



Version 1.1.0

- Documentation -

0. Change history

from	to	date	changes
1.0.0	1.0.1	08/05/2017	<ul style="list-style-type: none">• updated internal links.
1.0.1	1.1.0	22/05/2017	<ul style="list-style-type: none">• added <code>-riborase</code> option for most sensitive detection of rRNA fragments.• the option <code>-input</code> now accepts directories with input files in addition to standard file names.• automatic adjustment of FASTA headers with bad format in files provided via the option <code>-refseq</code>.• collapsed input FASTA files produced by <code>unitas</code> are sorted according to read counts.• minor improvements by adding the options <code>-quick</code>, <code>-slow</code>, <code>-skip_ncRNA</code>, <code>skip_mapping</code>• bug fixed: bad recognition of uncollapsed ELAND3 format.• bug fixed: bad detection of ping pong signature when not searching for phasiRNAs at the same time.

1. Scope

`unitas` is a convenient tool for small non-coding RNA (sncRNA, typically ~18-40 nt in length) annotation using Next Generation Sequencing data. `unitas` uses latest sequence information from publicly available online databases to annotate user input sequences. No installation, no further prerequisites; it runs out-of-the-box on all popular platforms (Linux, MacOS, Windows) and can be started with one simple command from the command line (terminal).

2. Getting started

All you need to start is a sequence file in FASTA or FASTQ format, or alternatively a map file in SAM or ELAND3 format (produced by SeqMap when using the option `/output_all_matches`, or by default by sRNAmapper), and a local copy of `unitas`. You can run the stand-alone `unitas` executable file (precompiled versions are available for Linux, MacOS and Windows) on your local machine without installation or any further requirements. Running the `unitas` Perl script 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 a free Perl distribution such as StrawberryPerl (www.strawberryperl.com) or ActivePerl (www.activestate.com/activeperl/downloads). When running `unitas` for the first time you need a connection to the internet in order to allow `unitas` to download a set of latest reference sequences. Subsequent runs can use previously downloaded data. You can start `unitas` from the command line (terminal) using the following command:

STAND-ALONE EXECUTABLES

Windows

```
unitas.exe -input sequence_or_map.file -species genus_species
```

MacOS or Linux

```
./unitas -input sequence_or_map.file -species genus_species
```

PERL SCRIPTS

Windows, MacOS or Linux

```
perl unitas.pl -input sequence_or_map.file -species genus_species
```

By default, `unitas` will not open more than one thread and multiple input files will be processed one after the other. We recommend to use the option `-threads [integer]` if more than one CPU core is available (which should be the case on most modern computers). During computation, `unitas` creates temporary files named, e.g., `-5388.unitas_pid` according to the process IDs of each forked process. `unitas` uses these temporary files to control the maximum number of parallel processes. Though this may not be the most sophisticated solution for this task, it is one that does not require the installation of additional Perl modules such as `Parallel::ForkManager` which are probably not part of a standard Perl distribution. You should not remove or rename these files while `unitas` is running. The files will be removed automatically when the annotation process for the input file is finished. You should also not run several instances of `unitas` in same directory.

3. Workflow

3.1. Prior to annotation

`unitas` creates a data dump folder on your local machine to store reference sequences downloaded from the internet. The folder will be named `UNITASrefdump_genus_species` (e.g. `UNITASrefdump_homo_sapiens` when

downloading human reference sequences). If present, subsequent annotation runs will use the information stored in this folder. You can force `unitas` to download the latest reference sequences using the option `-latest_ref`. This folder will contain a file named `db_versions.info` that provides information on database version or release date:

```
SeqMap version/date: ..... 1.0.13
Genomic tRNA database version/date: ... 30.05.2017 (dd.mm.yyyy)
piRNA cluster database version/date: .. 30.05.2017 (dd.mm.yyyy)
Ensembl version/date: ..... Release 88
EnsemblGenomes version/date: ..... Release 35
tRF-1 sequence data version/date: .... 30.05.2017 (dd.mm.yyyy)
tRNA-leader sequence data version/date: 30.05.2017 (dd.mm.yyyy)
SILVA rRNA (SSU) database version/date: Release 128.1
SILVA rRNA (LSU) database version/date: Release 128.1
miRBase database version/date: ..... Release 21
```

3.1.1 Input file check and conversion

`unitas` will check the format of the input file(s) which must be provided in FASTA, FASTQ, SAM or ELAND3 format. You should not use multiple files with identical names from different folders. Input files will be converted into FASTA format. `unitas` uses collapsed input files in FASTA format as this greatly reduces computation time. Collapsed input files contain only non-identical sequences with FASTA headers referring to read counts of each sequence. If necessary, `unitas` will automatically collapse the input files provided by the user, e.g.:

```
>SRR029124.2054760 WICMT-SOLEXA_309ETAAXX:2:43:1367:1703 length=18
AGCGGTAGGGATCCAAA
>SRR029124.2054761 WICMT-SOLEXA_309ETAAXX:2:43:1151:254 length=22
TAGCAGCACGTAATATTGGCG
>SRR029124.2054763 WICMT-SOLEXA_309ETAAXX:2:43:1785:135 length=18
AGCGGTAGGGATCCAAA
```

will be collapsed to:

```
>2
AGCGGTAGGGATCCAAA
>1
TAGCAGCACGTAATATTGGCG
```

In addition to reference sequences from online databases, the user can provide an arbitrary number of additional sequence files in FASTA format that will be used for sequence annotation (use the option `-refseq [sequence.file]`). The FASTA headers should have the following format:

```
>ncRNA_class|ncRNA_name
```

