

Code files:

Population analysis: Dataset S1 from publication is used as input

PopulationAnalysisbyLineage_RandomOnly.Rmd

- Calculates and compares the reproductive waiting time, percent fissions/fragmentations and number of middles for the 4 non-perturbed lineages (P, H, F, S).

DeathRatebyLineage_NonBottleneck.Rmd

- Calculates and compares the percent death for each non-perturbed lineage (P, H, F, S).

PopulationAnalysisbyLineage_BvFS.Rmd

- Calculates and compares the reproductive waiting time, and percent fissions/fragmentations for the F and S lineages vs the two bottleneck lineages (B1 and B2).

PopulationAnalysisbyLineage_BvFS_BinBirthDate.Rmd

- Calculates reproductive waiting time as a function of birth date for the F and S lineages vs the two bottleneck lineages.

DeathRatebyLineage_BFS.Rmd

- Calculates and compares the percent death of the F and S lineages vs the two bottleneck lineages (B1 and B2).

Genomic analysis

VCF_filtration_annotated.ipynb

- Code to make VCF files with all the variants that pass quality filtration and that are not found in repetitive regions.

Make_mega_df_annotated.ipynb

- Code to make the main dataframe that contains all SNPs that pass quality filtration and have an allele frequency lower than 0.5 in the population (so removing all SNPs that are fixed in the population or present in all individuals as heterozygous). Similar approach is used to get the INDELs.

family_plink_pca.ipynb

- This is the code used to run the PCA of PH and FS substrains.

Annotated_FST_code_pairwise_080625.ipynb

- Code for pairwise FST comparison. Using the same subset of variants as the PCA (using GQ > 20 to deal with the fact that the P and H libraries have lower depth).

Annotated_FST_code_across_genome_080625.ipynb

- Code for getting the FST scores across the genome between the PH, FS, and FS_B populations.

nucleotide.ipynb

- Calculate nucleotide diversity across the genome

Annotated_ROH_analysis.ipynb

- Code used to assess runs of homozygosity in the PH and FS substrains

Annotated_scaffold1_across_the_world.ipynb

- Code to assess runs of homozygosity in scaffold 1 using both our data and RNAseq data from different labs.

Annotated_code_unique_SNPs_090625.ipynb

- Used to assess PH vs FS specific SNPs and their allele frequencies etc. It also contains a calculation of heterozygosity in each group.

Annotated_make_mega_df_HEAD.ipynb

- Making the dataframe for the head dataset used to calculate the mutation rate etc. It uses the filtered VCF (using similar filtration parameters to those used in the Family dataset).

Annotated_isolate_denovo.ipynb

- This code isolates the putative de-novo mutations in the data. This is also where SNPs in repetitive regions are filtered out. Annotates which potential de novo mutations are passed on to the next generation.

Annotated_denovo_analysis_090625.ipynb

- This is the main analysis code for the putative de novo mutations. It also contains code for some of the filtration, such as filtering out putative de novo SNPs that were found within 150bs from one another (a read length from one another).