reallocate

version 1.1

- Documentation -

1. Scope

reallocate will process map files in order to reallocate read counts of multiple mapping sequences according to the transcription rate of genomic loci based on uniquely mapping reads. Map files must be in ELAND format and can be created using sRNAmapper which is provided along with the proTRAC software. reallocate will output a modified map file that contains two additional columns that refer to i) total number of genomic hits of a sequence and ii) read counts that are assigned to this locus. proTRAC 2.0.5 and later versions accept this format and utilize this information for cluster prediction. Generally, using reallocate will result in a higher amount of sequence reads that can be assigned to predicted piRNA clusters and may also alte number of predicted piRNA clusters (more true positives, less false positives). The latest reallocate version can be found at https://sourceforge.net/projects/protrac/files.

2. Getting started

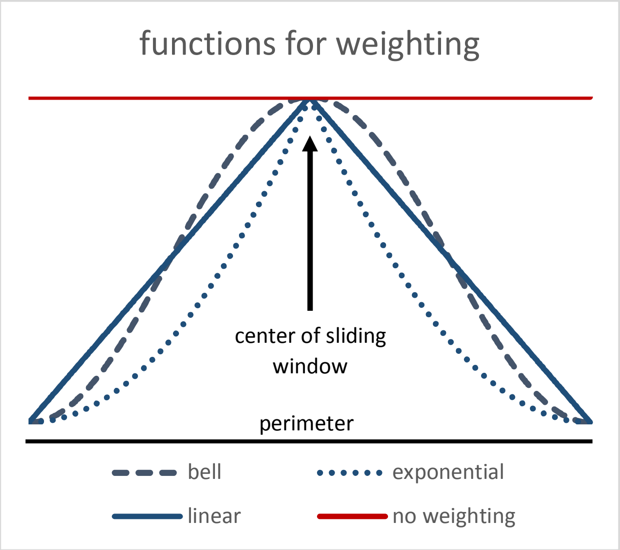
Running reallocate on your local machine requires the installation of a Perl interpreter. Perl is pre-installed on common Linux and Mac systems. For Windows you can download and install either StrawberryPerl (www.strawberryperl.com) or ActivePerl (www.activestate.com/activeperl/downloads). For information on how to create an appropriate map file take a look at the sRNAmapper documentation. Start reallocate from the command line using the following command:

```
perl reallocate.pl mapfile[file name] perimeter[bp] resolution[bp] shape[b,e,l,n] cutoff[value]
```

For example:

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

perimeter refers to the up/downstream region that will be considered for calculation of transcription rate. resolution refers to the space between the centers of the windows for that transcription rates will be calculated. shape refers to the shape of the function that determines how reads are weighted according to the distance from the center of the window. Valid values are b (bell-shaped), e (exponential), l (linear) or n (no weighting).
cutoff refers to the threshold value for the minimum number of assigned sequence reads for a locus to appear in the output file. Use -1 if you want to output all loci. Use 0 if you want to reject loci with no allocated sequence reads. Use higher thresholds as desired.
With the example above the output file will be named sRNA_data.map.weighted-10000-1000-b-0.



3. Output

A post-processed map file will add to columns a the input map file that refer to the number of total genomic hits produced by the sequence in question and the number of sequence reads allocated to the locus in question. For information on the ELAND format as output by `sRNAmapper` or `SeqMap` take a look at the `sRNAmapper` documentation.

As an example, the following ELAND output

```
Chr1   71      GATGGTCACAACGTCGATCGC        3   GATGGTCACAACGTCGATCGC        0   +
Chr1   238     TGTTACGGCTAGCTCAGTACGGC     15   TGTTACGGCTAGCTCAGTACGAA      2   +
Chr1   291     TACGCCAGCTCGACTCGCCTGTGCA   18   TACGCCAGCTCGACTCGCCTGTGCA    0   -
Chr1   470     TACGCTTATACGTCACAAA          8   TACGCTTATACGTCACAAA          0   +
Chr1   1390    AAATTGCGGTATTTTTCTTCTCGAT    5   AAATTGCGGTATTTTTCTTCTCGAT    0   +
Chr1   1466    TTCGAACGGCGGCGCGCGCGGCGC     7   TTCGAACGGCGGCGCGCGCGGCGC     0   +
Chr2   7913    GAGACTCTGATACGTCGTCG         1   GAGACTCTGATACGTCGTCG         0   +
Chr2   46887   GCGCGCGCGATCGATCGTTTTC       6   GCGCGCGCGATCGATCGTTTTC       0   +
Chr2   46946   TGTTACGGCTAGCTCAGTACGGC     15   TGTTACGGCTAGCTCAGTACGAA      2   +
Chr2   47090   TACGCCAGCTCGACTCGCCTGTGCA   18   TACGCCAGCTCGACTCGCCTGTGCA    0   -
Chr2   47204   CTATATCGTATAGCTAGCGATTAT     4   CTATATCGTATAGCTAGCGATTAT     0   +
Chr2   47204   AAAAAAAAAAAAAAAAAAAAGAGA     2   AAAAAAAAAAAAAAAAAAAAGAGA     0   +
```

will be processed to

```
Chr1   71      GATGGTCACAACGTCGATCGC        3   GATGGTCACAACGTCGATCGC        0   +   10      3
Chr1   238     TGTTACGGCTAGCTCAGTACGGC     15   TGTTACGGCTAGCTCAGTACGAA      2   +   2       10
Chr1   291     TACGCCAGCTCGACTCGCCTGTGCA   18   TACGCCAGCTCGACTCGCCTGTGCA    0   -   2       12
Chr1   470     TACGCTTATACGTCACAAA          8   TACGCTTATACGTCACAAA          0   +   1       8
Chr1   1390    AAATTGCGGTATTTTTCTTCTCGAT    5   AAATTGCGGTATTTTTCTTCTCGAT    0   +   1       5
Chr1   1466    TTCGAACGGCGGCGCGCGCGGCGC     7   TTCGAACGGCGGCGCGCGCGGCGC     0   +   1       7
Chr2   7913    GAGACTCTGATACGTCGTCG         1   GAGACTCTGATACGTCGTCG         0   +   1       1
Chr2   46887   GCGCGCGCGATCGATCGTTTTC       6   GCGCGCGCGATCGATCGTTTTC       0   +   1       6
Chr2   46946   TGTTACGGCTAGCTCAGTACGGC     15   TGTTACGGCTAGCTCAGTACGAA      2   +   2       5
Chr2   47090   TACGCCAGCTCGACTCGCCTGTGCA   18   TACGCCAGCTCGACTCGCCTGTGCA    0   -   2       6
Chr2   47204   CTATATCGTATAGCTAGCGATTAT     4   CTATATCGTATAGCTAGCGATTAT     0   +   1       4
Chr2   47204   AAAAAAAAAAAAAAAAAAAAGAGA     2   AAAAAAAAAAAAAAAAAAAAGAGA     0   +   10477   0
```

The two sequence reads highlighted in red both map to two genomic loci. Based on read counts of uniquely mapping reads within the relevant distance (blue=20 and orange=10) the read counts of the sequences highlighted in red are apportioned unequally to the two loci.
The obtained output file can be used as input file for proTRAC 2.0.5 and later versions.

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