PedMix2 Protocol for Large-Scale Sample Processing

Yiming Zhang, Haotian Zhang, and Yufeng Wu

1 Introduction

PedMix2 is an advanced probabilistic model-based software tool designed to infer the ancestral backgrounds of admixed individuals. This upgraded version builds upon its predecessors, 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). PedMix2 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.

This protocol provides a comprehensive, step-by-step guide to executing PedMix2 on large-scale data samples. For complete details on the use and execution of this protocol, please refer to **Zhang and Wu**.

2 Before you begin

As far as our current knowledge indicates, there are no existing methods capable of accurately inferring admixture proportions for more recent ancestors (say great-grandparents), solely based on an individual's genotypes. This challenge arises due to the technical complexity of losing half of parental DNA information during each meiosis process. Nevertheless, in the context of admixed populations, the reference allele frequency assumes a significant role in indicating the ancestral composition of admixed individuals.

PedMix2 capitalizes on this reference allele frequency information along with individual genotypes to deduce the admixture proportions of all ancestors within a given pedigree. For a comprehensive understanding of the algorithms and mechanisms employed by PedMix2, please refer to the paper [cite paper reference].

The ensuing protocol outlines the specific procedural steps for executing PedMix2 on a large-scale sample. Prior to delving into the operational details of PedMix2, it is important to be aware of certain prerequisites that must be satisfied.

2.1 Prerequisite

PedMix2 is implemented on Python, and Python version later than **3.8** has been showed to compile PedMix2 successfully.

NumPy: *NumPy is the fundamental package for scientific computing in Python. It provides support for arrays, matrices, and a wide variety of mathematical functions to operate on these arrays. [3]*

Numba: *Numba is a Just-In-Time (JIT) compiler for Python that specializes in optimizing code for numerical and scientific computations. It translates Python functions to optimized machine code at runtime using the industry-standard LLVM compiler library. Numba-compiled numerical algorithms in Python can approach the speeds of C or FORTRAN. [4]*

The **NumPy** version later than **1.21.5** and the **Numba** version later than **0.56.2** have been tested to complied PedMix2 successfully.

To install the prerequisites, run the following commands:

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

For users who prefer **Conda** environment, please run the following commands:

```
$ conda create -n pedmix2 python=3.10.12 numpy numba
$ conda activate pedmix2
```

2.2 Download

The source code and sample data of PedMix2 now available at: `https://github.com/biotoolscoders/PedMix2`.

2.3 Inputs

PedMix2 requires three distinct types of input files. In the current version, PedMix2 doesn't yet have the capability to work with the *.vcf* format. However, it's worth noting that this support will be incorporated in the upcoming upgrade.

2.3.1 Phased genotypes

The phased genotypes file for an individual holds the individual's SNP information. Each line in this file represents a sequence of 0s and 1s, indicating the SNPs on each haplotype. The reference allele shows as 0, and the alternative allele shows as 1. In this protocol, we present an example individual possessing 22 chromosomes. Consequently, there exist a total of 22 genotypes files.

Take *example/Geno_C1.dat* for example:

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

$$\tag{2}$$

2.3.2 Positions

The positions file encompasses the actual physical positions of each SNP. It can either be directly extracted from the *.vcf* file or computed from the (0,1) scaled position data provided by *macs*. It's important to note that, owing to the precision of *macs* output, certain SNPs might share the same position. A potential solution is to introduce a minor variation to these positions or to simply exclude them from the position list.

Take *example/Position_C1.dat* for example:

position 1: 1765.000353
position 2: 6150.001230
position 3: 9990.001998
position 4: 10480.002096
position 5: 12740.002548
position 6: 13290.002658
position 7: 17030.003406
position 8: 17610.003522
position 9: 22615.004523
position 10: 22850.004570

2.3.3 Allele frequency

The allele frequency file holds the allele details pertaining to reference populations. Each line in this file presents a collection of allele frequencies for each SNP, specific to a given reference population. These allele frequencies are typically derived from the genotypes of a selection of reference individuals within that particular population. Each frequency displayed in this file signifies the occurrence rate of the reference allele (referred to as *0's*) within a population.

