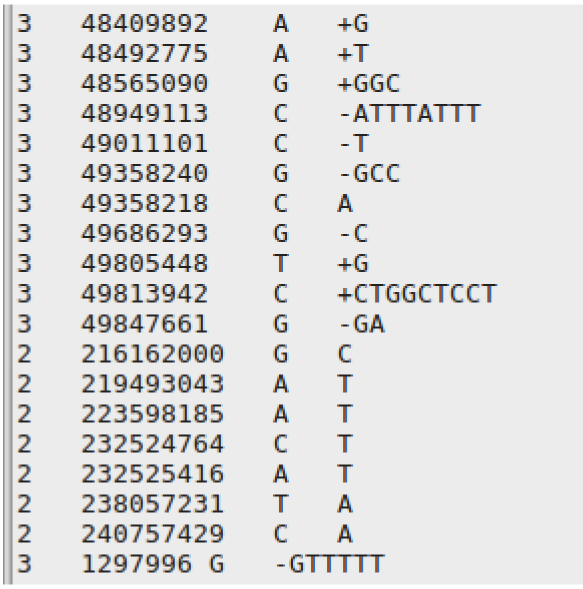
TransPEM manual

1 Introduction and input files

TransPEM (Software for Translation sequence into peptide based on mutation information) is a novel bioinformatics tool which was developed for providing mutated and wild type sequences based on bioinformatics analysis of single nucleotide polymorphisms. Protein sequences can serve as an input for proteomics analysis. This software provides fast and reliable extracting of coding sequences (CDS) from genome and their translation into proteins with respect to the information about mutations. It is suitable for any number of sequences and as an user-friendly software requires only the basic knowledge of a Linux operating system.

The software requires three main types of input files. The first required input file is a FASTA formatted file containing the genomic sequence. The second input file is a GTF file which is compatible with the fasta file. The last input is file which contains the information about mutations. The file must have defined structure (tab delimited format) and consists of four columns which are labeled as: 1. chromosome name, 2. position of mutation on chromosome, 3. reference allele at this position, 4. variant allele at this position (Figure 1). The software also accept insertion and deletion events. The insertion/deletion event must be listed in fourth column as +nucleotides/-nucleotides (Figure 1).



```
3   48409892    A   +G
3   48492775    A   +T
3   48565090    G   +GGC
3   48949113    C   -ATTTATTT
3   49011101    C   -T
3   49358240    G   -GCC
3   49358218    C   A
3   49686293    G   -C
3   49805448    T   +G
3   49813942    C   +CTGGCTCCT
3   49847661    G   -GA
2   216162000   G   C
2   219493043   A   T
2   223598185   A   T
2   232524764   C   T
2   232525416   A   T
2   238057231   T   A
2   240757429   C   A
3   1297996 G   -GTTTTT
```

Figure 1: Example of the input file, which represents information about mutation in nucleotide genomic sequence.

2 Prerequisites

TransPEM was created using Perl and is primarily designed for GNU/Linux but can run on other operating systems such as MS Windows or Mac OSX fulfilling requirements for launching.
The software prerequisites are Perl v5.18.2 or higher and samtools v0.1.19 or higher. The program further requires the following Perl modules, which are available on CPAN (Perl Archiving Perl Network, https://www.cpan.org):

- Getopt::Long
- Term::ANSIColor

3 Obtaining and installation

The latest source release is available on the web pages of RECAMO (http://www.recamo.cz/en/). To install the software from the source package unpack the directory by the command:

```
unzip TransPEM_v1.1.zip
```

4 Using TransPEM

Usage: perl main.pl [options]

The software can be launched directly from installation directory by the command:

```
perl mut_pep_extract.pl
```

In this case, the warning about the wrong number of input arguments and the help message is listed. To see the help message only use the command:

```
perl mut_pep_extract.pl --help
```

The software uses only the long switches (--) instead of the short ones (-). Options for software usage are divided on required params (which mast be listed) and other params which have default values.

Required params:

--input_reference <string>   Path to input file which contains the genome sequence.

--input_GTF <string>   Path to input GTF file compatible with genome sequence.

--input_MUT <string>   Path to input file with mutation information and defined format.

Other params:

--help Prints the help message and exit.

--out_WT_nucl <string> Name or path and name which software will use to write output file which will contain nucleotide CDS without mutation. The default option is "./WT_nucl.nucl".

--out_MUT_nucl <string> Name or path and name which software will use to write output file which will contain nucleotide CDS with mutation. The default option is "./MUT_nucl.nucl".

--out_WT_pep <string> Name or path and name which software will use to write output file which will contain peptide sequence without mutation. The default option is "./WT_pep.pep".

--out_MUT_pep <string> Name or path and name which software will use to write output file which will contain peptide sequence with mutation. The default option is "./MUT_pep.pep".

--out_log <string> Name or path and name which software will use to write logfile. The default option is "./logfile.log".

5 TransPEM output files

The TransPEM produces several output files in the directory in which it was launched. All the output files are described in the following subchapters.

5.1 Fasta output files

The software provides four output file in fasta format which contains nucleotide/peptide sequence for each transcript with mutation. Only sequences with mutation are translated from nucleotide to protein sequence. The head of output transcript is divided into sections by "_", which can be labeled as: 1. Gene name (from GTF file), 2. Gene identificator (from GTF file), 3. Transcript identificator (from GTF file), 4. Mutation description in protein sequence, 5. WT/MUT for labeling sequence without/with mutation.

5.2 Logfile

The file contains information about mutations which were processed by software. It has defined structure (tab delimited format) and consists of five columns which are labeled as: 1. Gene name, 2. Transcript name, 3. Mutation position in peptide sequence, 4. Aminoacid in peptide sequence without mutation, 5. Mutated aminoacid.

5.3 Error.log

The file contains information about mutations which cannot be processed by software. It has defined structure (tab delimited format) which is the same as input file with mutation information.