NovelSNPer - Manual

3. November 2011

Synopsis

NovelSNPer.pl [options] <*inputfile*>
NovelSNPer.pl -download [options]
statistics.pl <*inputfile*>

For example, use
NovelSNPer.pl -species human ExampleInput.txt

Installation

1. Install Perl.
2. Install Perl-modules by typing `ppm` and searching for the two modules DBI and DBD::mysql.
3. Install Bioperl 1.2.3
4. Install the APIs Ensembl-core, Ensembl-compara and Ensembl-variation
5. Download NovelSNPer.zip and extract it to an arbitrary directory.
6. If you use Linux, make the Perl-file executable, by typing the two commands in the shell:
   a) `dos2unix NovelSNPer.pl`
   b) `chmod 777 NovelSNPer.pl`
7. If the PERL5LIB environment is not set up, edit the file NovelSNPer.ini and enter the path of bioperl and the ensembl-modules.

Tutorial

Using the ensembl database

Write an input file in excel with the following columns:

- 1st column: Name of the polymporphism
- 2nd column: Chromosom number
- 3rd column: Start of the polymporphism
- 4th column: End of the polymporphism
- 5th column: Alternative allele
- 6th column: Reference allele

Attention: If there is an insertion between to basepairs the starting position is the higher number and the end position is the lower number. For example: There is an insertion between 24bp & 25bp. Then the start position is 25 and the end position is 24.

The table must be saved as textfile with tab separated values (.tsv format). For further understanding there exists the example file `ExampleInput.txt`.

To evaluate the file open the command shell an enter:
NovelSNPer.pl -species human ExampleInput.txt

The word behind `-species` declares the species to which the polymorphisms belong. There can be entered english names (e.g. human, mouse, bovine) or latin names (e.g. homo_sapiens, mus_musculus, bos_taurus).
Because of using the ensembl database you can only look for species which are in the ensembl database.

Description

NovelSNPer is a tool for identification and characterization of novel SNPs. It tells you for each SNP in two lists which were generated from next generation sequencing data

- whether the SNP ist novel or already known
- the rs-id if the SNP is known
- the functional class
- the codon and its amino acids, if it is in a coding region

- the next gene and the distance to it.

NovelSNPer connects to the Ensembl database and gets all genes and their transcripts in the region around the SNPs. Then it take the position of all exons and the start and end of each transcript of each gene. With this data, NovelSNPer determines whether the SNP is in a coding region or not and its codon, if it is in a coding region. To get the correct data, it is necessary to use the current version of Ensembl as reference-sequence for your assembly-program.

Also you get some statistic information about the output-file.

Input

Input arguments

The user can pass several arguments to NovelSNPer via command line. These arguments are:

-out [...] With this command you can specify the outputfile. (Default is NovelSNPer_)

-species [...] With this command you specify the species of your assembly. (Default is homo_sapiens)

-start [...] Starting position of the reference sequence in the chromosome. All positions in the inputfile must be written relative to the sequence. (If the positions in the inputfile are relative to the chromosome, use '-start 1'.)

-cs [...] If this argument is TRUE, the conservation score is calculated (very slowly). (Default is FALSE)

-domain [...] If this argument is TRUE, the Protein Domain Region is calculated. (Default is FALSE)

<**inputfile**> The last argument has to be the inputfile, where are written the SNPs.

Input file

The inputfile is a list of SNPs, which is saved as a tabular-separated-file (.tsv or .txt). The columns of the inputfile have to be in the following order:

1. **Name of SNP:** This can be an arbitrary alphanumeric word.

2. **Chromosome:** This column contains the number of the chromosome (a number or X or Y).

   All SNPs in one file must be from the same chromosome if the reference sequence is not identical with a chromosome sequence. If the reference sequence are chromosome sequences, the SNPs in the inputfile can be from different chromosomes.

3. **Position:** Position of the SNP, relative to the reference sequence.

4. **Alleles:** All alleles for this SNP, separated by "/". (e.g. A/G/T) Beware, to take the alleles from the main string of the chromosome. (Which is not necessarily the main string of the gene.)

5. **Reference Nucleotide:** The nucleotide of the reference sequence at this position at the main string.

NovelSNPer will detect automatically, if an input file has a header line or not.

Output

The output files contain a table in tabular-separated-file format. The columns of this file are:

1. **Name of SNP:** Same as in input file.
2. **Chromosome:** Same as in input file.
3. **relative Position:** Same as in input file.
4. **chromosomal Position:** Position of the SNP, relative to the chromosome.
5. **Alleles:** Same as in input file.
6. **Reference Nucleotide:** Same as in input file.
7. **Codon:** All possible codons for all known transcripts (separated by "/") if the SNP is in a coding region and NA for SNPs in a non-coding region.
8. **AminoAcid:** Amino acids for the given codon if the SNP is in a coding region and NA for SNPs in a non-coding region.
9. **Reference Amino Acid:** The amino acid for the reference sequence.
10. **Functional class:** All variation types for all known transcripts.
11. **Nearest gene:** The next gene.
12. **Distance:** The distance between the SNP and the nearest gene.
13. **Strand:** Strand of the gene, the SNP is associated with. (If -1, the codon of the SNP is encoded at the reverse string.)
14. **Status:** NOVEL, KNOWN or ALTERNATIVE: SNPs that are in the Ensembl database (same position and same alleles) are attributed the status KNOWN. If there is an SNP in the Ensembl database, which is as the same position, but has different alleles, our SNP is attributed as ALTERNATIVE. Otherwise they have the status NOVEL.
15. **rsID:** If the SNP has the status KNOWN or ALTERNATIVE, the rsID is shown. Otherwise this attribut is NA.
16. **Protein Domain Region:** Optional! Some protein-sequences hava known effects or are translated by different genes. If the SNP is in a coding region, this column shows the belonging Protein Domain Region.