

PedMix User Manual

Jingwen Pei, Yiming Zhang and Yufeng Wu

March 2, 2020

1 Introduction

PedMix is an HMM (Hidden Markov Model) based software tool, which is designed to infer the admixture proportion of parents, grandparents and even great grand parents from the genome of the focal individual. Standard methods for inferring the ancestry of focal individual implicitly assume the same admixture proportions in both parents. However, this is unrealistic for many human populations, especially for recently admixed populations. *PedMix* considers a single diploid individual from an admixed population with two ancestral populations A and B . We assume the phased genotype $G = (H_1, H_2)$ and the allele frequency in both ancestral populations at each SNP are given.

2 Prerequisite

g++ version later than 4.4.7 has been used to compile *PedMix* successfully. We suggest to enable `openmp` in compilation to enhance the performance of *PedMix*.

To obtain maximum likelihood estimates of admixture proportions, we apply the Boyden-Fletcher-Goldfarb-Shanno (BFGS) method of optimization, using an implementation of the limited-memory version of the algorithm (L-BFGS) written in C. L-BFGS library needs to be installed to compile *Pedmix*. Please download the source code of L-BFGS package from: <https://github.com/chokkan/liblbfgs>, and then compile with following commands:

```
$ ./configure
$ make
$ make install
```

PedMix requires the path where library and header are specified in `Makefile`. One can also install library and header under an alternative directory with specification of `configure` command. For more details of installation and method, please refer to website: <http://www.chokkan.org/software/liblbfgs>.

3 Download and Installation

Source code is now available: <https://github.com/pjweggy/PedMix>. After installation of L-BFGS library, please change the path of `$LBFGS_INC` and `$LBFGS_LIB` to your installation path of header and library in `Makefile` (first two lines).

`LBFGS_INC=<path to LBFGS INCLUDE DIR>`

LBFGS_LIB=<path to LBFGS LIB DIR>

To compile *PedMix*, run the following commands:

```
$ make clean
```

```
$ make
```

openmp is enabled with `-fopenmp` specified in `Makefile`.

4 Input Data format

4.1 Input file

The input of phased genotype data has the following format. Take `testPedMix.inp` for example:

```
//  
1 0.0000 0.5440 0.0002475900 1 0  
1 0.7200 0.0000 0.0001201200 0 0  
1 0.9500 0.0320 0.0003691200 0 0  
1 0.9140 0.2980 0.0001483800 0 1  
1 0.6620 0.0000 0.0021167700 0 0  
1 0.5000 0.0000 0.0000291600 0 0  
1 0.5000 0.0000 0.0000531000 0 0  
1 0.6060 0.0000 0.0013107000 0 0  
1 0.4280 1.0000 0.0000137400 1 1  
1 0.4280 1.0000 0.0000927000 1 1  
1 0.3660 0.9920 0.0001067100 1 1
```

The first column specifies the allele type that is used to count allele frequency. The second column specifies the allele frequency of such allele type in first ancestral population. The third column specifies the allele frequency in second ancestral population. The fourth column specifies the genetic distance (in centimorgan, cM) from current site to the next site. Thus, the last site should have distance 0.0, but it won't affect if greater than 0.0 since it will not be used. The last two columns specify two haplotypes (phased genotypes).

One can contain multiple locus in one input data. *PedMix* would consider multiple locus together to achieve maximum likelihood. Each loci should start with delimiter `//`. No newline is needed between two loci.

4.2 Parameter file

The parameter file contains phasing error rate, recombination rate, length of loci and BFGS step size. Take `parfile` for example:

```
phasing = 0.000001 #phasing error rate: number of phasing error occurrence per site
```

```

recombination = 0.00000001 #recombination rate: number of recombination event per site per
generation
length = 10000000 #length of loci: bp of a loci
step = 0.000000001 #BFGS step size: suggest  $1/\text{length} \times 10 \sim 1/\text{length} \times 100$ 

```

5 Usage

After successful compilation, now you can run *PedMix* to estimate admixture proportions of ancestors. The command line looks like:

```
$ ./PedMix -g 1 -p parfile testPedMix.inp
```

Options:

```
./PedMix <OPTIONS> <Input file name>
```

-g : number of generations to trace back, for example parents is '-g 1'.

-p parfile : input parameter file to *PedMix*, including phasing error rate, recombination rate, length and BFGS step size.

For more details of the parfile and the input file, please see section 4.

6 Tips

6.1 How to choose parameters?

In our paper, we discussed how to get the best performance of *PedMix* with parameters and data pre-process. For more details, please refer to our paper.

6.1.1 loci length and number of loci

In general, *PedMix* performs better when providing longer locus and more locus. *PedMix* computes likelihood for each loci independently, so the performance mainly depends the number of ancestral tracts in each loci. If an focal individual is well-admixed, but user provides a number of short locus, which might only contains 1 to 2 tracts, it will largely affect the accuracy. In this case, we suggest to concatenate multiple locus together.

6.1.2 phasing error

Phasing error can significantly affect the accuracy of estimation in ancestry inference of ancestors. Normally we observe an average phasing error rate as 1 over 50kbp. However, it is very high compared to human recombination rate. To gain high accuracy, we suggest to pre-process data based on the known knowledge (see *PedMix* paper for more details. For your convenience, we have provided a simple utility software tool to remove phasing errors. Based on our experience, this strategy can eliminate at least 2 out of every 3 phasing errors. Then you can use a smaller phasing error for inference.

6.2 Data filtering strategy: LD pruning and frequency-based pruning

In our method, we proposed a frequency-based pruning strategy to filter SNPs for whole genome (please see our paper for more details). LD-pruning strategy can also help to enhance the performance. But keep in mind to remove rare variants (frequency addition in both ancestral populations < 0.1 or 0.2) before either strategy. For your convenience, we have provided a simple utility software tool for performing data filtering based on allele frequency difference in ancestral populations.

6.3 Running Time

To make most use of computational resources, we suggest to use multiple threads for multi-locus data. *PedMix* is compiled via openmp, default version is set to use at most 22 threads at the same time. If your device can assign less than 22 threads, it should be OK to run the current version. If you want to compile a version with more than 22 threads, check file *PedMixTest.cpp* and find all following lines, change ‘22’ to the number you want.

```
#pragma omp parallel num_threads(22)
```

It’s efficient when the number of locus is the multiple of number of threads. User needs to balance between efficiency and accuracy.

7 How to cite

Our paper is currently under review. The preprint is available at bioRxiv: “Inferring the ancestry of parents and grandparents from genetic data”, Jingwen Pei, Rasmus Nielsen and Yufeng Wu, doi: <https://doi.org/10.1101/308494>, bioRxiv, 2018. If you have questions about *PedMix*, please contact Yufeng Wu through emails: yufeng.wu@uconn.edu.