A manual for the use of miRVaS

Authors: Sophia Cammaerts, Mojca Strazisar, Jenne Dierckx, Jurgen Del-Favero, Peter De Rijk

Version: 1.0.0
Date: March 11, 2015

Contact: peter.derijk@gmail.com, sophia.cammaerts@gmail.com

Table of contents

1. Introduction

miRVaS is a tool that predicts the impact of genetic variants on nearby miRNAs, by annotating the location of the variant relative to functional regions within the miRNA and by predicting the structural impact of the variant on the miRNA precursor. To assess the structural impact, different secondary structure representations are predicted by implementing RNAfold (1) v2.1.5: the centroid (CEN), maximal expected accuracy (MEA) and minimal free energy (MFE) structures. miRVaS generates tabular output, for fast screening, html output, for efficient visual comparisons, and separate images files. Images of predictions are generated using VARNA (2) v3.9. For background information, see (Cammaerts et al., submitted).

For referencing the use of the tool, you should cite: (Cammaerts et al., submitted). Since with miRVaS you are also using the underlying tools RNAfold (1) and VARNA (2), these should be cited as well.

2. Input files

miRVaS requires three types of input files: a variant file, a genome file and a miRNA database file. All input files have to be sorted using a natural sort on the genomic coordinates (chromosome, begin, end). An example of a natural sort on chromosomes is:
“chr1, chr2, chr10, chr11, chrX”,
while
“chr1, chr10, chr11, chr2, chrX”
is not.
Genomic coordinates in input and output files are in zero based, half open format. For an explanation of zero based coordinates, refer to https://www.biostars.org/p/84686/.

2.1. Variant file

The variant file is a tab-delimited file with header containing the genomic coordinates, type of variant (snp, ins or del), reference and alternative allele for the tested variants. An example of such a file can be found in Figure 1 and in the file supplemented with the software in the examples folder (file: “testset_input.txt”). Note that insertions and deletions that contain repeats of multiple bases have to be specified in full. Genomic coordinates in tab-delimited files need to be in zero based half open format. Alternatively, a VCF can also be used as input; miRVaS then converts the coordinates to the required format.

chromosome	begin	end	type	ref	alt
chr1	98511730	98511731	snp	C	T
chr1	98511733	98511733	ins		CCGCTGCCGCTGCTA
chr1	98511750	98511750	ins		TATATATATA
chr22	46509539	46509540	del	C	

Figure 1: Example of a tab-delimited variant file. Type can be snp (single nucleotide polymorphism), ins (insertion) or del (deletion). In case of insertions or deletions, the reference or alternative allele field needs to be left empty. Ins: empty reference allele field; del: empty alternative allele field.

