you may want to use a popular aligner such as Bowtie or STAR to map sequences in FASTA/FASTQ file to a genome. The map file must be in SAM format. You can convert BAM files to SAM files using SAM tools.

Download proTRAC (http://www.smallrnagroup.uni-mainz.de/software/proTRAC_2.3.1.pl) and start piRNA cluster prediction with the following command:
```
perl proTRAC_2.3.1.pl -map seq.fasta.collapsed.no-dust.map -genome genome.fasta
```

3. An exhaustive manual

3.1 Create a map file (mandatory)

There are two simple ways to create map files for proTRAC. You need a genome file in FASTA format and a file that contains your piRNA sequence reads in FASTA or FASTQ format. We recommend to use a collapsed format which does not contain redundant sequences. Instead, information on read counts for each sequence should be provided in the header (see below). If you intend to use sRNAmapper or SeqMap to create a map file in ELAND3 format you must collapse your sequence files.

FASTA

```
>34
TGGCACGTGCGATGGCAGTTCAACGTTCA
>23
TAAATCCCGTAGGCTTTTAGCATCGACG
>2
TTATTCGAAGCGCTTTACGACACGCGCGCA
```

or FASTQ

```
@34
TGGCACGTGCGATGGCAGTTCAACGTTCA
+
AAABBGHHHHHHHHHHHHHHKKLLLLLLLL
@23
TAAATCCCGTAGGCTTTTAGCATCGACG
+
@@19000ACCCCGHHHFFFBBBB(6321
```

You can convert your FASTA/FASTQ file to collapsed format using the `collapse` tool (Perl script) from the NGS TOOLBOX which you can find in the software section of this page (or download NGS TOOLBOX from https://sourceforge.net/projects/NGS_TOOLBOX/files/). We recommend to apply an additional filtering step prior to genomic mapping to remove low complexity reads from your datasets. Otherwise your map file may become extraordinary large due to multiple mapping of low complexity reads to low complexity regions in the genome. For this purpose you can use the `duster` tool from the NGS TOOLBOX which you can find in the software section of this page (or download NGS TOOLBOX from https://sourceforge.net/projects/NGS TOOLBOX/files/) or use a third party software like DUST (Morqulis et al. 2006, *J Comput Biol*). Alternatively you may use a genome with masked low complexity regions.
For mapping your piRNAs to the genome, we recommend the Perl script sRNAmapper.pl (which displaced the older piRmapper_1.0.pl) which you can find in the software section of this page (or download from https://sourceforge.net/projects/protrac/files). sRNAmapper will allow non-template nucleotides (default value =2) and internal mismatch downstreem of a defined 5' seed region (default length for seed region =18, default maximum internal mismatch =1). Output files are sorted according to chromosome coordinates by default. Start the mapping procedure from the command line or terminal with the following command:
```
perl sRNAmapper.pl -input piRNAs.fasta -genome genome.fasta -alignments best
```

With the above command, sRNAmapper will create a map file that is named piRNAs.fasta.map (you can give it a different name using the option `-output othername.map`) and that will have the following format:

```
Chr1    238     TGTTACGGCTAGCTCAGTACGGC       23      TGTTACGGCTAGCTCAGTACGAA       2       +
Chr1    291     TACGCCAGCTCGACTCGCCTGTGCA     23      TACGCCAGCTCGACTCGCCTGTGCA     0       -
```

The columns refer to: Chromosome, leftmost mapping position of the mapped sequence, reference sequence, FASTA/FASTQ header of the mapped sequence, mapped sequence, mismatch to reference sequence, strand. You can alternatively use the SeqMap software (Jiang and Wong 2008, *Bioinformatics*) to map your piRNAs using the following command:

```
seqmap 0 piRNAs.fasta genome.fasta output.map /output_all_matches
```

The `0` in the above command refers to the number of allowed mismatches. You may adjust this value as desired and/or add the option `/allow_insdel:n` where `n` refers to the maximum number of allowed insertions and deletions. The output format will be the same as above.