Where the ncRNA class (e.g. tRNA) is separated from its name (e.g. tRNA-Gly-GGT) with a pipe symbol. Do not use identical FASTA headers for different sequences. `unitas` will convert FASTA headers with bad format automatically using `'refseq'` as description for ncRNA class and the original FASTA header for the ncRNA name.

3.1.2 3' adapter recognition and trimming

In most cases, sequence reads from NGS datasets comprise 3' adapter sequences. When using datasets from NCBI's Sequence Read Archive it can be difficult to figure out which adapter was used to create the dataset in question. When using the option `-trimm`, `unitas` screens for the most frequent 3' sequence motifs (default length = 12 nt but is adjusted automatically when sequence reads in input file are shorter). A first round of adapter trimming is then performed based on the identified motif allowing 2 mismatches for 12 nt motifs, 1 mismatch for motifs ≤ 11 nt and 0 mismatch for motifs ≤ 8 nt. If the original motif is not found within a given sequence read, `unitas` truncates the motif sequentially by one 3' nt and checks for its occurrence at the very 3' end of the sequence read until the motif is found or the motif length falls below 6 nt. Following this first round of adapter trimming, `unitas` checks the positional nucleotide composition of the trimmed sequence reads and will remove further 3' nucleotide positions in case they exceed a specified nucleotide bias (default = 0.8, change this value with the option `-trim_maxfinalbias`). Recently, a nice tool for adapter prediction named `DNApi` has been published by Tsuji and Weng (<http://dx.doi.org/10.1371/journal.pone.0164228>). Noteworthy, `unitas` and `DNApi` reported identical

adapter sequences for the datasets that we have tested. However, if you prefer to use DNApi you can simply copy the software into the `unitas` folder. `unitas` will check for the presence of the DNApi Python script (`dnapi.py`) and call it with a system command. Note that there might be additional dependencies for DNApi, at least you should have Python installed on your computer. If, for what reason ever, adapter prediction with DNApi fails, `unitas` will continue adapter prediction with its own algorithm.

Relevant command line options:

option	default	explanation
<code>-trim</code>	<code>off</code>	Apply adapter recognition and trimming
<code>-trim_ignore_5p</code> [integer]	22	Ignore first <i>n</i> basepairs for 3' adapter prediction
<code>-trim_skip_reads</code> [integer]	10000	Skip the first <i>n</i> reads when searching for frequent motifs.
<code>-trim_check_reads</code> [integer]	500000	Check <i>n</i> reads for 3' adapter prediction.
<code>-trim_minlength</code> [integer]	10	Minimum length of trimmed read for further processing
<code>-trim_maxlength</code> [integer]	50	Maximum length of trimmed read for further processing
<code>-trim_maxfinalbias</code> [floating point, 0..1]	0.8	Maximum allowed 3' nucleotide bias after trimming

3.1.3 Low complexity filter

`unitas` can filter out low complexity sequences from the input file(s). This filter is an improved implementation of the `duster` tool provided as a part of the NGS TOOLBOX (<http://www.smallrnagroup.uni-mainz.de/software/TBr2.zip>). By default, it removes sequences that consist for $\geq 75\%$ (change default value with option `-dust_limit` [fraction]) of simple sequence repeats with a minimum motif size of 1 nt and a maximum motif size of 5 nt (change default value with option `-dust_max` [integer]). In addition, `unitas` will filter out reads that exhibit a high fraction of N positions (default = 0.5, change this value with the option `-dust_maxN` [floating point, 0..1]). The filtering step can be skipped using the option `-skip_dust`.

Relevant command line options:

option	default	explanation
<code>-skip_dust</code>	<code>off</code>	Do not filter low complexity reads
<code>-dust_limit</code> [floating point, 0..1]	0.75	Maximum allowed fraction of simple sequence motifs
<code>-dust_max</code> [integer]	5	Maximum length of a simple sequence motif
<code>-dust_maxN</code> [floating point, 0..1]	0.5	Maximum allowed fraction of N nucleotides
<code>-dust_ignore_2bases_bias</code>	<code>off</code>	Ignore bias for 2 bases in sequence reads

