ReMo-SNPs program version 1.0 30/09/2014

GENERAL USAGE NOTES

ReMo-SNPs is a computational program developed to aid researchers in the process of selecting functional SNPs in the human genome for association analyses in user-specified regions and/or motifs genome-wide. An additional useful feature of the ReMo-SNPs program is the automatic selection of genotyped markers in the user-provided material. This enables the researcher to directly use the ReMo-SNPs output data in a following association study.

The user defines which human genomic regions and/or motifs the program should search through. In our example we have used experimentally validated transcription factor binding regions and *in silico* identified binding motifs. However, the flexibility of the ReMo-SNPs program makes it easy to adapt to different projects and research questions without limitation to transcription factors.

The ReMo-SNPs program can be run in three different modes: long, medium, and short. When using ReMo-SNPs to select SNPs for an association study, the user will choose the long run, which generates genotyped, genome-wide data on markers in the user-specified motifs and/or regions. In addition, a list with interesting markers, for which the program was unable to find genotype data, is also provided.

The medium and short runs generate descriptive statistics of the SNPs located in the regions and/or motifs of interest. The medium run provides descriptive statistics on SNPs located in motifs. The short option provides information on which SNPs are located in motifs and regions of interest respectively. For all three options the program provides information about how many times a SNP is found in each position of the motif, and how many motifs contain one, two, three, or more SNPs.

PROGRAM INSTALLATION

The ReMo-SNPs Perl script is freely available at:

Web site: http://www.neuro.ki.se/ReMo-SNPs/

E-mail: sbuerven@gmail.com and graae.lisette@gmail.com

Using ReMo-SNPs requires the download of several publicly available data files. Each file, or category of files, should be saved in a separate folder on the local hard drive.

If the user does not yet have the *Perl interpreter* installed, it can be downloaded at: http://www.perl.org/get.html

A relevant *HapMap file* can be downloaded at: http://hapmart.hapmap.org/BioMart/martview. Our example uses genomic Build 36. The resulting text file has three columns: chromosome, position, and marker ID.

FASTA files; ReMo-SNPs requires one FASTA file for each chromosome of interest. The IUPAC-masked files, which provide information regarding the position of SNPs, can be downloaded from the genome browser at: http://genome.ucsc.edu/. In our example we used SNP129-FASTA, hg18 build 36.1, March 2006. It is absolutely crucial that the FASTA-files and the HapMap-file are based on the same genomic build.

LD files, containing pairwise linkage disequilibrium data, can be downloaded from http://hapmap.org/ bulk data download / LD data. These files are compressed (.gz) and should not be unpacked for the ReMo-SNPs analysis.

COMMAND LINE OPTIONS

On the command line the user specifies the required information for each type of run. For a long run all information, a) - n), should be provided. For medium and short runs, the user should specify the information stated in a) – h) below.

a) perl -w ReMo.SNPs.pl

b) -- HapMap [path and name of the HapMap-data file]

c) -- Motifs [path and name of the motif-file]

d) -- FASTAdir [path and name of the folder with FASTA-files]

e) -- bed [path and name of the BED file containing the region data]

f) -- regionScore [value for the score-threshold; this command is optional]

g) -- combo [AND, OR, or SCORE, for type of combination of regions and motifs]

h) -- typeOfRun [long, medium or short. The default value is 'long' for a full run.]

i) -- map [path and name of the MAP file]

j) -- LDdir [path and name of the folder with the .gz LD files; do not unzip these files for analysis]

k) -- r^2 [between 0.0 and 1.0; threshold for inclusion of proxy markers]

l) -- log [file name for the log-file (the default name is ReMo.SNPs.log)]

m) -- out [file name for the out-file (the default name is ReMo.SNPs.out)]

n) >name.of.screenoutput.file.txt, optional command to re-direct the script's output if the user wants to save the information written in the terminal window

EXAMPLE FILES

The command line for the example data and the information in the output files are described below. The generated example output files for each type of run (long, medium, and short) are found in the Output folder.

