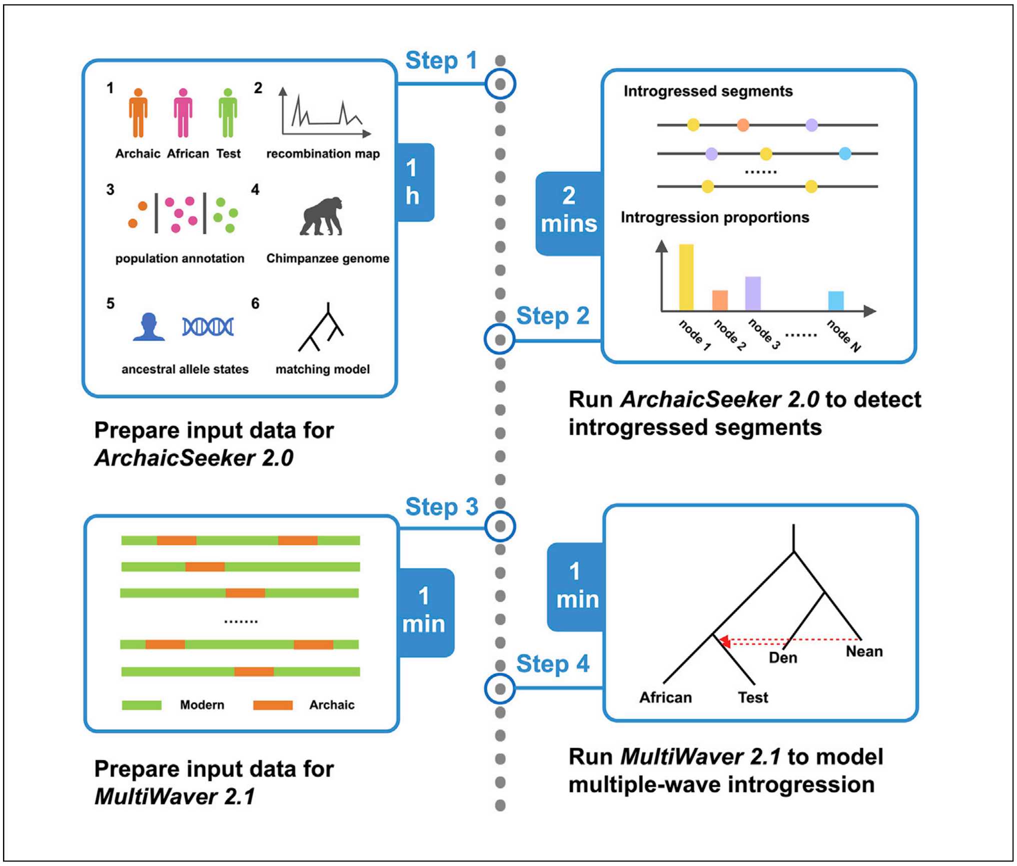
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Detecting archaic introgression and modeling multiple-wave admixture with ArchaicSeeker 2.0


Step 1
1 Archaic African Test
2 recombination map
3 population annotation
4 Chimpanzee genome
5 ancestral allele states
6 matching model
1 h
Prepare input data for *ArchaicSeeker 2.0*
2 mins
Introgressed segments
Introgression proportions
node 1 node 2 node 3 ...... node N
Step 2
Run *ArchaicSeeker 2.0* to detect introgressed segments
Step 3
Modern Archaic
1 min
Prepare input data for *MultiWaver 2.1*
1 min
Step 4
African Test Den Nean
Run *MultiWaver 2.1* to model multiple-wave introgression

ArchaicSeeker 2.0 is designed to identify the sequences derived from known or unknown archaic hominins and to further model the multiple-wave gene flows. The main functional modules involve seeking introgressed sequences, determining corresponding ancestries, and reconstructing the admixture history. The protocol below describes the analytic steps for the application of ArchaicSeeker 2.0 into analysis of example data of the Han Chinese population.