You can alternatively use a more popular aligner such as Bowtie or STAR. Doing so, output your map file in SAM format. BAM files can be converted into SAM files with SAM tools. You should make sure to allow multi-mapping. Note: Currently, SAM files are not accepted by reallocate (but we are working on that). If you want to reallocate sequence read counts according to estimated local transcription rates using the reallocate.pl Perl script (see 3.2), you have to use sRNAmapper or SeqMap as described above.

3.2 Post process your map file (optionally, proTRAC 2.0.5 and later)

Typically, a large proportion of sequences will produce more than one genomic hit. In practice this means that the real origin of a sequence with multiple genomic hits remains unknown. Therefore, when searching for genomic piRNA clusters, proTRAC by default apportions the read counts for the sequence in question equally to each hit locus. However, you have the option to process your map file in order to allocate read counts of multiple mapping sequences according to the transcription rate of the genomic region (calculated on the basis of uniquely mapping sequence reads). To this end, you may want to use the Perl script `reallocate` which you can find in the software section of this page (or download from https://sourceforge.net/projects/protrac/files). Start the post processing procedure from the command line or terminal with the following exemplary command:

```
perl reallocate.pl mapfile.map 10000 1000 b 0
```

`10000` in the above command refers to the up- and downstream region that is considered for calculation of transcription rates. `1000` refers to the sliding window increment. `b` refers to the form of the function that is used to weight the sequence read counts according to the distance between the center of the window and the hit locus. `0` represents the threshold value for the minimum number of allocated reads per locus. A threshold of `0` will remove hits with no allocated read counts. Use higher values as desired. Use `-1` to keep all hits.

3.3 Run proTRAC

Once you have created a map file you can start proTRAC from the command line or terminal using the following command:

```
perl proTRAC_2.3.1.pl -map piRNAs.fasta.map -genome genome.fasta
```

3.4 Include RepeatMasker annotation (optionally)

You can optionally include RepeatMasker annotation in proTRAC image output files. Therefore, you need a RepeatMasker annotation that can be obtained by running RepeatMasker on the genome you use for mapping. If you do not want to install the RepeatMasker software on your local machine you can find Repeatmasker outputs (assembly_name.out.gz) for selected model organisms on the RepeatMasker website (http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html). Make sure that the chromosome or scaffold names in the RepeatMasker output (query sequence column) correspond to the FASTA headers in the genome file. RepeatMasker output files should look like this:

```
   SW  perc perc perc  query    position in query           matching    repeat          position in  repeat
score  div. del. ins.  sequence  begin     end    (left)    repeat      class/family    begin  end (left)  ID

 1741   7.8  5.5 14.0  Bf_V2_1       1     358 (7763997) +  hAT-N3_BF   DNA/hAT-Tip100    410  735 (1077)   1
   94   2.3  0.0  0.0  Bf_V2_1     848     936 (7763419) +  (TA)n       Simple_repeat       1   89    (0)   2
 1123   3.7  0.7  0.0  Bf_V2_1    1236    1369 (7762986) C  CR1-7_BF    LINE/CR1          (1) 4088   3954   3
  293  27.7  7.7  7.3  Bf_V2_1    2023    2295 (7762060) +  hATw-2_BF   DNA/hAT-hATw     5524 5797  (408)   4
```

To include RepeatMasker annotation add the option:

```
-repeatmasker RepeatMasker_output.out
```

3.5 Include Ensembl Gene sets (optionally)

You can optionally include information from Ensembl Gene sets in proTRAC image output files. Gene sets can be downloaded from http://www.ensembl.org/info/data/ftp/index.html. Make sure that the chromosome or scaffold names in the Gene sets file (first column) correspond to the FASTA headers in the genome file. Gene Set files should look like this:

```
#!genome-build Zv9
#!genome-version Zv9
#!genome-date 2010-04
#!genome-build-accession NCBI:GCA_000002035.2
#!genebuild-last-updated 2014-02
7       ensembl  gene     100958   101715   .       +       .       gene:id [...]
7       ensembl  exon     100958   100975   .       +       .       gene:id [...]
7       ensembl  CDS      100958   100975   .       +       0       gene:id [...]
```

