tSNPannot User Manual

Summary

Description 1

Prerequisites 2

  Operating System 2

  Perl 2

  Similarity search programs 2

  Connection 2

  Disk space 2

Installation 3

  Perl packages 3

  Similarity search programs 3

  Program 3

Program step and operation 4

Launch program 6

  Usage 6

  Option detail 6

Input files format 10

Understanding the output files 12

  Core information 12

  Warnings 15

  Verbose mode 15

  Rejection Mode 17

  Usage tips 18

The configuration file 20

  File structure 20

  Parameters definition 20

Description

This application allows you to annotate SNPs present in species without a sequenced genome. Transcribed sequences or parts of transcribed sequences are typically used to identify SNPs.

In tSNPannot, annotations are constructed from similarities with the proteome of one or more near species selected from the Ensembl project.

The annotations contain :

- ✓ SNP consequences : amino acid change in the sequence featuring the SNP (optionally also virtual consequences of the SNP on the genomes of reference species).
- ✓ the distances between SNP and the containing exon limits (useful for primer design).
- ✓ similar SNPs found in the genomes of reference species.

Prerequisites

Operating System

Linux

Perl

Perl interpreter :

version 5.10.x
You can check what Perl version is installed on your system by issuing the command '*perl -v*' in your terminal.

Perl packages :

BioPerl 1.6.1
Try::Tiny >=0.09 for exception handling.
Archive::Extract >=0.56 to extract protein banks.
Net::FTP to download protein banks.
XML::Writer >= 0.612 to write output file.
Ensembl to query then Ensembl project database for genomic and transcriptomic information.
Ensembl-variation to search for information on a SNP in a reference genome.

Similarity search programs

tSNPannot can use either of :

- NCBI blast 2.2.24+ or later version accessible by the command 'blastall',
- fasta 35.4.12. accessible by the command 'fast35',
- blast2 2.2.21.20090809 (on synaptic) accessible by the command 'blast2'.

Caution : Currently only fasta can handle frameshift in sequence alignment.

Connection

An Internet connection is necessary to download the Ensembl protein banks and to retrieve information about reference genomes and their variations.

Disk space

The disk space required depends on banks that you will use. The application itself is only 500 KB.

Installation

Perl packages

BioPerl 1.6.1

Use synaptic for a simple and graphical installation of this package (keyword : bioperl).

If you do not use Ubuntu, see http://www.bioperl.org/wiki/Installing_Bioperl_for_Unix for installation instructions.

Try::Tiny, Archive::Extract, Net::FTP and XML::Writer

If you don't have one of these packages, use CPAN for its install : in terminal write ‘cpan’, then write ‘install PACKAGE_NAME’.

Ensembl and Ensembl-variation

The files and installation instructions are available on this URL http://www.ensembl.org/info/docs/api/api_installation.html.

Similarity search programs

Fast-All

The files and installation instructions are available on this URL http://fasta.bioch.virginia.edu/fasta_www2/fasta_down.shtml.

NCBI Blast

The files are available on this URL ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/.

Program

Copy the application directory into a convenient location and add it to the PATH environment variable.

Note : add folders to your PATH. In terminal write “echo $SHELL”, according to the results :

- csh or tcsh : add ‘setenv PATH <application_directory>:$PATH’ to ~/.tcshrc
- bash : add ‘export PATH=<application_directory>:$PATH’ to ~/.bashrc

Program step and operation

Annotation in tSNPannot is done in eigth steps.

