# RNAfoldEff 1.0 user manual

## 1. About RNAfoldEff

RNAfoldEff is an easy-to-use application for windows OS, which annotate the functionality of variants (SNPs and Indels) in human regulatory RNAs (non-coding RNAs as well as UTRs in mRNAs) based on how they alter the RNA secondary structure. Taking a VCF (high-throughput model) or user specified files with sequences and variants (low-throughput model) in our pre-defined format as input, RNAfoldEff will evaluate how the RNA function is altered due to variants through calculating the structural change. RNAfoldEff also annotates the transcript id, RNA type, conserved scores *et al.* to provide more information associated with variants.

## 2. Methods

RNAfoldEff uses human genome and GENCODE as references to tell whether the variants in a VCF file locate in regulatory RNAs. It performs equilibrium partition function by calling RNAstructure and generates a base-pair-probability matrix to predict RNA secondary structure. The correlation coefficient (range:0-1) between the matrix of the reference sequence and the variant sequence is calculated to evaluate the structural change due to the variant (formula in Fig 1).Variant with low coefficient values indicate potential functional damaging.

$$R = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{(\sum\sum (A_{mn} - \bar{A})^2)(\sum\sum (B_{mn} - \bar{B})^2)}}$$

Fig. 1. Computational formula for correlation coefficient, where 'A' and 'B' are reference and variant base-pair-probability matrix respectably, m and n are row and column respectably.

## 3. Availability

RNAfoldEff is available via https://github.com/wksofia/RNAfoldEff for free.

## 4. Installation

Required database and third party tools:

RNAfoldEff requires several databases and tools to work. Table 1 lists all the required materials.

| Database/ tool | Description | Format | Source |
|---|---|---|---|
| Human genome | Necessary for VCF annotation | Fasta | ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/ |
| phonCons | Optional for VCF annotation | wigFix | http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way/ |
| phyloP | Optional for VCF annotation | wigFix | http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phyloP100way/ |
| Gencode | Necessary for VCF annotation | GTF | ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/ |
| RNAstructure | RNA structure partition function | - | http://rna.urmc.rochester.edu/RNAstructure.html, included in RNAfoldEff. |

Table1. Materials required in RNAfoldEff.

- Please make sure you have java and C++ environment on your machine. It requires at least 2Gb memory to run in high-throughput model.
- Files in the folder "data_tables" (see https://github.com/wksofia/RNAfoldEff) are thermodynamic parameters required for RNAstructure. In order to successfully use the data tables in the RNAstructure repository during calculations, an environment variable called DATAPATH must be set to the location of the data_tables folder on the current machine.
- To decrease compute complexity in high-throughput model, Human reference genome and transcriptome annotations in GENCODE and conserved score files are required to be processed to pick up the information of regulatory RNAs before annotating VCF files. We provide processed files generated from GENCODE 19 and GRCh37, so you are no necessary to download the files in these versions any more. Users who need other versions of human reference genome or GENCODE could make their own files by tools implemented in RNAfoldEff.

## 5. USAGE

### Preparation for the input:

In high-throughput model, RNAfoldEff takes VCF files generated from NGS data as input. For detailed description for this format please check:

http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41In low throughput model, RNAfoldEff accepts RNA sequences in fasta format and a variant file as input. In the variant file, variants (format: [site:REF:ALT] ) in the same sequence were tab-

delimited in a line. Each line corresponds to variants in one RNA sequence sequentially in the RNA sequence fasta file. Here is the example:

Example: (Each row of variants correspond to one RNA sequence)

| Sequences file: fasta | Variants file: [site: REF:ALT][tab] [site: REF:ALT].. |
|---|---|
| >1 | 22:G:A 59:C:T |
| ATCG… | 141:C:T 135:C:T 117:G:A 103:C:T |
| >2 | … |
| CATG… | |
| … | |

Table 2. Input format for low throughput model.

RNAfoldEff can perform well in both command-line and graphical user interface.

**Use in Command-line:**

Table 3 lists all the RNAfoldEff optional parameters.