3.1.4 Reference sequence download

`unitas` uses publicly available reference sequences for the species in question. Automatic sequence download is supported for all species that are included in the current Ensembl release (744 species or strains, 13-MARCH-2017). For other species you should select a closely related species or use the option `-species x` and provide separate reference sequence file(s) with the option `-refseq [sequence.file]` (see 3.1.1 for correct format). `unitas` connects to the Mainz University server to get the latest list of supported species. If that fails, `unitas` uses an internal list that may vary depending on the `unitas` version you use. Latest download links are also fetched from the Mainz University server or alternatively taken from an internal list. If download of the reference sequence from the respective database fails (in case of server downtime or outdated links [links on the smallRNAGroup server are updated weekly]), `unitas` downloads the required reference sequences directly from the Mainz University server where they are updated weekly. The following online sources are used for annotation (BE FAIR: Please cite the according references in addition to `unitas` when making use of `unitas` sequence annotation):

database name	data	reference
GtRNAdb	genomic tRNA sequences	Chan PP, Lowe TM. GtRNAdb: A database of transfer RNA genes detected in genomic sequence. <i>Nucleic Acids Res.</i> 2009 37 :D93-D97.
SILVA rRNA database	genomic rRNA sequences	Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. <i>Nucleic Acids Res.</i> 41 :D590-D596.
piRNA cluster database	genomic piRNA cluster sequences	Rosenkranz D. piRNA cluster database: a web resource for piRNA producing loci. <i>Nucleic Acids Res.</i> 2016 44 :D223-D230.

Ensembl	ncRNA/cDNA collection (miRNA sequences will be removed from ncRNA data)	Yates A, Akanni W, Amode MR, Barrell D, Billis K et al. Ensembl 2016. <i>Nucleic Acids Res.</i> 2016 44:D710-716.
miRBase	miRNA (mature/precursor) sequences	Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. <i>Nucleic Acids Res.</i> 2014 42:D68-D73.

3.1.5 SeqMap download and compilation

`unitas` uses SeqMap to perform many (not all) probe to reference mapping steps. The SeqMap source code (C++) is downloaded and compiled to an executable file via the g++ compiler which is part of most operating systems. The executable file is named `seqmap.exe` and will appear in the same directory as `unitas`, the source code is stored in the data dump folder (see 3.2.3). If compilation fails, `unitas` will download precompiled SeqMap executable from the Mainz University sever (On linux and Mac you may have to manually change file permissions in order to allow execution of the SeqMap file!). You can also manually download the files from here: <http://www-personal.umich.edu/~jianghui/seqmap/>. Save a local copy of SeqMap to the same directory as `unitas` and rename it into `seqmap.exe` (always use the extension `.exe` since `unitas` will do a system call). Please do not forget to cite the SeqMap paper when making use of `unitas` sequence annotation: Jiang H, Wong WH SeqMap: Mapping Massive Amount of Oligonucleotides to the Genome. 2008 *Bioinformatics* 24:2395-2396.

3.2 The annotation process

`unitas` can process an arbitrary number of sequence files in parallel. By default, multiple input files are processed one after another. For multithreading use the option `-threads [integer]`. For each input dataset, `unitas` creates a results dump folder, which is named according to the current date and input file name (plus consecutive numbering), e.g.: `WIZ_01-01-2017_input.fasta_#1`. Result files and temporary files are stored in this folder. By default, temporary files (SeqMap reports, SeqMap input/output files other reference sequence files) are removed after annotation. To keep temporary files use the option `-keep_temp`.

3.2.1 miRNA annotation

Mature miRNA sequences from miRBase are loaded into memory. Matches to miRNAs from the species in question and matches to miRNAs from different species will be reported separately. In the first step, `unitas` searches for perfect full length matches to mature miRNAs of the species in question (canonical miRNAs). Subsequently, non-matching sequences are mapped to miRNA precursor sequences of the species in question in order to identify non-canonical miRNAs. Still not matching sequences are trimmed at the 3' end to check whether they represent 3' tailed canonical or non-canonical miRNAs. By default, `unitas` allows up to 2 non-template 3' nucleotides. This value can be changed using the option `-tail [integer]`. Then, the same procedure is performed for the remaining sequences and miRNAs from other species. In summary, the order of miRNA annotation is:

1. untailed canonical miRNA from the species in question
2. untailed non-canonical miRNA from the species in question
3. 3' tailed canonical miRNA from the species in question
4. 3' tailed non-canonical miRNA from the species in question
5. untailed canonical miRNA from other species
6. untailed non-canonical miRNA from other species
7. 3' tailed canonical miRNA from other species
8. 3' tailed non-canonical miRNA from other species

Examples:

1. **UUCAAGUAAUCCAGGAUAGGCU**
2. **UUCAAGUAAUCCAGGAUAGGCUUU**
3. **UCAAGUAAUCCAGGAUAGGCU--G**
4. **UUCAAGUAAUCCAGGAUAGGCU--GA**

```

      G      U      C      --G  CA
5' GUG CCUCGU CAAGUAAUC AGGAUAGGCU  UG  G
   ||| ||||| ||||| ||||| ||||| |||  G
3' CGC GGGGCA GUUCAUUGG UCUUAUCCGG  AC  U
      A      C      U      GUA  CC

```

This means that, e.g., when using human input sequences, a sequence is annotated as 3' tailed canonical human miRNA instead of untailed canonical mouse miRNA, even though both alternatives are possible.

unitas outputs tab-delimited miRNA annotation tables (e.g. `unitas.miR-table_Human.txt` and `unitas.miR-table_non-Human.txt`) that contains the following information: miRNA name (e.g. miR-376b-3p), miRNA sequence, total number of reads, number of reads without tailing (+N0-3'), number of reads with one non-template 3' nucleotide (+N1-3') [... number of additional columns depends on the number of allowed non-template 3' nucleotides], number of A-tailed reads, number of T-tailed reads, number of G-tailed reads, number of C-tailed reads [... number of additional columns depends on the number of allowed non-template 3' nucleotides, e.g. AA-tailed reads etc.].