Command line:

LONG:
$ perl -w ReMo.SNPs.pl --HapMap ../HapMap_file/mart_export.txt --Motifs ../motif_file/GR.motif.txt --FASTAdir ../FASTA_files/ --bed ../bed_file/GR.BED --combo AND --typeOfRun long --map ../map_file/BP.data.map --LDdir ../LD_files/ --r2 0.89 --log ReMo.SNPs.long.log >screenoutput.ReMo.SNPs.long.txt

MEDIUM:
$ perl -w ReMo.SNPs.pl --HapMap ../HapMap_file/mart_export.txt --Motifs ../motif_file/GR.motif.txt --FASTAdir ../FASTA_files/ --bed ../bed_file/GR. BED --combo AND --typeOfRun medium --log ReMo.SNPs.medium.log >screenoutput.ReMo.SNPs.medium.txt

SHORT:
$ perl -w ReMo.SNPs.pl --HapMap ../HapMap_file/mart_export.txt --Motifs ../motif_file/GR.motif.txt --FASTAdir ../FASTA_files/ --bed ../bed_file/GR.BED --combo AND --typeOfRun short --log ReMo.SNPs.short.log >screenoutput.ReMo.SNPs.short.txt

Output files:

ReMo.SNPs.out: This file is created in Step 7. It shows all interesting genotyped markers from the candidate list in Step 4 and the genotyped LD-markers from Step 6.
lddata.txt-file: This file is created in Step 6 when the script searches for proxy markers for those markers that have not been genotyped. It contains 11 columns with the following information: chromosomal position of marker 1, chromosomal position of marker 2, population code, rs-number for marker 1, rs-number for marker 2, D-prime, R-square, LOD, fbin, rs-number of the candidate SNP, and chromsome.
genotyped.lddata.txt-file: This file is also created in Step 6, when the program identifies SNPs from the lddata.txt-file that have been genotyped in the material. It contains the same columns as the lddata.txt-file.
list.of.markers.with.no.genotype.and.no.proxy.out: This file shows the interesting SNPs from the candidate list that should be analyzed based on their location but have not been genotyped and have no good LD-SNP.
motifsnplist.txt and regionsnplist.txt: these files are created in Step 4 if one has chosen the short run. They show all the SNPs found to be located in the motif of interest genome-wide and the specified regions of interest, respectively.

Log file example output:
This is ReMo.SNPs.pl
Analysis started with the following arguments:
(In Step 1:)
Currently working on Chromosome A
Sequence length, and line counter
The sequence length shows how many letters the FASTA-file contains and the line counter corresponds to the number of rows the FASTA-file had before the program made one row of it.

After giving this information for all chromosomes, the script provides information for each chromosome on how many SNPs are found to be located in the motif of interest.
(In Step 2:)
A list of all the motif-snps is printed, and at the end a summary of how many markers have been found in the motif of interest genome-wide.
(In Step 3:)
Information on how many markers the program found in genomic regions of interest is given.
(In Step 4:)
The number of total candidate SNPs is given.
(In Step 5:)
The number of markers that have been read from the MAP file is printed.
(In Step 6:)
The number of interesting SNPs with genotypes is given.

Terminal window example output:
Step 1 …
My motif is X character long
The original motif is ABC....
The reverse complement is ABC …
The IUPAC-motif is: [ABC...][ABC...]...
The reverse IUPAC-complement is: [ABC...][ABC...]...
Problem SNP found: chr / bp / motif-length
Position 1 had A mutations
Position 2 had B mutations
….
There were C motifs with 1 SNP(s)
There were D motifs with 2 SNP(s)
...
Step 2 …
No rs-number was found for the following sequence: XXX at position YYY on chromosome ZZ
Step 3 …
Step 4 …
If a medium-run is chosen the following will be printed in this step:
HapMap: position in bp and rs-number Problem: position in bp and motif-length
The following SNPs may be problematic because they are located in motifs with more than one SNP:
Step 5 …
Step 6 …
Step 7 …