Rui Zhang, Kai Yuan, Shuhua Xu

xushua@fudan.edu.cn

Highlights

ArchaicSeeker 2.0 can identify introgressed sequences from archaic hominins

Modeling complex introgression history without massive computer simulation

Overcomes the limitations of existing methods in inferring multiple introgression

Zhang et al., STAR Protocols 3, 101314
June 17, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.xpro.2022.101314

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Detecting archaic introgression and modeling multiple-wave admixture with ArchaicSeeker 2.0

Rui Zhang,[1,5] Kai Yuan,[1] and Shuhua Xu[2,3,4,6,*]

[1]Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China

[2]State Key Laboratory of Genetic Engineering, Center for Evolutionary Biology, Collaborative Innovation Center of Genetics and Development, School of Life Sciences, Fudan University, Shanghai 200438, China

[3]Human Phenome Institute, Zhangjiang Fudan International Innovation Center, and Ministry of Education Key Laboratory of Contemporary Anthropology, Fudan University, Shanghai 201203, China

[4]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

[5]Technical contact

[6]Lead contact

*Correspondence: xushua@fudan.edu.cn

https://doi.org/10.1016/j.xpro.2022.101314

SUMMARY

ArchaicSeeker 2.0 is designed to identify the sequences derived from known or unknown archaic hominins and to further model the multiple-wave gene flows. The main functional modules involve seeking introgressed sequences, determining corresponding ancestries, and reconstructing the admixture history. The protocol below describes the analytic steps for the application of ArchaicSeeker 2.0 into analysis of example data of the Han Chinese population. For complete details on the use and execution of this protocol, please refer to Yuan et al. (2021).

BEFORE YOU BEGIN

Download the software ArchaicSeeker 2.0

Timing: 3 min

1. Download the toolkit ArchaicSeeker 2.0 from the following GitHub website: https://github.com/Shuhua-Group/ArchaicSeeker2.0. It includes the source code of ArchaicSeeker 2.0, a folder named as WaveEstimate for modeling the multiple-wave admixture with MultiWaver 2.1, a folder named as examples for illustrating the usages, the Manual for ArchaicSeeker 2.0 in a pdf format, a LICENSE file, and a README file.

Download the dependent libraries

Timing: 2 h

2. Download the following dependent libraries:
   a. nlopt, nonlinear optimization library (https://nlopt.readthedocs.io/en/latest/).
   b. Boost Iostreams Library, boost library used to input and output compressed files (https://www.boost.org/).
   c. zlib, an open library for file compression and decompression (https://zlib.net/).

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Altai Denisovan genomes	(Meyer et al., 2012)	http://cdna.eva.mpg.de/neandertal/altai/Denisovan
Altai Neanderthal genomes	(Prufer et al., 2014)	http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal
Yoruba in Ibadan, Nigeria (YRI) genomes	The 1000 Genome Project (The 1000 Genomes Project Consortium, 2015)	http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
Test data	The Simons Genome Diversity Project (Mallick et al., 2016)	https://www.ncbi.nlm.nih.gov/bioproject/PRJEB9586/
HapMap genetic map	The International HapMap Consortium (The International HapMap Consortium, 2007)	https://github.com/Shuhua-Group/ArchaicSeeker2.0/tree/master/WaveEstimate/MWexamples/genetic_map_b37
Chimpanzee reference genome	The Chimpanzee Sequencing and Analysis Consortium (The Chimpanzee Sequencing and Analysis Consortium, 2005)	https://drive.google.com/drive/folders/115LSXmYDlitNKDO58SgxbEYlNd4EG1WK?usp=sharing
Ancestral allele states of the human genome	Ensembl database	https://drive.google.com/drive/folders/115LSXmYDlitNKDO58SgxbEYlNd4EG1WK?usp=sharing
Software and algorithms		
GCC	The GNU Compiler Collection	http://gcc.gnu.org
ArchaicSeeker 2.0	(Yuan et al., 2021)	https://github.com/Shuhua-Group/ArchaicSeeker2.0, https://doi.org/10.5281/zenodo.5526693
MultiWaver 2.1	(Yuan et al., 2021)	https://github.com/Shuhua-Group/ArchaicSeeker2.0/WaveEstimate/MultiWaver2.1, https://doi.org/10.5281/zenodo.5526693
Other		
Linux server	N/A	N/A