2.2. Genome file

The genome file is a FASTA file containing the sequence of each chromosome on a single line. The hg19 and hg38 genome files in the correct format can be downloaded from the website (genome_hg19.ifas, genome_hg38.ifas). If the genome file given is a plain FASTA file (with possibly multiple lines per sequence, not correctly sorted), it will be converted. The genome file can also contain a minigenome, if the coordinates of the variant file and miRNA database file correspond to the positions in the genome file. An example of part of such a file is shown in Figure 2.

```
>chr1
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
>chr2
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Figure 2: Example of the format to be used for the genome file.

2.3. miRNA database file

The miRNA database file contains information on annotated miRNA hairpins, the mature sequences and the loop regions. Two miRNA database files are supplemented with the software based on miRBase (3) v20 and v21 (mirna_hg19_mirbase20.tsv, mirna_hg38_mirbase21.tsv). In these files, the hairpin and mature miRNA sequences are extracted from the miRBase database file in gff3 format. The loop region is defined by structure prediction of the hairpin and is derived from the miRBase “miRNA.str” file.

The miRNA database file has to be in tab-delimited format (with a header), with the following fields:
chromosome, begin, end, strand, name, mature1start, mature1end, loopstart, loopend, mature2start, mature2end
where (chromosome, begin, end) refers to the genomic position of the hairpin in zero based half-open format, and (mature1start, mature1end), (loopstart, loopend), (mature2start, mature2end) indicate the genomic positions of the functional elements of the miRNA precursor. mature1 is the first mature miRNA sequence in the genomic reference sequence (the 5p mature in a positive strand miRNA gene, 3p in a negative strand one).

3. Database conversion tool “miRVaS convert”

Next to using the available miRNA database files and genome files presented on the website, it is also possible to create

- miRNA database files from different miRBase gff3 files (e.g. other species, newer versions)
- miRNA database files from files with gene coordinates of novel miRNA genes that are not present in miRBase

To create such files, a database conversion tool called “miRVaS convert” is present within the miRVaS graphical user interface (GUI) (Figure 3).

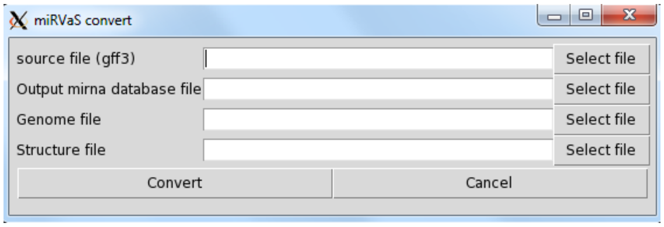


Figure 3: miRVaS convert GUI.

To use this conversion tool, you need to supply three files:

- the source file: the miRBase database file in gff3 format for the relevant species (download at: ftp://mirbase.org/pub/mirbase/); or the novel miRNA gene database file, in the same format;
- the genome file of the relevant species, with sequence per chromosome on one or on multiple lines
- the structure file: the miRBase miRNA.str file (download at: ftp://mirbase.org/pub/mirbase/); or a similar file containing hairpin structure predictions for novel miRNA genes;

For the correctness of the output all the input files need to be based on the same genome build and the same miRBase version (if used).

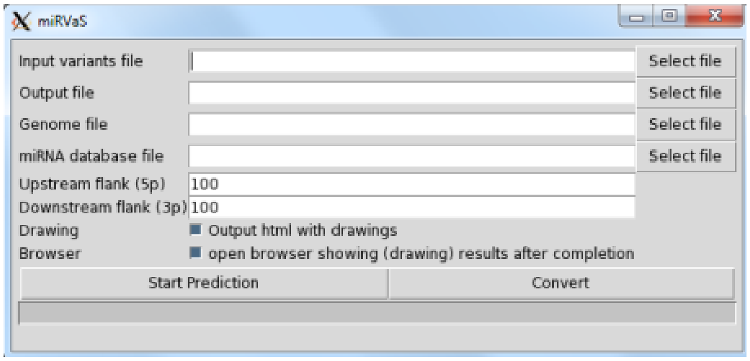
If the genome file is not in the single line format, the conversion tool will format the file to single line format. The tool will make a backup copy of the original genome file named filename.old.

The miRVaS convert tool uses the three input files to make a new miRNA database file: hairpin and mature sequences are extracted from the source file and loop regions are determined from the structure file. In case the structure file is not available, the database conversion tool will predict MFE structures using RNAfold based on the source file and determine loop regions from the predictions.

4. Running miRVaS

4.1. Via GUI

When using the GUI of miRVaS (Figure 4), specify the input files and the name of the output file in the designated fields. Make sure the directory where the output file will be made does not contain files or directories that conflict with the intended output (outputfile.txt, outputfile.html, outputfile.struct). If the genome FASTA file is in multiple line format, miRVaS will automatically format the file to a single line format (similar to miRVaS convert tool).



Input variants file

Output file

Genome file

miRNA database file

Upstream flank (5p) 100

Downstream flank (3p) 100

Drawing ■ Output html with drawings

Browser ■ open browser showing (drawing) results after completion

Figure 4: miRVaS GUI.

Four parameters can be specified:
Upstream flank (5p): Number of bases upstream of the hairpin that need to be included in the prediction.
Downstream flank (3p): Number of bases downstream of the hairpin that need to be included in the prediction.
Drawing: Output html with drawings: Uncheck box if no html output is required. In that case, only the output table is generated.
Browser: Check box if html output needs to be opened automatically when the run is finished. This option only applies when the Drawing option is enabled.

Default GUI parameters are: up- and downstream flanks of 100 nt, generating visual output and opening html automatically when the run is finished.

When all files and parameters are specified, press the “Start Prediction” button. The blue bar will indicate the progress.

4.2. Command line use

To use miRVaS via command line, use:

```
mirvas ?options? variantfile outputfile.txt genomefile.ifas miRNAdatabasefile.tsv
```

Make sure the directory where the output file will be made does not contain files or directories that conflict with the intended output (outputfile.txt, outputfile.html, outputfile.struct).
Parameters between “?” are optional. The different options are:
-flank5p num: length of upstream flank
-flank3p num: length of downstream flank
-drawings num: to disable, use “0”; to enable, use “1”
If no parameters are specified, miRVaS will run with flanks of 100 nt and generate visual output.

E.g:

```
mirvas -flank5p 80 -flank3p 40 -drawings 0 var.txt var_fl100.txt genome_hg19.ifas
mirna_hg19_mirbase20.tsv
```

With this command, miRVaS will run predictions for variants from the “var.txt” input file with upstream flanks of 80 nt and downstream flanks of 40 nt. Only the tabular output will be generated and saved in a file named “var_fl100.txt”.

5. Output files

miRVaS generates three types of output: a tab-delimited output file, an html file and a directory containing all separate html files and image files.

5.1. Tab-delimited output

The output file (name given as parameter, e.g. outputfile.txt) will contain the results in tab-delimited format. It contains:

- genomic coordinates of the variant;
- location of the variant relative to the functional regions within miRNA gene tested (seed, mature, hairpin arm, loop, flanks, downstream or upstream);
- the miRNA gene that is located nearby and for which the impact is tested. When a variant is located near several miRNAs (e.g. a variant located between two clustered miRNA genes), location of the variant for the different genes is separated in different rows
- highest structural impact per structure representation (CEN/MEA/MFE) for the tested miRNA;
- impact per structure representation (CEN/MEA/MFE): all structural changes are annotated;
- conservation (see below);
- delta free energies of reference and variant predictions

5.2. Overview html file

An html file containing links to all structure predictions for all variants is generated (when visual output is enabled) using the same base filename as the tab-delimited file, but with the extension .html (e.g. outputfile.html)

5.3. Structures directory

A directory with the same base filename but with the extension .struct is generated containing all separate html files and svg and png images per variant for each wild-type and variant prediction, and dot plots. The overview html file links to individual files in this directory. When moving the html file, the directory should be moved with it.

The filename of the separate image files and html files in this directory is composed of: the specific miRNA gene, the variant position, reference and alternative alleles, the type of structure prediction and a short text indicating a wild-type or variant prediction (e.g. hsa-mir-146a_chr5_159912417-159912418_snp_C_G_MEA_Ref.png). When variants contain long stretches of inserted or deleted bases, the name of these files would become too long, so then the reference or alternative allele is represented by the number of bases in it, instead of by the exact sequence. If a file already exists (e.g. an insertion of the same size, but different sequence) generating the same filename is avoided by adding a number. For this reason it is also best to clear the output directory (esp. outputfile.struct).