To include Ensembl Gene Sets information add the option:
`-geneset GeneSets.gtf`

4. How proTRAC works

Initially proTRAC reads the provided map file and saves information about the total number of mapped reads and the number of reads and genomic hits for each sequence. For the prediction of a piRNA cluster proTRAC needs a threshold value for mapped reads per kb. There are three different ways to set this value:

1. Directly set a value with the option -dens [read counts per kb]
2. Use the option `-pdens [p-value]`. **This is the default and recommended method**. Applying this option proTRAC will scan the map file with a sliding window and save read counts per kb for each sliding window. Running proTRAC with the option `-pdens 0.01` (=default) will set the threshold for mapped reads per kb to the value that corresponds to the top 1% of sliding windows in terms of read coverage. In other words. A maximum of 1% of the genome can be annotated as piRNA clusters. In mammals, piRNA clusters typically represent 0.1% to 0.3% of the whole genome. This option will ensure that the sensitivity/specificity of proTRAC will remain mostly equal irrespective of the total number of mapped sequence reads.
3. Use the option `-pdens [p-value]` in combination with the option `-est [integer]`. Instead of scanning the map file with a sliding window, proTRAC will create [integer] random sliding windows based on the data obtained from the map file and save read counts per kb. This approach may be faster especially for large map files.

Once a threshold for mapped reads per kb is defined proTRAC again reads the map file with a sliding window. The default sliding window size is 5kb and can be adjusted with the option `-swsize [bp]`. The default sliding window increment is 1kb and can be adjusted with the option `-swincr [bp]`. If a genomic locus exhibits an appropriate sequence read coverage thus represents a putative piRNA cluster proTRAC starts to analyze the locus and the reads that map to this locus in detail and checks whether

1. the locus exhibits the minimum size of a typical piRNA cluster. The default value is 5kb and can be adjusted with the option `-clsize[bp]`.
2. a defined minimum fraction of mapped reads exhibits the typical size of piRNAs. The default minimum size is 24 nt (can be adjusted with the option `-pimin [nt]`), the default maximum is 30 nt (can be adjusted with the option `-pimax [nt]`). The default minimum fraction is 0.75 and can be adjusted with the option `-pisize [0..1]`. If you do not want proTRAC to consider length of mapped sequences as a criterion start proTRAC with the option `-pisize 0.0`.
3. a defined minimum fraction of mapped reads exhibits U at position 1 (1U) or A at position 10 (10A). The default minimum fraction is 0.75 and can be adjusted with the option `-1Tor10A [0..1]`. If you do not want proTRAC to consider positional nucleotide frequencies of mapped sequences as a criterion start proTRAC with the option `-1Tor10A 0.0`. The option `-1Tand10A` (note the 'and' instead of 'or') allows to bypass the option -1Tor10A if both the fraction of 1U and 10A reads is above a stated minimum. The default value for this option is 0.5.
4. a defined minimum fraction of reads maps to the main strand of the putative piRNA cluster. For mono-directional clusters the main strand is simply represented by the strand with the majority of

mapped sequence reads. However, piRNA clusters may also be organized in a bi-directional manner. Hence, proTRAC splits each cluster between each pair of mapped reads to check whether the main strand switches. The default value for the minimum fraction of reads mapped to the main strand is 0.75 and can be adjusted with the option `-clstrand [0..1]`. In case of bi-directional piRNA clusters the user can set the minimum fraction of sequence reads that map to each arm of the cluster with the option `-clsplit [0..1]`. The default value for `-clsplit` is 0.1. If you do not want proTRAC to consider strand bias as a criterion start proTRAC with the option `-clstrand 0.5`.