*USAGE*:

Reference processing:

Java –jar RNAfoldEff.jar [–P] [genomePath] [gencodePath][outputPath]<options>

High throughput:

Java –jar RNAfoldEff.jar [–VCF] [VCFpath] [outputPath][–R] [ncTranscriptomePath]<options>

Low throughput:

Java –jar RNAfoldEff.jar [sequencesPath] [VariantsPath] [outputPath]<options>

| **Parameter** | **Description** |
|---|---|
| -S[phastConsPath][phylopPath][scoreOutput] | Preprocess the conserved scores. Used in "reference processing" model. |
| -L[start-end] | Range of RNAs' length. Default value: 0-500. Used in high/low throughput model. |
| -PC [ncPhastConsPath] | PhastCons annotation. Used in high throughput model. |
| -PP [ncPhylopPath] | PhyloP annotation. Used in high throughput model. |
| -H | Print whole parameters. |

Table.3. RNAfoldEff optional parameters.

**Graphical user interface**: RNAfoldEff GUI is easy to use. Figure 2 shows the annotation and supplementary tool interface.

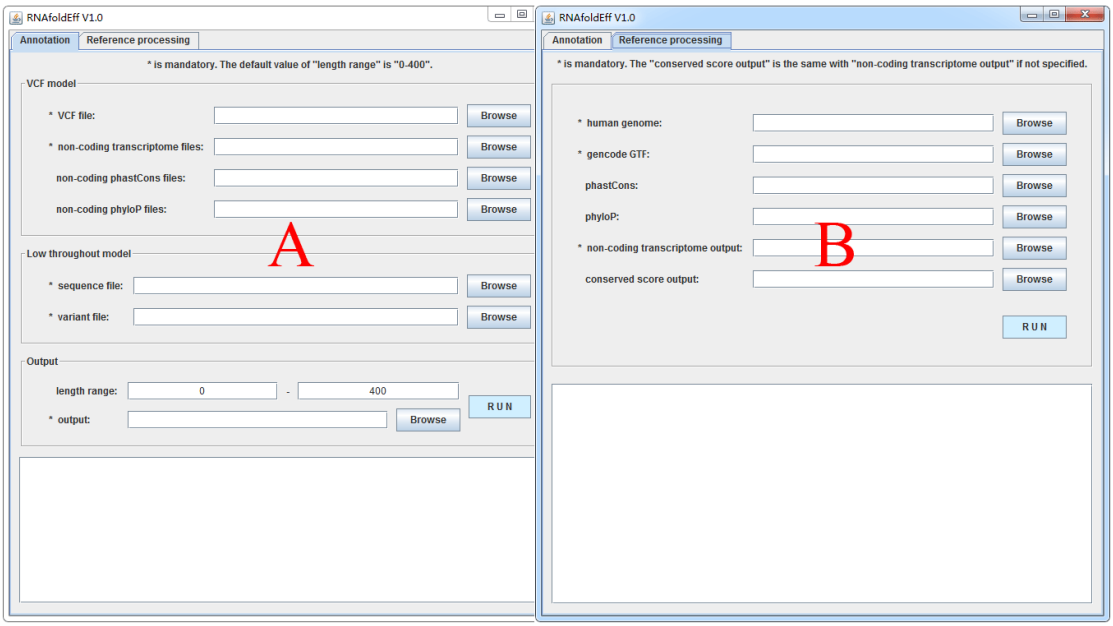


Fig 2. RNAfoldEff GUI. A. The annotation interface. B. The supplementary tool interface.

## 6. OUTPUT

The results are output into a tab-delimited txt file named by your_result.txt, which can be open and edit in OFFICE Excel.

Example:

```
#########################################################################
# RNAfoldEff v1.0
# Time = 2014/2/24-23:34:26
# Parameter = -VCF F:\test\sample.vcf F:\test\out  -L 200-300 -PC F:\test\non-coding phastCons -PP F:\test\non-coding phyloP -R F:\test\ncT
# POS     CHROM   STRAND  TRANSCRIPT          TYPE        REF:ALT  phastCons phyloP   R
#########################################################################
982941    chr1    +       ENST00000478677.1 retained_intron     87:T:C   0.0       0.109    0.9997878790727076
982994    chr1    +       ENST00000478677.1 retained_intron     140:T:C  0.405     0.501    0.9994089661834598
1220954   chr1    +       ENST00000467651.2 processed_transcript 181:G:A null      null     0.9975670080366864
1242468   chr1    -       ENST00000438966.1 processed_transcript 119:G:A 0.0080    0.525    0.9649210972352719
1254841   chr1    -       ENST00000526332.1 5-UTR       144:C:G  0.699     -0.147   0.9027582919678467
1335790   chr1    +       ENST00000570344.1 3-UTR       106:A:G  0.0       -0.317   0.8175585980452919
1663851   chr1    -       ENST00000246421.4 3-UTR       46:G:C   0.0       -1.37    0.7617124508140477
1663861   chr1    -       ENST00000246421.4 3-UTR       36:G:A   0.0       -0.774   0.9999996378181255
```

