User`s Guide for mPopTag

Program: mPopTag (Multi-population tag SNP Picker)
Version:1.0
Sept. 2006
by Zongli Xu, Norman L. Kaplan, Jack A. Taylor
National Institute of Environmental Health Sciences (NIEHS)

Contents

1. Legal information

mPopTag License Agreement:

mPopTag, including its source code and documentation is freely distributed under the following license terms. Installation of the program on any computer or any use of the program implies that the user and the user's organization agree to the following terms:

This software is provided on an "as is" basis, with no warranty of any type, including, but not limited to, warranty of suitability for any particular purpose or ability to function correctly on any type of computer.

You may redistribute mPopTag. However, the entire package, including documentation, software, this license, and source code, must be preserved.

You may modify mPopTag, and distribute your results, but you must (a) preserve all copyright notices, license agreements and credits in software and documentation, (b) add your own notice which makes it clear immediately that it is a modified version, (c) also distribute the unmodified version along with your modified version, (d) distribute the modified version under this licensing agreement, and (e) notify the copyright holders of mPopTag that you are distributing a modified version, and supply us a full copy of source code.

2. Introduction

The Multi-Population Tag SNP Picker (mPopTag) is a tool to select or evaluate linkage disequilibrium (LD) tag SNPs for multiple populations. The program analyzes patterns of linkage disequilibrium (LD) (measured by composite linkage disequilibrium (CLD) or r^2) between polymorphic sites in a genome region for multiple populations, and uses an agreedy algorithm to select a single set of near-minimum number of LD tag SNPs for multiple populations.

3. How it runs

On the command line, it runs by typing:

./mpoptag

4. Parameters

All parameters required by mPopTag can be set in a self-explanatory parameter file *params*. The name of the parameter file should not be changed. In the file, sentences beginning with the sign "#" are annotated lines of the immediately following parameters. For each parameter line, words or phases before the sign ":" are the key words for that parameter. No space is allowed before key words. These key words should not be changed . User specifyd parameter values should be put after the colon sign.

Example of a parameter file:

```
#Number of populations and the file names for each population data
n_pop: 3, pop1, pop2, pop3
#diretory for input file
input_dir: sample_input
#directory for output
output_dir: sample_output
#LD method 1:CLD, 2:r^2
LD_method: 1
#LD threshold
cutoff_LD: 0.8
#cutoff value of minor allele frequence for common SNP
maf: 0.05
#Minimum number of SNPs tagged by a tag SNPs, must great or equal to 1
minimum: 2
#pre-included tag SNPs list; 0: no, 1: yes; provide file name if yes
include_snp: 1, include.csv
#SNP list excluded from tag SNPs; 0: no, 1: yes; provide file name if yes
exclude_snp: 1, exclude.csv
#SNP design score; 0: no, 1: yes; provide file name if yes
score: 1, score.csv
#evaluation of a list of tag SNPs in included file? 0:no  1:yes
evaluation: 0
```

The above parameter file already includes simple explanation for each parameter. Below is a more detailed explanation.

n_pop
The first parameter value is the number of populations that a user has specified, followed by file names of SNP genotype data for each population. The number of file names must equal to the number of populations. If the number of populations is 1, it is equivalent to selecting LD tag SNPs from a single population. All parameter values are delimited by a comma. The program assumes each file name is a population name, and therefore labels the output contents using these file names.

The genotype files follow a comma-delimited text format that contain the columns listed below:

gene_name,SNP_identifier ,SNP_genotype_list,MAF

For example:

gene_1,snp_1,-1,-1,0,…,1,1,9,1,0.13
gene_1,snp_2,1,0,0,…,-1,-1,9,-1,0.32
…
gene_2,snp_1,0,-1,1,…,1,1,-1,1,0.07
…
gene_n,snp_1,-1,-1,0,…,1,0,0,0,0.43

No "-" is allowed in any gene name. SNP genotype should be coded as -1, 0, 1 and 9 for homozygote common, heterozygote, homozygote rare and missing genotype, respectively.

input_dir
Name of the directory that contains all input files. This directory must be created before running the program, and all input files should be put in the directory.

output_dir
Name of the directory that the program will use for all output files. mPopTag will place all output files into the directory.