6. Interpretation of output

6.1. Location annotation (“mir_location” column)

The location of the variant relative to the miRNA gene tested is annotated as follows:

region(reference +/- start:stop). Location annotations and examples are illustrated in Figure 5 (section 6.2).

The first part of the annotation, the region, refers to the functional region in which the variant is located:

- mature
  - mature5p: variant located in miR-5p sequence
  - mature5p()seed: the variant is located in the 5p mature miRNA, within nt 2-7, which is defined as the seed sequence
  - mature3p: variant located in miR-3p sequence
  - mature3p()seed: the variant is located in the 3p mature miRNA, within nt 2-7, which is defined as the seed sequence
- loop
- arm: hairpin region outside loop and mature regions
  - arm5p: 5p portion of the hairpin arm,
  - arm3p: 3p portion of the hairpin arm,
- flank: variant located outside hairpin, but within the flank size specified by the user
  - flank5p: upstream of the 5p portion of the hairpin
  - flank3p: downstream of the 3p portion of the hairpin
- downstream/upstream: variant located outside specified flank size

These regions are determined by the miRNA database files that are given as input and based on the flank sizes that the user specifies. It should be noted that when a given mature miRNA region overlaps with the loop region in the specified miRNA database file, the tool will annotate the variant with “mature” and not with “loop”.

The second part of the annotation, the part between parentheses, indicates the position of the variant in the given region relative to the flanking functional segment (called the reference region).
Different reference regions are:

- a: arm
- l: loop
- m: mature

The start and stop part states how far the variant is located from this reference region. For a SNP or a one-base deletion, this would be just one position (e.g. mature3p(a+9)), while a start and stop position is given for large deletions and insertions. In case of an insertion, the positions of the bases between which the insertion takes place are indicated (e.g. mature3p(a+10:11)).
When a variant is located at the end or start of a region, an “e” is appended to the base position (e.g. mature3p(a+22e)).
When a variant is located more than 2000 nt outside any miRNA hairpin, the location annotation is “no miRNA within 2000 nt” in the html output. In the txt output, the mir_location and mir_name fields are left empty in this case.

In the generated images, the variant is marked in red. For insertions and deletions, the two bases surrounding the inserted sequence or the deleted sequence are also highlighted, to clearly indicate in both the wild-type and the variant miRNA where the sequence is inserted or deleted.

6.2. Structural change annotation (“impact” columns)

Structural changes are annotated in the same way as location annotations. In case of changes in multiple regions, annotations are done for each region and then concatenated, the order is relative to the orientation of the gene. Structural changes are highlighted in the images by black bases.

When the tested variant lies outside the specified flank size of any miRNA hairpin, the structural change annotation in the html file is “variant not in analysed region”. In the txt output, the structural fields are left empty in this case. When the variant is located within the specified flanks of a miRNA, but the variant does not result in any predicted structural change, the structural change annotation is “nochange” (in html and txt output).