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2121118 | chr1 | - | ENST00000378543.2 | 3-UTR | 101:C:T | 0.0 | -0.253 | 0.9923522108709878 |
| 2564465 | chr1 | - | ENST00000288709.6 | 5-UTR | 17:T:C | 0.0 | -3.105 | 0.9912623496694996 |
| 3397188 | chr1 | + | ENST00000413250.2 | 3-UTR | 40:T:C | 0.0 | -1.652 | 0.9615100410096327 |
| 6087411 | chr1 | + | ENST00000435937.1 | 5-UTR | 55:T:C | 0.0 | -4.576 | 0.735477294666612 |
| 6165419 | chr1 | - | ENST00000475121.1 | 3-UTR | 12:C:T | 0.0010 | 0.21 | 0.6453036423660876 |
| 6659505 | chr1 | - | ENST00000463043.1 | 5-UTR | 166:G:A | 0.0 | -5.599 | 0.9943686941196597 |
| 6693097 | chr1 | + | ENST00000472925.1 | 3-UTR | 278:A:G | 0.0 | -2.49 | 0.9922184008670606 |
| 7760892 | chr1 | - | ENST00000602916.1 | lincRNA | 139:A:G | 0.0 | -1.488 | 0.9999921289967388 |
| 7760985 | chr1 | - | ENST00000602916.1 | lincRNA | 46:G:A | 0.0 | -0.217 | 0.99376080780487 |
| 7761001 | chr1 | - | ENST00000602916.1 | lincRNA | 30:G:A | 0.0020 | -0.141 | 0.9852309987643402 |
| 9039704 | chr1 | - | ENST00000583026.1 | misc_RNA | 230:G:A | 0.201 | 0.225 | 0.8395769912370391 |
| 9148470 | chr1 | - | ENST00000487835.1 | 5-UTR | 47:T:G | 0.03 | -1.744 | 0.9998953977352849 |
| 9148470 | chr1 | - | ENST00000464985.1 | 5-UTR | 33:T:G | 0.03 | -1.744 | 0.9996072799319039 |
| 9427842 | chr1 | + | ENST00000357898.3 | 3-UTR | 211:A:G | 0.0 | -0.169 | 0.9010501022795621 |

Here R column is the coefficient value of the structural matrix of the reference sequence and the variant sequence. The smaller values mean that it is more likely the variant will alter the structure and thus exert an effect on the functionality of the corresponding RNA.

## 7. Contact

Qian Zhao (zhaoqian@fuwaihospital.org)

WenKe Li (wk1lian@126.com)

If you find any bugs, please let us know.

## 8. FAQs

### 1. What's the meaning of "3' exceed" and "5' exceed" in annotation log?

These means the variant spans outside the RNA sequence. "3' exceed" means the variant exceed RNA 3' terminal, while "5' exceed" exceed 5' terminal. It may occur in high throughput model annotation.

### 2. What's the meaning of "unmatched" in annotation log?

The mark "unmatched" means neither of the allelic bases in variant file matches with the reference base in RNA sequence. You should check the variant file and sequence file.

### 3. Why the coefficient value in results shows "NaN"?

When the RNA sequence has exceptional bases or it is too short, RNAstructure will fail to perform equilibrium partition function and output "NaN".

### 4. Why RNAfoldEff costs more than 2 GB memory on my computer?

It depends on the Java platform's garbage collection mechanism. Java usually won't release useless variable when you have enough memory. However, 2 GB is enough for RNAfoldEff. And you can control the max memory of JVM by Java parameter: -Xmx2000m.