Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/At1_Genome/Analysis/2023-02-07.EffectivePopSize

PDF Version generated by

Katarina Stuart (z5188231@ad.unsw.edu.au)

on

Aug 23, 2023 @09:41 AM NZST

Table of Contents

2023-02-07.EffectivePopSize

Katarina Stuart (z5188231@ad.unsw.edu.au) - Aug 18, 2023, 9:17 AM GMT+12

Analysis for effective population size

Tassie devil comparison: https://academic.oup.com/mbe/article/36/12/2906/5551343

https://onlinelibrary.wiley.com/doi/full/10.1002/ece3.5888

Explore PSMC and MSMC and SMC++

SMC++:https://github.com/popgenmethods/smcpp

MSMC:https://www.nature.com/articles/ng.3015
https://pubmed.ncbi.nlm.nih.gov/31975167/

https://github.com/stschiff/msmc-tools/blob/master/msmc-tutorial/guide.md
https://github.com/stschiff/msmc2

PSMC:

https://github.com/lh3/psmc

https://www.nature.com/articles/nature10231

Methods Reference Papers:

https://www.nature.com/articles/s41559-022-01829-5#Sec11

Population genomics of the critically endangered kākāpō - ScienceDirect

repeats: Repetitive genomic regions and the inference of demographic history - PMC (nih.gov)

Myna data:

Following https://github.com/drewschield/venom_population_genomics

Individuals: starling with, then cut down

J725 NZ_LEI-19_M0019

J728 NZ_GBI-13_M0133

J737 NZ_NAP-05_M0155

J745 IND_MA_979_M0213

J748 IND_MA_3731_M0307

J760 IND_TN_3711_M0306

J768 IND_MP_3758_M0324

J775 AUS_CAI_11601

J783 AUS_MEL_12408

J785 SA_M0256

J797 FIJI_923_M0203

STEP 1: Prepeare files for PSMC

1) Call consensus variants per individual and index output. This will also mask sites overlapping repeat annotations.

2) generate diploid sequences

3) Convert to psmcfa file format

Redo with repeats masked

see kakapo paper for estimations: Population genomics of the critically endangered kākāpō - ScienceDirect

```
#!/bin/bash -e

#SBATCH --job-name=2023_03_01.psmc_preprocessing.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=10GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task
#SBATCH --array=1-18

SAMPLE=$(sed
"${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/psmc/psmc_individuals_subset_round2.txt)
echo "working with sample:" $SAMPLE

#load modules
module load psmc/0.6.5-gimkl-2018b
module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1
module load SAMtools/1.15.1-GCC-11.3.0
module load tabix/0.2.6-GCCcore-9.2.0
module load BCFtools/1.15.1-GCC-11.3.0
module load BBMap/39.01-GCC-11.3.0

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta
BAM_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/mapped_reads
REP_BED=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/repeats/repeatmasker/AcTris_vAus2.0repeatlib_hardmask/AcTris_vAus2.0.fasta_rm.bed

#when redo for all, add repreat reagion maskng
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/psmc/input

bcftools mpileup -C 50 -q 20 -Q 25 -Ou -f ${GENOME} ${BAM_DIR}/${SAMPLE}.sorted.dup.bam | bcftools call -c  -Oz -o ./${SAMPLE}.vcf.gz

bcftools filter --SnpGap 10 -i "DP>=5 & DP<=50"  ./${SAMPLE}.vcf.gz   -Oz -o ./${SAMPLE}.filter.vcf.gz

bcftools view --exclude-types indels  ./${SAMPLE}.filter.vcf.gz -T ^${REP_BED} | bcftools sort --temp-dir ./tmp_${SAMPLE} -Oz -o ./${SAMPLE}.sort.vcf.gz

tabix -p vcf ./${SAMPLE}.sort.vcf.gz

bcftools view ./${SAMPLE}.vcf.gz | vcfutils.pl vcf2fq | gzip > ./${SAMPLE}.fastq.gz

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/psmc/
fq2psmcfa -q 20 ./input/${SAMPLE}.fastq.gz > ./${SAMPLE}.psmcfa
```

STEP 2: Run PSMC

Time segment patterns:

4+5*3+4 (default)

4+10*3+6+8

4+25*2+4+6

4+30*2+4+6+10

```
#!/bin/bash -e

#SBATCH --job-name=2023_03_02.psmc_main.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=2GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task
#SBATCH --array=1-18

SAMPLE=$(sed
"${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/psmc/psmc_individuals_subset_round2.txt)
echo "working with sample:" $SAMPLE

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/psmc

#load modules
module load psmc/0.6.5-gimkl-2018b

psmc -N30 -t5 -r5 -p "4+30*2+4+6+10" -o ./results/pattern4.${SAMPLE}.diploid.psmc ./${SAMPLE}.psmcfa
```

for parameters: Temporal Dynamics of Avian Populations during Pleistocene Revealed by Whole-Genome Sequences - PMC (nih.gov)

"Based on the results from the pilot runs, we chose the final settings for the PSMC to be “N30 –t5 –r5 –p 4+30∗2+4+6+10” for all species."