5. the observed sequence read coverage is caused by a low number of sequences with an exceptionally high read count which would be atypical for genuine piRNA clusters. To this end proTRAC calculates the amount of mapped reads for the locus in question that is attributed to the top fraction of mapped sequences in terms of read counts. The parameters can be set with the option `-distr [int-int]` and the default is `-distr 1-90` which means that the top 1% of sequences mapped to this locus (sequences with highest read counts) should not account for more than 90% of all reads mapped to this locus. If you do not want proTRAC to consider how read counts distribute among different mapped sequences start proTRAC with the option `-distr 1-100`.
6. any sliding window with size n bp (default=1000) comprises more than i % (default=50) of total mapped reads of the cluster. The parameters can be set with the option `-spike [int-int]`. E.g. `-spike 50-1000` means that each 1000bp sliding window must comprise less than 50% of the reads in the cluster.
7. (NOT SET PER DEFAULT) the putative piRNA cluster exhibits a minimum number of hits at different sites. This number can be set with the option `-clhits [int]` and is `0` by default. Using this option may enhance specificity since genuine piRNA clusters usually exhibit a large number of sites corresponding to piRNAs.
8. (NOT SET PER DEFAULT) the putative piRNA cluster exhibits a minimum number of mapped sequence reads. This number can be set with the option `-clhitsn [int]` and is `0` by default. This option can be used if it is desired that smaller piRNA clusters should exhibit a higher amount of mapped reads per kb follows from the generally applied threshold.

5. Command line options

ftp = floating point number
int = integer
0..1 = floating point number between 0 and 1 (Type 1.0 for 1 and 0.0 for 0).

`-genome OR -g [file]`	Name of the file that contains the genomic sequence that was used for mapping the sequence reads.
`-map OR -m [file]`	Name of the file that contains mapped reads in ELAND3 format. Use SeqMap with option `/output_all_matches` or sRNAmapper to create an appropriate file.
`-format [SAM/ELAND3]`	Specify the input format. Allowed values are SAM and ELAND3. This is only required if the input file contains less than 1000 hits. proTRAC will check the first 1000 alignment lines in the map file and reliably determine the input file format.
`-help OR -h`	Shows this information.
`-repeatmasker [file]`	Name of the file that contains the RepeatMasker annotation. Make sure that the names for the chromosomes/scaffolds are identical in your RepeatMasker and genome file.
`-geneset [file]`	Name of the file that contains gene annotation (GTF-file from Ensembl database). Make sure that the names for the chromosomes/scaffolds are identical in your GTF- and genome file.
`-swsize [int]`	Size of the sliding window (default=5000)
`-swincr [int]`	Increment of the sliding window (default=1000)
`-nohc`	Do not consider total number of genomic hits for the sequence in question for calculation of hit counts.
`-norc`	Do not consider number of reads for the sequence in question for calculation of hit counts.
`-norpm`	Do not normalize the hit count with the total number of mapped sequence reads (reads per million). Normalization will make the values comparable across different piRNA libraries and/or organisms.

