Manual for pedigreeSA v1.0 Software for estimating maximum composite likelihood pedigree from genome-wide SNP data

Amy Ko
amyko@berkeley.edu

December 19, 2016

Contents

1 Description 1

2 Installation 2

3 Usage 2
3.1 Running the software 2
3.2 Run example 3

4 Options 4

5 Output files 5
5.1 fam 5
5.2 lkhd 6
5.3 Visualizing the data 6

6 Recommendations for preprocessing the data 6

1 Description

The latest versions of our software and manual can be downloaded at `https://github.com/amyko/pedigreeSA`.

pedigreeSA estimates the pedigree of a sample of individuals from genome-wide single nucleotide polymorphism (SNP) data. Simulated annealing is used to find the maximum composite likelihood pedigree. We assume that the individuals are not inbred and that there are no cyclic relationships such as double first cousins. To estimate population allele frequencies, the method requires several individuals to be present in the data. The method uses only autosomal diploid chromosomes and so any sex chromosomes must be removed from the input files.

2 Installation

Download the software from `https://github.com/amyko/pedigreeSA`. The download should include two folders: data and src. Go into the source folder (src) and make sure that the script ”pedi” is executable:

```
cd pedigree-master/src
chmod +x pedi
```

3 Usage

3.1 Running the software

The progra takes PLINK-formatted TPED and TFAM files as input. The alleles in the TPED file must be represented by characters A,C,T, and G. The TPED is a white-space delimited file that contains the following columns: 1) chromosome, 2) marker, 3) genetic distance (cM), 4) physical distance (bp), and genotypes of each individual. The TPED file does not contain a header file. For example, the following TPED file contains the genotypes of 2 individuals at 3 markers.

```
1   snp1   0.026   2014219   A A A T
1   snp2   0.032   2449448   T T T T
1   snp3   0.047   3652230   T A T A
```

The TFAM file contains 1) family ID, 2) individual ID, 3) paternal ID, 4) maternal ID, and 5) sex (1=male, 2=female). Missing paternal or maternal ID is encoded as ”0”.

```
fam1   indiv1   0   0   1
fam1   indiv2   0   0   2
fam1   indiv3   0   0   1
```

We can run the program by providing an option file which contains the name of the input files and other various parameters (See Section 4 for the format of the option file).

```
./pedi myOptionFile
```

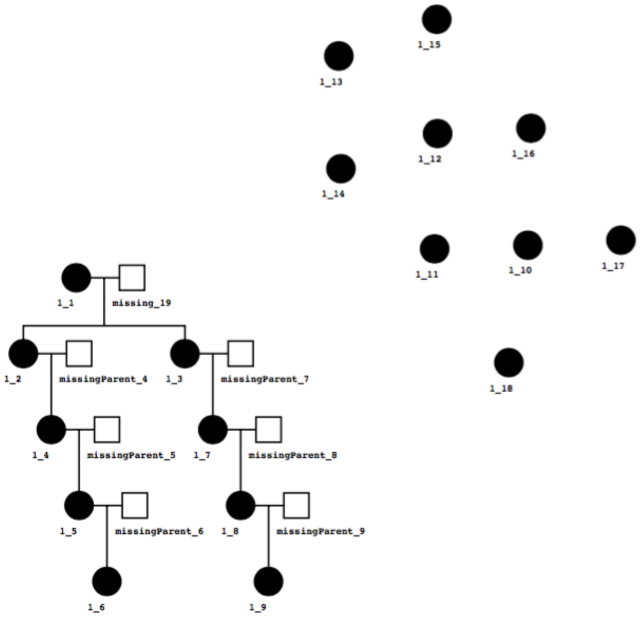

Figure 1: Pedigree drawn by CraneFoot

3.2 Run example

Here we describe how to run the program on the provided example data. The example data is located in pedigree-master/data.

```
cd data
```

The data folder contains test files containing 18 individuals (test.tped and test.tfam); reference population files (refpop.tped and refpop.tfam); and options file that specifies various parameters for the program (options). From the data folder, we can run the program by calling pedi:

```
../src/pedi options
```

This creates two types of output files: named test.*i*.fam and test.*i*.lkhd, where *i* is the a particular run (See Section 5 for more detail). To visualize the estimated pedigree, we can use the pedigree-drawing software Cranefoot [1], where test.tfam is the input. An example of the the pedigree diagram drawn by Cranefoot is shown in Fig 1.

4 Options

This section describes various options you can include in the option file. Each line in the option file corresponds to a particular option described below (order of options does not matter). The first column is the value of the option and the second column is the option name. The option name is always preceded by "#" (e.g. #fileName). The first and second columns are separated by a white space delimiter.

#fileName Name of tped and tfam files that contains the genotype data for the sample. For example, if the files are named "myFile.tped" and "myFile.tfam", then the value for the fileName option should be "myFile".