In addition, unitas outputs files (e.g. `unitas.miR-modifications_Human.txt` and `unitas.miR-modifications_non-Human.txt`) that summarize information on non-template 3' ends in the following simple format (tail, read counts):

```
3'-tailings (non-template nucleotides)
A      33596
AA     3143
AAA    855
AAC    114
AAG    347
AAT    547
[...]
```

This file also summarizes internal modifications that were identified during miRNA annotation in the following format:

```
Internal modifications (?->[ATGCN] = mapped sequence exceeds precursor sequence)
G->T   53898
C->T   13998
T->G   13514
T->C   10506
A->G   8916
T->A   6270
[...]
```

It further gives you information on the positions where modifications occur like this:

```
Positions where internal modifications occur
1      2481
2      1638
3      3030
4      1589
5      1706
6      3580
[...]
```

Finally, it provides a table that combines modifications and positions in the following style:

```
Modifications per position
      G->T   C->T   T->G   T->C   A->G   T->A   A->C   G->A   C->A   G->C
1     425   108   251   403   103   247   153   110   184   258
2     332   42    51   109   156   40    360   193   43    202
3     1047  100   69    36    90    11    175   771   167   499
4     177   263   85    39    104   17    153   60    495   67
5     607   116   216   194   62    74    70    124   84    98
6     398   211   1351  414   218   101   97    139   133   72
[...]
```

Relevant command line options:

option	default	explanation
--------	---------	-------------

<code>-skip_miR</code>	off	Skip miRNA annotation
<code>-tail [integer]</code>	2	Maximum number of non-template 3' nucleotides
<code>-intmod [integer]</code>	1	Maximum number of internal modifications

3.2.2 ncRNA/mRNA fragments

Input sequences that cannot be annotated as miRNA sequence are mapped to ncRNA/cDNA sequences in sense orientation to identify fragments of larger ncRNA classes or fragments of mRNA. By default the number of allowed mismatches is 1 and the number of allowed insertions/deletions is 0. You can change the default value with the option `-mismatch [integer]` and `-insdel [integer]`, respectively. The maximum possible number of allowed mismatches is 5, the maximum possible number of allowed insertions/deletions is 3. Note that increasing the number of allowed mismatches and insertions/deletions will significantly increase memory usage and computation time. When allowing mismatches, only the best alignments in terms of mismatch counts will be considered for sequence annotation. Read counts of sequences that match different reference sequences (with equal alignment quality) are apportioned accordingly. As a result, the values for read counts per class may be decimal numbers rather than integers.

Mapping to cDNA is performed by default, but can be skipped using the option `-skip_cDNA`. Considering tRNA-derived sequences, `unitas` further classifies sequences into 5' tRFs, 5' tR-halves, 3' tRFs, 3' CCA-tRFs, 3' tR-halves, tRF1, tRNA-leader and misc-tRFs based on alignment position and sequence length (read counts for genomic and mitochondrial tRNA-derived sequences are listed separately). In addition, `unitas` creates a table (`unitas.tRF-table.txt`) that lists which fragments are processed from which tRNAs.

`unitas` outputs FASTA files for each class of annotated sequences (miRNAs, tRNA, rRNA etc.). Note that one sequence can occur in more than one output file if the sequence maps to different reference sequences. Sequences in the FASTA output files are sorted according to read counts in descending order, e.g.:

```
>5044
TCCCTGGTGGTCTAGTGGTTAGGATTCGGCGCT
>781
GCCCGGATAGCTCAGTCGGTAGAGCATCAGAC
>774
CGCGGGAGACCGGGGTTTCGATTCCCCGACGGG
>647
GCGCCGCTGGTGTAGTGGTATCATGCAAG
[...]
```

For sequences that map to genomic piRNA clusters, the FASTA headers contain additional information that refers to the genomic coordinates of the piRNA producing locus, e.g.:

```
>46:Chr6 4863485-4881004
TAGGCGAATCTAGGTTATTTCAACGATGCA
```

The FASTA files are stored in the results dump folder (see 3.2) and are named e.g. `unitas.miR.Human.fas` when using human input sequences. Each FASTA file is accompanied by a file that contains information on length distribution and positional nucleotide composition for the sequence reads assigned to this class, e.g.:

```
LENGTH DISTRIBUTION
length reads
18      271
19      263
20      173
21      902
22     1675
23      729
24     311
[...]
```

```
POSITIONAL NUCLEOTIDE COMPOSITION
pos.   A       C       G       T
1     1963   2406   11144   7520
2      472   16896   4775    890
3     2277   13295   5735   1726
```

```

4      1613   12796   3578   5046
5      667    4171    7075   11120
6      2124   1965    17564  1380

```

[...]