`-dens [fpt]`	Define an absolute minimum number of (normalized) read counts per kb.
`-pdens [0..1]`	Define a p-value for minimum number of (normalized) read counts per kb. A p-value of 0.01 means that the (normalized) read counts in a sliding window must belong to the top 1% of all sliding windows.
`-est [int]`	Use that option together with `-pdens`. Estimate the required minimum number of (normalized) read counts in a sliding windows based on n random 1kb sliding windows (faster). Without that option proTRAC will scan the map file and calculate the required minimum number of (normalized) read counts in a sliding window based on the observed distribution. We recommend not to use this option.
`-pisize [0..1]`	Fraction of (normalized) read counts that have the typical piRNA size (default=0.75).
`-pimin [int]`	Define the minimum length of a typical piRNA (default=24).
`-pimax [int]`	Define the maximum length of a typical piRNA (default=32).
`-1Tor10A [0..1]`	Fraction of (normalized) read counts that have 1T (1U) or 10A (default=0.75).
`-1Tand10A [0..1]`	If the fraction of (normalized) read counts with 1T (1U) OR 10A is below the value defined by `-1Tor10A`, accept the sliding window if BOTH the 1T (1U) and the 10A fraction reach this value (default=0.5).
`-distr [ftp-ftp]`	To avoid false positive piRNA cluster annotation of loci with one or few mapped sequences represented by exceptionally many reads. You can e.g. type `-distr 10-75` which means that the TOP 10% of mapped sequences account for max. 75% of the piRNA clusters (normalized) read counts (default=1-90). Otherwise the locus is rejected.
`-spike [ftp-ftp]`	Somewhat similar to `-distr`. With e.g. 50-1000 you allow max. 50% of reads in a cluster to be within any 1000bp sliding window. Otherwise the locus is rejected (default=50-1000).
`-clsize [int]`	Set the minimum size for a piRNA cluster (default=5000).
`-clhits [int]`	Minimum number of sequence hit loci per piRNA cluster (default=0).
`-clhitsn [ftp]`	Minimum number of normalized sequence read counts per piRNA cluster (default=0).
`-clstrand [0..1]`	Fraction of (normalized) read counts that map to the main strand (default=0.75).
`-clsplit [0..1]`	Minimum fraction of (normalized) read counts on the smaller arm of a bi-directional piRNA cluster. Otherwise the cluster will be annotated as mono-directional (default=0.1).
`-nohtml`	Do not output .html files for each piRNA cluster.
`-notable`	Do not output a summary table.
`-nofaspi`	Do not output a FASTA file comprising mapped piRNAs for each cluster.
`-nofascl`	Do not output a FASTA file comprising all piRNA cluster sequences.
`-nomotif`	Do not search for transcription factor binding sites.
`-flank [int]`	Include n bp of flanking sequence in output files.
`-pti`	Output a file that contains information on mapped sequence reads in a tab-delimited table that comprises sequence, reads, genomic hits e.g: TGGGCACGCAAATTCGAGTATCG 12 4

6. Output files

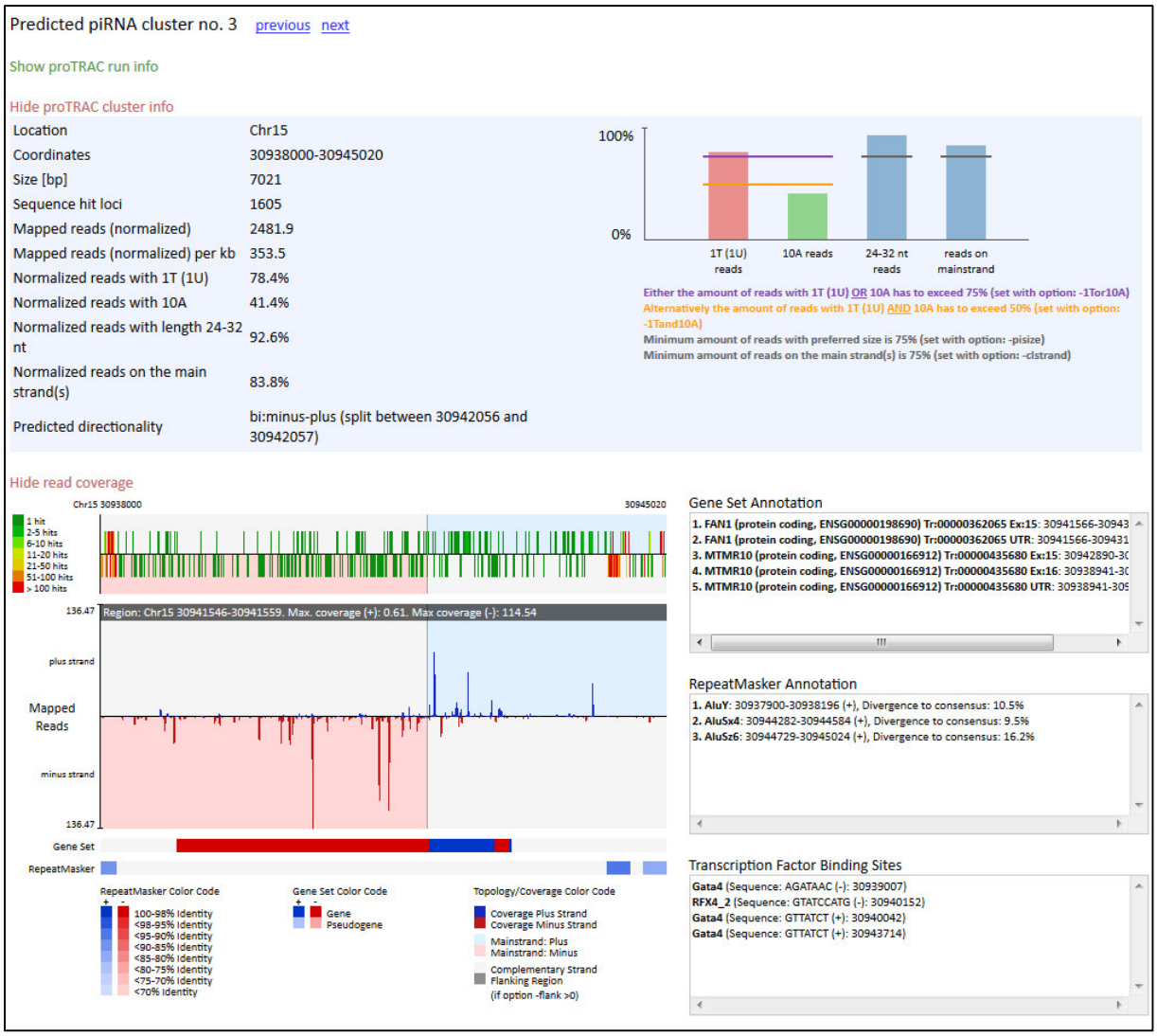
5.1 Image files (only versions 2.0 - 2.2.0)