MATERIALS AND EQUIPMENT

- In this protocol, we used the Han from the Simons Genome Diversity Project (SGDP) (Mallick et al., 2016) as the test non-African population, the Yoruba in Ibadan, Nigeria (YRI) from the 1000 Genome Project (KGP) as the representative population of Africans, the Altai Neanderthal and the Altai Denisovan as the archaic hominins, and the Chimpanzee genome as the outgroup population (see Key resources table).
- The sequence data of Chromosomes 21 and 22 were used to illustrate the usage.
- All tests in this protocol were conducted on 64-core Intel Xeon CPU E7-4850 v4 2.10 GHz Linux servers.

STEP-BY-STEP METHOD DETAILS

Install ArchaicSeeker 2.0 and MultiWaver 2.1

Timing: 4 min

1. Compile the ArchaicSeeker *2.0* (Troubleshooting 1 and 2).

First modify the parameters specified by CFLAG (-I and -L) in the makefile into the path of installed libraries, and then type the following commands to compile:

```
> cd ArchaicSeeker2.0

> make clean

> make all
```

Note: This is an optional step. We have provided the static version of ArchaicSeeker 2.0 and it is executable in Linux Systems (Troubleshooting 3).

2. Compile the getAS2Seg and MutliWaver 2.1 (Troubleshooting 4).

```
> cd WaveEstimate/getAS2Seg
> rm getAS2Seg
> g++ getAS2Seg.cpp -o getAS2Seg
> cd ../MultiWaver2.1
> make clean
> make
```

Note: This is an optional step. We have provided the static version of getAS2Seg and Multi-Waver 2.1 and they are executable in Linux Systems.

Prepare input data for ArchaicSeeker 2.0

⏱ Timing: 1 h

⚠ CRITICAL: The software ArchaicSeeker 2.0 takes the following six files respectively regarding to the information of several VCF files (archaic hominins, African population, and the test non-African population), recombination map, population annotation, outgroup population, ancestral allele states, prior matching model as input (Troubleshooting 7).

3. Prepare the configuration file about the phased genomic VCF files.

Here is the VCF configuration file ''vcf.par'' in the examples folder:

```
vcf
examples/Neanderthal.KGP3.combined.21.vcf.gz
examples/Neanderthal.KGP3.combined.22.vcf.gz
examples/Denisova.KGP3.combined.21.vcf.gz
examples/Denisova.KGP3.combined.22.vcf.gz
examples/YRI.chr21.vcf.gz
examples/YRI.chr22.vcf.gz
examples/Han.chr21.vcf.gz
examples/Han.chr22.vcf.gz
```

⚠ CRITICAL: The VCF configuration file should begin with the ''vcf'' and arrange the paths to the genomic VCF files in the following rows.

Note: Input genomic VCF files should be phased and at least include the archaic hominins, African, and the test non-African population (Troubleshooting 5 and 6).

Note: Physical positions cannot be duplicated and should be strictly increasing in one single VCF File. Only biallelic SNPs are allowed and the reference allele in the input VCF files should be same as that of ancestral allele states.

Note: The order of input VCF files can be arbitrary and random. Nonetheless, we recommend that for the same chromosome, make the VCF file with fewer SNPs in front as our algorithm can take the intersection of the input VCF files automatically.

4. Prepare the configuration file regarding the recombination maps.

Here is the recombination configuration file “remap.par” in the examples folder:

```
remap                                         contig
examples/genetic_map_chr21_combined_b37.txt   21
examples/genetic_map_chr22_combined_b37.txt   22
```

⚠ CRITICAL: The recombination configuration file consists of two columns and should start with the “remap contig”. The paths to the recombination file and corresponding Chromosome ID are arranged in the following rows.