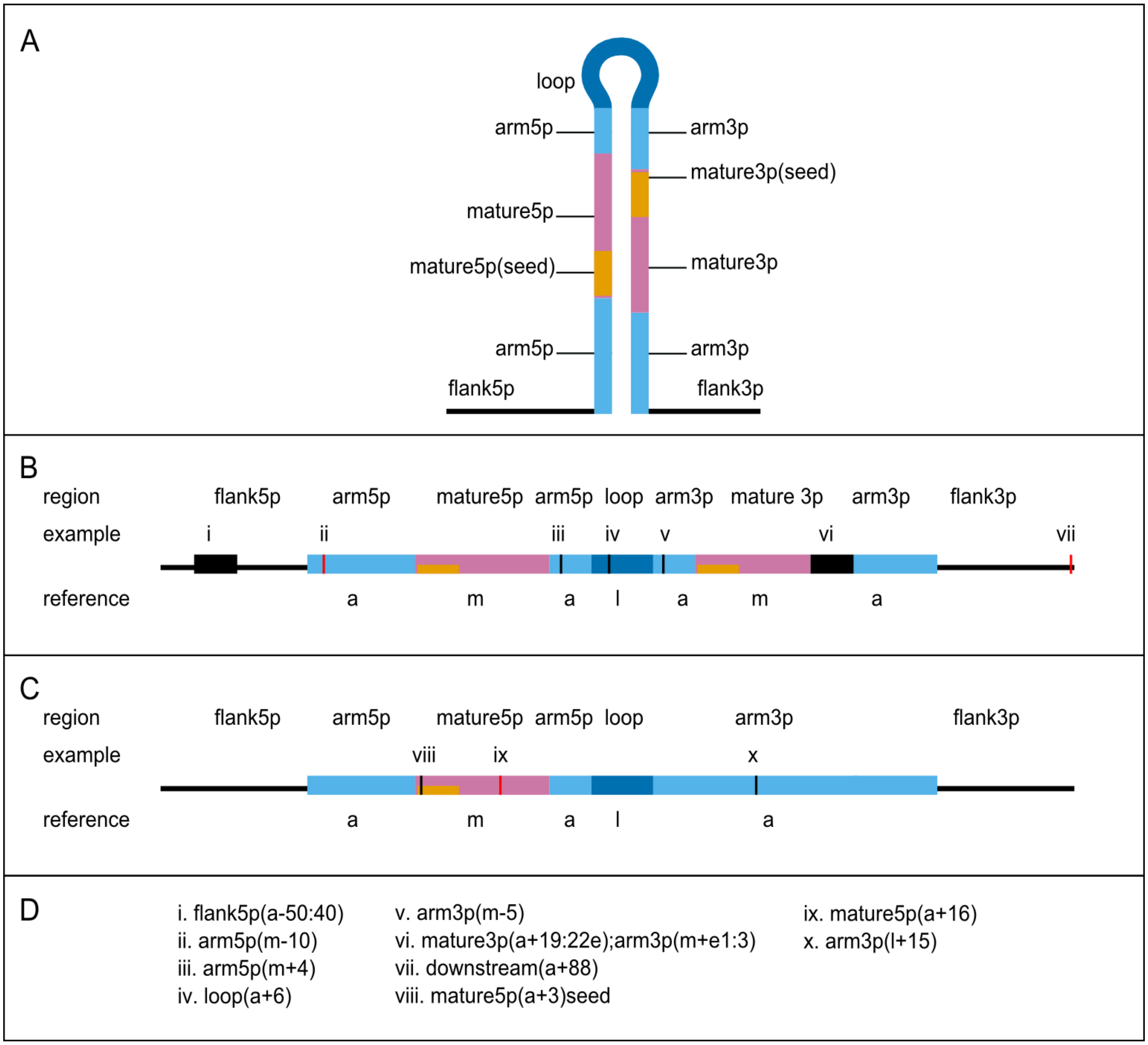


Figure 5: miRVaS annotation scheme. Image taken from Cammaerts et al., submitted.

Examples of sequence and structural changes (examples taken from: Cammaerts et al., submitted):
i) a structural change in the 5p flanking region covering bases 50-40 upstream of the hairpin arm
ii) a sequence change in the lower 5p hairpin arm, 10 bases upstream of the 5p mature
iii) a structural change in the upper 5p hairpin arm, 4 bases downstream of the mature 5p
iv) a structural change in the loop region, 6 bases downstream of the upper 5p arm
v) a structural change in the upper 3p hairpin arm, 5 bases upstream of the mature 3p
vi) a structural change spanning the end of the mature 3p (bases 19-22 downstream of the upper 3p arm, base 22 is the last base of the 3p mature sequence) and the beginning of the lower 3p arm (1st-3rd base downstream of the mature 3p)
vii) a sequence change outside the given flank region, 88 bases downstream of the lower 3p hairpin arm
viii) a structural change within the seed sequence, 3rd base downstream of the lower 5p hairpin arm
ix) a sequence change within the 5p mature sequence, 16 bases downstream of the lower 5p hairpin arm
x) a structural change within the 3p hairpin arm, 15 bases downstream of the calculated loop region.

6.3. Highest structural impact (“highest impact” columns)

This column denotes the “most important” region with a structural impact. Importance of regions is ranked as follows: seed > mature > arm > loop > flank. The usefulness of this parameter was demonstrated on a test set (Cammaerts et al., submitted).

6.4. Conservation

These columns specify whether the structure prediction with the tested flank size retains the structure of the hairpin. When the structure of the hairpin region is the same for predictions with the tested flank size and the hairpin without flanks, the structure prediction is flagged “conserved”. In other cases, the structure predicted in flagged as “changed”, followed by the number of bases that have different structure in the prediction with given flank sizes.

6.5. Filtering

If you have a list of genetic variants near miRNA genes, several approaches might help to reduce this list to the variants that, based on the predictions, might be more interesting to functionally validate. To check the relevance of the structural changes induced by the variant and the variant location, you can for instance look at:

- variant location: is the variant located within the mature miRNA sequence or other relevant regions
- what is the highest impact of structural changes, e.g. is it located within the hairpin
- is the (highest) structural change the same for different structure representations (CEN, MEA, MFE)
- is the (highest) structural change the same for predictions with different flank size
- conservation: is the hairpin structure using a specific flank size conserved and if not, how many bases are changed

7. Example: running the test set

As an example, we will describe how to run miRVaS and how to interpret the output for the test set that was used in the study (Cammaerts et al., submitted).

Following files can be found in the miRVaS download:

- input variant file (“testset_input.txt”)
- genome file (“genome_hg19.ifas”)
- miRNA database file (“mirna_hg19_mirbase20.tsv”)

Next, start the miRVaS GUI. Specify the location of the input files. Specify where the results should be saved and with which name (e.g. “testset_fl100.txt”).
Specify the flank sizes: for this example, use 100nt for upstream and downstream flank size. Start the miRVaS run. The blue progress bar will indicate when miRVaS is ready.

Open the resulting html file (named: “testset_fl100.html”). A summary of all variants, with their location, the highest structurally changed region and the conservation of the hairpin, are described. The links on this page will direct you to the structure prediction of a specific variant.

Consider the first two results in the table of the html file. They describe the effect of the variant chr1:98511730-98511731 C>T on two different miRNAs. The variant is located 824 bases upstream of the hairpin of hsa-mir-2682 and is also located in the flanking region of hsa-mir-137, 4 bases upstream of the hairpin.

Because the input file was run with 100 nt of upstream flanks, the impact of the variant on the secondary structure for hsa-mir-2682 cannot be predicted, since this variant is located more than 100 nt upstream from the miRNA. The structural change field says “variant not in analysed region” to indicate this.

However, the structure predictions can be done for hsa-mir-137, since the variant is located within the flanking region that was specified. The structural changes field says: flank, flank, arm. This is the highest impact change for CEN, MEA and MFE predictions respectively. The conservation for CEN, MEA, MFE predictions is: changed 9, changed 9, changed 14; so e.g. for centroid predictions the hairpin including 100 nt flanks has 9 bases that have a structural change (paired vs unpaired base) within the hairpin compared to the secondary structure prediction of the mir-137 hairpin without flanks.

Go to the structure predictions for hsa-mir-137 with this variant. In this page you can see the centroid prediction for wild-type miRNA (to the left) and for the variant miRNA (to the right). It is also possible to inspect the MEA and MFE structure representations. The genetic variant is coloured in red in the images, the black bases indicate bases that change structure (paired vs unpaired bases) due to introduction of the variant. All structural changes are annotated in this page. For instance, in the MFE prediction, the highest region with structural change is the hairpin arm: there are structural changes in the 5p and 3p flanks, but also in the 5p hairpin arm (47 nt before the loop and bases 45-40 before the loop) and in the 3p hairpin arm (bases 12- 21 downstream of the mature sequence, base 21 is the last base of the hairpin arm).

In the text file output (“testset_fl100.txt”) it is also possible to look at the difference in free energy between the wild-type and variant miRNAs in the fields ref_Cen_deltaG and var_Cen_deltaG to assess whether the variant might affect the stability of the miRNA.

8. References

1. Lorenz,R., Bernhart,S.H., Höner Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol. AMB*, **6**, 26.

2. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinforma. Oxf. Engl.*, **25**, 1974–1975.

3. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–73.