From proTRAC version 2.1 image files are replaced by HTML files. If you want proTRAC to output image files you have to use the option `-image`. If doing so, proTRAC will output one separate image file for each predicted piRNA cluster (1.png, 2.png ...). The image file is separated into different sections. At the top you will find the distribution of normalized hit counts (blue: hits on the plus strand, red: hits on the minus strand). The main strand is indicated by blue or red background shading. In case of bi-directional clusters (see above) the main strand switches inside the cluster. The position of identified transcription factor binding sites will be indicated with green lines. Additional information for identified binding sites is provided on the right side. Below you will find a second chart that shows the same locus. Each hit is displayed with a bar of equal size where the color refers to the number of genomic hits produced by the sequence in question. Dark green indicates a hit to a single copy locus whereas dark red indicates a hit to a multi-copy locus (e.g. a transposon). The color code legend is placed on the left side. If you included RepeatMasker and/or GeneSet annotation you will find according tracks between the two charts. Transposons are indicated with blue (plus strand) and red (minus strand) bars with color saturation that corresponds to its divergence from the consensus sequence. Each bar contains a number that refers to the list of transposon names in the RepeatMasker annotation section. This

section also comprises the according color code legend. The track for GeneSet annotation comprises coding sequences (blue: plus strand, red: minus strand). As for the RepeatMasker track, numbers in bars refer to a list in the GenSet annotation section. Finally, you will find some cluster data (coordinates, normalized hit count etc) on the bottom of the image. This data is also included in the tabular output of proTRAC (results.table).

