Random Mutation Table Documentation

Marius Kühl

March 5, 2020

Contents

1 Example

```
#meta−information
fasta=example.fa
md5=fcd082b26602f0114096fc4c91fcc726
species_name=Example Species
assembly_name=example_assembly_v1
titv=1
in_block=100
du_block=2500

#standard
std
it None
None

#range definitions
chr 2
it 0.00001
1−END sn 0.05 in 0.001 inl 1000
chr 5
70−12038 None
13023−END de 0.001 del 100 du 0.001 dul 100
chr 3
it 0.00001
```

2 General

This file format allows the user to create specific recreatable patterns in which mutations can be generated at given rates and lengths across specific chromosomes. RMT also features meta-information about the genome, specific interchromosomal translocation (IT) rates for each chromosome in the reference, standard values for unspecified areas and blocking positions from mutating entirely.

RMT files are structured in 3 sections: Meta-Information (optional), Standard and Range-Definitions (optional). Meta-Information provides optional settings and check marks for the end user of the RMT file. The Standard section provides detailed information about mutation rates and applies when no range definitions are set at all, or whenever ranges or chromosomes are missing. The Range-Definitions are the key feature of the RMT file format and Mutation-Simulator. This allows for declaring specific areas on the genome, where the mutation rates differ from the standard to achieve the most natural simulation possible. For declaring mutation rates, RMT files utilize a two letter acronym format similar to the options used in ARGS.

To add comments or custom header information to the RMT file, the # sign is used and can be placed anywhere in the file. Mutation-Simulator's RMT parser will not read any notations following a number sign.

3   Meta-Information

Table 1: METAINF Keywords

Keyword	Description
fasta	Name of the Fasta file, this RMT is created for
md5	MD5 hash of the corresponding Fasta file
titv	Transition / Transversion ratio (SNPs)
species_name	The species name for the VCF file
assembly_name	The assembly name for the VCF file
sn_block	Number of bases blocked after a SNP event
in_block	Number of bases blocked after an insertion event
de_block	Number of bases blocked after a deletion event
du_block	Number of bases blocked after a duplication event
iv_block	Number of bases blocked after an inversion event
tl_block	Number of bases blocked after a translocation event

The keywords can be included in the optional METAINF section of RMT files.

METAINF can hold multiple key value pairs in variable order (see Table 1). Each key used must be assigned to a value with a = symbol. Key value pairs must be separated with a newline. If the name of the referenced Fasta file does not match the value of the "fasta" keyword (see Table 1 line 1) in the RMT file, Mutation-Simulator launches a warning. This also applies to the MD5 hash (see Table 1 line 2), but this can be circumvented by executing Mutation-Simulator with the "–ignoremd5" option. If the "species_name" and "assembly_name" options are missing in the RMT file, the VCF file will state "Unknown". The block keywords (see Table 1 lines 6-11) function is equivalent to the block options in ARGS mode. The default value for missing block keywords is 0.

4   Standard

Table 2: Options for RMT

Option	Description
None	Sets all rates to 0
sn	Rate of SNPs
in	Rate of insertions
inl	Maximum length of inserts
de	Rate of deletions
del	Maximum length of deletions
iv	Rate of inversions
ivl	Maximum length of inversion
du	Rate of duplications
dul	Maximum length of duplications
tl	Rate of translocations
tll	Maximum length of translocations

The table shows all possible options for RMTs. Those can be used in the STD and RD section of RMTs.

The only mandatory section is Standard (STD), which must begin with a "std" keyword and a newline. STD contains detailed information about mutation rates per type and always applies when either areas on the chromosome are not defined, a chromosome is not mentioned or the Range Definitions (RD) section is missing entirely. Right below the "std" keyword, a new keyword "it" must be stated. The standard rate for interchromosomal translocations can be set as a float in the range 0-1 separated from the keyword with a space symbol. This rate will apply to chromosomes that are missing from the third section of the RMT or when a chromosome entry has no "it" keyword. To block chromosomes from interchromosomal translocations entirely the value must be set to "None", as 0 will still allow Mutation-Simulator to select this chromosome as a partner for another chromosome.

The line below can hold multiple keyword value pairs for each mutation type (see Table 2). Those keywords must be separated from the value with a space symbol, as well as the pairs themselves.

If the "None" keyword is used, no other keyword can be selected (see Table 2 line 1). It will prevent all standardized regions on the genome from mutating. All other keywords can be combined as needed. If a rate keyword is used, except "sn" (see Table 2 line 2), its corresponding length keyword must be specified too.

5 Range Definitions

RMT is designed to be tailored to specific genomes or Fasta files, which can be done in the RD section. Each region on individual chromosomes that should deviate from the mutation rates stated in STD can be redefined here.

To accomplish this, the keyword "chr" followed by a space character and the number of the chromosome in the reference Fasta can be added below the STD section. One or more chromosomes can be added in no specific order. If the rate of interchromosomal translocations shall differ from STD, it must be stated in the first line below the "chr" statement. Also single chromosomes can be blocked for interchromosomal translocation with an "it None" statement. Each following line can contain a range in the format "start-stop" with the same keyword value pairs as in the STD section (see Table 2 lines 1-11). These keyword value pairs must be in the same line, separated with a space symbol. If multiple ranges are defined, they should be stated in ascending numerical order, each in a new line. Overlapping should be avoided to guarantee correct results. RMT uses 1 based positions and the defined ranges are inclusive. The last specified range can be defined with an "END" statement, replacing the "stop" in the "start-stop" expression to specify the last base of a chromosome.