Appendix 2 ancestryinfer pipeline

Output of *mixnmatch* simulations <u>or</u> 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	<pre>genome1=xiphophorus_birchmanni_10x_12Sep2018_yD AA6.fasta</pre>	Always
genome2=	User provided fasta file for species 2	<pre>genome2=Xmalinche_dovetail_assembly.fa</pre>	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
<pre>prop_genome_g enome1_parent =</pre>	Expected proportion of the genome derived from the parent species listed under genome1	<pre>prop_genome_genome1_parent=0.5</pre>	If not provided, AncestryH MM will attempt to estimate (may increase run time)
<pre>number_indiv_ per_job=</pre>	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	<pre>program_path=/home/groups/schumer/shared_bin/An cestry_HMM_pipeline</pre>	If not provided the program

provide_AIMs= Coordinates and identities of ancestry informative sites that distinguish the two parent species provide_Count S= Counts of parental allele frequencies at ancestry informative sites (and recombination provide_AIMs= Coordinates and identities of ancestry informative sites (and recombination provide_AIMs= Coordinates and identities of ancestry informative sites (and recombination provide_AIMs= Coordinates and programs are in the working directory Required unless provided genomes provided genomes on the same coordinate system at coordinate system at can be audetected ScyDAA6-2-HRSCAF-26 59896 T C ScyDAA6-2-HRSCAF-26 63105 G A ScyDAA6-2-HRSCAF-26 66290 C A ScyDAA6-2-HRSC	ripts and ograms e in the orking rectory equired aless ovided enomes are a the same fordinate stem and an be auto extected enot ovided,
provide_AIMs= Coordinates and identities of ancestry informative sites that distinguish the two parent species provide_Count S= Drovide_Count S= Drovide_Count	ripts and ograms e in the orking rectory equired aless ovided enomes are a the same fordinate stem and an be auto exected enoted ovided,
provide_AIMs= Coordinates and identities of ancestry informative sites that distinguish the two parent species	ograms e in the orking rectory equired aless ovided enomes are a the same ordinate stem and n be auto etected not ovided,
provide_AIMS=	e in the orking rectory equired aless ovided enomes are a the same fordinate estem and an be auto etected enoting ovided,
provide_AIMS=	equired equired enomes are a the same eordinate estem and n be auto etected enoting the enoting the enoting the enomes are a the same eordinate estem and n be auto etected enoting the en
provide_AIMs=	rectory equired aless ovided enomes are a the same ordinate stem and n be auto etected not ovided,
provide_AIMs=	rectory equired aless ovided enomes are a the same ordinate stem and n be auto etected not ovided,
Coordinates and identities of ancestry informative sites that distinguish the two parent species	equired aless ovided enomes are a the same ordinate stem and n be auto etected not ovided,
identities of ancestry informative sites that distinguish the two parent species ScyDAA6-2-HRSCAF-26 58345 T C on the same coordinate species ScyDAA6-2-HRSCAF-26 58976 T A scyDAA6-2-HRSCAF-26 59896 T C scyDAA6-2-HRSCAF-26 69164 G A system and scyDAA6-2-HRSCAF-26 69164 G A scyDAA6-2-HRSCAF-26 69105 G A scyDAA6-2-HRSCAF-26 65532 G A scyDAA6-2-HRSCAF-26 66290 C A scyDAA6-2-HRSCAF-26 66290 C A scyDAA6-2-HRSCAF-26 68233 T C scyDAA6-2-HRSCAF-26 68233 T C scyDAA6-2-HRSCAF-26 78398 G A scyDAA6-2-HRSCAF-26 78398 G A scyDAA6-2-HRSCAF-26 78399 G A s	ovided enomes are a the same coordinate stem and an be auto extected not ovided,
ancestry informative sites that distinguish the two parent species ScyDAA6-2-HRSCAF-26 58976 T A coordinate species ScyDAA6-2-HRSCAF-26 59896 T C scyDAA6-2-HRSCAF-26 60164 G A system are can be audetected ScyDAA6-2-HRSCAF-26 663105 G A can be audetected ScyDAA6-2-HRSCAF-26 66290 C A scyDAA6-2-HRSCAF-26 66290 C A scyDAA6-2-HRSCAF-26 66290 C A scyDAA6-2-HRSCAF-26 70398 G A scyDAA6-2-HRSCAF-26 70398 G A scyDAA6-2-HRSCAF-26 70398 G A scyDAA6-2-HRSCAF-26 73869 G A Provide_count series frequencies at ancestry informative sites (and recombination ScyDAA6-2-HRSCAF-26 163722 129 3 0 54 0.00000078 that provided ancestry ScyDAA6-2-HRSCAF-26 166535 6 0 0 6 0.00000754 provided ancestry	ovided enomes are a the same fordinate stem and in be auto etected
informative sites that distinguish the two parent species ScyDAA6-2-HRSCAF-26 58976 T A ScyDAA6-2-HRSCAF-26 59896 T C ScyDAA6-2-HRSCAF-26 69164 G A ScyDAA6-2-HRSCAF-26 69105 G A ScyDAA6-2-HRSCAF-26 65532 G A ScyDAA6-2-HRSCAF-26 66290 C A ScyDAA6-2-HRSCAF-26 66290 C A ScyDAA6-2-HRSCAF-26 68233 T C ScyDAA6-2-HRSCAF-26 68233 T C ScyDAA6-2-HRSCAF-26 68233 T C ScyDAA6-2-HRSCAF-26 7398 G A ScyDAA6-2-HRSCAF-26 73869 G A ScyDAA6-2-HRSCAF-26 73869 G A ScyDAA6-2-HRSCAF-26 73869 G A ScyDAA6-2-HRSCAF-26 73869 G A ScyDAA6-2-HRSCAF-26 166158 135 5 0 54 0.000000754 Example format: ScyDAA6-2-HRSCAF-26 166535 6 0 0 6 0.000000754 provided ancestry informative sites (and recombination	nomes are a the same fordinate stem and n be auto etected