5.2 HTML files

From proTRAC version 2.1 image files are replaced by HTML files. In the cluster info section you can find some basic data on the predicted piRNA cluster. The plot on the right side shows parameter values (e.g. reads with 1U, reads with typical piRNA length etc.) that are important for piRNA cluster prediction. The lines indicate the applied (default or user set) thresholds. The read coverage section starts with a topology plot where each hit is indicated by a bar. The color of the bar refers to the number of genomic hits produced by the sequence in question. Many adjacent red or yellow bars can indicate the presence of a multi copy locus such as tRNA genes or transposons. Below the topology plot you find the sequence read coverage plot (blue: reads on the plus strand, red: reads on the minus strand). The main strands are indicated by light blue or light red background color. You can get exact values of mapped reads by moving the mouse cursor over a desired position. Below the coverage plot you will find Gene Set and RepeatMasker annotation. You can include Ensembl GeneSets annotation files (.gtf) using the option `-geneset [file.gtf]`. You can include RepeatMasker annotation using the option `-repeatmasker [file.rm]`. A list of Gene Set and RepeatMasker elements for the predicted piRNA cluster can be found on the left side. Move the mouse cursor over a desired element to highlight it in the annotation bar. Transcription factor binding sites are listed below Gene Set and RepeatMasker annotation. Move the mouse cursor over putative binding site and you will see its position inside the piRNA cluster.



5.2 FASTA files with hit loci sequences (one for each piRNA cluster)

proTRAC will output one separate FASTA file for each predicted piRNA cluster (1.fasta, 2.fasta ...) that comprises mapped sequences in genome coordinates order (one sequence may occur several times in one file). The FASTA header comprises the following information: Chromosome, starting coordinate of the hit locus, read count of the mapped sequence, total number of genomic hits for the mapped sequence, the normalized hit count for this locus (usually reads/hits, but can differ if you processed the map file with reallocate), strand. In case you used the option -flank [value] the FASTA header might end with 'FLANK' if the hit locus is not within the predicted piRNA cluster but within the flanking region.

```
>Location:Chr15  Coordinate:92566255      Reads:1  Hits:2   Allocated_reads:0.985261992396835 Strand:-
GGAAGGCTGAGCTTCAGAGTGATGT
>Location:Chr15  Coordinate:92566258      Reads:2  Hits:2   Allocated_reads:1.97052398479367  Strand:-
TGTGGAAGGCTGAGCTTCAGAGTGA
>Location:Chr15  Coordinate:92566298      Reads:1  Hits:2   Allocated_reads:0.944347645866361 Strand:-
AGCAAAGAACTTGGCTGCATCGTGCC
>Location:Chr15  Coordinate:92566300      Reads:1  Hits:4   Allocated_reads:0.585764890726597 Strand:-
TAGCAAAGAACTTGGCTGCATTGTG
>Location:Chr15  Coordinate:92566307      Reads:1  Hits:6   Allocated_reads:0.20698211463811  Strand:-
TAGCAAAGAACTTGGCTG
>Location:Chr15  Coordinate:92566309      Reads:1  Hits:1   Allocated_reads:1        Strand:-
TCGTCACACATTAGCAAAGAACTTGGC
```

5.3 FASTA file with piRNA cluster sequences

proTRAC will output one FASTA file that comprises the sequences of all predicted piRNA clusters. The FASTA header comprises the following information: Chromosome, cluster coordinates, included flanking sequence ('+-0 bp' if you do not use the -flank option), predicted directionality of the cluster.

```
>Chromosome_1 15411306-15416703 (+-0 bp) directionality: mono:plus
ATTTCTGTGTGGAGTTTGCAGGTTCTCCCCATGTTCAAGTGTATGAGTGTAAATGGGTGTTTCCCAGTACTGGGTTGCAGCTGGAA
GGGCATCCGCTGTGTAAAACATATGCTGGATAAGTTGGCGGTTCATTCTGCTGTGGCAACCCCTGATAAATAAAGGGACTGAGCTG
AAAAAGAAAATGAATGAATATATATATATTCTTGAATATGTAAAACAATGTATATGAACTCAAAAGGTTAATCAAGTGTTATGAAA
TTAAATTTAGATTAATATTTAATATAAGAGGTATAATAACAAACATTTGCTTTACAGTGTTAGCTATAATATATATATATATATAT
ATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATTTCTGGCTTAAGCGTA
```