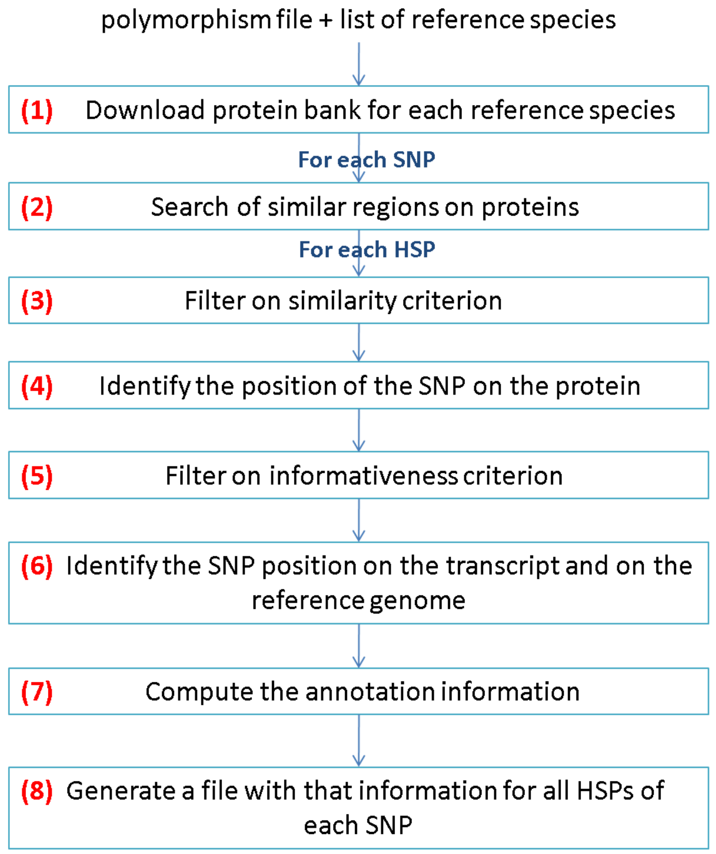

polymorphism file + list of reference species
(1) Download protein bank for each reference species
For each SNP
(2) Search of similar regions on proteins
For each HSP
(3) Filter on similarity criterion
(4) Identify the position of the SNP on the protein
(5) Filter on informativeness criterion
(6) Identify the SNP position on the transcript and on the reference genome
(7) Compute the annotation information
(8) Generate a file with that information for all HSPs of each SNP

Figure 1 Steps to the annotation of SNPs from closely related species.

An SNP annotation is generated by comparing the SNP with the transcripts and the genomes of reference species. However, phylogenic distance between reference species and studied specied can compromise the relevance of the comparisons at the genetic level. So, the software first searches for similarities at the protein level.

(1) tSNPannot downloads the necessary proteins banks from Ensembl. Banks are downloaded only if they have never been, or if a newer version exists on Ensembl.

(2) The software searches for similarities with reference proteins.

(3) For each identified HSP, the conserved fraction and identical fraction must be greater than or equal to the provided conserved and identical values (see these options). Furthermore, amino acids around the SNP must be conserved (see conservsize option).

(4) tSNPannot calculates the position of the SNP on the reference protein.

(5) In this step the software rejects HSPs when the protein does not provide information about SNP :

the SNP corresponds to a gap or an insert on the reference protein,

the SNP does not exist in the region of the HSP.

(6) tSNPannot calculates the position of the SNP on the transcript and the genome of the reference species.

(7) The software determines :

the distances to the SNP-featuring exon limits,

the similar SNPs found in the genomes of reference species,

the consequence of the SNP on the studied sequence,

the consequences of the SNP when projecting on the genome of reference species.

(8) tSNPannot generates an output file for these annotation.

Launch program

Usage

```
tSNPannot --species LATIN_NAME [LATIN_NAME ...]
           --snpfile FILE_PATH
           --outfile OUTPUT_FILE
          [--blastprog EXECUTABLE_NAME]
          [--conserved FRAC_CONSERV]
          [--identical FRAC_IDENTICAL]
          [--conservsize NB_AA]
          [--shiftpenal PENAL]
          [--blastopt PROG_OPTIONS]
          [--outtype OUTPUT_FORMAT]
          [--tmp TEMP_DIRECTORY]
          [--threadid THREAD_ID]
          [--parallel]
          [--rejection]
          [--verbose]
          [--help]
```

Options between [] are optional. All other options are mandatory.

The 'threadid' option must not be issued directly by user. It is for internal use only.

Option detail

species :

The program uses the comparison with the gene products of these species to generate the annotation of each sequence.

Example of correct syntax for this option : -species 'Homo sapiens'

If you want more than one species, add latin name of each species after the last one.

Example : -species 'Danio rerio' 'Tetraodon nigroviridis'

Caution : The given species must exist in the Ensembl database (see the current list of species on http://www.ensembl.org/info/about/species.html).

Note : Not all species present in Ensembl have variation information. Without this information, the software will not be able to determine :

whether the reference features an equivalent to the SNP,

whatever effect the SNP has on the reference transcripts.

snpfile :

Path to the file containing the sequences to annotate.

Caution : These sequences must represent cDNA molecules.

See ‘Input files format’ for more information.

Outfile :

Path to the file that will contain results of the annotation.

Blastprog :

Default : fastx35.

Program name to use to find similar proteins.