In manual allele frequency calculations, it's crucial to employ soft counts for each allele. This practice prevents instances of encountering percentages of either 0% or 100% in the file, which can sometimes arise due to counting approaches.

Take *example/AF_C1.dat* for example:

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

$$\tag{4}$$

3 Usage

The fundamental concept of PedMix2 revolves around deducing the configuration of founders. This inferred data then forms the basis for calculating admixture proportions across all ancestors within a pedigree. PedMix2 demonstrates a notably heightened accuracy when employed for inference within fewer than 3 generations. However, as we trace back through more recent ancestors, accuracy is influenced by the scarcity of available information for these individuals, resulting in a gradual decline in accuracy with each subsequent generation.

For using PedMix2, first simply type:

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

-g	Number of generations since admixture (Default = 8)
-b	Number of blocks per chromosome (Default = 20)
-r	Recombination Rate (per base pair per generation) (Default = $1e^{-8}$)
-c	Number of chromosomes (Default = 22)
-p	Number of reference panels (Default = 2)
-s	Number of random starting points (Default = 5)
-F	Allele frequency file prefix
-P	Position file prefix
-G	Genotypes file prefix
-o	Output path prefix
-a	Admixture proportion analysis (Optional)
-S	Random Seed (Default = None)

Table 1: the arguments for running PedMix2

Table 1 outlines the adjustable parameters that grant control over PedMix2's functionality. In addition to the fundamental configurations, we offer two distinct methods for displaying the results:

1. When the ***-a*** option is omitted, PedMix2 generates results in the format of founder configurations. This choice provides users with the most likely founder configurations of the extant individual.
2. By selecting the ***-a*** option, PedMix2 initiates an admixture proportion analysis based on the inferred founder configuration. The value assigned to ***-a*** specifies how many generations of ancestors' admixture proportions PedMix2 should compute.

Within this protocol, the command line for executing the analysis on large-scale sample data appears as follows:

```
$ python PedMix2.py -F Large_Sample/AF/AF_Chro -P Large_Sample/POS/position_Chro
-G Large_Sample/Geno/Gen8_Chro -o Large_Sample/result_FC -S 55
```

In this case, PedMix2 will infer the founder's configuration in 8 generations ago (-g 8) using 22 chromosome (-c 22). There are 2 reference panels (-p 2), and PedMix2 divides each chromosome with 20 blocks (-b 20). The recombination rate is 10^{-8} (-r 1e-8), and PedMix2 will run the local search algorithm 5 times with different random start points (-s 5). All input files' names are "Prefix"+"chromosome serial number".

For admixture proportions analysis, the command line for executing the PedMix2 shows as follows:

```
$ python PedMix2.py -F Large_Sample/AF/AF_Chro -P Large_Sample/POS/position_Chro
-G Large_Sample/Geno/Gen8_Chro -o Large_Sample/result_AD -a 4 -S 55
```

In this case, PedMix2 will generate the calculated admixture proportions for the ancestors of the extant individual (included), spanning up to 4 generations back (-a 4).

4 Expected outcomes

When PedMix2 is executed using default configurations and with admixture proportion analysis enabled, the software will print out five distinct founder configurations. It will subsequently select the one with the highest probability and save the results in a file named "result_AD.txt."

Fig. 1 shows the contents of the results for this large-scale example.

To conduct a comparative analysis between the results and ground-truths, we have included an Excel file containing the ground-truth data, configuration accuracy rates, and admixture proportion accuracy rates. Please consult this file to assess PedMix2's performance.

