Manual for *ArchaicSeeker 2.0*

ArchaicSeeker is a series of software for detecting archaic introgression sequences and reconstructing introgression history. The latest version of this series, *ArchaicSeeker 2.0*, has the following three notable improvements compared with the original version of this software. First, it can automatically determine the boundary of each introgressed sequence. Next, it is capable of tracing both known and unknown ancestral sources of a given introgressed sequence. Finally, it has the ability to reconstruct the introgression history with more sophisticated introgression models. Our new method consists of three modules, namely, seeking introgressed sequences, matching segments to proper ancestries, and reconstructing introgression history.

1 Getting Started

1.1 Library Dependency

To compile *ArchaicSeeker 2.0*, the following software / libraries are required,

g++, C++ source code compiler.

nlopt, nonlinear optimization library. (https://nlopt.readthedocs.io/en/latest/)

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

First, you need to modify the parameters specified by CFLAG (-I and -L) in the makefile into the path of installed library. Next, it's very convenient to compile the software from the source code by the following commands:

```
$tar –zvxf ArchaicSeeker2.tar.gz
$cd ArchaicSeeker2
$make all
```

1.3 Static Version

We also provide a static version of *ArchaicSeeker 2.0*, which was pre-compiled on Linux Redhat System and static linked to dependent libraries. This version is executable in Linux Systems.

2. Input Files

ArchaicSeeker 2.0 requires genomic information of Archaic Hominins, Africans, Test Non-Africans and outgroup (chimpanzee), recombination information, ancestral state information and several corresponding configuration files as input. In this section, we will explain file format of each input file.

2.1 Phased Genomic VCF Files

Our method requires haplotype genomic information as the input. The input files should be in VCF Version 4.2 and only bi-allelic SNPs allowed in the input files. Input haplotype genomes should include at least three population / group, Archaic Hominins, like Neanderthal or Denisovan, Africans, like YRI from the 1000 Genome Project, and the test Non-Africans.

Input data could be either splitted by chromosomes or combined together in a single file. Our software could get the intersection of data, automatically. Please put reference populations and target population into different VCF files. Phasing is required for the target population but is not necessary for the reference population.

The reference allele in the input VCF files should be same as that of reference genome.

To specify the paths to each input genomic files, a VCF configuration file is required. This configuration file contains every path to the input genomic VCF files. The header line of this file should be "vcf" and one vcf file path per line. ("vcf.par" in the example folder)

2.2 Recombination Maps

Recombination map consists three columns, physical position, recombination rate and genetic map coordinate. One recombination file contains recombination information from one chromosome. A recombination configuration file contains path to each recombination map and chromosome ID is required. The header line is "remap contig". In the following lines, the first column is the path to the recombination map and the second is the contig / chromosome ID. ("remap.par" in the example folder).

2.3 Population Annotation File

To specify the population information, a population annotation file is required as input. The header line of this file is "ID Pop ArchaicSeekerPop". In the following lines, the fist column is sample

ID, which corresponds to the ID in vcf files; the second column is population ID, which corresponds to the population in “model file (2.6)”; the third column is ArchaicSeeker population ID and it should be “African”, “Test” or “Archaic”, which corresponds to African reference, test non-African and archaic reference. (“pop.par” in the example folder)

2.4 Outgroup Genomic Files

In our method, we built-in a model calibration step to adjust the branch length of introgression model. An outgroup genome is required to set the ancestry / root of the model. Here, we recommend to use chimpanzee reference genome as the outgroup input. The format of the input genome should be in FASTA format and one chromosome / contig per file.

An outgroup configuration file is used to specify the path to each FASTA file. The header line is “outgroup contig”. In the following lines, the first column is the path to the outgroup FASTA genome and the second is the contig / chromosome ID. (“outgroup.par” in the example folder).

2.5 Ancestral State Files

To demine the allele state of each SNP, ancestral states of human genome are required. Like the outgroup genomic files, the information should be in FASTA format and a configuration file, containing path to each FASTA is required. The header line of the configuration file is “ancestor contig”. In the following lines, the first column is the path to the outgroup FASTA genome and the second is the contig / chromosome ID. (“anc.par” in the example folder).

The Outgroup Genomic Files and Ancestral State Files can be downloaded from https://drive.google.com/drive/folders/115LSXmYDlitNKDO58SgxbEYlNd4EG1WK?usp=sharing.

2.6 Model File

In this file, you should input the fitting model of your input genomic data. This model is in “Newick” tree format and each leaf node should be population ID, which presents in the population annotation file. (“model.txt” in the example folder)

3. Output Files

There are two types of output files. Introgression segment file (*seg) and individual introgression summary file (*sum).

In the segment file, there are 8 columns. They are haplotype ID, chromosome / contig ID, segment start / end in base pair, segment start / end in genetic distance (cM), best matched ancestry and best matched divergence time.

In the summary file, we calculated the introgression proportion from different ancestry for each individual.

4. Software Arguments

-a / --alpha <float> [0.02]

Introgression proportion argument. This argument is used to set the initial value of introgression proportion. The default value is 0.02.

-T / --introT <float> [2000]

Introgression time (in generation) argument. This argument is used to set the initial value of introgression time. The default value is 2000.

-e / --emit <float> [0.99]

Emission probability argument. This argument is used to set the initial value of HMM emission probability matrix. The default value is 0.99.

-o / --out <string>

Output prefix. The argument to specify the output prefix. No default values.

-v / --vcf <string>

VCF file configuration file arguments. No default values.

-p / --pop <string>

Population annotation file arguments. No default values.

-r / --remap <string>

Recombination configuration file arguments. No default values.

-X / --outgroup <string>

Outgroup genome configuration file arguments. No default values.

-A / --anc <stirng>

Ancestral state configuration file arguments. No default values.

-m / --model <string>

To specify the fitting model file. No default values.

-h / --help

Print this help.

5. Examples

We also provided an example in the software. It includes chr21 and chr22 of Denisovan, Neanderthal, African (YRI from the 1000 Genome Project) and Han (from SGDP). To run with this example, just execute the command line in the "run.sh".

```
./ArchaicSeeker2 -v examples/vcf.par -r examples/remap.par -m examples/model.txt -X examples/outgroup.par -p examples/pop.par -A examples/anc.par -o examples/Han
```

6. Questions and Trouble-shooting

Questions and suggestions are welcome, feel free to contact

Kai Yuan

yorkklause@gmail.com