LD_method
Statistical method used to measure the LD relationship between SNPs. Parameter value 1: composite linkage disequilibrium; 2: r^2. If the number of valid genotype pairs between 2 SNPs (a valid genotype pair means one DNA sample without missing genotype for both SNPs) is less than 2, 0 will be assigned as the value of LD between the 2 SNPs.

maf
A cutoff value of minor allele frequency used to distinguish between common and rare SNPs.

cutoff_LD
A threshold value to claim LD between SNPs using a specified LD measure statistics.

minimum

Minimum number of SNPs required to be tagged by each tag SNP across multiple populations. It must great or equal to 1. The parameter value will substantially influence the number of tag SNPs. For example, if set a value of 2 to the parameter to exclude singleton tag SNPs (tag SNPs that only tag themselves), the number of tag SNPs will be reduced down to half for most genes while tagging proportions were not decreased much.

include_snp
Is there a list of required tag SNPs? 1=yes and 0=no. If yes, specify the name of a file that has a list of pre-included tag SNPs. Values of the two parameters are delimited with a comma.

The file follows a comma-delimited format that contains the columns listed below

gene_name, SNP_identifier

For example:

gene_1,SNP_3
gene_1,SNP_7
…
gene_n,SNP_k

If a user wants to select tag SNPs using LD information from both HapMap and resequencing data, the user can use the program by first selecting tag SNPs using HapMap data, and then list HapMap tag SNPs in the *include_snp* file, then continue on to select tag SNPs using resequencing data. This way the program will include all HapMap tag SNPs as tag SNPs, and select more tag SNPs to tag the rest of SNPs that can not be tagged by HapMap tag SNPs alone.

exclude_snp
Is there a list of SNPs needing to be excluded from tag SNPs? 1:yes; 0:no. If yes, specify the name of a file that has a list of the undesired SNPs. Values for the two parameters are delimited by a comma. The format of a *exclude_snp* file is the same with a *include_snp* file

score
Is there a SNP design score file? 1:yes; 0: no. If yes, specify the name of a file that has a list of SNP design scores. Values of the two parameters are delimited by a comma. A SNP design score can be any score that reflect the probability of successfully typing a SNP in a certain assay. The program preferentially selects SNPs with higher scores when there are multiple SNPs that can tag same number of SNPs.

The file follows a comma-delimited format that contains the columns listed below

gene_name, SNP_identifier,score

For example:

gene_1,SNP_3,0.387
gene_1,SNP_7,0.897
…
gene_n,SNP_k,0.656

Evaluation
Evaluation of a list of tag SNPs in included file and do not select more tag SNPs? 0:no 1:yes

5 Output

mPopTag will print the total number of tag SNPs and a list of tagging proportions for each population into stand screen output and also place all output files into the user's specified output directory. Genotype and LD figures are helpful to visually check and optimize tag SNPs. Below is the detailed explanation for each output file.

multipop_tags.txt
This is the main output file which has a list selected tag SNPs by the program and the list of SNPs that are tagged by tag SNP in each population. The file has format and columns below:

gene_name,number of tagged SNPs,tag SNP identifier <pop_1> tagged SNP list in population 1… <pop_n> tagged SNP list in population n

For example:

gene_1,8,205861 <pop_1> 205861 <pop_2> 205861,205190,205995 <pop_3> 205861,205995
…
gene_n,2,168697 <pop_1> <pop_2> 168697 <pop_3> 168697

untagged_snp.csv
This file list SNPs that are not tagged by tag SNPs as listed in file *multipop_tags.txt*. The file follws a comma-delimited format that contains the columns listed below.

gene_name,population_name,SNP_identifier

For example:

Gene_1,pop_1,02187
Gene_1,pop_1,03243
…
Gene_n,pop_k,93456

6. Installation

It can run on Linux operating system.

Questions can be addressed to Zongli Xu (xuz@niehs.nih.gov).

7. References

1. Zongli Xu, Norman L. Kaplan, Jack A. Taylor. LD tag SNPs selection for candidate gene association studies using HapMap and gene resequencing data. *European Journal of Human Genetics*, in press (2007).