Manual for ArchaicSeeker 1.0

This document describes how to use ArchaicSeeker, a program to detect archaic-like chromosomal segments in non-African modern human genomes.

This tool was originally developed to search archaic-like segments chromosomal segments in Tibetan genomes. Please refer to our paper which will appear soon, it will be updated in our publication list which can be found via the link here http://www.picb.ac.cn/PGG/index.php.

1 Getting Started

After extracting the .tar file, compile it with:

```
$make
```

Note that you must have “zlib” (http://www.zlib.net/) installed before the compile step.

2 Input Files

Three types of populations are required by ArchaicSeeker as input.

(1) Test population, the non-African modern human population of your interest.

(2) Archaic populations, the archaic hominins which are suspected to have contributed ancestry (gene flow) to the test population.

(3) African population, as a control which is assumed to be free of gene flow from archaic hominins.

All of the above data should be in “vcf” format (http://samtools.github.io/hts-specs/VCFv4.2.pdf) and phased (haplotype known). Input files could be either compressed (compressed with gzip or bgzip) or in plain format.

3 Output Files

3.1 Summary File

This file stores the total length of archaic segments account for whole genome.

```
[Individual ID]_[Haplotype ID]      [Length]
…
```

For example,

```
ID1_1     10000
ID1_2     12000
```

…
It means the total length of archaic-like sequence in the first haplotype of ID1 is 10kb and that of ID1's second haplotype is 12kb.

3.2 Segment File

This file stores the start and end position of each archaic segments.

```
[Chromosome ID]   [Start Position]       [End Position]  [Individual ID]_[Haplotype ID]
…
```

For example,

```
1     1000     2000     ID1_1
1     2000     3000     ID1_2
…
```

It means that chr1:1000-2000 of ID1's first haplotype is an archaic-like segment.

4 Command Line

```
./ArchSeeker -Afr <Afr_file> -Arch <Arch_file> -Test <Test_file> -BinSize <int> -AfrRank <int> -Summary <Sum_file> -Seg <Seg_file> -MinBinSites <int>
```

-Afr <Afr_file>	Path of African population input file.
-Arch <Arch_file>	Path of Archaic population input file.
-Test <Test_file>	Path of non-African modern human population input file.
-BinSize <int>	Bin size; Default = 10k.
-AfrRank <int>	Rank of similarity between non-African and African; Default = 0, means minimum of that similarity. [0, NumAfrHap – 1]
-Summary <Sum_file>	Path of output summary file.
-Seg <Seg_file>	Path of output segment file.
-MinBinSites <int>	Minimum number of sites in a bin; Default = 10.
-gzInput	Whether input files are compressed.