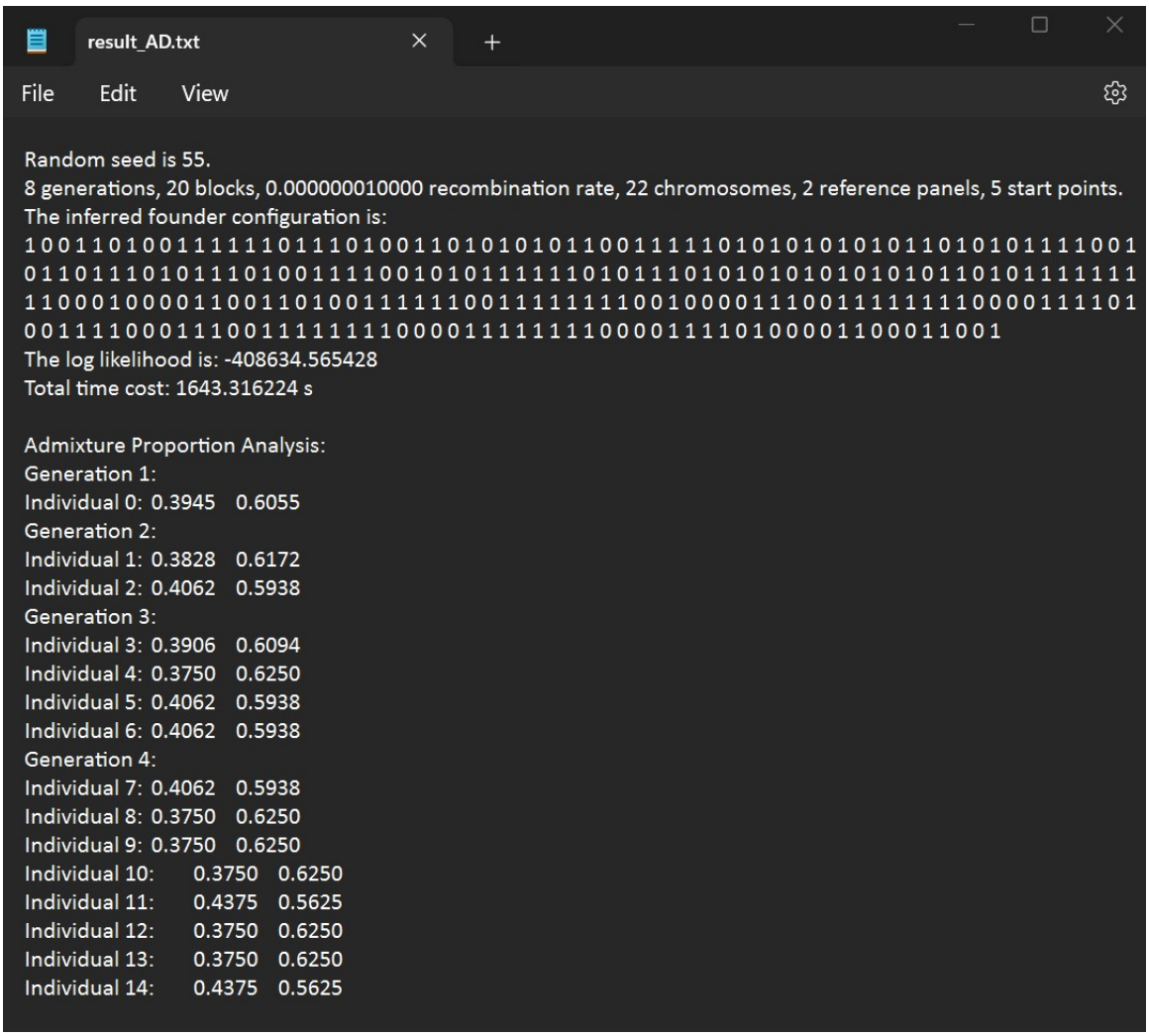

Figure 1: Results for this large-scale example experiment

5 Overheads

We evaluate the running time of PedMix2 under different numbers of generations g and numbers of blocks n_b. Among all the simulation parameters, g and n_b have the largest impact on running time. Our experiments are run on a machine with the Linux and an Intel(R) Core(TM) i9-9900K CPU (3.60 GHz).

