Manual for *SimAncestry*

SimAncestry is a script to extract archaic introgressed sequences from *ms* [1] simulation output. The script *SimAncestry* is only a simple script to analyze the tree structure outputted by the software *ms* [1]. It cannot apply to the real data as it was not designed to work with SNP data. The algorithm used in SimAncestry is as follows. For each non-African node, we make a transversal up to the root. If this node coalesces to archaic lineage before any African lineages, we say this node is introgressed. This script is used only to obtain the ground truth for introgressed sequences.

1 Getting Started

1.1 Library Dependency

To compile *simAncestry*, the following software / libraries are required,
g++, C++ source code compiler.
Boost Iostreams Library, boost library used to input and output compressed files. (https://www.boost.org/)
zlib, an open library for file compression and decompression. (https://zlib.net/)

1.2 Compile

It's very convenient to compile the software from the source code by the following commands:

$tar –zvxf simAncestry.tar.gz
$cd simAncestry
$make all

2. Software Arguments

--file <string>
      Path to the ms output file.
--out <string>
      Output prefix. The argument is to specify the output prefix. Introgressed sequences of each replicated simulation will be outputted to different files.
--archaic <num>
      Individual number of archaic hominins. For example, "101,102", "105-109", "101,102,105-109".
--african <num>
      Individual number of African references. For example, "101,102", "105-109", "101,102,105-109".
--test <num>

Individual number of tested modern humans. For example, "101,102", "105-109", "101,102,105-109".

--thread <num> [1]

Number of threads used in the analysis. Default value 1.

3.Examples

./simAncestry --file ms.log.gz --african 1-200 --test 201-400 --archaic 401,402 --out ms.output

1   Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-338 (2002).