Note: The recombination map file consists of three columns representing the physical position (unit: bp), recombination rate (unit: cM/Mb), and genetic distance (unit: cM).

Note: Recombination information of different chromosomes should be put separately in different files.

5. Prepare the population annotation files.

Here show the several lines of the population annotation file “pop.par” in the examples folder:

```
ID              Pop             ArchaicSeekerPop
DenisovaPinky   Denisova        Archaic
AltaiNea        Neanderthal     Archaic
NA18486         YRI             African
NA18488         YRI             African
……              ……              ……
NA19257         YRI             African
Han-3           Modern          Test
Han-1           Modern          Test
Han-2           Modern          Test
```

⚠ CRITICAL: The header line of the population annotation files should be the “ID Pop ArchaicSeekerPop”.

Note: This file demonstrates the individual ID and corresponding population label in the prior matching model as well as the population label specified by ArchaicSeeker 2.0, one individual per line.

Note: The population label specified by ArchaicSeeker 2.0 should be one of the "Archaic", "African", and "Test", which respectively corresponds to the archaic hominins, African population, and test non-African population.

6. Prepare the configuration file regarding the outgroup genomic files.

Here we show the outgroup configuration file "outgroup.par" in the examples folder:

```
outgroup                          contig

examples/chr21.hg19.chimp.fa.gz   21

examples/chr22.hg19.chimp.fa.gz   22
```

⚠ CRITICAL: The outgroup configuration file starts with "outgroup contig". The following lines specify the path to each outgroup genomic file and corresponding chromosome ID, one file per line.

Note: The genomes of the outgroup population are used to determine the root of the model. The Chimpanzee reference genome is recommended to represent the outgroup population. The input genome should be in FASTA format and one chromosome per file.

7. Prepare the configuration file regarding the ancestral allele state files.

Here is the ancestral allele states configuration file "anc.par" in the examples folder:

```
ancestor                                  contig

examples/homo_sapiens_ancestor_21.fa.gz   21

examples/homo_sapiens_ancestor_22.fa.gz   22
```

⚠ CRITICAL: The ancestral allele states configuration file starts with "ancestor contig". The following lines specify the path to each ancestral allele states file and corresponding chromosome ID, one file per line.

Note: The ancestral state files are used to determine the ancestral allele state of each SNP. Same as the above outgroup genomic file, the information should be in FASTA format and one chromosome per file.

8. Prepare the matching model file.

Here we show the matching model file "model.txt" in the examples folder and the corresponding tree topology (Figure 1):

((YRI:100,Modern:100):557.5,(Denisova:340,Neanderthal:300):237.5);

Note: The matching model describes the prior phylogenetic relationship information of input populations. The model is in the "Newick" tree format and the leaf node is denoted by the population ID. The unit of time could be in generations, one thousand years, or others specified by the users. In the above example file, the unit is in one thousand years.

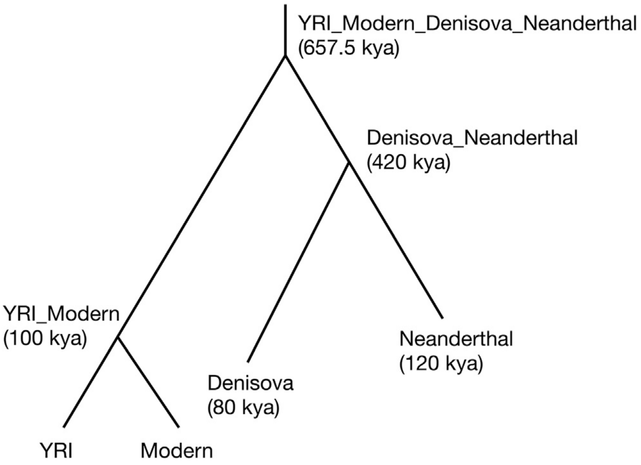


Figure 1. Corresponding tree topology of the example matching model

Note: With the addition of Vindija Neanderthal, the matching model could be as follows: ((YRI:100,Modern:100):557.5,(Denisova:340,(Vindija:82.5,Neanderthal:20):280):237.5);.

