parMix User Manual

Yiming Zhang and Yufeng Wu

November 20, 2021

1 Introduction

parMix is an HMM (Hidden Markov Model) based software tool, which is designed to jointly infer parental ancestry and call parental genotypes from data of a small number of children. parMix is an extended version of our previous tool PedMix [1] (which can estimate the admixture proportion of recent ancestors from a single child). And different from PedMix, parMix can provide fine-scale inference, namely parental ancestry and genotypes at each single nucleotide polymorphism site. In current version, parMix considers a single diploid family with more than one child, and each child belongs to an admixed population with two ancestral populations A and B. We assume the phased genotypes of these children are given, as well as the allele frequency and LD (Linkage Disequilibrium) frequency are known in both reference ancestral populations at each SNP.

2 Prerequisite

Python3 version in 3.8.5 has been used to compile parMix successfully, and numpy version later than 1.20.3 is required for running parMix. For installing numpy, simply run the follow command:

```
$ pip3 install numpy
```

3 Download

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

4 Inputs

4.1 Input Files

parMix needs six different input files, and they are shown as follows.

1. Phased genotypes file of children, and every two lines indicate one child's genotypes. Take *example/testgenotypes.dat* for example:

$$\textit{first child} \begin{cases} 0\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 0 \\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1 \end{cases} \tag{1}$$

$$\textit{second child} \begin{cases} 1\ 0\ 0\ 0\ 0\ 1\ 0\ 1\ 1\ 0 \\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 1 \end{cases} \tag{2}$$

$$\textit{third child} \begin{cases} 0\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 0 \\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1 \end{cases} \tag{3}$$

2. Position file of SNPs. The position of each SNP is scaled in {0,1} based on its physical position. Take *exam-*

ple/testposition.dat for example:

position 1:	0.010712233500000
position 2:	0.110770420400000
position 3:	0.210804650400000
position 4:	0.317549716400000
position 5:	0.441551113400000
position 6:	0.543687953700000
position 7:	0.644049178800000
position 8:	0.744166260200000
position 9:	0.844255575300000
position 10:	0.944401871100000

3. Allele frequency files of population A and B. Take *example/testAFA.dat* for example:

SNP 1:	0.6138613861386139
SNP 2:	0.3861386138613861
SNP 3:	0.4108910891089109
SNP 4:	0.3168316831683168
SNP 5:	0.9950495049504950
SNP 6:	0.2326732673267327
SNP 7:	0.8069306930693070
SNP 8:	0.2079207920792079
SNP 9:	0.2128712871287129
SNP 10:	0.9950495049504950

4. LD frequency files of population A and B. Each line in the file indicates the frequencies of the two adjacent SNPs with (0,0), (0,1), (1,0) and (1,1). Take *example/testLDA.dat* for example:

SNPs	**(0,0)**	**(0,1)**	**(1,0)**	**(1,1)**
SNP 0-1:	0.2500000000000000	0.2500000000000000	0.2500000000000000	0.2500000000000000
SNP 1-2:	0.0049019607843137	0.6078431372549019	0.3823529411764706	0.0049019607843137
SNP 2-3:	0.3823529411764706	0.0049019607843137	0.0294117647058824	0.5833333333333334
SNP 3-4:	0.1225490196078431	0.2892156862745098	0.1960784313725490	0.3921568627450980
SNP 4-5:	0.3137254901960784	0.0049019607843137	0.6764705882352942	0.0049019607843137
SNP 5-6:	0.2303921568627451	0.7598039215686274	0.0049019607843137	0.0049019607843137
SNP 6-7:	0.0588235294117647	0.1764705882352941	0.7450980392156863	0.0196078431372549
SNP 7-8:	0.0196078431372549	0.7843137254901961	0.1911764705882353	0.0049019607843137
SNP 8-9:	0.2058823529411765	0.0049019607843137	0.0098039215686275	0.7794117647058824
SNP 9-10:	0.2107843137254902	0.0049019607843137	0.7794117647058824	0.0049019607843137

4.2 Input Parameters

There are 4 parameters need to be specified when running parMix. The number of children, the total number of base pairs before trimming, the recombination rate, and the phasing error rate. Moreover, the genotyping error rate is 10^{-8} in default, and can be modified by users.

5 Usage

For using parMix to infer ancestry and genotypes of parents, first simply typing:

```
$ parMix -h
```

-c	Number of Children
-p	Phasing Error Involved. 1: Without Phasing Error, Run Independent; 2: With Phasing Error
-b	Number of Base Pairs (Before Trimming)
-P	Position File Path
-C	Children Genotypes File Path
-A	Allele Frequencies File Path of Population A
-B	Allele Frequencies File Path of Population B
-D	LD File Path of Population A
-E	LD File Path of Population B
-R	Recombination Rate
-r	Phasing Error Rate (Used only when "-p" is "1")
-e	Genotyping Error Rate (Default is 10^{-8})

Table 1: the optional arguments for running parMix

It is worth mentioning that when you choose to run parMix in *"no phasing error"* mode, the input genotypes file should be modified to haplotypes file since there is no phasing error and parMix can infer the genotypes and ancestry for each parent independently.

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

```
$ parMix -c 3 -p 2 -b 259000000 -P example/testposition.dat -C example/testgenotypes.dat
  -A example/testAFA.dat -B example/testAFB.dat -D example/testLDA.dat -E example/testLDB.dat
  -R 0.00000001 -r 0.000002
```

In this case, parMix will infer the genotypes and ancestry of parents based on three children's genotypes (-c 3) under the phasing error involved mode (-p 2), and the number of base pairs before trimming is 2.59×10^8 (-b), the recombination rate is 10^{-8} (-R), the phasing error rate is 2×10^{-6} (-r), and the genotyping error rate is 10^{-8} (-e, in default) as well.

Or you can just use the shell file to compile parMix:

```
$ ./Command.sh
```

and the optional arguments can be modified in the shell file (**Command.sh**).

6 Outputs

The outputs of parMix are located in **Results** folder. There are three output files:

1. Result1Recom.dat: The inferred **Recombination** and **Phasing** sequences.
2. Result2Ance.dat: The inferred **Ancestry** sequence of parents.
3. Result3Geno.dat: The inferred **Genotypes** sequence of parents.

Moreover, the output files under the no phasing error model are similar to the files under the model with phasing errors, and they are "Result1RecomInd.dat", "Result2AnceInd.dat" and "Result3GenoInd.dat".

7 How to cite

The paper, "**Joint Inference of Ancestry and Genotypes of Parents from a Small Number of Children**" by Yiming Zhang and Yufeng Wu, is currently 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 parMix.

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.