#refPopFileName Name of tped and tfam files from which population allele frequencies and two-locus haplotype frequencies will be estimated. The set of markers in this file should be the same as those in #fileName. (default=#fileName)

#ageFileName Name of the file that contains the age information for the sample. The first column is the family ID; second column is the individual ID; and the third column is the age. The columns are separated by white space and the family ID and the individual ID must match those in #fileName. If the age file is not given or the file does not contain age information for some individuals, the ages for those individuals are set as missing. (default=N/A)

#errorRate Genotype error rate. (default=0.01)

#maxGen Maximum number of generations spanned by the pedigree. For example, if the sample contains a first cousin relationship, then the number of generations spanned by the pedigree is 3 because the common ancestors of first cousins go back up to their grandparents. Our method supports up to 5 generations. (default=5)

#maxSampleDepth Maximum depth for the sample individuals (1=present generation, 2=parent generation, 3=grandparent generation, etc). #maxSampleDepth must be less than or equal to #maxGen. (default=#maxGen)

#conditional Indicator (0 or 1) for whether to condition on LD markers. (default=1)

#back The maximum recombination distance between the current marker and the previous marker to condition on when computing likelihoods. (default=0.04)

#startTemp The starting temperature for simulated annealing. (default=100)

#tempFact The factor by which to decrease the current temperature. (default=1.01)

#iterPerTemp The number of iterations for each temperature. (default=40000)

#maxIter The maximum number of iterations before stopping simulated annealing. (default=30000000)

#conv The stop threshold for stopping simulated annealing. If the likelihood value for the current pedigree is within #conv of the likelihood value of the pedigree 100,000 iterations ago, the algorithm is terminated. (default=0.0001)

#poissonMean The mean of the Poisson distribution in regularization term. (default=number of samples)

#numRun The number of independent runs. (default=3)

#numThreads The number of threads to use. (default=2)

5 Output files

There are 2 output files for each run of the program: fam and lkhd. For example, for a test file named myFile and run number 2, the two output files are named myFile.2.fam and myFile.2.lkhd.

5.1 fam

This file describes the pedigree of the estimated pedigree. There are 5 columns in the file:

Name Name of the individual. For sampled individuals, the name is "FID_IID" from the input file (e.g. familyA_indiv6). Missing individuals are named "missing_x" or "missingParent_x", where *x* is an individual ID.

FATHER Name of the father

MOTHER Name of the mother

SEX Sex of the individual (1=Female, 7=Male, 4=Unknown).

SAMPLED Sample status of the individual (000000=sampled, 999999=unsampled).

The values for SEX and SAMPLED are such that the FAM file can be used by the pedigree drawing program CRANEFOOT [1] to visualize the estimated pedigree (See Section 5.3).

5.2 lkhd

This file contains the composite likelihood values over the course of the simulated annealing algorithm. The file has two columns, where the first column is the iteration number and the second column is the composite likelihood of the pedigree at that iteration.

NumIter Number of iterations

Likeliihood Composite likelihood value

We recommend that you examine the likelihood values from multiple runs to check whether the composite likelihood values from all the runs converge to the same or a similar value. If not, consider increasing maxIter or iterPerTemp described in the options.

5.3 Visualizing the data

The estimated pedigree encoded in FAM file can be drawn by the software CraneFoot [1], which can be downloaded `http://www.finndiane.fi/software/cranefoot/`. CraneFoot is not included in our package and must be downloaded separately. Cranefoot takes a configuration file (config.txt) as an argument.

```
./cranefoot config.txt
```

where the content of config.txt is given by:

```
PedigreeFile writeYourFullDirectoryHere/test.0.fam
PedigreeName test
NameVariable NAME
FatherVariable FATHER
MotherVariable MOTHER
ShapeVariable SEX
ColorVariable SAMPLED
TextVariable NAME
```

This creates post script file test.ps containing the pedigree diagram given by test.0.fam.

6 Recommendations for preprocessing the data

The current version of our program does not support missing genotypes. Therefore, markers with missing genotypes must be removed from the data before running the program. Incorporation of missing data will be implemented in the next version of the software.

Furthermore, we recommend that you prune markers to reduce the effects of LD to prevent overestimation of relatedness. The easiest way to do this is using PLINK [2].

```
plink –tfile myInputFile –indep-pairwise 50 5 .05
```

The above command prunes the set of markers in myIntputFile.tped at $r^2 = .05$ with a window size of 50 SNPs and step size of 5. It outputs ”plink.prune.in” which contains the filtered markers. Appropriate values for r^2 and the window size depend on the genome length. For human genomes, our simulations showed that $r^2 = .05$ and window size equivalent of .04 cM work well.

Now we can extract the markers contained in ”plink.prune.in” from the original set of markers:

```
plink –tfile myInputFile –extract plink.prune.in –make-bed –recode –tab
–transpose –out myInputFile.LDpruned
```

which creates myInputFile.LDpruned.tped and myInputFile.LDpruned.tfam files, where LD markers are pruned away.

References

[1] Makinen VP, Parkkonen M, Wessman M, Groop PH, Kanninen T, Kaski K. High-throughput pedigree drawing. Eur J Hum Genet. 2005;13(8): 987-9. doi:10.1038/sj.ejhg.5201430.

[2] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. doi:10.1186/s13742-015-0047-8.