An SNP annotation is generated by comparing the SNP with the transcripts and the genomes of reference species. However, phylogenic distance between reference species and studied specied can compromise the relevance of the comparisons at the genetic level. So, the software first searches for similarities at the protein level. The program used for the searches step is specified by this option.

Currently, program name must be one of blast2, blastall or fastx35.

conserved :

Default : 0.7.

To infer an annotation, tSNPannot only retains the HSPs in which the conserved fraction is greater than or equal to the specified value. HSPs are the result of alignment between a the proteins of a reference species and your sequence.

identical :

Default : 0.55.

To infer an annotation, tSNPannot only retains the HSPs where identical fraction is greater than or equal to the specified value. HSPs are the result of alignment between a the proteins of a reference species and your sequence.

conservsize :

Default : 6 amino acids.

The minimum number of amino acids that must be conserved on both sides of the alignment around the amino acids corresponding to the SNP. HSPs that fail to comply are discarded.

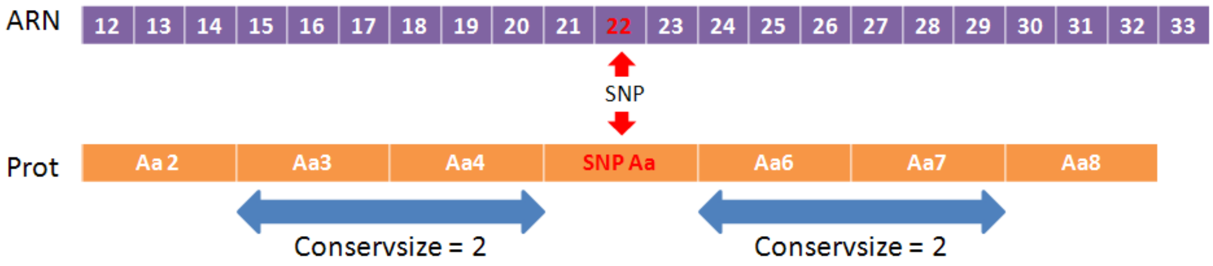


Figure 2 Example of conservsize application. In this example with a conservsize of 2, the amino acids 3, 4, 6 and 7 of the protein must be conserved with regard to the reference protein.

shiftpenal :

Default : 10.

Frameshift penalty to use when searching for similar proteins. This option is only valid with fastx.

blastopt :

Additional parameter string that will be appended to the blastprog command upon invocation.

See –blastprog option and refer to the documentation of the selected program.

Example of right syntax for this option :

```
-blastprog blastall -blastopt -blastopt " '-f 10 -X 10' "
```

Note : depending on the selected blastprog, some options are used internally and therefore not directly issuable by the user :

Blastx : query, db, out, outfmt.

Blast2 : p, i, d, o, m, F (filter false).

Blastall : p, i, d, o, m, F (filter false).

Fastx : Q, H, m, h, O.

outtype :

Default : txt.

Output file format : plain text (txt) or XML (xml).

See 'Understanding the output files' for more information.

tmp :

Default : /tmp.

Path to the folder where temporary files will be created.

parallel :

Requests parallel execution of the program. This option increases computation speed. Its use is restricted to infrastructures with SGE scheduler.

The number of CPUs to use is parametered in configuration file (See 'The configuration file').

threadid :

Internal option. It is automatically used to differentiate each execution context when using the parallel option.

Caution : You must not use this option.

rejection :

Adds information to the output file. These are information about the similar proteins that were not used for the annotation because they were deemed uninformative.

The verification of informativeness excludes the HSPs where the SNP position corresponds to a gap or an insert on the reference transcript. It also excludes HSPs where border regions are not conserved (see option conservsize).

See 'Understanding the output files' for more information.

verbose :

Adds information to the output file. These information contain query string and hit string of each HSP. They also mention consequences of your SNP on the reference genome.

See 'Understanding the output files' for more information.

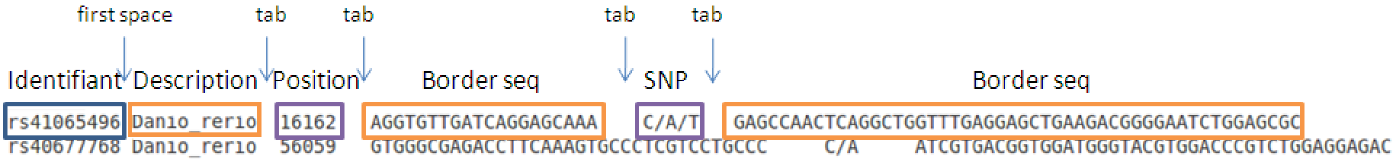
help :

Displays help and exits.

