Appendix 1: *mixnmatch* simulator user manual

Getting Started

Install

Option 1 – install dependencies:

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

To install dependencies, follow instructions outlined in:

```
installation_instructions.txt
```

Option 2 – load docker file for dependencies:

With docker:

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

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

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

With singularity:

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

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

singularity run mixnmatch-ancestryinfer-image_mixnmatch-ancestryinfer-
docker.sif bash
```

Setting parameters in the configuration file

There are several example configuration files available from github:

1) Example of a basic macs-based simulation using a user-provided ancestral sequence

```
hybrid_simulation_configuration_ancestral_seq_example_parallel.cfg

hybrid_simulation_configuration_ancestral_seq_example_nonparallel.cfg
```

2) Example of a macs-based simulation with drift between the hybridizing populations and the reference parental populations

```
hybrid_simulation_configuration_sourcepopdrift_example.cfg
```

3) Example of a simulation with user-provided reference genomes

`hybrid_simulation_configuration_usergenomes.cfg`

Parameter descriptions:

Note: some parameters should only be set if you are using macs to simulate parental genomes and some should only be set if you are using your own genomes. See below sections for a list of each.

Parameter	Description	Example	Include if
genome1=	User provided fasta file for species 1 or ancestral sequence	genome1=group1_par1.fa	use_ancestral=1 or use_macs=0
genome2=	User provided fasta file for species 2	genome2= group1_par2.fa	use_macs=0
use_ancestral=	Treat genome1 as an ancestral sequence in simulations (options are 0 - no or 1 - yes)	use_ancestral=1	Optional to include if use_macs=1
mixture_prop_par1=	Expected proportion of the genome derived from parent species 1	mixture_prop_par1=0.5	Always required
rec_rate_Morgans_kb=	Recombination rate in Morgans per kb to simulate	rec_rate_Morgans_kb=0.00002	Always required
num_indivs=	Number of individuals to simulate	num_indivs=50	Always required
gens_since_admixture=	Generations since initial admixture to simulate	gens_since_admixture=50	use_macs=0 or SELAM_param_file=

chr_to_simulate=	Chromosome to use for simulations	chr_to_simulate=group1	use_macs=0
poly_perbp_par1=	Per-basepair polymorphism rate in parent species 1	poly_perbp_par1=0.001	use_macs=0
poly_perbp_par2=	Per-basepair polymorphism rate in parent species 2	poly_perbp_par2=0.001	use_macs=0
rate_shared_poly_at_aims=	Rate of shared polymorphisms between species	rate_shared_poly_at_aims=0.01	use_macs=0
read_type=	Type of reads to simulate (paired end - PE or single end - SE)	read_type=PE	Always required
read_length=	Length of reads to simulate	read_length=100	Always required
per_bp_indels=	INDEL rate per basepair to simulate in reads	per_bp_indels=0.006	Always required
sequencing_error=	Sequencing error rate per basepair to simulate	sequencing_error=0.005	Always required
number_reads=	Number of reads to simulate	number_reads=100000	Always required
parental_drift=	Simulate drift from the source parental populations (0 - no, 1 - yes). **Note: this must be paired with the appropriat**	parental_drift=0	Required if the following parameters are defined: macs_par1_aims_pop= macs_par2_aims_pop=

	e macs command		
macs_par1_aims_pop=	Parent 1 population to sample for AIMs generation	macs_par1_aims_pop=3	Required if parental_drift=1
macs_par2_aims_pop=	Parent 2 population to sample for AIMs generation	macs_par2_aims_pop=4	Required if parental_drift=1
aim_freq_cutoff=	Frequency difference required between parental populations to treat a site as ancestry informative	aim_freq_cutoff=0.9	Always required
cross_contam=	Contamination rate to simulate in hybrids. Contamination reads are drawn from the parental haplotypes at the observed mixture proportion.	cross_contam=0.02	Optional
job_submit_cmd=	Use slurm resource management system or run individuals sequentially	job_submit_cmd=sbatch or job_submit_cmd=bash	Always required. If sbatch is specified, users must provide a job submission header in job_header=
use_macs=	Indicate whether to use macs with seq-gen to simulate sequences	use_macs=1	Always required

	(0 - no, 1 - yes)		
use_map=	Use a macs-formatted local recombination map in simulations of parental and hybrid populations and to set recombination priors (0-no, 1 - yes)	use_map=1	Always required
SELAM_param_file=	Provide a parameter file for SELAM describing the hybrid population history. If this is left blank the program assumes a neutral demographic history.	SELAM_param_file=selam_demography_params.txt	Only required if users wish to simulate a particular demographic history in hybrids
SELAM_selection_file=	Provide a SELAM formatted file indicating which site experience selection. See SELAM documentation.	SELAM_selection_file=selam_selection.txt Example: $cat selam_selection.txt S A 0 0.1 1 1 0.9	Optional
macs_params=	Command to be used for macs simulations of parental species demographi	macs_params=200 10000000 -I 2 100 1000 -t 0.001 -h 1e2 -r 0.001 -ej 2 2 1 -R recombination_map_for_macs.txt macs_params=80 10000000 -I 4 20 20 20 20 0 -t 0.001 -h 1e2 -r 0.001 -ej 0.05 3 2 -ej 0.051 4 1 -ej 2 2 1 -R recombination_map_for_macs.txt	Required if use_macs=1

	c history. See ms/macs documentation for simulation options.		
program_path=	Path to the install location of the simulator program. If left blank, the program assumes a global install.	program_path=/home/bin	Optional
par1_for_aims=	Number of parent 1 haplotypes to use to define ancestry informative sites	par1_for_aims=20	Required if use_macs=1
par2_for_aims=	Number of parent 2 haplotypes to use to define ancestry informative sites	par2_for_aims=20	Required if use_macs=1
seq_params=	Provide base composition and transition transversion parameters for seq-gen sequence generation		Optional if use_macs=1
job_header=	Provide cluster-specific job submission parameters to be used	job_header=#!/bin/sh #SBATCH --ntasks=1 #SBATCH --cpus-per-task=1 --mem=64000 #SBATCH --time=02:00:00	Always required

	to submit simulation jobs		
num_indiv_per_job=	Set how much parallelization to perform by setting the number of individuals to run per submitted job. To run a different job for each individual set this parameter to 1	num_indiv_per_job=2	Always required

We also provide *macs* and SELAM documentation in the git repository for *mixnmatch* for convenience.

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