

ancestryinfer 3way pipeline

Data can be input into the *ancestryinfer* pipeline to run local ancestry inference following (Corbett-Detig & Nielsen, 2017).

Install

Option 1 – install dependencies:

```
git clone https://github.com/Schumerlab/ancestryinfer.git
```

To install dependencies, follow instructions outlined in:

```
installation_instructions.txt
```

Test that the install and pipeline are working:

```
cd ancestryinfer
```

```
perl Ancestry_HMM_parallel_v6.pl  
hmm_configuration_file_nonparallel.cfg
```

Option 2 – load docker file for dependencies:

```
docker pull schumer/mixnmatch-ancestryinfer-image:mixnmatch-  
ancestryinfer-docker
```

```
docker run -it mixnmatch-ancestryinfer-image bash
```

Test that the install and pipeline are working:

```
cd ancestryinfer
```

```
perl Ancestry_HMM_parallel_v6.pl  
hmm_configuration_file_nonparallel.cfg
```

Setting parameters in the configuration file

There are example configuration files available on github.

Example files for 2-way local ancestry inference:

```
hmm_configuration_file_parallel.cfg
```

```
hmm_configuration_file_nonparallel.cfg
```

Example files for 3-way local ancestry inference:

hmm_configuration_file_3way_parallel.cfg

hmm_configuration_file_3way_nonparallel.cfg

Parameter descriptions:

Parameter	Description	Example	Include if
genome1=	User provided fasta file for species 1	genome1=xiphophorus_birchmanni_10x_12Sep2018_yDAA6.fasta	Always
genome2=	User provided fasta file for species 2	genome2=Xmalinche_dovetail_assembly.fa	Always
genome3=	User provided fasta file for species 3	genome2=Xvariatus_10x_assembly.fa	Required if performing 3 way ancestry calling
read_type=	Indicated whether data is paired end or single end	read_type=PE	Always
read_list=	Provide list (including full paths) to the reads to be analyzed	read_list=combined_all_call_hybrids_read_list example list format for paired end data (single end file should contain one line per individual): ./reads/CALL1_read1.fq.gz\t./reads/CALL1_read2.fq.gz ./reads/CALL2_read1.fq.gz\t./reads/CALL2_read2.fq.gz ./reads/CALL3_read1.fq.gz\t./reads/CALL3_read2.fq.gz	Always
read_length=	Provide expected read length. If read lengths of input samples differ, use the longest read lengths.	read_length=150	Always
mapping_quality=	Required mapping quality for a read to be retained	mapping_quality=20	Optional, if not specified the mapping quality threshold used is 30
prop_genome_genome1_parent=	Expected proportion of the genome derived from the parent species listed under genome1	prop_genome_genome1_parent=0.5	If not provided, AncestryHMM will attempt to estimate (may increase run time)
prop_genome_genome2_parent=	Expected proportion of the	prop_genome_genome2_parent=0.5	If not provided, AncestryHMM

	genome derived from the parent species listed under genome2		will attempt to estimate if 3 way ancestry calling is being performed (may increase run time)
<code>number_indiv_per_job=</code>	Parallelize jobs such that each job processes this number of individuals. Low numbers mean high parallelization and high number mean low parallelization.	<code>number_indiv_per_job=1</code>	Always
<code>program_path=</code>	Path to the program install folder	<code>program_path=/home/groups/schumer/shared_bin/Ancestry_HMM_pipeline</code>	If not provided the program will assume necessary scripts and programs are in the working directory
<code>provide_AIMs=</code>	Coordinates and identities of ancestry informative sites that distinguish the two parent species	<code>provide_AIMs=Xbirchmanni10xgenome_ancestry_informative_sites_filterF1</code> Example list format for 2 or 3-way ancestry calling: <pre> ScyDAA6-2-HRSCAF-26 58345 T C ScyDAA6-2-HRSCAF-26 58976 T A ScyDAA6-2-HRSCAF-26 59896 T C ScyDAA6-2-HRSCAF-26 60164 G A ScyDAA6-2-HRSCAF-26 63105 G A ScyDAA6-2-HRSCAF-26 65532 G A ScyDAA6-2-HRSCAF-26 66290 C A ScyDAA6-2-HRSCAF-26 68233 T C ScyDAA6-2-HRSCAF-26 70398 G A ScyDAA6-2-HRSCAF-26 73869 G A </pre>	Required unless provided genomes are on the same coordinate system and can be auto detected <i>Always required for 3 way ancestry calling</i>
<code>provide_counts=</code>	Counts of parental allele frequencies at ancestry informative sites (and recombination rates between adjacent sites if available)	<code>provide_counts=Xbirchmanni10xgenome_Xmalinche_observed_parental_counts_filterF1</code> Example format for 2-way ancestry calling: <pre> ScyDAA6-2-HRSCAF-26 163722 129 3 0 54 0.00000078 ScyDAA6-2-HRSCAF-26 166158 135 5 0 54 0.00001374 ScyDAA6-2-HRSCAF-26 166535 6 0 0 6 0.00000754 </pre> Columns are:	If not provided, the program will assume that provided ancestry informative sites are fixed between species <i>Always required for 3 way ancestry calling</i>

