Manual for LOCO-LD Software

Introduction

The LOCO-LD software is an implementation of the spatial localization method introduced in the following paper:

- **Enhanced Localization of Genetic Samples through Linkage Disequilibrium Correction** (2013) Yael Baran, Ines Quintela, Angel Carracedo, Bogdan Pasaniuc and Eran Halperin. The American Journal of Human Genetics.

LOCO-LD geographically localizes human samples given their genotypes, using a model trained on other genotypes for which location data is given.
The software enables the user to either use their own training data, or to use an existing model pre-trained on the PORPES European samples. The latter option is appropriate when the samples to be localized are assumed to be of European descent.

LOCO-LD is implemented in both C and MATLAB. This manual provides usage instructions for both implementations.

Input Files

LOCO-LD uses the following input files:

- Genotype file - Contains a genotype matrix, whos entries are either of 0/1/2 (number of reference alleles) or 9 (missing genotype). The number of rows equals the number of samples genotyped, the number of columns equals the number of SNPs, and the entries per row are space-delimited. SNPs should appear ordered by physical position.
  Example:
  2 9 2 2 2
  2 2 2 2 2
  1 2 1 1 9
  is a genotypes file for three samples and 5 SNPs.

- SNP position file - Contains the physical positions of the SNPs: space-delimited chromosome and bp position. The SNP order corresponds to their order in the genotype file.
  Example:
  1 742429
  1 767376

```
1 769185
1 775852
1 782343
```

is a position file for 5 SNPs.

- SNP allele file - Contains the reference and non-reference alleles of the SNPs: space-delimited base of each. The SNP order corresponds to their order in the genotype file.
  Example:

  ```
  C T
  A G
  G A
  T C
  G A
  ```

  is an allele file for 5 SNPs. Together with the above genotype file, we can conclude that the first sample has the genotype TT in the first SNP.

- Location file - Contains the estimated geographic coordinates of the training samples: space-delimited longitude and latitude. The sample order corresponds to their order in the genotype file.
  Example:

  ```
  45.320000 16.100000
  43.949493 20.615723
  49.736607 15.376984
  ```

  is a location file for 3 samples.

Training a model

LOCO-LD can be used to train a spatial model given a group of samples for which both genotypes and locations are known. We recommend imputing sporadically missing SNPs in the training genotypes prior to localization; this both decreases running time and has been shown to improve localization accuracy.

An additional parameter is the window size (in SNPs). Based on our experiments, we recommend setting this parameter to 50 when the data includes no, or a small fraction of, missing genotypes, and to 10 otherwise.

Finally, when run in default mode (”transformation mode”), the program uses 9/10 of the training samples to infer the model parameters, and the rest of the samples to infer the transformation which fits the estimated locations to the final geographic assignments (see equations 7,8 in the LOCO-LD paper). We recommend using the default transformation mode, however the user may choose to use all training samples to infer the model parameters.

C usage

Model training is performed by the program locold_train.
The following files are required:

- a genotype file (of the training samples)
- a location file (of the training samples)

The program takes a parameter file of the following format:

number of SNPs in training set
number of samples in training set
name of genotype file
name of location file
window size
trans/notrans (yes/no to transformation mode)
miss/nomiss (the genotypes include/do not include sporadically missing data)
name of the output model file

The program is executed with the line
locold_train <parameter file name>
Sample train parameter file can be found in example/paramfile_train, along with the relevant input files.

MATLAB usage

Model training is performed by the function locold_train().
The following files are required:

- a genotype file (of the training samples)
- a location file (of the training samples)
- a SNP position file (of the training dataset)
- a SNP allele file (of the training dataset)
- a SNP position file (of the samples to be localize)
- a SNP allele file (of the samples to be localize)

Training a model with window size 50 on the provided example files:

```
locold_train('example/train.geno','example/train.loc',
'example/train.pos','example/train.allele',
50,'example/train.model.w50.mat');
```

Training a model with window size 10 on the provided example files, without inferring the transformation:

```
locold_train('example/train.geno','example/train.loc',
'example/train.pos','example/train.allele',
10,'example/train.model.w10.notrans.mat','notrans');
```

Localizing samples

Given a model trained on samples from the relevant geographic region, LOCO-LD can be used to localize additional samples given their genotypes.

C usage

Sample localization is performed by the program locold_localize.
The following files are required:

- a genotype file (of the samples to be localize)
- a LOCO-LD model file
- a SNP position file (of the training dataset)
- a SNP allele file (of the training dataset)
- a SNP position file (of the samples to be localize)
- a SNP allele file (of the samples to be localize)

The program takes a parameter file of the following format:

number of SNPs in training set
number of SNPs in the samples to be localized
window size (on which the model was trained)
number of samples to be localized
position file for training set
allele file for training set
position file for localization set
allele file for localization set
trans/notrans (apply/do not apply transformation; transformation can be applied only if the model was trained with the "trans" option)
name of model file
name of genotype file
name of output file

The program is executed with the line
locold_localize <parameter file name>
Sample localization parameter file can be found in example/paramfile_localize, along with the relevant input files.

MATLAB usage

Sample localization is performed by the function locold_localize().
The following files are required:

- a genotype file (of the samples to be localize)
- a LOCO-LD model file
- a SNP position file (of the samples to be localize)
- a SNP allele file (of the samples to be localize)

Localizing a set of samples using the model in example/model.w50.mat trained above:

```
locold_localize('example/loc.geno','example/loc.pos','example/loc.allele',
'example/train.model.w50.mat','example/loc.estimate.w50');
```

Localizing a set of samples using the model in example/model.w10.notrans.mat trained above:

```
locold_localize('example/loc.geno','example/loc.pos','example/loc.allele',
'example/train.model.w10.notrans.mat','example/loc.estimate.w10.notrans','notrans');
```

Using the trained European model

LOCO-LD is provided with a pre-trained European model inferred on the POPRES European samples.

C usage

A model file trained on the inputed POPRES European samples with window size 50 is available, along with the position and allele files, are available for download from LOCO-LD's website (model.popres.euro.w50.trans.tgz). Note that the positions in the .pos file should correspond to genome build 36.3 in order to match the trained positions.

MATLAB usage

A model file trained on the inputed POPRES European samples with window size 50 is available for download from LOCO-LD's website (model.popres.euro.w50.trans.mat).
Download the appropriate file and place it in the "trained" directory.
Note that the positions in the .pos file should correspond to genome build 36.3 in order to match the trained positions.

An example for localizing samples with the pre-trained model:

```
locold_localize('example/loc.geno',
example/loc.pos’,example/loc.allele’,
’trained/model.popres.euro.w50.trans.mat’,’example/loc.estimate.w50.withpopres’);
```

Contact

Comments and suggestions for improvement as well as bug reports will be gratefully received.
Please contact Yael Baran at yaelbara@post.tau.ac.il.