Each data point in Fig 2 represents the running time for a single local search starting point. For example, the total running time of PedMix2 under the default setting is around 27 minutes (8 generations, 20 blocks per chromosome, 22 chromosomes, 2 reference panels, and 5 local search starting points).

6 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.

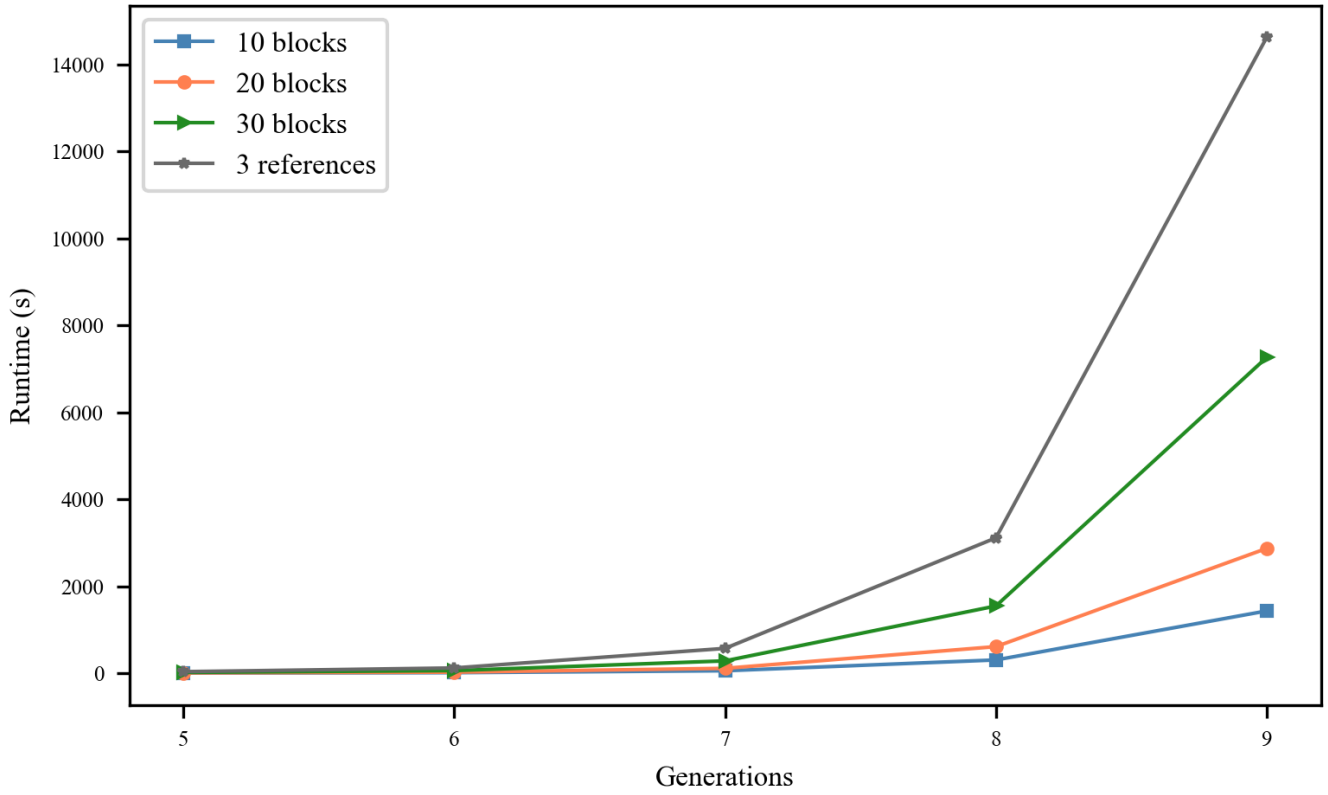

Figure 2: Running time of PedMix2 under varying parameters (22 chromosomes, 10 starting points)

[3] Numpy. `https://numpy.org/`.

[4] Numba. `https://numba.pydata.org/`.