Input files format

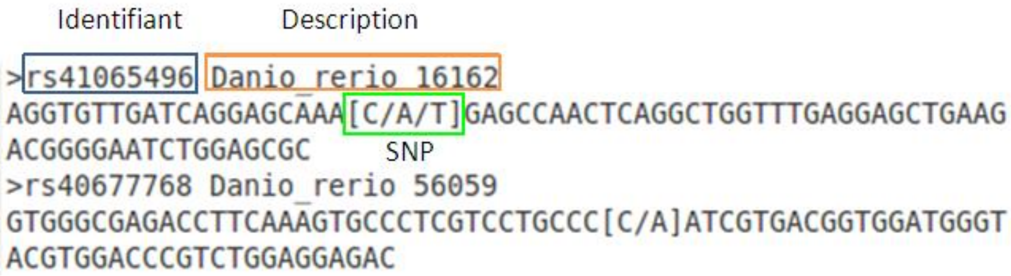
The input file contains the SNPs to annotate. Each SNP is indicated with its border sequences. The sequences (SNP and border sequences) must represent desoxyribonucleic molecules.

You can use three different file formats : tabular, fasta or fasta usable by blast.



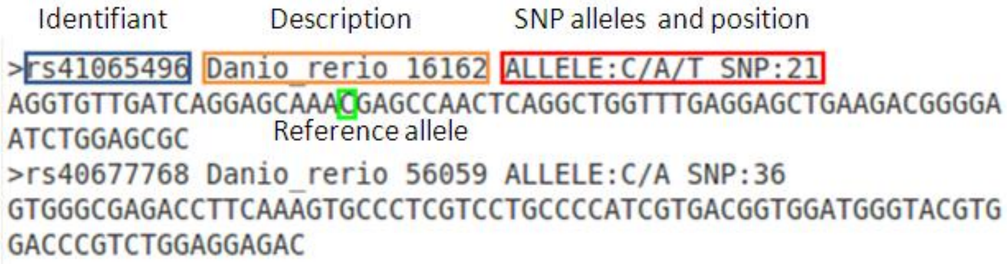
```
rs41065496 Danio_rerio	16162	AGGTGTTGATCAGGAGCAAA	C/A/T	GAGCCAACTCAGGCTGGTTTGAGGAGCTGAAGACGGGGAATCTGGAGCGC
rs40677768 Danio_rerio	56059	GTGGGCGAGACCTTCAAAGTGCCCTCGTCCTGCCC	C/A	ATCGTGACGGTGGATGGGTACGTGGACCCGTCTGGAGGAGAC
```

Figure 3 Input file in tabular format.



```
>rs41065496 Danio_rerio 16162
AGGTGTTGATCAGGAGCAAA[C/A/T]GAGCCAACTCAGGCTGGTTTGAGGAGCTGAAG
ACGGGGAATCTGGAGCGC
>rs40677768 Danio_rerio 56059
GTGGGCGAGACCTTCAAAGTGCCCTCGTCCTGCCC[C/A]ATCGTGACGGTGGATGGGT
ACGTGGACCCGTCTGGAGGAGAC
```

Figure 4 Input file in fasta format.



```
>rs41065496 Danio_rerio 16162 ALLELE:C/A/T SNP:21
AGGTGTTGATCAGGAGCAAACGAGCCAACTCAGGCTGGTTTGAGGAGCTGAAGACGGGGA
ATCTGGAGCGC
>rs40677768 Danio_rerio 56059 ALLELE:C/A SNP:36
GTGGGCGAGACCTTCAAAGTGCCCTCGTCCTGCCCCATCGTGACGGTGGATGGGTACGTG
GACCCGTCTGGAGGAGAC
```

Figure 5 Input file in fasta format usable by blast.

For the first two formats, the blanks and dashes are deleted when the program treats the sequence.

The 'Position' tag represents the SNP position on the nucleic sequence where it was identified. tSNPannot does not use this information. This value is not mandatory (Example : rs41065496 Danio_rerio AGGT...).

The 'Border seq' tag corresponds to the 5' and 3' sequence of the SNP. These sequences can be of any length. However, if one is shorter than the conservsize option specifies, the SNP will never be annotated (See the conservsize option).