Finally, the file `unitas.sncRNA-annotation.txt` provides a table that summarizes sequence annotation (normalized reads per class) and lists the number of normalized sequence reads per each individual reference sequence, e.g.:

```

low_complexity          1244
miRNA                   80727
  miRNA:homo_sapiens   76338
  miRNA:other          4389
rRNA                   100575.918129203
  genomic_rRNA         100367.418129203
  Mt_rRNA              208.5
tRNA                   22867.0863247866
  genomic_tRNA         22846.0863247866
    5'tR-halves        12672
    5'tRFs             4800.75299145299
    3'tR-halves         64
    3'tRFs             23.7142857142857
    3'CCA-tRFs         67
    tRF-1              21
    tRNA-leader        4
    misc-tRFs         5218.61904761903
  Mt_tRNA              21
    5'tR-halves        1
    5'tRFs             4
    3'tR-halves        1
    3'tRFs             2
    3'CCA-tRFs         0
    tRF-1              0
    tRNA-leader        0
    misc-tRFs         13
pseudogene             2.18972986748216
ribozyme               8
antisense              73.9023104237751
snRNA                  849.6666666666667
lincRNA                1202.5266023332
misc_RNA               807.009174311925
protein_coding         4682.66446764933
snRNA                  4349.3477084787
scaRNA                 45
no annotation          37381
  mapped to piRNA producing loci 1265

```

```

TRANSCRIPT_CLASS      TRANSCRIPT_NAME      NORMALIZED_COUNT
3prime_overlapping_ncRNA  CTD-2196E14.9      3.57337335772603
3prime_overlapping_ncRNA  RP11-473I1.9       2.0014450867052
3prime_overlapping_ncRNA  RP11-373L24.1     1.30392156862745
3prime_overlapping_ncRNA  RP11-134G8.10     0.333333333333333
3prime_overlapping_ncRNA  CTD-2651B20.1     0.0206557649767618
3prime_overlapping_ncRNA  LINC00846          0.00144508670520231
IG_V_pseudogene          IGHVII-28-2        0.0606060606060606
IG_V_pseudogene          IGHVII-30-43       0.0417693316175626
IG_V_pseudogene          IGHVII-20-3        0.0113937832958541
IG_V_pseudogene          CH17-60O17.11     0.00377588629090349
IG_V_pseudogene          IGHVII-44-2D       0.00377588629090349
Mt_rRNA                  MT-RNR1            7953.666666666667
Mt_rRNA                  MT-RNR2            3363.69603174603
Mt_tRNA                  MT-TP              8870.833333333333

```

[...]

The list that gives you the normalized read count per transcript is sorted alphabetically according to transcript classes (capital letters first). The order of transcripts within one class depends on the number of normalized read count.

Relevant command line options:

option	default	explanation
<code>-mismatch [integer]</code>	1	Maximum number of mismatch to reference
<code>-insdel [integer]</code>	0	Maximum number of insertions/deletions
<code>-riborase</code>	off	Map sequences to a complete collection of rRNA sequences from NCBI nucleotide database. This is recommended to identify rRNA degradation products in species with low quality ncRNA annotation.

3.2.3 Searching piRNA candidates

Piwi interacting RNAs (piRNAs) typically pass the described pipeline without producing a match to any known ncRNA or cDNA. Thus, non-annotated sequences (saved in the output file `unitas.no-annotation.fas`) are subsequently mapped to known piRNA producing loci of the species in question (if available). This step is performed by default, but can be skipped using the option `-skip_piR`. Searching for piRNA candidates is particularly insightful when handling data derived from germ line tissues, or tissues that were shown to contain noteworthy amounts of piRNA-like molecules (e.g. brain or epididymis). In this case, we further recommend using sequences without annotation (saved in the output file `unitas.no-annotation.fas`) for subsequent piRNA cluster prediction with proTRAC.

When using map files as input you can also search for ping-pong signatures within your sequence data. When using the option `-pp`, `unitas` will check and report the 5' overlaps of mapped sequence reads and calculate a Z-score for the enrichment of 10 bp overlaps according to Zhang et al. 2011, *Mol Cell* 44:572-584.

