Manual for CLAPPER v0.9
Software for estimating maximum composite likelihood pedigree from genome-wide SNP data

Amy Ko
amyko@berkeley.edu

August 23, 2017

Contents

1 PLEASE NOTE!

CLAPPER is designed to estimate outbred pedigrees and therefore will not necessarily work well for highly inbred pedigrees. We suggest you first check the level of inbreeding in your samples before proceeding with pedigree estimation.

Furthermore, likelihood computation implemented in CLAPPER is sensitive to linkage disequilibrium (LD) so we suggest you prune the markers before using our software. For human data, we found 10,000 to 20,000 SNPs to work well. Alternatively, you can provide likelihood values directly as input (See 4.2 for details).

Future releases of the program will aim to account for inbreeding and LD.

2 Description

CLAPPER 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 complex 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.

3 Installation

The latest versions of our software and manual can be downloaded at `https://github.com/amyko/clapper`. The download should include two folders: data and src. Go into the source folder (src) and make sure that the script "clapper" is executable:

```
cd clapper-master/src
chmod +x clapper
```

4 Usage

4.1 Input Files (.tped and .tfam)

The program takes PLINK-formatted TPED and TFAM files as input.

4.1.1 yourFileName.tped

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 (Morgan), 4) physical distance (bp),

and genotypes of each individual. The TPED file does not contain a header. 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
```

N.B. All markers should be polymorphic in the reference population (see #refpop in Section 5) even if they may not be polymorphic in the samples.

4.1.2 yourFileName.tfam

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". Here's an example of a TFAM file for 3 individuals:

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

CLAPPER will then use the TPED and TFAM files to compute the marginal and pairwise likelihoods for the samples, which in turn will be used in simulated annealing to estimate the pedigree.

4.2 Alternative Input Files (.marginal and .pairwise)

Alternatively, you can provide files containing marginal and pairwise likelihoods directly as input using pairwise likelihood computation method of your choice (e.g. [3, 4]). If you choose this option, you still need to provide the TFAM file described in 4.1.2, along with marginal (*.marginal*) and pairwise (*.pairwise*) likelihood files described below.

4.2.1 yourFileName.tfam

See 4.1.2 for format.

4.2.2 yourFileName.marginal

The marginal likelihood file contains the log marginal likelihood of each sample per line. The order of the likelihood values must match the order of the individuals in the TFAM file. Therefore, the number of lines in this file should be equal to the number of individuals in the sample. For example, the marginal likelihood file for three individuals would look something like this:

-1762.60
-1784.04
-1782.60

where the first line corresponds to the first individual in the TFAM file, the second line corresponds to the second individual, and so on.

4.2.3 yourFileName.pairwise

The pairwise likelihood file contains the log pairwise likelihood values for all pairs of individuals for each possible relationship type. The relationship type is denoted by " > ", followed by three integers: 1) number of generations between individual 1 and the common ancestor, 2) number of generations between individual 2 and the common ancestor, and 3) the number of common ancestors between individual 1 and 2. For example, an avuncular relationship is denoted by "> 2 1 2." The table below shows the relationship keys for all pairwise relationships that span up to 5 generations.

Relationship	Key
Unrelated	0 0 0
Parent-offspring	1 0 1
Full siblings	1 1 2
Half siblings	1 1 1
Grandparent-grandchild	2 0 1
Full avuncular	2 1 2
Half avuncular	2 1 1
Full cousins	2 2 2
Great-grandparent - great-grandchild	3 0 1
Full avuncular once-removed	3 1 2
Half avuncular once-removed	3 1 1
Full cousins once-removed	3 2 2
Half cousins once-removed	3 2 1
Full second cousins	3 3 2
Half second cousins	3 3 1
Great-great-grandparent - great-great-grandchild	4 0 1
Full avuncular twice-removed	4 1 2
Half avuncular twice-removed	4 1 1
Full cousins twice-removed	4 2 2
Half cousins twice-removed	4 2 1
Full second cousins once-removed	4 3 2
Half second cousins once-removed	4 3 1
Full third cousins	4 4 2
Half third cousins	4 4 1

Following the keys shown this table, the pairwise file must contain *all possible pairwise relationships for the maximum number of generations you specify* in the option file (See Section 5). For example, if the maximum generation you specify

is 2, then the pairwise file must contain values corresponding to unrelated, parent-offspring, full siblings, and half siblings.

After you specify the relationship type, the subsequent lines correspond to the log likelihood values for $n(n-1)/2$ pairs of samples. The first two columns of the file are the indices of the two individuals, where the index $0, 1, 2, ...$ correspond to $indiv1, indiv2, indiv3, ...$, respectively, specified in the TFAM file. For example, the following shows the pairwise likelihood values for "unrelated" and "full-sibling" relationships for 3 individuals.

> 0	0	0
0	1	-3546.65
0	2	-3545.20
1	2	-3487.71
> 1	1	2
0	1	-3332.10
0	2	-3300.49
1	2	-3658.31

4.3 Run Command

We can run the program by providing an option file which contains the name of the input files and other various parameters (See Section 5 for the format of the option file).

```
./clapper myOptionFile
```

4.4 Run example

Here we describe how to run the program on the provided example data. The example data is located in clapper-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 "clapper":

```
../src/clapper 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 6 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.

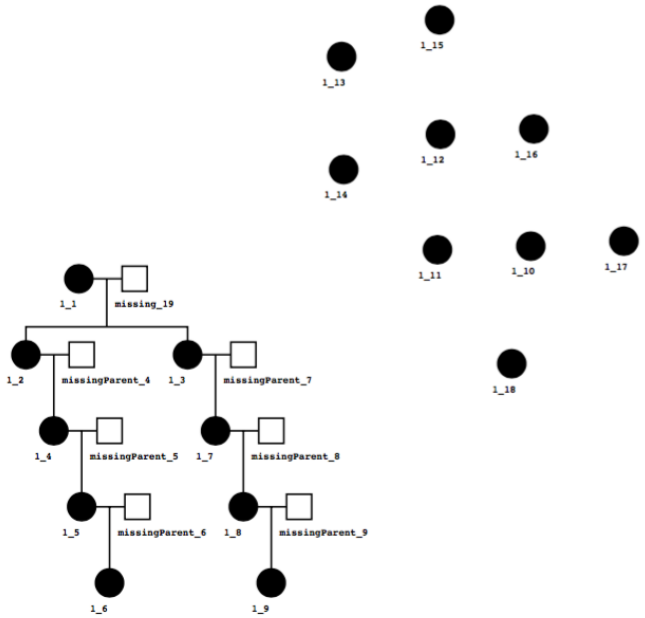

Figure 1: Pedigree drawn by CraneFoot

5 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)

#computeLikelihood Indicator (0 or 1) denoting whether to compute marginal and pairwise likelihoods internally. If it is set to 0, then the the user must provide *.marginal* and *.pairwise* files (See 4.2 for format of these files). If set to 1, the likelihood files are generated from the genotype data given by *.tped* files. (default = 1)

#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 the *.tfam* file. 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. This option is ignored if #conditional is set to 0. (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)

#beta The power to which the regularization term is raised. (default=30)

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

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

6 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.

6.1 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 6.3).

6.2 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.

6.3 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.

7 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. One 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.

[3] Albrechtsen A, Sand Korneliussen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R. Relatedness mapping and tracts of relatedness for genome?wide data in the presence of linkage disequilibrium. Genetic epidemiology. 2009 Apr 1;33(3):266-74.

[4] Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin rapid analysis of dense genetic maps using sparse gene flow trees. Nature genetics. 2002 Jan 1;30(1):97-101.