**Tutorial of Simulation Tool**

This document is a tutorial for using PedMix Simulation Scripts.

1. **Generate raw data.**  
   One can us ‘ms’ by Hudson1 or ‘MaCS’2 to generate some samples in two or more populations. For example, if we want to generate ‘nsam’ individuals, and ‘nsam1’ of them were came from population A and the others (‘nsam2’) were came from population B, we can use the command as follow.

*“ms nsam nreps -t -r nsites -I nsam1 nsam2 -ej t 1 2”*

Please read the tutorial of ms to learn the meaning of the parameters and commands.

However, ‘ms’ cannot handle very long length SNPs. Thus, for whole genome simulation, we recommend use ‘MaCS’ instead of ‘ms’. ‘MaCS’ and ‘ms’ used the same model so usage should be similar.

1. **Processing the raw data.**  
   For using this tool to generate admixed individuals’ haplotypes, we need to make some input file for this tool. There are mainly two type of files. The first type of file is the positions of SNPs. This file includes the position information of each SNP, and the file name should be renamed as ‘prefix\_Recomrdc\_reshape.dat’, like ‘sim\_1000\_Recomrdc\_reshape.dat’ in this example. The second type of file is the individuals’ haplotypes. This file includes the haplotypes of one individual. For example, if we generated 22 chromosomes for one individual, then this file will have 44 haplotypes. Moreover, if we generated 500 individuals for 2 populations, then the first 250 individuals were belonged to population A and the last 250 individuals were belonged to population B. ‘sim\_1000\_indv1\_gen0\_Seq.dat’ is an example for one sample with 22 chromosomes.

‘simdata\_ms.py’ in this simulation tool includes some functions about the usage of ‘ms’, the usage of ‘macs’ and the information extraction from raw data files. One can use these functions in ‘main\_multi2.py’ for own purpose.

1. **Generate admixed samples.**  
   We can simply run this this tool to generate admixed individuals’ haplotypes. The command is shown as follow.

*‘python main\_multi2.py -prefix -n\_threads’*

where prefix is the prefix of file’s name, and it is ‘sim\_1000’ in this example. N-threads is the number of threads, and larger number of threads leads to higher computation speed.

After this step, many files will be generated. For example, ‘sim\_1000\_indv1\_gen10\_Seq.dat’ shows the haplotypes of one individual after 10 generations random mating. ‘sim\_1000\_indv1\_gen10\_AP.dat’ shows the ancestral population of this individual. ‘sim\_1000\_indv1\_gen10\_Rec.dat’ shows the parent information of this individual. Moreover, ‘sim\_1000\_Parent.dat’ shows that which two individuals were selected as parents for random mating in each generation. Usually, we simulate 10 generations random mating actions, but one can modify ‘Ng’ in code to get a result with different number of generations.

1. **Sample individuals.**

Once the random mating process was done, we can generate the inputs of PedMix using ‘SampleData.py’, and the command is shown as follow.

*‘python SampleData.py -prefix’*

However, due to some drawbacks of this tool, we recommend users generating the inputs of PedMix in their own way.

Please feel free to contact us if you have any questions.

Contact: yiming.zhang.cse@uconn.edu

Reference

1. Hudson, Richard R. "Generating samples under a Wright–Fisher neutral model of genetic variation." *Bioinformatics* 18.2 (2002): 337-338.
2. Chen, Gary K., Paul Marjoram, and Jeffrey D. Wall. "Fast and flexible simulation of DNA sequence data." Genome research 19.1 (2009): 136-142.