Note: For estimating the archaic introgression proportion in the non-African modern human population, the above model example can be used. Actually, the value of these divergence times is not so important, since our algorithm can automatically correct the branch length of the matching model based on the input data.

Run ArchaicSeeker 2.0 to detect archaic introgression segments

Timing: 2 min

The arguments of ArchaicSeeker 2.0 are listed in Table 1.

9. Run ArchaicSeeker 2.0 to detect the archaic introgressed segments.

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

Prepare input data for MultiWaver 2.1

Timing: 1 min

10. Convert the information output by ArchaicSeeker 2.0 (Troubleshooting 8).

Here is an example to model the introgression pattern from the Denisova:

```
> cd WaveEstimate

> getAS2Seg/getAS2Seg MWexamples/genetic_map_b37/ ../examples/Han.seg MWexamples/Han
Denisova
```

Table 1. Arguments of ArchaicSeeker 2.0

Arguments	Type	Description	Note
`-v/ -vcf`	string	VCF configuration file	required
`-r/ -remap`	string	recombination configuration file	required
`-p/ -pop`	string	population annotation file	required
`-X/ -outgroup`	string	outgroup genome configuration file	required
`-A/ -anc`	string	ancestral state configuration file	required
`-m/ -model`	string	matching model file	required
`-o/ -out`	string	output prefix	required
`-a/ -alpha`	double	introgression proportion	optional, default = 0.02
`-T/ -introT`	integer	introgression time (in generation)	optional, default = 2000
`-e/ -emit`	double	emission probability parameter	optional, default = 0.99
`-h/ -help`	n/a	print help message	optional

⚠ CRITICAL: The script "getAS2Seg" takes four parameters as input respectively representing the folder path to the genetic map files, the file path to the segment file output by ArchaicSeeker 2.0, the prefix of the output file, and specific archaic hominin.

Note: In order to model the multiple-wave admixture model from a certain archaic hominin, firstly use the executable script "getAS2Seg" in the WaveEstimate folder to convert the information output by ArchaicSeeker 2.0 to the input format of MultiWaver 2.1.

Note: MultiWaver 2.1 is a modified version of MultiWaver 2.0 (Ni et al., 2019) and is specifically used to estimate the number of archaic introgression waves utilizing the information output by ArchaicSeeker 2.0.

Run MultiWaver 2.1 to model multiple-wave introgression

⏲ Timing: 1 min

The arguments of MultiWaver 2.1 are listed in Table 2.

11. Infer the multiple-wave introgression model (Troubleshooting 9 and 10).

```
> MultiWaver2.1/MultiWaver2.1 -i MWexamples/Han.seg -l 0.00015 -o MWexamples/Han
```

EXPECTED OUTCOMES

The output information of ArchaicSeeker 2.0 includes two files representing the inferred segments of archaic introgression (*.seg) and the summary of introgression proportion on individual level (*.sum).

The first file with the suffix "seg" contains eight columns representing the haplotype ID [ID], the Chromosome ID [Contig], the physical position of start point [Start (bp)], the physical position of endpoint [End (bp)], the genetic distance of start point [Start (cM)], the genetic distance of endpoint [End (cM)], best-matched ancestry [BestMatchedPop], and best-matched