that distinguish the two parent species ScyDAA6-2-HRSCAF-26 58976 T A coordinate system are scyDAA6-2-HRSCAF-26 66164 G A system are scyDAA6-2-HRSCAF-26 665532 G A can be audetected scyDAA6-2-HRSCAF-26 66290 C A scyDAA6-2-HRSCAF-26 66290 C A scyDAA6-2-HRSCAF-26 66290 C A scyDAA6-2-HRSCAF-26 68233 T C scyDAA6-2-HRSCAF-26 68233 T C scyDAA6-2-HRSCAF-26 70398 G A scyDAA6-2-HRSCAF-26 166535 6 0 0 6 0.00000078 that same stry will assure that scyDAA6-2-HRSCAF-26 166535 6 0 0 6 0.000000754 provided ancestry	n the same fordinate stem and in be auto etected
the two parent species ScyDAA6-2-HRSCAF-26	ordinate stem and n be auto rected
species ScyDAA6-2-HRSCAF-26 59896 1	n be auto etected not ovided,
ScyDAA6-2-HRSCAF-26 63105 G A can be au detected ScyDAA6-2-HRSCAF-26 65532 G A detected 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 Provide_count Counts of parental allele frequencies at ancestry informative sites (and recombination ScyDAA6-2-HRSCAF-26 163722 129 3 0 54 0.00000078 ScyDAA6-2-HRSCAF-26 166158 135 5 0 54 0.000000754 provided ancestry Counts of parental allele frequencies at ancestry informative sites (and recombination ScyDAA6-2-HRSCAF-26 166158 135 5 0 54 0.000000754 provided ancestry	n be auto stected not ovided,
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 ScyDAA6-2-HRSCAF-26 73869 G A Provide_count S= Counts of parental allele frequencies at ancestry informative sites (and recombination ScyDAA6-2-HRSCAF-26 163722 129 3 0 54 0.00000078 ScyDAA6-2-HRSCAF-26 166158 135 5 0 54 0.00000754 provided ancestry Counts of parental allele frequencies at ancestry ScyDAA6-2-HRSCAF-26 166158 135 5 0 54 0.000000754 provided ancestry Counts of parental allele frequencies at ancestry ScyDAA6-2-HRSCAF-26 166535 6 0 0 6 0.000000754 provided ancestry	not ovided,
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 Provide_count Counts of parental allele frequencies at ancestry informative sites (and recombination ScyDAA6-2-HRSCAF-26 163722 129 3 0 54 0.000000754 provided ancestry Counts of parental allele frequencies at ancestry informative sites (and recombination ScyDAA6-2-HRSCAF-26 163722 129 3 0 54 0.00000078 provided ancestry ScyDAA6-2-HRSCAF-26 166158 135 5 0 54 0.000000754 provided ancestry	not ovided,
ScyDAA6-2-HRSCAF-26 68233 T C ScyDAA6-2-HRSCAF-26 70398 G A ScyDAA6-2-HRSCAF-26 73869 G A Provide_count S= Counts of parental allele frequencies at ancestry informative sites (and recombination ScyDAA6-2-HRSCAF-26 163722 129 3 0 54 0.000000754 provided ancestry Counts of parental allele frequencies at ancestry informative sites (and recombination ScyDAA6-2-HRSCAF-26 166158 135 5 0 54 0.000000754 provided ancestry	ovided,
provide_counts of parental allele frequencies at ancestry informative sites (and recombination ScyDAA6-2-HRSCAF-26 73869 G A provide_counts 73869 G A provide_counts Stbirchmanni10xgenome_Xmalinche_o bserved_parental_counts_filterF1 Example format: the progratile will assure that provided 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 0.000000754 provided ancestry	ovided,
provide_counts of parental allele frequencies at ancestry informative sites (and recombination Counts of parental allele frequencies at ancestry informative sites (and recombination Counts of parental counts = Xbirchmanni10xgenome _ Xmalinche_o provided. The provided the progration will assure that provided ancestry Counts of parental allele beeved_parental_counts_filterF1 Example format: the progration will assure that provided ancestry	ovided,
parental allele frequencies at ancestry informative sites (and recombination bserved_parental_counts_filterF1 bserved_parental_counts_filterF1 provided, the progr will assure that provided ancestry scyDAA6-2-HRSCAF-26 166732 129 3 0 54 0.00000078 ScyDAA6-2-HRSCAF-26 166535 6 0 0 6 0.00000754 provided ancestry	ovided,
frequencies at ancestry informative sites (and recombination Example format: Example format: Example format: ScyDAA6-2-HRSCAF-26 163722 129 3 0 54 0.00000078 will assure that provided ancestry	
ancestry informative sites (and recombination Datample formatic	
informative sites (and combination ScyDAA6-2-HRSCAF-26 163722 129 3 0 54 0.000000078 ScyDAA6-2-HRSCAF-26 166158 135 5 0 54 0.000001374 provided ancestry	
(and recombination ScyDAA6-2-HRSCAF-26 166158 135 5 0 54 0.00001374 provided ancestry	
recombination Scydaa6-2-HRSCAF-26 166535 6 0 0 6 0.00000754 provided ancestry	
	ovided
rates between Columns are: informati	cestry
	formative
adjacent sites if Chromosome site allele1 count parent1 sites are	es are
available) available allele2_count_parent1 allele1_count_parent2 fixed	red
allele2_count_parent2 recombination_rate between	tween
species	
per_site_erro Per-site error per_site_error=0.02 Always	
r= ref site effor r	
HMM (i.e. due	
to sequencing	
error,	
contamination,	
etc)	
gen_initial_a Estimated gen_initial_admix=20 If not	
generation of provided,	
	ovided,
MM will	ovided, ncestryH
attempt to	ovided, ncestryH M will
estimate	ovided, ncestryH M will tempt to
(may	ovided, ncestryH M will tempt to
	ovided, ncestryH M will tempt to timate
Increase i	ovided, ncestryH M will tempt to timate

