Appendix 2 *ancestryinfer* pipeline

Output of *mixnmatch* simulations *or* your own 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_v5.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_v5.pl
hmm_configuration_file_nonparallel.cfg
```

Setting parameters in the configuration file

There is are example configuration files available on github:

```
hmm_configuration_file_parallel.cfg

hmm_configuration_file_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
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 ./reads/CALL1_read2.fq.gz ./reads/CALL2_read1.fq.gz ./reads/CALL2_read2.fq.gz ./reads/CALL3_read1.fq.gz ./reads/CALL3_read2.fq.gz	Always
read_length=	Provide expected read length	read_length=150	Always
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)
number_indiv_per_job=	Parallelize jobs such that each job processes this number of individuals. Low numbers mean high parallelization and high number mean low parallelization.	number_indiv_per_job=1	Always
program_path=	Path to the program install folder	program_path=/home/groups/schumer/shared_bin/Ancestry_HMM_pipeline	If not provided the program

			will assume necessary scripts and programs are in the working directory
provide_AIMs=	Coordinates and identities of ancestry informative sites that distinguish the two parent species	provide_AIMs=Xbirchmanni10xgenome_ancestry_informative_sites_filterF1 Example list format: 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	Required unless provided genomes are on the same coordinate system and can be auto detected
provide_counts=	Counts of parental allele frequencies at ancestry informative sites (and recombination rates between adjacent sites if available)	provide_counts=Xbirchmanni10xgenome_Xmalinche_observed_parental_counts_filterF1 Example format: 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 Columns are: Chromosome site allele1_count_parent1 allele2_count_parent1 allele1_count_parent2 allele2_count_parent2 recombination_rate	If not provided, the program will assume that provided ancestry informative sites are fixed between species
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=	Estimated generation of initial admixture	gen_initial_admix=20	If not provided, AncestryHMM will attempt to estimate (may increase run time)

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 computational speed)	max_alignments=2000000	Optional
retain_intermediate_files=	Keep all intermediate files. Warning: setting this to 1 results in a high space footprint for a large run; only recommended for troubleshooting.	retain_intermediate_files=0	Options are 1 to keep or 0 to delete.
posterior_thresh=	Posterior probability threshold to use for identifying ancestry transition intervals	posterior_thresh=0.9	Recommended 0.8-1
job_submit_command=	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	job_submit_command=bash or job_submit_command=sbatch	Always required

	parallel on a slurm cluter		
slurm_command_map= slurm_command_variant_call= slurm_command_hmm=	If running on a slurm cluster, provide cluster specific parameters for queues, time & memory	slurm_command_map=#!/bin/sh #SBATCH --ntasks=1 #SBATCH --cpus-per-task=1 #SBATCH -p schumer --mem=64000 #SBATCH --time=02:30:00 slurm_command_variant_call=#!/bin/sh #SBATCH --ntasks=1 #SBATCH --cpus-per-task=1 #SBATCH -p schumer --mem=64000 #SBATCH --time=05:00:00 slurm_command_hmm=#!/bin/sh #SBATCH --ntasks=1 #SBATCH --cpus-per-task=1 #SBATCH -p schumer --mem=64000 #SBATCH --time=03:00:00	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 mixnmatch/ simulate_admixed_genomes_v6.pl
hybrid_simulation_configuration.cfg
```

where path is the path to your simulator install