Table 2. Arguments of MultiWaver 2.1

Arguments	Type	Description	Note
`-i/ -input`	string	ancestral tracks file	required
`-l/ -lower`	double	lower bound to discard short tracks	optional, default = 0
`-b/ -bootstrap`	Integer	Number of bootstrapping	optional, default = 100
`-a/ -alpha`	double	a significant level to reject the null hypothesis in LRT	optional, default = 0.001
`-e/ -epsilon`	double	epsilon to check whether a parameter converges or not	optional, default = 1.0e-6
`-p/ -minProp`	double	minimum survival proportions for a wave at the final generation	optional, default = 0.05
`-m/ -maxIter`	integer	maximum number of iterations to scan for waves of admixture events	optional, default = 10000
`-t/ -thread`	Integer	Number of threads	optional, default = 1
`-o/ -output`	string	Prefix of output	required
`-h/ -help`	n/a	print help message	optional

divergence time [BestMatchedTime]. Here show the first several lines of "Han.seg" in the examples folder:

ID contig	Start (bp)	Start (bp)	End (bp)	Start (cM)	End (cM)	BestMatchedPop	BestMatchedTime
Han-1_1	21	16318684	16448591	3.59117	3.74739	Neanderthal	1e-128
Han-1_1	21	16489322	16520650	3.96965	4.00165	Denisova	211.474
Han-1_1	21	16968574	16973470	4.95173	4.95568	Denisova_Neanderthal	87.8493
Han-1_1	21	17059207	17064026	5.16393	5.19943	Neanderthal	0.450056
Han-1_1	21	17776840	17828310	6.78231	6.85602	Neanderthal	1e-128

The second file with the suffix "sum" demonstrates the estimated introgression proportion originating from each tree node, one individual per line. Here show the "Han.sum" in the examples folder (Due to the space limitations, we omit the unit (cM) after each node, use "Den" to represent Denisova, use "Nean" to present Neanderthal, and keep three decimals for the introgression proportion.):

ID	YRI_Modern_Den_Nean	YRI_Modern	YRI	Modern	Den_Nean	Den	Nean
Han-1	0.322%	0%	0.002%	0.187%	0.029%	0.161%	1.043%
Han-2	0.212%	0%	0%	0.218%	0.001%	0.264%	0.934%
Han-3	0.277%	0%	0%	0.101%	0.023%	0.146%	0.765%

The output information of the script getAS2Seg contains three columns representing the genetic distance of the start point (unit in Morgan), the genetic distance of endpoint (unit in Morgan), and the population label ("Modern" or "Archaic"). Each row represents a sequence segment. Here we show the first several lines of "Han.seg" in the folder MWexamples:

```
0                         0.03969647197121846382 6     Modern
0.03969647197121846382 6  0.04001650432907340182 4     Archaic
0.04001650432907340182 4  0.10752631667002321425       Modern
0.10752631667002321425    0.10843499206584480965       Archaic
0.10843499206584480965    0.1280532388696065671        Modern
0.1280532388696065671     0.12819095752416279077       Archaic
```

The output information of MultiWaver 2.1 includes two files respectively demonstrating the parameters of the admixture model (*.sum) and corresponding bootstrapping details (*.bootstrap).

The first one (*.sum) contains five columns: type of dataset [DataSet], supporting ratio [SupportRatio], number of supporting times [SupportNum], number of introgression wave from certain archaic hominin [NumOfArchaic], estimated introgression time (proportion) [Time (Generation)]. Here show the ''Han.sum'' in the folder WaveEstimate/MWexamples/:

Dataset	SupportRatio	SupportNum	NumOfArchaic	Time (generation)
CompleteData	–	–	1	1729.88 (0.00202044)
BootstrapData	100%	100	1	1139.16~2956.12 (0.00202044~0.00202044)

The second file (*.bootstrap) contains three columns representing the index number of bootstrapping time [NumBootstrap], number of introgression wave from certain archaic hominin [NumArchaic], and estimated introgression time (proportion) [Time]. Here show the ''Han.bootstrap'' in the folder WaveEstimate/MWexamples/:

NumBootstrap	NumArchaic	Time
0	1	1756.5 (0.00202044)
1	1	1318.61 (0.00202044)
2	1	2491.55 (0.00202044)
97	1	1355.2 (0.00202044)
98	1	1640.67 (0.00202044)
99	1	1799.86 (0.00202044)

LIMITATIONS

Our tool is especially suitable for estimating the archaic hominins introgression proportion in the non-African modern human populations. One of the basic assumptions of ArchaicSeeker 2.0 is that African populations did not receive any gene flow from archaic hominins. Still, several studies have recently identified the gene flow from Neanderthal to Africans (Chen et al., 2020). Under the circumstance, this might affects the power of the estimation on introgression proportion from archaic hominins in modern humans. Besides, since the divergence time of sequenced Altai Denisovan and the historically introgressed one is relatively large, the estimation on the introgression proportion from the Denisovan could be underestimated like other methods.

