MQscore_SNP 1.1

Software for multipoint parametric linkage analysis of quantitative traits using SNPs in large pedigrees without loops

DESCRIPTION

There are many program packages, which use Elston-Stewart algorithm to compute exact likelihood function (Linkage, FastLink, Vitesse, Mendel's module for model-based analysis). In principle, this algorithm allows calculation of likelihood of arbitrary large pedigrees. However underflow problems can occur when pedigree size is large. We used special algorithm of pedigree peeling which solves the underflow problem (Axenovich TI, Aulchenko YS. "Solution for underflow problem in linkage and segregation analysis" Comput Biol Chem 2006; 30(5): 382-385).

Package MQscore_SNP is designed for model based parametric linkage analysis of quantitative trait and SNP markers. The MQscore_SNP can calculate likelihood of arbitrary large pedigrees without loops based on the mixed major gene and polygene model of trait inheritance. We assumed diallelic autosomal QTL placed either near marker locus (programs MQS_2point and MQS_2point_max), or between the adjacent marker loci (program MQS_3point). To calculate the likelihood of an extended pedigrees under mixed model we used hypergeometric approximation (Lange, 1997) and finite polygene model (Fernando *et al.*, 1994; Stricker *et al.*, 1995).

Different variants of logarithm of odds (lod) score can be calculate with the help of MQscore_SNP:

MQS_2point program assumes that all model parameters are fixed; the lod score compares null hypothesis about independent segregation of SNP and QTL and alternative hypothesis about close linkage between QTL and each marker locus;

MQS_2point_max estimates all model parameters under null and alternative hypothesis, lod score compares null hypothesis about independent segregation of SNP and QTL and alternative hypothesis about close linkage between QTL and each marker locus;

MQS_3point program assumes fixed values of all model parameters, lod score compares null hypothesis about independent segregation of two adjacent SNPs and QTL and alternative hypothesis about QTL located between two SNP loci.

PACKAGE STRUCTURE

All executable files are in subdirectory BIN.

Preparing of pedigree and phenotype data: peeling_2006_4.pl and recode_ped_2006.pl

Parametric linkage analysis: MQS_2point.exe or MQS_2point_max.exe or MQS_3point.exe

Detail descriptions of peeling_2006_4.pl and recode_ped_2006.pl are available in PedPeel and recode_ped directories by address http://mga.bionet.nsc.ru/soft/.

Source codes for MQS_2point, MQS_2point_max and MQS_3point are available in SOURCE directory. Programs may be recompiled to run under Windows or Linux.

INPUT FILES:

1) Initial data files contain information about a) pedigrees and phenotypes, b) marker genotypes and c) genetic map (necessary only for MQS_3point program). Names of all data files are arbitrary.

Input_pedigree_file is a text file of standard LINKAGE format. Separator may be defined as comma or tab or space.

The file must contain columns with pedigree ID (ped), personal ID (id), father's ID (fa), mother's ID (mo), and sex (sex) defined for every person. There may be more columns, for example, for several traits. The file must have a header line, which should contain abovementioned keywords. One pedigree may be entered after another, each pedigree with its own pedigree ID. The personal IDs are coded as integers from 1 to total number of persons in the sample and must be continuous across all pedigrees. The parents come before offspring. The pedigree ancestors who have no parents in analysed sample should be described prior to the offspring. Gender should be coded as '1' for male and '2' for female.

This is a typical output from our pedigree verification and recoding program, recode_ped_2006.pl. (see recode_ped directory in http://mga.bionet.nsc.ru/soft).

Input_genotype_file is a text file. Separator may be defined as comma or tab or space. Number of marker loci in analysis should be specified on the first line. Every next line contains all genotypes for every pedigree member (including ungenotyped persons) (without the personal ID, so the line starts with the genotype of the corresponding person in the pedigree file). Number of row in the file equals to number of pedigree member + 1. Definition of genotypes: 0 – unmeasured genotype, 1 – MM, 2 – Mm and 3 – mm. Symbols M or m are assigned to allele arbitrary. So, frequency of M allele may be more or less than 0.5.

Input_map_file is a text file. The file has a header line, which contains names of columns. Every following line contains information about markers: ordered number, marker

name, physical location (bp), genetic position (cM). The order of columns is important. Physical location is not used in the analysis. If this information is absent, the second column of *input_map_file* should be filled by any integer values, for example, 0.

Maps for every chromosome are presented in separate map files.

2) Command file *par.inp* (name is fixed) contains information about type of analysis, model of the trait inheritance, type of parameter's estimation to be undertaken (example of *par.inp* is below).

3) Command file *task.inp* (name is fixed) contains the name of analyzed trait (corresponding to that specified by pedigree file), *input_genotype_file* and *output_result _file*. Additionally for MQS_3point program it contains name of *Input_map_file* and for MQS_2point_max it contains name *output_model_file.*

OUTPUT FILES:

The names of output files are defined in *task.inp* by user.