3.2.4 Searching phased small RNAs (phasiRNAs)

When using map files as input you can search for phased small interfering RNAs (phasiRNAs). Use the option `-phasi [n]` to search for phasiRNAs with length *n*. You can use the option multiple times in one command, like this:

```
-phasi 21 -phasi 24
```

`unitas` will scan the map file with a sliding window (default size = 1 kb) and save coordinates and read counts of putative phasiRNAs and all other mapped sequence reads. The coordinates of putative phasiRNAs (that is all reads with correct length) from plus and minus strand are unified to check for phasing on both strands. Each sliding window is subsequently analyzed for the presence of phasiRNAs with significant read count over the background. Significance is calculated based on the following formula:

$$p = 1 - \left(\sum_{k=0}^j \binom{n}{k} q^k (1-q)^{n-k} \right)$$

in which *j* refers to the observed number of reads with length *i* in a specified phase, *n* refers to the total number of reads with length *i* and *q* is given by $1/i$ and refers to the probability of a read to be located in a given phase, assuming that a sequence read can map to any position within the sliding window with equal probability. `unitas` employs strict Bonferroni correction for multiple testing by means of the number of analyzed sliding windows, that is the number of sliding windows with at least one mapped sequence. Adjacent sliding windows (default < 1 kb) with significant enrichment for phasiRNAs are merged to phasiRNA clusters. Besides significant enrichment of phasiRNAs there is a series of further thresholds (see below for command line options) that aim to reduce the false positive prediction rate. `unitas` produces two output files in addition to the obligatory FASTA and .info file. A file named `unitas.phasiRNA.sorted.fas` comprises the same sequences that are stored in the FASTA file, but sorted according to the identified phasiRNA clusters. Note, that in this file sequences may occur multiple times when they can be assigned to more than one phasiRNA cluster. A file named `unitas.phasiRNA.align` comprises information on predicted phasiRNA clusters with aligned phasiRNAs. The alignment file will look like this:

```
cluster_ID:1 location:chr01 start:2301 end:2386 size:86 bp non-identical_sequences:5 reads:61 reads_normalized:50
2/2 ATCGATAACTTGGTTCAGCAA
```


of these columns depends on the number of input samples (`unitas` output folders). The next two columns refer to the average rpm values for samples in group 1 and samples in group 2. In this case the values are identical because there is only one sample per group (note that you should really use more than one replicate per group for this kind of analysis). The next columns refer to the raw p-value and the alpha-corrected p-value. The final column gives you information to which ncRNA class this small RNA belongs (can be more than one class).

The statistical test performed for detection of differentially expressed small RNAs in principal bases on the assumption that the number of sequence reads observed in biological replicates for one condition (variance across samples, individuals, RNA extraction- or library preparation procedures) is roughly normally distributed. To calculate the variance for one condition you should not use less than two replicates for one condition (one group). Otherwise, there is no possibility to get a reasonable estimation of the biological variance occurring between samples of the same condition. However, `unitas` will also perform tests with only one replicate, in this case using predefined minimum values of variance (standard deviation) depending on dataset size and the observed rpm value for each sequence. The predefined minimum values correspond to the standard deviation one would expect to observe when the identical dataset would be resampled randomly (e.g. doing resequencing of the same library). They are calculated according to the formula:

$$\sqrt{\frac{r}{10^6}}^{-1} \cdot a^{-1}$$

where r represents the total number of reads for a replicate and a represents the rpm value for the sequence in question. These minimal values for standard deviation are also employed when providing more than one replicate per group, which is relevant for those cases where the rpm values for one sequence in different replicates are by hazard (almost) identical. When you have no possibility to use more than one replicate for a condition (group), we recommend that you give an estimation on the deviation that you *would* expect across replicates (rather than relying on the minimum standard deviation values). You can do this using the option `-diffexpr_estdev [0..1]`. As an example, you may estimate that when you had two additional replicates, the rpm value for a given small RNA sequence could be 20% higher or 20% lower compared to the replicate that you have. In this case, you would have to use the option like this:

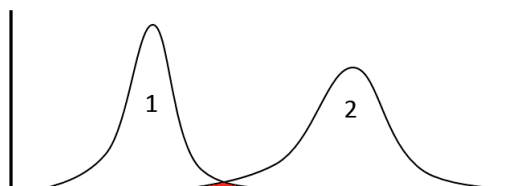
```
-diffexpr_estdev 0.2
```

Doing so, `unitas` will assume that rpm values can be 20% higher or lower. You can think of it as adding two virtual replicates to the group that has only one replicate, one comprising read counts 20% higher, one comprising read counts 20% lower).

Having the expectation value (average read counts across replicates) and standard deviation of rpm values across replicates, the probability density function of rpm values can be displayed according to the formula:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the expectation (or mean) value and σ is the standard deviation. When we now compare the probability density function for the same sequence across two different conditions (1 and 2) we should observe two graphs that look like this:



In theory, these graphs represent the distribution of rpm values that one would expect when we have an infinite number of replicates for each condition (x-axis: reads or rpm, y-axis: number of replicates). The shared area under

both graphs (red), in relation to the remaining integrals of both graphs, represents the probability to observe the same rpm values in both conditions. Accordingly, we can assume that read counts for one sequence are significantly different across two conditions when the shared area is below a critical value (e.g. 0.05, with an overall integral of 1).