<pre>focal_chrom_l ist=</pre>	Provide a list of	focal_chrom_list=mychrs.txt	Not required
ISC=	chromosomes to run (other	Example:	
	chromosomes will not be run)	ScyDAA6-2-HRSCAF-26 ScyDAA6-7-HRSCAF-50	
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_alignment s=	Limit analysis to a maximum number of alignments (for computational speed)	max_alignments=2000000	Optional
retain_interm ediate_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_thr esh=	Posterior probability threshold to use for identifying ancestry transition intervals	posterior_thresh=0.9	Recommend ed 0.8-1
job_submit_co mmand=	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	<pre>job_submit_command=bash or job_submit_command=sbatch</pre>	Always required

	parallel on a slurm cluter		
slurm_command _map= slurm command	If running on a slurm cluster, provide cluster	<pre>slurm_command_map=#!/bin/sh #SBATCHntasks=1 #SBATCHcpus-per-task=1 #SBATCH -p schumer mem=64000 #SBATCHtime=02:30:00</pre>	Required if running on a cluster
_variant_call = slurm command	specific parameters for queues, time &	<pre>slurm_command_variant_call=#!/bin/sh #SBATCH ntasks=1 #SBATCHcpus-per-task=1 #SBATCH -p schumermem=64000 #SBATCHtime=05:00:00</pre>	
_hmm=	memory	<pre>slurm_command_hmm=#!/bin/sh #SBATCHntasks=1 #SBATCHcpus-per-task=1 #SBATCH -p schumer mem=64000 #SBATCHtime=03:00:00</pre>	

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