For the 'SNP' each allele is separated by a slash and the first allele is the most frequent. In fasta format the SNP is separated from the rest of the sequence by [ ].

The 'SNP alleles and position' tag must be the last information on the headline. In this tag only one space exists, between ALLELE:X/Y and SNP:POSITION.

Note : tSNPannot uses the fasta format usable by blast internally. You can use the script snp2BlastSnp to convert the first two formats in the fasta format usable by blast. (Example : snpFile2BlastFile -i myFile)

Understanding the output files

Two types of output files can be generated: plain text and XML.

Core information

Each sequence in the polymorphism file produces one entry in the output file. This sequence is identified by the 'query id' tag on XML file or the 'SNP sequence id' header in TXT file.

XML format :

```
<query id="rs41065496">
   <desc>
      <text>16162 </text>
      <length>152</length>
      <snp_alleles>C/A/T</snp_alleles>
      <snp_position>52</snp_position>
   </desc>
   <hsp id="ENSDARP00000018993_2">
   <subject>
      <species>Danio rerio</species>
      <protein>ENSDARP00000018993</protein>
      <gene_name>f7</gene_name>
      <gene_desc>coagulation factor VII [Source:ZFIN;Acc:ZDB-GENE-010814-
1]</gene_desc>
   </subject>
      <alignment>
         <properties>
            <frac_identical>1</frac_identical>
            <frac_conserved>1</frac_conserved>
            <evalue>2e-15</evalue>
            <bits>80.1</bits>
         </properties>
      </alignment>
      <annotation status="valid">
         <snp_genome_location>
            <region>1</region>
            <strand>1</strand>
            <position>16162</position>
         </snp_genome_location>
         <distance_exon_limits>
            <five_prime_limit>47</five_prime_limit>
            <three_prime_limit>116</three_prime_limit>
         </distance_exon_limits>
         <ref_related_snps>
            <related_snp id="rs41065496">C/A</related_snp>
         </ref_related_snps>
         <variation>
            <on_snp_seq>
               <allele desc="C/A" aa="Arg/Arg">SYNONYMOUS_CODING</allele>
               <allele desc="C/T" aa="Arg/*">STOP_GAINED</allele>
            </on_snp_seq>
         </variation>
      </annotation>
   </hsp>
</query>
```

TXT format :

```
###########################################################
# SNP sequence id : rs41065496
##########################################################
Description : 16162  ALLELE:C/A/T SNP:52 LENGTH:152
 ---> Subject protein id ENSDARP00000018993
        Subject description
                Species : Danio rerio
                Gene name : f7
                Details : coagulation factor VII [Source:ZFIN;Acc:ZDB-GENE-010814-1]
        HSP properties
                Conserved fraction : 1
                Identical fraction : 1
                Bit score : 80.1
               E-value : 2e-15
        SNP position on reference genome
                Region : 1 Strand : 1 Nucleic Acid number : 16162
        Distance to the limit of the exon
                5' limit : 47
                3' limit : 116
        SNP related to the reference genome
                rs41065496 C/A
        SNP consequences on translated query sequence
                C/A(Arg/Arg) => SYNONYMOUS_CODING
                C/T(Arg/*) => STOP_GAINED
//
```

Each HSP between your sequence and the reference species's proteins provides the following information.

'<subject>' or 'Subject description'

Gives the gene information about the protein implicated in the HSP.

'<alignment><properties>' or 'HSP properties'

Provides the information about the HSP quality.

'<snp_genome_location> ' or 'SNP position on reference genome'

This tag gives the SNP position on the reference genome. This position is a projection deduced from the alignment.

'<distance_exon_limits>' or 'Distance to the limit of the exon'

This field indicates the distance between SNP and the limits of its exon. The 5' and 3' orientation is determinated in relation to your initial sequence.

'<ref_related_snps>' or 'SNP related to the reference genome'

This field indicates if there exists a known SNP on the reference genome which corresponds to the input SNP in alignment. If this is not the case, or if Ensembl does not store the variation data about this species, this field contains 'There is no reference SNP'.

The alleles are indicated for the strand upon which your SNP is projected.

Note : The list of species for which Ensembl stores variation data http://www.ensembl.org/info/docs/variation/index.html.

Note : Several SNPs can exist on one position.