Relevant command line options:

option	default	explanation
<code>-diffexpr</code>	<code>off</code>	Search for differentially expressed small RNAs
<code>-diffexpr_p [floating point, 0..1]</code>	0.05	Critical p-value for differential expression
<code>-diffexpr_acorrection [0;1;2]</code>	1	Chose how to correct alpha error. 0: no correction. 1: Bonferroni-Holm procedure. 2: Strict Bonferroni correction.
<code>-diffexpr_estdev [floating point, 0..1]</code>	0	Estimate standard deviation when you have only one replicate for a condition.
<code>-g1 [folder name]</code>	n.a.	Specify <code>unitas</code> output folder for group 1
<code>-g2 [folder name]</code>	n.a.	Specify <code>unitas</code> output folder for group 2

4. Command line options

Allowed values for options: [s]=string, [i]=integer, [f]=floating point number

<code>-i OR -input [s]</code>	Name of the input file. You can use multiple input files at once, e.g.: <code>-i file1.fas -i file2.fq -i file3.sam</code> Alternatively you can provide the name of a directory that contains all of your input files, e.g.: <code>-i path/to/your/input/files</code>
<code>-s OR -species [s]</code>	Species name. Use the binominal nomenclature with genus and species separated by an underline, e.g.: <code>-s mus_musculus</code>
<code>-latest_ref</code>	Force <code>unitas</code> to download the latest database sequences. Otherwise <code>unitas</code> will use previously downloaded sequences (if available) from the <code>UNITAS_refdump</code> folder.
<code>-refdump</code>	Create a reference sequence dump folder, download species-specific reference sequences and QUIT. This folder and sequences herein can be used for later offline <code>unitas</code> runs.
<code>-refseq [s]</code>	Use an arbitrary number of additional files that contain reference sequences in FASTA format. FASTA headers should have the following format: <code>>ncRNA_type ncRNA_name</code>
<code>-riborase</code>	Find all rRNA-like sequences in your input file by using a complete NCBI rRNA nucleotide collection in addition to the available ncrRNA data for your species.
<code>-trim</code>	Use this option to search for and remove any 3' adapter sequence in input files. There is no need to provide a specific sequence.
<code>-trim_minlength [i]</code>	Minimum allowed read length [nt] after removal of adapter sequence. Default=10.
<code>-trim_maxlength [i]</code>	Maximum allowed read length [nt] after removal of adapter sequence. Default=50.
<code>-trim_ignore_5p [i]</code>	Ignore first n basepairs for 3' adapter prediction. Default=22.
<code>-trim_skip_reads [i]</code>	Skip the first n reads when searching for frequent motifs. Default=10000
<code>-trim_check_reads [i]</code>	Check n reads for 3' adapter prediction. Default=500000
<code>-trim_maxfinalbias [f]</code>	Maximum allowed 3' nucleotide bias after trimming. If this value is exceeded, <code>unitas</code> will trim further 3' positions. Default=0.8.
<code>-quick</code>	Allow less mismatch. Equivalent to: <code>-tail 1 -intmod 0 -mismatch 0 -insdel 0</code> Faster, less memory usage, less sensitive.

-slow Allow more mismatch. Equivalent to:
-tail 3 -intmod 2 -mismatch 2 -insdel 1
Slower, more memory usage, more sensitive.

-tail [i] Maximum number of allowed non-template 3' nucleotides when mapping user sequences to reference miRNA (miRNA-precursor) and piRNA cluster sequences. Default=2. Summed values for -tail and -intmod must not exceed 5.

-intmod [i] Maximum number of allowed internal mismatches when mapping user sequences to reference miRNA (miRNA-precursor). Default=1. Summed values for -tail and -intmod must not exceed 5.

-mismatch [i] Maximum number of allowed mismatches when mapping user sequences to reference ncRNA sequences (does not apply to miRNAs and putative piRNAs). Default=1.

-insdel [i] Maximum number of insertions/deletions when mapping user sequences to reference ncRNA sequences (does not apply to miRNAs and putative piRNAs). Default=0.

-threads [i] Maximum number of parallel threads. Applies only if using more than one input file. Default=1.

-memory [i] Limit memory usage [GB] for SeqMap mapping. Default is no limit. When using a limit, unitas will quit before the limit will be exceeded.

-less_memory Will use modified parameters for SeqMap mapping to reduce memory usage. Will increase running time.

-even_less_memory Will use modified parameters for SeqMap mapping to reduce memory usage even more. Will increase running time.

-phasi [i] Search for phased siRNAs (phasiRNAs or tasiRNAs) with specified length = [integer]. When using this option, the input file must be a map file in either SAM or ELAND3 format. You can use this option multiple times, e.g: -phasi 21 -phasi 24.

-na_for_phasi Only sequences without annotation will be considered for phasiRNA prediction.

-pp Analyze 5' overlaps of mapped reads to identify so-called ping-pong signatures (preference for 10 bp overlaps) which is a footprint of secondary piRNA biogenesis. When using this option, the input file must be a map file in SAM or ELAND3 format.

-keep_temp Do not remove temporary files from data dump folder.

-skip_mapping Skip mapping to reference sequences. Equivalent to:
-skip_miR -skip_cDNA -skip_ncRNA -skip_piR
Use this when you want to trim 3' adapters and/or want to search for phasiRNAs only.