Since the methodology of ArchaicSeeker 2.0 to identify the introgressed segments is mainly based on the Hidden Markov Model (HMM), the accuracy of parameters estimation could be affected by the small marker size of the Test population. Therefore, we recommend using the high coverage sequence data to perform the inference. Besides, the performance on inferring the archaic introgressed history could be affected if the sample size or the number of detected introgressed segments of the Test population is too small. Still, the time-consuming of ArchaicSeeker 2.0 can be relatively long if the sample size of the Test population is very large.

TROUBLESHOOTING

Problem 1

During the compilation of ArchaicSeeker 2.0, the following error message is output (step 1):

```
g++ -c data.cpp -I/home/unix/kyuan/link/102.InstalledSoftwares/include/ -L/home/unix/
kyuan/link/102.InstalledSoftwares/lib/ -lnlopt -lzIn file included from data.cpp:9:
gzfstream.hpp:12:11: fatal error: boost/iostreams/categories.hpp: No such file or directory
```

Potential solution

Check whether the Boost Iostreams library is installed properly. We here provide an instruction if the user has the sudo permission, https://www.howtoinstall.me/ubuntu/18-04/libboost-iostreams-dev/.

Modify the parameters specified by CFLAG (-I and -L) in the makefile into the path of installed libraries.

Use the static version of ArchaicSeeker 2.0.

Problem 2

During the compilation of ArchaicSeeker 2.0, the following error message is output (step 1):

```
In file included from matching.cpp:8:0: matching.hpp:17:20: fatal error: nlopt.h: No such file or directory.
```

Potential solution

Check whether the nlopt library is installed properly. Here is a documentation for installing the nlopt: https://nlopt.readthedocs.io/en/latest/NLopt_Installation/. Also, the user can refer to the instruction for solving this problem: https://blog.csdn.net/hunter___/article/details/103610841.

Use the static version of ArchaicSeeker 2.0.

Problem 3

When using the static version of ArchaicSeeker 2.0, the following error message is the output (step 1):

```
./ArchaicSeeker2: error while loading shared libraries: libnlopt.so.0: cannot open shared object file: No such file or directory
```

Potential solution

First, the user is suggested to check whether the nlopt library is installed properly. Next, the user needs to tell the dynamic linker where to look for the libraries and specify the file path of libnlopt.so.0 in the profile document. Here is an instruction for solving such problems: https://stackoverflow.com/questions/5357869/error-while-loading-shared-libraries.

We also provide another static version of ArchaicSeeker 2.0 executable file which is independent of the shared library libnlopt.so but with a larger memory size (https://github.com/Shuhua-Group/ArchaicSeeker2.0/tree/master/libnlopt.so.0_free).

Problem 4

During the compilation of MultiWaver 2.1, the following error message is output (step 2):

```
g++ -c -o EMExp.o EMExp.cppg++ -c -o MultiWaver.o MultiWaver.cpp g++ -c -o ParamExp.o ParamExp.cpp g++ -c -o Utils.o Utils.cpp Utils.cpp:11:10: fatal error: boost/math/distributions/chi_squared.hpp: No such file or directory
```

Potential solution

Check whether the Boost Iostreams library is installed properly and the compiler can find these installed libs during the compilation. Here is an instruction if the user has the sudo permission: https://www.howtoinstall.me/ubuntu/18-04/libboost-math-dev/.

Use the static version of MultiWaver 2.1.