		Chromosome site allele1_count_parent1 allele2_count_parent1 allele1_count_parent2 allele2_count_parent2 recombination_rate Example format for 3-way ancestry calling: ScyDAA6-2-HRSCAF-26 230490 12 0 0 11 94 0 4.1e-06 ScyDAA6-2-HRSCAF-26 240248 220 6 0 54 74 0 8.78e-06 ScyDAA6-2-HRSCAF-26 240783 218 6 0 54 59 0 4.26e-06 ScyDAA6-2-HRSCAF-26 241074 217 5 0 54 27 0 5.82e-06 Chromosome site allele1_count_parent1 allele2_count_parent1 allele1_count_parent2 allele2_count_parent2 allele1_count_parent3 allele2_count_parent3 recombination_rate	
per_site_error =	Per-site error parameter for HMM (i.e. due to sequencing error, contamination, etc)	per_site_error=0.02	Always
gen_initial_admix_p1_p2=	Estimated generation of initial admixture of genome 1 and genome 2 species	gen_initial_admix_p1_p2=20	If not provided, AncestryHMM will attempt to estimate (may increase run time)
gen_initial_admix_p3=	Estimated generation of initial admixture of genome 3 species	gen_initial_admix_p3=60	If not provided and 3 way admixture is indicated, AncestryHMM will attempt to estimate
focal_chrom_list=	Provide a list of chromosomes to run (other chromosomes will not be run)	focal_chrom_list=mychrs.txt Example: ScyDAA6-2-HRSCAF-26 ScyDAA6-7-HRSCAF-50	Not required
rec_M_per_bp=	Estimated recombination rate in Morgans/bp	rec_M_per_bp=0.00000002	Always; Use an estimate for a related species if not available
max_alignments =	Limit analysis to a maximum number of alignments (for	max_alignments=2000000	Optional

	computational speed)		
<code>retain_intermediate_files=</code>	Keep all intermediate files. Warning: setting this to 1 results in a high space footprint for a large run; only recommended for troubleshooting.	<code>retain_intermediate_files=0</code>	Options are 1 to keep or 0 to delete.
<code>posterior_thresh=</code>	Posterior probability threshold to use for identifying ancestry transition intervals	<code>posterior_thresh=0.9</code>	Recommended 0.8-1
<code>job_submit_command=</code>	Option to run sequentially if using Docker image for dependencies or from a desktop computer. Set bash to run sequentially and sbatch to run in parallel on a slurm cluster	<code>job_submit_command=bash</code> or <code>job_submit_command=sbatch</code>	Always required
<code>slurm_command_map=</code> <code>slurm_command_variant_call=</code> <code>slurm_command_hmm=</code>	If running on a slurm cluster, provide cluster specific parameters for queues, time & memory	<code>slurm_command_map=#!/bin/sh #SBATCH --ntasks=1 #SBATCH --cpus-per-task=1 #SBATCH -p schumer --mem=64000 #SBATCH --time=02:30:00</code> <code>slurm_command_variant_call=#!/bin/sh #SBATCH --ntasks=1 #SBATCH --cpus-per-task=1 #SBATCH -p schumer --mem=64000 #SBATCH --time=05:00:00</code> <code>slurm_command_hmm=#!/bin/sh #SBATCH --ntasks=1 #SBATCH --cpus-per-task=1 #SBATCH -p schumer --mem=64000 #SBATCH --time=03:00:00</code>	Required if running on a cluster

Examples

Several example files are available with the git repository including example configuration files

Running the pipeline

After setting the parameters in the configuration file and loading required dependencies, simply run:

```
perl $PATH/Ancestry_HMM_parallel_v6.pl  
hmm_configuration_file_3way_parallel.cfg
```

where \$PATH is the path to your simulator install (e.g.
/home/groups/schumer/shared_bin/Ancestry_HMM_pipeline)