Maximum likelihoods for null and alternative hypothesis and Lod Score are shown in *output_result_file* (its name is specified by the third line of *task.inp*). Estimated parameters (MQS_2point_max) are shown in *output_model_file* (its name is specified by the fourth line of *task.inp*).

PLATFORM AND LANGUAGE

Perl (http://www.perl.org/)

Fortran77, we used the GNU Compiler Collection (GCC): http://www.mingw.org/ (Windows) and http://gcc.gnu.org/ (Linux)

ANALYSIS STEP BY STEP

STEP 1:

Copy MQS_2point.exe or MQS_2point_max.exe (and maxpar.inp, if MQS_2point_max is used) or MQS_3point.exe and peeling_2006_4.pl files from BIN directory to directory defined by user and containing *initial_pedigree_file*, *initial_marker_file* (and *initial map_file*, if MQS_3point is used).

STEP 2:

PREPARE THE DATA:

Edit the text file named *task.inp*. Include name of analyzed trait in this file (example *trait_name*).

COMMAND LINE: perl peeling_2006_4.pl *input_pedigree_file*

OUTPUT FILES: *iway.dat*, *trait_name*

If user plans to analyze several traits, the files with each additional trait may be made by simple consecutive writing all values in one line. Separator may be defined as comma or tab or space.

WARRNING: missing data is coded by -999.

STEP 3:

FORM *par.inp*:

In general, model is defined via ordered set of 7 parameters

1. q = Pr(A) – frequency of A allele
2. mu1 = Pr(x|AA) – mean for genotype AA
3. mu2 = Pr(x|AB) – mean for genotype AB
4. mu3 = Pr(x|BB) – mean for genotype BB
5. theta – recombination fraction
6. sigma – environmental component of variance
7. pga – effect of single polygene allele

The values of the parameters and type of analysis are passed to the program using *par.inp* file. The first line contains six names: num ICP PR ST GL GR

The structure of each of seven next lines (one for q, mu1, mu2, mu3, theta, sigma, pga in this order) is follows:

- numeric parameter ID,
- ICP (ICP=1 if parameter is estimated; else, if fixed, ICP=0),
- (starting/fixed) parameter value,
- initial step of maximization,
- lower limit of parameter,
- upper limit of parameter.

Example of *par.inp:*

num	ICP	PR	ST	GL	GR	
1	0	0.40000E+00	0.10000E+00	0.00000E+00	1.00000E+00	q
2	0	0.80000E+00	0.10000E+00	-1.00000E+01	2.50000E+01	mu1
3	0	0.00000E+00	0.10000E+00	-1.00000E+01	2.50000E+01	mu2
4	0	-0.70000E+00	0.10000E+00	-1.00000E+01	2.50000E+01	mu3
5	0	0.001	0.10000E-02	0.00000E+00	0.50000E+00	theta
6	0	1.20000E+00	0.10000E+00	0.00000E+00	1.00000E+02	sigma
7	0	1.10000E+00	0.10000E+00	-1.00000E+01	1.50000E+01	pga

					|_ upper limit of parameter
				|_ lower limit of parameter	
			|_ initial step of maximization		
		|_ initial (or fixed) value of parameter			
	|_ indicator of parameter estimation				
_ number of parameter						

All parameter values are fixed under linkage analysis performed with MQS_2point and MQS_3point programs. For MQS_2point_max, where all parameters but theta are estimated, ICPs for all lines except line 5 must be marked by 1.

STEP4:

Add lines to the file *task.inp*. The first line of this file had been formed at the STEP 2. It includes the name of analyzed trait. Add the name of *input_genotype_file* as the second line of the file; the name of *output_result _file* as the third line of the file and name of *input_map_file* (for MQS_3point) or *output_ model_ file* (forMQS_2point_max) as the fourth line of the file.

COMMAND: MQS_2point.exe or MQS_2point_max.exe or MQS_3point.exe. The program will check the file names defined in *task.inp* and perform analysis.

RESOURCES: sizes of arrays is defined in *mas_size.inc*.

DIAGNOSTICS: if arrays sizes should be grater then declared size, ERROR is indicated and program is stopped. It is recommended to increase the array size in *mas_size.inc* and to recompile the program.

SEE ALSO:

COPYRIGHT:

GNU General Public License as published by the Free Software Foundation, either version 2 of the License, or (at your opinion) any later version. This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.

BUGS:

AUTHOR: Tatiana Axenovich (aks@bionet.nsc.ru)

VERSION: 3.0 2008

CREATED: 10.05.2008

LAST MODIFIED: 02.11.2009

LANGUAGE: Fortran77

OS: WINDOWS, LINUX