-skip_miR Skip search for miRNA reads, search for other ncRNAs only.

-skip_cDNA Skip search for reads that map to Ensembl cDNA sequences.

-skip_ncRNA Skip search for reads that map to ncRNA collection.

-skip_piR Skip search for reads that map to known piRNA producing loci.

-skip_dust Skip filtering low complexity sequences.

-dust_limit [f] Threshold for fraction of a sequence that consists of simple sequence repeats When filtering low complexity reads from input files. Default=0.75.

-dust_max [i] Maximum motif size for simple sequence repeats. Default=5.

-dust_maxN [f] Maximum fraction of Ns in a read. Default=0.5.

-no_html Do not output html files with visualization of sncRNA annotation and analysis.

-diffexpr Search for differentially expressed sRNAs across different probes. Define 2 groups with a recommended minimum of 2 probes per group using the options -g1 and -g2 as described below. When using this option unitas will do the comparison and quit. Change the name of the output file using the option -diffexpr_out.

-diffexpr_p [f] Critical p value for differential expression analysis. Default=0.05.

-diffexpr_acorrection [i] Correction of alpha error (multiple testing). Allowed values are 0 (=no correction), 1 (=Bonferroni-Holm, default) and 2 (=Bonferroni).

```

-diffexpr_estdev [f]    If one group comprises only one replicate, the
                        variance can be estimated by means of this factor
                        which is 0.15 by default. That means if sequence A
                        has 100 reads in the given replicate, unitas
                        virtually adds two replicates with read count values
                        115 and 85 (+- 15%).
-g1 [s]                unitas output folder for probe assigned to group 1.
                        Use this option multiple times for every probe that
                        you want to assign to group 1. E.g:
                        -g1 UNITAS_04-01-2017_SRR123456.fasta_#1
-g2 [s]                Probes assigned to group 2 (see above).
-diffexpr_out [s]     Specifies file names where the results of the
                        differentially expressed RNAs analysis will be
                        written to. Extensions .up1 and .up2 will be added
                        to the name. Default file names are:
                        'unitas.diff-expr.up1' and
                        'unitas.diff-expr.up2'.
                        (~.up1->higher expression in group 1, ~.up2->higher
                        expression in group 2).
-silent                Less console (STDOUT) output during annotation.
-supp_spec             Prints the complete list of supported species and
                        quits.
-h OR -help           Shows usage information and quits.
-show_options         Shows available options and quits.

```

5. Troubleshooting

We really tried hard to design `unitas` in a way that ensures compatibility with most platforms. Although we have tested `unitas` thoroughly on several different systems, we cannot guarantee that it will run on every machine on this planet. Here are some solutions for problems that may occur:

Command perl is not available.

You try to run the `unitas` Perl script but Perl is not installed on your computer (most likely you are on a Windows machine). You should try one of the following:

1. Use the precompiled executable file for your system instead of running the Perl script.
2. Download and install either Strawberry Perl or ActivePerl, both are freely available.

```

Can't locate LWP/Simple.pm in @INC [...]
Can't locate Archive/Extract.pm in @INC [...]
Can't locate File/Copy.pm in @INC [...]
Can't locate File/Path.pm in @INC [...]
Can't locate Getopt/Long.pm in @INC [...]

```

You try to run the `unitas` Perl script but the module `LWP::Simple` (or one of the other modules) is missing on your computer for some reasons. You should try one of the following:

1. Use the precompiled executable file for your system instead of running the Perl script.
2. Upgrade your Perl distribution with the following two commands:

```

cpan
upgrade

```
3. Install the `LWP::Simple` module (or any other missing module) with the following two commands:

```

cpan
install LWP::Simple

```

Failed to load current species list from smallRNAgroup-SERVER. Use internal list.

This message will occur if you are not connected to the internet or the `smallRNAgroup-SERVER` is down. In the latter case, `unitas` will use an internal list of supported species which is maybe not up to date. If you are not connected to the internet `unitas` can only use previously downloaded sequence data and will quit in case that there is no reference dump folder for the species in question in the current working directory.

```

Unable to create temporary file with list of data dump folders.
Unable to create dump folder for reference sequences.

```

This is most likely because you are on a Mac or Linux machine and do not have permission to write into the current working directory. You can change permissions with the following command (may require password):

```

chmod -R 755 /path/to/folder

```

```
No success when building SeqMap from source. Will download pre-compiled SeqMap version from smallRNAgroup-SERVER.
```

`unitas` tried to compile the downloaded SeqMap source code with `g++` but failed. Maybe the `g++` compiler is missing on your computer. `unitas` will download a precompiled SeqMap executable that should run on your system. However, it is likely that you have to change file permissions for that executable file manually (allow execution). Check subsequent `unitas` messages.

6. Citation policy

Sequence annotation with `unitas` is not possible without having reference sequences. `unitas` uses reference sequences from different public online databases. When you use `unitas` for sequence annotation please cite the papers listed in 3.1.3. as well as the SeqMap paper from Jiang and Wong (see 3.1.4) in addition to `unitas`.

7. Contact

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

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