MLRtagging: Genotype Tagging
Based on SNP Multivariate Linear Regression

Jingwu He

Last updated: December 7, 2005

MLRtagging software package implements a novel genotype tagging method based on multivariate linear regression (MLR) analysis. This software can be used for tag selection and genotype prediction. The stepwise tag selection algorithm (MLRsta) selects positions of the given number of tags based on a genotype sample population. The MLR SNP prediction algorithm (MLRprediction) predicts a complete genotype based on the values of its tag SNPs, tag positions among all SNPs, and a sample of complete genotypes. The description of the algorithms and experimental results are in [1].

If you use MLRtagging Package, please cite [1].

Downloading and Installing

All relevant files including this pdf file are included in the tar files: available at http://alla.cs.gsu.edu/~software/tagging. Download this tar file to your machine then extract the files from the archive.

tar -xvf MLRtagging.tar

Currently, there is only Linux version available.

The package contains the following files:

1. *taggingReadme.pdf*: Readme file
2. *MLRsta*: Binary code for tag selection
3. *MLRprediction*: A Binary code for SNP prediction
4. *genoInput.txt*: Sample input of a genotype population sample: 129 offspring genotypes each with 103 SNPs from Daly et al. [2].
5. *tagGenoIndividual.txt*: Sample input of unknown genotype with typed tags
6. *tagFile.txt*: Sample input of tag positions

Running the Program

For running MLRsta, type

./MLRsta genoInput.txt 2 tagFile.txt

- First parameter = the file name of a genotype sample population
- Second parameter = desired number of tags k
- Third parameter = the name of output tag file (it contains selected k tag positions)

For running MLRprediction, type

./MLRprediction genoInput.txt tagFile.txt tagGenoIndividual.txt G fullIndividual.txt

- First parameter = the file name of a genotype sample population
- Second parameter = the name of input tag file (it contains selected k tag positions)
- Third parameter = the name of input file of a tag-restricted genotype with only tag SNP values
- Fourth parameter = the name of output file of a complete genotype

File Formats

genoInput.txt and tagGenoIndividual.txt contain the following lines:

- The number of genotypes
- The number N of SNPs in each genotype
- Description of data (can be empty)
- The first genotype represented by a sequence of 0/1/2's without gaps, 0 stands for homozygous major allele, 1 stands for homozygous minor allele, and 2 stands for heterozygous SNP.
- …
- The last genotype

tagFile.txt consists of k+3 lines:

- The number of tags
- Description of data (can be empty)
- Description of data (can be empty)
- The position of the first tag (a number in the range from 0 to N-1, where N is the number of SNPs.)
- …
- The last tag

Support

Please, direct all questions and comments to Jingwu He at the Department of Computer Science, Georgia State University, Atlanta, GA 30318.
Email: jingwu@cs.gsu.edu

References

[1] J.He and A.Zelikovsky (2005). MLR-Tagging: Genotype Tagging Based on SNP Multivariate Linear Regression. Submitted to Journal of Bioinformatics.
[2] Daly,M., Rioux,J., Schaffner,S., Hudson,T. and Lander,E. (2001). High-resolution haplotype structure in the human genome. Nat Genet 29:229-232.