**TODO for the Next Week:**

1. Have the Reference tables to run ABC-RF

2. Run 10,000 simulations with higher values of Ne;

3. Increase the mutation rate or the genome size to have at least 1,000 markers.

**GENERAL ADVICE:**

Create small datasets to check if step of the pipeline and function are working the right way.

**IMPROVEMENTS**

**Pipeline version 1**

**SLiM Model**

1. Work with strong selection values now (toy model) to minimizes the lost of the beneficial mutation;
2. Report the allele frequency at sample t1 and t2 in the SLiM output (remember it is not the initial allele frequency at 10\*Ne, it is the allele frequency at ti);
3. Output “Alleles” and “Status" in SLiM;
4. Check all the possibilities of the selected site being monomorphic in SLiM;
5. To avoid having monomorphic markers in each sample of individuals, try to implement something similar as the sample vcf output (vcf only contain SNPs segregating in the sample); OR output all the mutation only for polymorphism inside the sample of individuals (maybe right after sampling the genomes);
6. Another way to minimize the chance to lost beneficial mutations before the sampling is to work with standing genetic variation model- where the fitness of some mutation type change after some time - this is more realistic but we can work with it latter;
7. Check if the code to add one m2 mutation in SLiM is correct or if I need to fix it - the way it is programmed now SLiM sample one genome (haplotype) at generation = 10\*Ne and add one mutation of type m2 that has selection coefficient = 0.1; probably a best way to program it in a way the beneficial mutation has more chance to come is to try to add it from 10\*Ne to 10\*Ne + (t1+INT12) if any beneficial not exist (with conditional);
8. Try to implement the parameter for the hitchhiking in later models;
9. Check stack policy rules and set the appropriate one - remove stack police to remove bias and better represent reality;
10. Better way to implement custom outputs with the new version of SLiM2 (2.5) since it now allows you to write custom functions.

**R integration**

1. The proper use of folders and folder hierarchy in *"SLiM\_in\_R"* folder:
   1. Datafolder: it should only contain datasets;
   2. Resultsfolder: it should contain all the results and files generated during the workflow (e.g. SLiM infiles, SLiM outputs; egglib datasets and outputs);
   3. src folder: it should contain source code, scripts and softwares.
2. Remove the prior sampling in the “create\_slim\_infile” function and do this part separately in the pipeline - it will allow use different distributions for the parameters;
3. Use the number of simulations of the first script to all other scrips to avoid put it manually - maybe I can include a data= argument that allow to add the dataset (or table) that has all the necessary information;
4. Specify the seed value in slim script instead at the running script - it allow you to 1) minimize the steps and 2) keep track on simulations/seed easily;
5. For the loci-specific reference table:
   1. take random polymorphic sites only in chromossome 2;
   2. take the beneficial locus only if polymorphic.
6. Add conditional to runEggLib function to deal with:
   1. SFS summary statistics (add SFS-bin when SFS is in arguments for GSS);
   2. Different argument than “all" for the select argument (select-num OR select-freq).

**Pipeline**

1. Remove monomorphic sites before run the summary statistics in egglib;
2. We need 2 matrices for the ABC-RF:
   1. One with global parameters and global summary statistics;
   2. Another with loci-specific parameters and summary statistics (global and locus-specific).
3. Report the SFS error to Stephane de Mita;

1 - SFS calculation

MacBook-Pro-de-Vitor:egglib-inputs vitorpavinato$ /Users/vitorpavinato/anaconda/bin/python summstats.py input-file=input\_egglib\_1\_testdata.txt output-file=output\_egglib\_1\_testdata.txt LSS=He,Dj,WCst WSS=He,Dj,WCst,S,thetaW,D,Da,ZZ GSS=He,Dj,WCst,S,thetaW,D,Da,ZZ,SFS wspan=1 SFS-bins=19 select=all

[summstats] starting

warning: 'selection' column is present in input file but won't be used

[summstats] sample structure

number of populations: 2

number of samples: 20

pop1: 10

pop2: 10

an error occurred: unsupported operand type(s) for /: 'NoneType' and 'float'

[summstats] failed

2 - select=list error

MacBook-Pro-de-Vitor:egglib-inputs vitorpavinato$ /Users/vitorpavinato/anaconda/bin/python summstats.py input-file=input\_egglib\_3\_testdata.txt output-file=output\_egglib\_3\_testdata.txt LSS=He,Dj,WCst WSS=He,Dj,WCst,S,thetaW,D,Da,ZZ GSS=He,Dj,WCst,S,thetaW,D,Da,ZZ wspan=1 select=list

[summstats] starting

[summstats] sample structure

number of populations: 2

number of samples: 20

pop1: 10

pop2: 10

window: contig=chr1 first=1 center=1 last=2 num=2

window: contig=chr1 first=2 center=3 last=4 num=3

window: contig=chr1 first=3 center=4 last=5 num=3

window: contig=chr1 first=4 center=5 last=6 num=3

window: contig=chr1 first=6 center=7 last=8 num=3

window: contig=chr1 first=8 center=9 last=10 num=3

window: contig=chr1 first=9 center=10 last=10 num=2

window: contig=chr2 first=1 center=1 last=2 num=2

window: contig=chr2 first=2 center=3 last=4 num=3

window: contig=chr2 first=3 center=4 last=5 num=3

window: contig=chr2 first=4 center=5 last=6 num=3

window: contig=chr2 first=6 center=7 last=8 num=3

window: contig=chr2 first=8 center=9 last=9 num=2

an error occurred: list index out of range

[summstats] failed

3 - Some GSS are repeated before GSS columns in the output;

4- select=rand and select-num failed:

MacBook-Pro-de-Vitor:egglib-inputs vitorpavinato$ /Users/vitorpavinato/anaconda/bin/python summstats.py input-file=input\_egglib\_1\_testdata.txt output-file=output\_egglib\_2\_testdata.txt LSS=He,Dj,WCst WSS=He,Dj,WCst,S,thetaW,D,Da,ZZ GSS=He,Dj,WCst,S,thetaW,D,Da,ZZ wspan=1 select=rand select-num=8

[summstats] starting

warning: 'select-num' is specified but won't be used

warning: 'selection' column is present in input file but won't be used

[summstats] sample structure

number of populations: 2

number of samples: 20

pop1: 10

pop2: 10

window: contig=chr1 first=2 center=3 last=4 num=3

window: contig=chr2 first=1 center=1 last=2 num=2

window: contig=chr2 first=1 center=2 last=3 num=3

window: contig=chr2 first=3 center=4 last=5 num=3

window: contig=chr2 first=4 center=5 last=6 num=3

window: contig=chr2 first=5 center=6 last=7 num=3

window: contig=chr2 first=7 center=8 last=9 num=3

window: contig=chr2 first=8 center=9 last=9 num=2

an error occurred: list index out of range

[summstats] failed

5 - Mutation with same position caused a problem in egglib

1. Have the reference tables to test the ABC-RF and have a meeting with JJM.