5.4 Summary table of the proTRAC results

proTRAC will output one summary table that is devided into several sections. At the top you will find information on proTRAC version and the parameters set for the analysis. This is followed by a TAB-delimited list of predicted piRNA clusters where each line comprises the following information: Location, coordinates, size[bp], absolute hits, normalized hits, normalized hits per kb, normalized hits with 1T, normalized hits with 10A, normalized hits with typical piRNA size, normalized hits on the predicted main strand, predicted directionality, putative transcription factor binding sites. At the bottom you will see the total size of the predicted piRNA clusters (% of the genome), the number of non-identical sequences inside piRNA clusters (% of all non-identical sequences in the map file) and the number of sequence reads inside piRNA clusters (% of all sequence reads in the map file).

```
================================ proTRAC ===================================
VERSION: 2.3.1
LAST MODIFIED: 22.MARCH 2017

Please cite:
Rosenkranz D, Zischler H. proTRAC - a software for probabilistic piRNA cluster
detection, visualization and analysis. 2012. BMC Bioinformatics 13:5.

Contact:
David Rosenkranz
Institute of Organismic and Molecular Evolutionary Biology
Dept.  Anthropology, small RNA group
Johannes Gutenberg University Mainz
email: rosenkrd@uni-mainz.de

You can find the latest proTRAC version at:
http://sourceforge.net/projects/protrac/files
============================================================================

PARAMETERS:
Map file: ..............piRNA.map
Genome file: ............Danio_rerio.Zv9.dna
RepeatMasker annotation: n.a.
GeneSet:.................n.a.

Significant (p<=0.01) hit density will be calculated based
on observed hit distribution.

Sliding window size: ..................................... 5000 bp
Sliding window increament: ............................... 1000 bp
Normalize each hit by number of genomic hits: ............ yes
Normalize each hit by number of sequence reads: .......... yes
Normalize values (-> per million mapped reads): .......... yes
Min. fraction of hits with 1T(U) or 10A: ................. 0.75
Alternatively: Min. fraction of hits with 1T(U) and 10A: .... 0.5
Min. fraction of hits with typical piRNA length: ......... 0.75
Typical piRNA length: .................................... 24-32 nt
Min. size of a piRNA cluster: ............................. 5000 bp.
Min. number of hits (absolute): .......................... 0
Min. number of hits (normalized): ........................ 0
Min. fraction of hits on the mainstrand: ................. 0.5
Top fraction of mapped sequences (in terms of read counts): . 1%
Top fraction accounts for max. n% of sequence reads: ..... 90%
Max. percentage of reads within any x bp sliding window: .... 50%
Size of sliding window with max x % of reads (see above): ... 1000 bp
Min. fraction of hits on each arm of a bidirectional cluster: 0.1
Output image file for each cluster: ...................... yes
Output a summary table: .................................. yes
Output a FASTA file for each cluster (piRNA sequences): ..... yes
Output a FASTA file comprising cluster sequences: ......... yes
Search DNA motifs in clusters: ........................... yes
Output flanking sequences: +/- ........................... 0 bp
Output ~.pTi file: ....................................... no
============================================================================

[LIST OF PREDICTED CLUSTERS AND RELATED INFORMATION]

Total size of 75 predicted piRNA clusters: 795922 bp (0.056%)
Non identical sequences that can be assigned to clusters: 437960 (76.32%)
Sequence reads that can be assigned to clusters: 690440 (70.07%)
```

6. Contact

If you have any questions or comments or found any bugs in the software please do not hesitate to contact:

David Rosenkranz
Institute of Organismic and Molecular Evolutionary Biology
Dept. Anthropology, small RNA group
Johannes Gutenberg University Mainz, Germany
Email: rosenkranz@uni-mainz.de
Web: http://www.smallRNAgroup.uni-mainz.de

7. Citation

If you use the proTRAC software for your publication please cite the following paper:
Rosenkranz D, Zischler H. proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis. 2012. *BMC Bioinformatics* 13:5.