PPmeter

version 0.4

- Documentation -

1. Scope

PPmeter is is a tool to quantify and compare the amount of ongoing ping-pong amplification. Since the number of ping-pong pairs within a given datasets depends on dataset size and grows non-linearly, other methods must be applied when comparing the ping-pong footprint across different datasets. PPmeter generates pseudo-replicates by repeated bootstrapping (default=100) of a fixed number of sequence reads (default=1000000) from a set of original sRNA sequence datasets. PPmeter then calculates the ping-pong signature of each pseudo-replicate and counts the number of sequence reads that participate in the ping-pong amplification loop. The obtained parameter - ping-pong reads per million bootstrapped reads [**ppr-mbr**] - is directly comparable across different datasets.

2. Getting started

Running PPmeter 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). To rum PPmeter 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 while maintaining information on sequence read counts in the FASTA/FASTQ header:

FASTA

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

or FASTQ

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

Then type:

```
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. Although we recommend to sort out low-complexity reads, this step is not mendatory to run PPmeter.

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.

With the above command, sRNAmapper will create a map file that is named seq.fasta.collapsed.no-dust.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 seq.fasta.collapsed.no-dust 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.

Alternatively you may want to use a popular aligner such as Bowtie, bwa 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. ***Note that `PPmeter` requires SAM files that are sorted according to chromosome coordinates in ascending order!*** *Use SAM tools to make sure that your SAM file is a sorted SAM file.*

Now having an appropriate map file, you can start `PPmeter` with the following command:

```
perl PPmeter.pl –i map [-option (value)]
```

3. Output

`PPmeter` will create one output file that contains the results for each analyzed file. Essentially, it lists all the 5' overlaps for each pseudo-replicate dataset separately in one line, followed by corresponding ping-pong Z-score, number of ping-pong sequences (non-identical) and number of ping-pong reads for the given pseudo-replicate.
It also calculates the average number for a given 5' overlap across all pseudo-replicates together with the standard deviation, and also the average ping-pong Z-score, number of ping-pong sequences (non-identical) and number of ping-pong reads. We refer to the latter value as ping-pong reads per million bootstrapped reads [**ppr-mbr**]. In a narrow sense, this is the average number of ping-pong reads (reads with 10 nt 5' overlap with another mapped read) from all (default=100) pseudo-replicate datasets. This value is directly comparable across different datasets and can serve to quantify the amount of ping-pong amplification in a specific sRNA dataset.
For readability, we recommend to copy-paste the content of the output file into a spreadsheet using a software of your choice. ping-pong Z-scores are calculated according to Zhang et al. 2012 Mol Cell 44(4):572-584.

4. Test for significant differences of ping-pong signatures

We have implemented a simple Mann-Whitney U-test that allows the user to find out if the values for 10 nt 5' overlaps significantly differ between two datasets, or two groups of datasets. To account for biological variation we strongly recommend to use at least two sRNA datasets per group, while the two groups can contain different numbers of sRNA datasets. To run this test in addition to the regular analysis, use the following exemplary command:

```
perl PPmeter.pl –i map1 –i map2 –i map3 –i map4 [-option (value)] –c –g1 map1 –g1 map2 –g2 map3 –g2 map4
```

In the example above you have four sRNA datasets (in four map files: map1-map4). map1 and map2 represent biological replicates for condition A, while map3 and map4 represent biological replicates for condition B. Values for 10 nt 5' overlaps will be merged for each group and compared to the other group applying a standard Mann-Whitney U test. The results of this test (Z-score and p-value) will appear at the very bottom of the output file.

5. Command line options

```
[s] = string, e.g. a file name.
[i] = integer

-h OR -help             Will print this information.
-i OR -input [s]        sRNA map file in SAM or ELAND3 (sRNAmapper) format. Use
                        this option multiple times for multiple input files:
                        -i input_1.map -i input_2.map [...]. Input files must
                        be sorted by chromosomal position in ascending order.
-f OR -format [s]       Specify the input format. Allowed values are 'SAM' and
```

```
                        'ELAND3'. This is only required if on of the input files
                        contains less than 1000 hits.
-o OR -output [s]       Name of the output file (default='PPmeter_results.txt').
-t OR -threads [i]      Number of maximum allowed parallel threads.
-b OR -bootstraps [i]   Number of bootstrap pseudo-replicate datasets that will
                        be created for each input dataset (default=100).
-d OR -depth [i]        Number of sequence reads for each pseudo-replicate
                        (default=1000000).
-c OR -compare          Compare ping-pong signatures of pseudo-replicates across
                        different input files. This will perform a statistical
                        test (Mann-Whitney U test) to check whether the ping-
                        pong signatures are different between two input files.
                        By default all pairwise comparisons will be conducted.
                        Use the option -g1 AND -g2 to build two groups of input
                        files (e.g. each group comprising biological
                        replicates). Groups do not have to comprise the same
                        number of input files.
-g1 OR -group1 [s]      Specify files that will constitite group 1 for
                        statistical comparison of ping-pong signatures. This
                        option can only be used together with -c AND -g2.
-g2 OR -group2 [s]      Specify files that will constitite group 2 for
                        statistical comparison of ping-pong signatures. This
                        option can only be used together with -c AND -g1.
-less_memory            Will use a different algorithm for bootstrapping that
                        uses less memory but might be considerably slower.
-more_output            Temporary map files for bootstrap replicates and ping-
                        pong results for each temporary map file will not be
                        removed. Sequences producing 10 nt 5' overlaps will
                        be saved for each bootstrap replicate in a separate
                        FASTA file together with information on sequence read
                        length distribution and positional nucleotide
                        composition in an additional text file.
-silent                 Less output to STDOUT.
```

6. Cite PPmeter and contact

If you use this software please cite the following research article:

Julia Jehn, Daniel Gebert, Frank Pipilescu, Sarah Stern, Julian Simon Thilo Kiefer, Charlotte Hewel, David Rosenkranz. Conserved and dynamic expression of piRNAs and PIWI genes in germline and soma of mollusks. *Communications Biology* 2018. *In review*.

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 Evolution
Anthropology, small RNA group
Johannes Gutenberg University Mainz, Germany
Email: rosenkranz@uni-mainz.de
Web: http://www.smallRNAgroup-mainz.de