Problem 5

The human assembly version of input VCF genomic data is not in GRCh37 but the assembly version of currently available archaic hominins data is in GRCh37 (step 3).

Potential solution

Use the software liftOver (Kuhn et al., 2013), GenomeWrap (McLean et al., 2019), or other tools to do the conversion. ArchaicSeeker 2.0 requires that all the input data are in the same version (all GRCh37 or all GRCh38). The choice of the human assembly version does not make any difference to our algorithms.

Problem 6

The genotype data of input VCF is unphased (step 3).

Potential solution

Use the software SHAPEIT (O'Connell et al., 2014), Beagle (Browning et al., 2018), or other tools to do the phase.

Problem 7

ArchaicSeeker 2.0 report error message: Aborted (core dumped) (steps 3–8).

Potential solution

- Check whether the input arguments and corresponding parameters are correct.
- Check whether there are repeated physical positions in one single VCF file.
- Check whether the order of physical position is strictly increasing in one single VCF file.
- Check whether the header line and file path in each configuration file are correct.
- Check whether the format of each input file conforms to the standard.

Problem 8

The script getAS2Seg fails (step 10).

Potential solution

- Check whether the four input parameters of getAS2Seg are correct. Specifically, check whether the file path to the genetic map, the path to segment file output by ArchaicSeeker 2.0, the prefix of output file, and the specified archaic hominin are correct.
- Check whether the format of each input file is correct.

Problem 9

The script MultiWaver 2.1 fails (step 11).

Potential solution

- Check whether the input parameters of MultiWaver 2.1 are correct. Specifically, check whether the path to input ancestral tracks file and prefix of output are correct.
- Check whether the format of the input segment file conforms to the standard.

Problem 10

The running time of MultiWaver 2.1 is too long (step 11).

Potential solution

Filter some segments less than a certain length, like 0.00015 Morgan.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Shuhua Xu (xushua@fudan.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The datasets and code used during this study are available at https://github.com/Shuhua-Group/ArchaicSeeker2.0 or https://doi.org/10.5281/zenodo.5526693.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (NSFC) grants (32030020, 32041008, and 31961130380), the Strategic Priority Research Program (XDPB17, XDB38000000) of the Chinese Academy of Sciences (CAS), the UK Royal Society-Newton Advanced Fellowship (NAF\R1\191094), and the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01). The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

S.X. conceived and supervised the project. R.Z. prepared the protocol. K.Y. and R.Z. wrote the computer code. R.Z. prepared the draft of the manuscript and the additional materials. S.X. revised the manuscript. All authors read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A one-penny imputed genome from next-generation reference panels. Am. J. Hum. Genet. *103*, 338–348.

Chen, L., Wolf, A.B., Fu, W., Li, L., and Akey, J.M. (2020). Identifying and interpreting apparent neanderthal ancestry in African individuals. Cell *180*, 677–687.e616.

Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Brief. Bioinform. *14*, 144–161.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature *538*, 201–206.

McLean, C.Y., Hwang, Y., Poplin, R., and DePristo, M.A. (2019). GenomeWarp: an alignment-based variant coordinate transformation. Bioinformatics *35*, 4389–4391.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., De Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic denisovan individual. Science *338*, 222–226.

Ni, X., Yuan, K., Liu, C., Feng, Q., Tian, L., Ma, Z., and Xu, S. (2019). MultiWaver 2.0: modeling discrete and continuous gene flow to reconstruct complex population admixtures. Eur. J. Hum. Genet. *27*, 133–139.

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. *10*, e1004234.

Prufer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai mountains. Nature *505*, 43–49.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

The Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature *437*, 69–87.

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

Yuan, K., Ni, X., Liu, C., Pan, Y., Deng, L., Zhang, R., Gao, Y., Ge, X., Liu, J., Ma, X., et al. (2021). Refining models of archaic admixture in Eurasia with ArchaicSeeker 2.0. Nat. Commun. *12*, 6232.