Supplementary Information

TAGster: Efficient Selection of LD Tag SNP in Single or Multiple Populations

---Tutorial

Contents:

1. Tag SNPs selection using genotype data in HapMap format

- Download genotype data from HapMap website (www.hapmap.org) for each gene in each population, and store them in a user defined directory, for example "indir/hapmap". User can give the files any name, for example: "gene1_pop1", "gene2_pop2".
- In the same directory, create a file to specify the downloaded file list and related gene and population information. The file can be given any name, for example "list". This file contains a list of genes and associated file names with one row for each file in tab delimited format. For example:

pop1	gene1	gene1_pop1
pop2	gene1	gene1_pop2
pop3	gene1	gene1_pop3
pop1	gene2	gene2_pop1
pop2	gene2	gene2_pop2
pop3	gene2	gene2_pop3

- Edit parameter file *paraconv.txt*
  - Set the value for "-format" to 1;
  - Specific input directory for HapMap file, for example:
    -hapmap_dir: indir/hapmap
  - Specific the name of the file that has the genotype file list for example:
    -hapmap_list: list
- Run

  Convert

  to convert genotype file from HapMap format to genotype format required by TAGster. Record the output file names, these names need to be specified at parameter file *params.txt* later.

- Edit parameter file *params.txt*

- Specify the number of populations and the names for genotype files (in genotype format, such as the output file from genotype conversion utility) for each population data. For example

  -n_pop: 1, african
- Specify input file directory. For example:

  -input_dir: indir
- Specify output file directory. For example:

  -output_dir: outdir

- Run

  TAGster

  to select tag SNPs,

2. Tag SNPs selection using genotype data in Prettybase format

- Download genotype data in prettybase format from Seattle SNP website (http://pga.mbt.washington.edu/) for each gene, and store them in a user defined directory, for example "indir/prettybase".
- In the same directory, create a file to list the downloaded files used for tag SNP selection. The file can be given any name. For example "list". The content of the file looks like the following:

  gene1.prettybase.txt
  gene2.prettybase.txt
- If a gene was not resequenced in EGP Panel 2, user need to create another Individual population information file. For example file "pid.txt", This file has the following tab delimited format: "population individual_id". The program will use "population" specified in the file as the file names of genotype output files.

  Example: population information file

  african D001
  african D002
  …
  ceu E101
  ceu E102

- Edit parameter file *paraconv.txt*
  - Set the value for "-format" to 2;
  - Specific the file name for the population information file. For example:

    -ind_list: p2pid.txt
  - Specific input directory for prettybase files, for example:

    -pretty_dir: indir/prettybase
  - Specific file name which contain the list of input files. For example:

    -pretty_list: list
- Run

  convert

to convert genotype file from prettybase format to genotype format required by TAGster. Record the output file names, these names need to be specified at parameter file *params.txt* later.

- Edit parameter file *params.txt*
  - Specify the number of populations and the names for genotype files (in genotype format, such as the output file from genotype conversion utility) for each population data. For example

    -n_pop: 1, african
  - Specify input file directory. For example:

    -input_dir: indir
  - Specify output file directory. For example:

    -output_dir: outdir
- Run

  TAGster

  to select tag SNPs,

3. Tag SNPs selection using both HapMap and Seattle SNP data

To select tag SNP based on both HapMap and resequencing data, user can select tag SNPs using HapMap gentoype data first, then select more tag SNPs based on resequcing data to cover LD information in resequencing data. The following is the more detailed steps:

- Select tag SNPs using HapMap genotype file.
- Copy tag SNP output file "multipop_tags.txt" from output directory (for example: ourdir) to input directory (for example: indir) and rename it as "include.txt". It can be any names, but it needs to be specified in parameter file *params.txt* later.
- Download genotype file for the same sets of genes from Seattle SNPs website, store them in directory "indir/prettybase". Edit gene list file. Be sure to use the same gene names and the same population names for the same gene or the same population between HapMap related files and Seattle SNP related files.
- Edit parameter file *paraconv.txt* and the run *convert* to convert genotype from prettybase format to genotype format.
- Change SNP identifies from Seattle SNP id to rs number that were used in HapMap genotype file. It can be accomplished by doing the following.
  - Download file rsEGP.txt and Perl utility snprspos.pl, store them into the directory at where the converted genotype files (from prettybase format) are located.
  - For each converted genotype file, type the following command to replace Seattle SNP id to rs number.

    Perl snprspos.pl genotype_file_name

- Edit parameter file *params.txt* and specify the required tag SNP file. For example:

  –include 1, include.txt
- Run

to select tag SNPs,

4. Multiple population tag SNPs selection

In parameter file *params.txt*, specify multiple populations to the parameter *-n_pop*, for example:

-n_pop: 4, african,asian,ceu,hisp

5. Multiple SNP bin tag SNPs selection

In parameter file *params.txt*, specify the minimum number of SNPs tagged by a tag SNPs to a value of greater or equal 2, for example:

-minimum: 2

6. Create LD and genotype figures

In parameter file *params.txt*, specify the figure output parameter to 1, 2 or 3. For example

-figure: 3

To create both LD and genotype figures.