PedMix2 User Manual - v0.1

Yiming Zhang and Yufeng Wu

1 Introduction

PedMix2 is a probabilistic model-based software tool designed to make the ancestry inference of ancestors of an admixed individual. PedMix2 is an upgraded version of our previous tools PedMix [1] (which can estimate the admixture proportion of recent ancestors from a single child), and parMix [2] (which can infer parental ancestry and call parental genotypes from data of a small number of children). In the current version, PedFAM uses the extant individual's genotypes and the reference allele frequencies to infer the configuration founders (we define founders as the first-generation ancestors of the extant individual that are not admixed), then use this inferred configuration to calculate the admixture proportions of all ancestors of the extant individual.

2 Prerequisite

Python version in 3.8.5 has been used to compile PedMix2 successfully. The numpy version later than 1.22.1 and the numba version later than 0.56.2 are required for running PedMix2. numba is a decorator for accelerating the PedMix2. To install the prerequisites, run the following commands:

```
$ python -m pip install --upgrade pip
$ pip install numpy numba
```

3 Download

Source code now available: https://github.com/biotoolscoders/PedMix2.

4 Inputs

4.1 Input Files

PedMix2 needs three different input files, and they are shown as follows.

1. Phased genotypes file of the extant individual. Take *example/Geno_C1.dat* for example:

$$Genotypes \begin{cases} first\ haplotype: & 0\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 0 \\ second\ haplotype: & 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1 \end{cases} \tag{1}$$

(2)

2. Position file of SNPs. The physical position of each SNP. Take *example/Position_C1.dat* for example:

$$\begin{aligned} \textit{position 1:}\ & 1765.000353 \\ \textit{position 2:}\ & 6150.001230 \\ \textit{position 3:}\ & 9990.001998 \\ \textit{position 4:}\ & 10480.002096 \\ \textit{position 5:}\ & 12740.002548 \\ \textit{position 6:}\ & 13290.002658 \\ \textit{position 7:}\ & 17030.003406 \\ \textit{position 8:}\ & 17610.003522 \\ \textit{position 9:}\ & 22615.004523 \\ \textit{position 10:}\ & 22850.004570 \end{aligned}$$

3. Allele frequency file of reference populations. Take *example/AF_C1.dat* for example:

$$\textit{Two references} \begin{cases} \textit{population } A: & 0.990050\ 0.472637\ 0.997512\ 0.781095\ 0.997512\ 0.519900\ 0.997512 \\ \textit{population } B: & 0.997512\ 0.982587\ 0.942786\ 0.997512\ 0.594527\ 0.002488\ 0.987562 \end{cases} \tag{3}$$

(4)

5 Usage

For using PedMix to infer the founders' configuration, first simply type:

```
$ python PedMix2.py -h
```

-g	Number of generations since admixture
-b	Number of blocks per chromosome
-r	Recombination Rate (per base pair per generation)
-c	Number of chromosomes
-p	Number of reference panels
-s	Number of random starting points (Default = 5)
-F	Allele frequency file prefix
-P	Position file prefix
-G	Genotype file prefix
-o	Output path prefix
-S	Random Seed (Default = None)

Table 1: the arguments for running PedMix2

For example, the command line for running the example data set looks like this:

```
$ python PedMix2.py -g 5 -b 30 -r 1e-8 -c 1 -p 2 -s 6-F example/AF_C -P example/Position_C
-G example/Geno_C -o example/result -S 55
```

In this case, PedMix2 will infer the founder's configuration in 5 generations ago (-g 5) using 1 chromosome (-c 1). There are 2 reference panels (-p 2), and PedMix2 divides each chromosome with 30 blocks (-b 30). The recombination rate is 10^{-8} (-r 1e-8), and PedMix2 will run the local search algorithm 6 times with different random start points (-s 6). All input files' names are "Prefix"+"chromosome serial number".

6 Outputs

The PedMix2 will output 6 different founder configurations and pick one with the highest probability, then output.

The optimal founders' configuration in the example will be:

```
$ 1 0 0 1 1 0 1 0 0 1 0 1 0 1 1 0 1 0 1 0 1 0 0 1 1 0 1 0 1 0 1 0
```

with the log probability as -641.060866.

7 How to cite

The paper, "**A general approach for inferring the ancestry of ancestors of an admixed individual**" by Yiming Zhang, Haotian Zhang, and Yufeng Wu, is under review, and we will keep updating the information. Please feel free to contact *Yiming Zhang* via yiming.zhang.cse@uconn.edu or *Yufeng Wu* via yufeng.wu@uconn.edu if you have any questions about PedMix2.

References

[1] Jingwen Pei, Yiming Zhang, Rasmus Nielsen, and Yufeng Wu. Inferring the ancestry of parents and grandparents from genetic data. *PLoS computational biology*, 16(8):e1008065, 2020.

[2] Yiming Zhang and Yufeng Wu. Joint inference of ancestry and genotypes of parents from children. *Iscience*, 25(8):104768, 2022.