'<variation><on_snp_seq>' or 'SNP consequences on translated query sequence'

For each allele of the SNP this field gives the consequences on your sequence :

- SYNONYMOUS_CODING : the amino acid coded by the reference allele is the same that the amino acid coded by the new allele. This tag is also used when the amino acid coded is STOP for each allele.
- NON_SYNONYMOUS_CODING : the two amino acids are different.
- STOP_LOST : the reference corresponds to a STOP and the other allele corresponds to an amino acid.
- STOP_GAINED : the reference corresponds to an amino acid and the other allele corresponds to a STOP.
- UNCHARACTERIZED : the SNP's codon contains an uncharacterized nucleic acids (example : NT[G/C]). In this case, instead of the aa's name the output file contains the codon.

If there is no HSP with sufficient similarity between your sequence and the reference species proteins, the query entry has the following form.

XML format :

```
<query id="rs41065496">
   <desc>
      <text>16162 </text>
      <length>152</length>
      <snp_alleles>C/A/T</snp_alleles>
      <snp_position>52</snp_position>
   </desc>
</query>
```

TXT format :

```
#######################################################
# SNP sequence id : rs41065496
######################################################
Description : 16162  ALLELE:C/A/T SNP:52 LENGTH:152//
```

Warnings

Two differents warnings can occur in the output file.

XML format :

```
<on_snp_seq>
   <warning>
      <stop_before>There is a stop codon before the SNP</stop_before>
      <frameshift_before>This consequences are determined without taking into account
frameshift</frameshift_before>
   </warning>
   <allele desc="A/T" aa="Gly/Gly">SYNONYMOUS_CODING</allele>
</on_snp_seq>
```

TXT format :

```
SNP consequences on translated query sequence
        WARNING : This consequences are determined without taking into account frameshift
        WARNING : There is a stop codon before the SNP
        A/T(Gly/Gly) => SYNONYMOUS_CODING
```

The warnings indicate that a stop or a frameshift was detected before the SNP on the query alignment. tSNPannot only uses the SNP and the other nucleic acids of the codon featuring it to compute the consequences of the SNP.

Verbose mode

With the option -verbose you can obtain more information.

Information about HSP

The output file gives more details about the alignment (query and subject sequence, SNP position, …).

XML format :

```
<description>
   <query_start>1</query_start>
   <query_end>150</query_end>
   <query_strand>1</query_strand>
   <hit_start>20</hit_start>
   <hit_end>69</hit_end>
   <query_string>AVFVHRDEAHEVLIRSKRANSGWFEELKTGNLERECLEEKCSYEEARE
VF</query_string>
   <hit_string>AVFVHRDEAHEVLIRSKRANSGWFEELKTGNLERECLEEKCSYEEAREVF</
hit_string>
   <snp_location>18</snp_location>
</description>
```

TXT format :

```
HSP description
        Query  1
        AVFVHRDEAHEVLIRSKRANSGWFEELKTGNLERECLEEKCSYEEAREVF     150
        Sbjct   20
        AVFVHRDEAHEVLIRSKRANSGWFEELKTGNLERECLEEKCSYEEAREVF     69
SNP position on reference genome
        Region : 1 Strand : 1 Nucleic Acid number : 16162
SNP position on alignment
        Character number : 18
```

Note : on the query string the \ indicates an insertion and / a deletion.

Information about consequences of the SNP

The output file gives the consequences of the alleles of your SNP on the reference genome. Each allele is applied on genome and the program computes its consequences.

XML format :

```
<on_ref>
   <mutation desc="C/T">
      <consequence
transcript_id="ENSDART00000013209">STOP_GAINED</consequence>
      <consequence
transcript_id="ENSDART00000109479">DOWNSTREAM</consequence>
```

```
   </mutation>
   <mutation desc="C/A">
      <consequence
transcript_id="ENSDART00000013209">SYNONYMOUS_CODING</consequence>
      <consequence
transcript_id="ENSDART00000109479">DOWNSTREAM</consequence>
   </mutation>
</on_ref>
```

TXT format :

```
SNP consequences on reference transcripts
        Change in genome C/T
                Change in transcript ENSDART00000013209 : STOP_GAINED
                Change in transcript ENSDART00000109479 : DOWNSTREAM
        Change in genome C/A
                Change in transcript ENSDART00000013209 : SYNONYMOUS_CODING
                Change in transcript ENSDART00000109479 : DOWNSTREAM
```

Caution : This option significantly increases the computation time.

Note : If Ensembl does not store the variation data about the reference species, these informations can not be computed and consequences fields are empty in the output files.

Rejection Mode

A HSP can be rejected for the annotation because the SNP position corresponds to a gap or an insert on the reference, or because the border regions are not conserved. If however you want to view these HSPs, for example in the goal to refine the value of conservsize option, you could use the -rejection option.

XML format :

```
<hsp id="ENSDARP00000023791_3">
   <subject>
         <species>Danio rerio</species>
         <protein>ENSDARP00000023791</protein>
         <gene_name>f7i</gene_name>
         <gene_desc>coagulation factor VIIi [Source:ZFIN;Acc:ZDB-GENE-021206-
10]</gene_desc>
   </subject>
   <alignment>
      <properties>
         <frac_identical>0.68</frac_identical>
         <frac_conserved>0.88</frac_conserved>
```

```
          <evalue>1.1e-08</evalue>
          <bits>57.7</bits>
       </properties>
   </alignment>
   <annotation status="rejected">
       <reasons_for_rejection>The border regions of SNP are not
conserved.</reasons_for_rejection>
   </annotation>
</hsp>
```

TXT format :

```
 ---> Subject protein id ENSDARP00000023791
        Subject description
                 Species : Danio rerio
                 Gene name : f7i
                 Details : coagulation factor VIIi [Source:ZFIN;Acc:ZDB-GENE-021206-10]
        HSP properties
                 Conserved fraction : 0.88
                 Identical fraction : 0.68
                 Bit score : 57.8
                 E-value : 1.1e-08


        ----------------------------------------------------------------
        Rejected : The border regions of SNP are not conserved.
        -----------------------------------------------------------------
//
```

In the XML format, the 'status' attribute in annotation markup indicates if the HSP is 'rejected' or 'valid'.

A message gives the reason of rejection :

- The border regions of SNP are not conserved.
  or
- Indicated position corresponds to an insert.
  or
- Indicated position corresponds to a gap.

Usage tips

Choose your output format

For a human-readable output, use the plain text (txt) output. This format is more compact and most readable.

In contrast, the XML is designed for automatic processing. If you want to extract information with scripts or generate statistics, this format is more suitable.

Choose your options

Use the verbose option only if you need it. Use it in a second pass only on the sequences of interest.

The rejection option is useful to refine the length of the conservation area, but greatly increases the size of output files. Use it on a small sample of your data. However this option should never be used to produce final results.

The configuration file

The configuration file is located in the ‘config’ folder inside the application folder.

File structure

In this file, each line which describes a parameter has this structure :

For simple parameter : parameter = 'value'

or

for list : parameter @= 'value1' 'value2'

Comments start with a ‘#’ and stop at the end of line.

At the installation, the default configuration file contains the following elements.

```
############################################
#General
############################################
temporary_folder = '/tmp'
blast_programs @= 'blastall' 'blastx' 'blast2' 'fastx35'
default_align_prog = 'fastx35'
db_directory = '/tmp'
############################################
#Ensembl database connection
############################################
ensembl_db_user = 'anonymous'
ensembl_db_host = 'ensembldb.ensembl.org'
max_connection_attempt = '3'
time_between_attempt = '10' #time between new attempt in seconds
############################################
#Cluster parallel use
############################################
sge_param = '-q workq -sync y -l h_vmem=2G -l mem=2G'
nb_process = '20'
############################################
#Parallel log files
############################################
error_file = 'error'
stdout_file = 'out'
```

Parameters definition

blast_programs	Programs used to search for similar proteins. You can restrict them by removing some programs from the list. You can also change the version of fastx.
default_align_prog	The default program used to search for similar proteins.
db_directory	Folder where protein banks files will be stored. tSNPannot downloads and updates itself these banks.
ensembl_db_user	The user name for the connection to the databse.
ensembl_db_host	The database to connect to. If you have a local mirror of this database you can replace this value by the address of this mirror.
error_file	The name of the error file for the parallel execution.
max_connection_attempt	Number of connection attempt with the Ensembl database. If after this number of attempts, the connection still does not the program stops.
nb_process	In parallel execution the maximum number of processes is limited by this parameter.
sge_param	Parameters for the qarray command
stdout_file	The name of the file where the standard output of each process is written for the parallel execution.
temporary_folder	Folder where the temporary files and folders are created during the execution.
time_between_attempt	Number of seconds to wait after each failed attempt to connect to the Ensembl database. (see max_connection_attempt)