" We therefore used age of sexual maturity multiplied by a factor of two as a proxy for generation time. This transformation has proved applicable to other avian systems where detailed information on age-specific rates of survival and reproductive output, necessary for estimating generation time, has been obtained "

Grab mutation rate from: **Table S1**

also mutation rate: Direct estimate of the rate of germline mutation in a bird - PMC (nih.gov)

PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers - PMC (nih.gov)

Speciation and population divergence in a mutualistic seed dispersing bird | Communications Biology (nature.com)

STEP 3: bootstrap PSMC (got 50 rounds with 100 wall hrs)

Validating the main result:

1) Generate 'split' psmcfa inputs for bootstrapping

2) Perform bootstrapping analysis (100 bootstrap replicates per sample)

12 hrs = 8 complete bootstrapping rounds. Will do more in proper run through.

```
#!/bin/bash -e
```

```
#SBATCH --job-name=2023_03_22.psmc_bootstrap.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-100:00:00
#SBATCH --mem=10GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task
#SBATCH --array=1-2

SAMPLE=$(sed "${SLURM_ARRAY_TASK_ID}q;d"
/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/psmc/psmc_individuals_subset_bootstrap.txt)
echo "working with sample:" $SAMPLE

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/psmc

#load modules
module load psmc/0.6.5-gimkl-2018b

#Step 1: split files
splitfa ./${SAMPLE}.psmcfa > ./${SAMPLE}-split.psmcfa

#Step 2: bootstrapping
seq 100 | xargs -i echo psmc -N30 -t5 -r5 -b -p "4+30*2+4+6+10" -o ./results/bootstrap/${SAMPLE}.round-{}.psmc ./${SAMPLE}-split.psmcfa | sh
```

STEP 4: Combine main and bootstrap results and plot

Validating the main result:

1) Concatenate main and bootstrap results files

2) Plot combined results

tools/epstopdf.pl at master · wlz0726/tools (github.com)

PSMC journal club walkthrough - Harvard FAS Informatics

plotting: combining two plots · Issue #26 · lh3/psmc (github.com)

if I loop this over different pattrns need to change output file names

For multiple samples:

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_10.psmc_plot.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/psmc/results

#load modules
module load psmc/0.6.5-gimkl-2018b

#Concatenate the individual .psmc files you wish to plot:
cat pattern4.J725.diploid.psmc pattern4.J728.diploid.psmc pattern4.J737.diploid.psmc pattern4.J745.diploid.psmc pattern4.J748.diploid.psmc pattern4.J760.diploid.psmc
pattern4.J768.diploid.psmc pattern4.J775.diploid.psmc pattern4.J783.diploid.psmc pattern4.J785.diploid.psmc pattern4.J797.diploid.psmc > combined.psmc

#Plot using multiline mode:
psmc_plot.pl -u 2.3e-09 -g 2 -M 'J725 (NZ LEI),J728 (NZ GBI),J737 (NZ NAP),J745 (IND MA),J748 (IND MA),J760 (IND TN),J768 (IND MP),J775 (AUS CAI),J783 (AUS
```

MEL),J785 (SA),J797 (FIJI)' combined_AcTris combined.psmc
perl epstopdf.pl combined_AcTris.eps

#INDIA
cat pattern4.J742.diploid.psmc pattern4.J743.diploid.psmc pattern4.J745.diploid.psmc pattern4.J748.diploid.psmc pattern4.J750.diploid.psmc pattern4.J752.diploid.psmc pattern4.J755.diploid.psmc pattern4.J757.diploid.psmc pattern4.J760.diploid.psmc pattern4.J766.diploid.psmc pattern4.J768.diploid.psmc > combined_india.psmc

psmc_plot.pl -u 2.3e-09 -g 2 -M 'J742 (IND MA),J743 (IND MA),J745 (IND MA), J748 (IND MA), J750 (IND MA),J752 (IND MA),J755 (IND TN),J757 (IND TN),J760 (IND TN),J766 (IND MP),J768 (IND MP)' combined_india_AcTris combined_india.psmc
perl epstopdf.pl combined_india_AcTris.eps

#High POPS (NZ NAP, AUS CAI, SA, FIJI)
cat pattern4.J731.diploid.psmc pattern4.J732.diploid.psmc pattern4.J737.diploid.psmc pattern4.J772.diploid.psmc pattern4.J773.diploid.psmc pattern4.J775.diploid.psmc pattern4.J785.diploid.psmc pattern4.J791.diploid.psmc pattern4.J792.diploid.psmc pattern4.J793.diploid.psmc pattern4.J794.diploid.psmc > combined_outliers.psmc

psmc_plot.pl -u 2.3e-09 -g 2 -M 'J731 (NZ NAP),J732 (NZ NAP),J737 (NZ NAP), J772 (AUS CAI),J773 (AUS CAI), J775 (AUS CAI), J785 (SA) ,J791 (SA), J792 (SA), J793 (FJI), J794 (FIJI)' combined_outliers_AcTris combined_outliers.psmc
perl epstopdf.pl combined_outliers_AcTris.eps

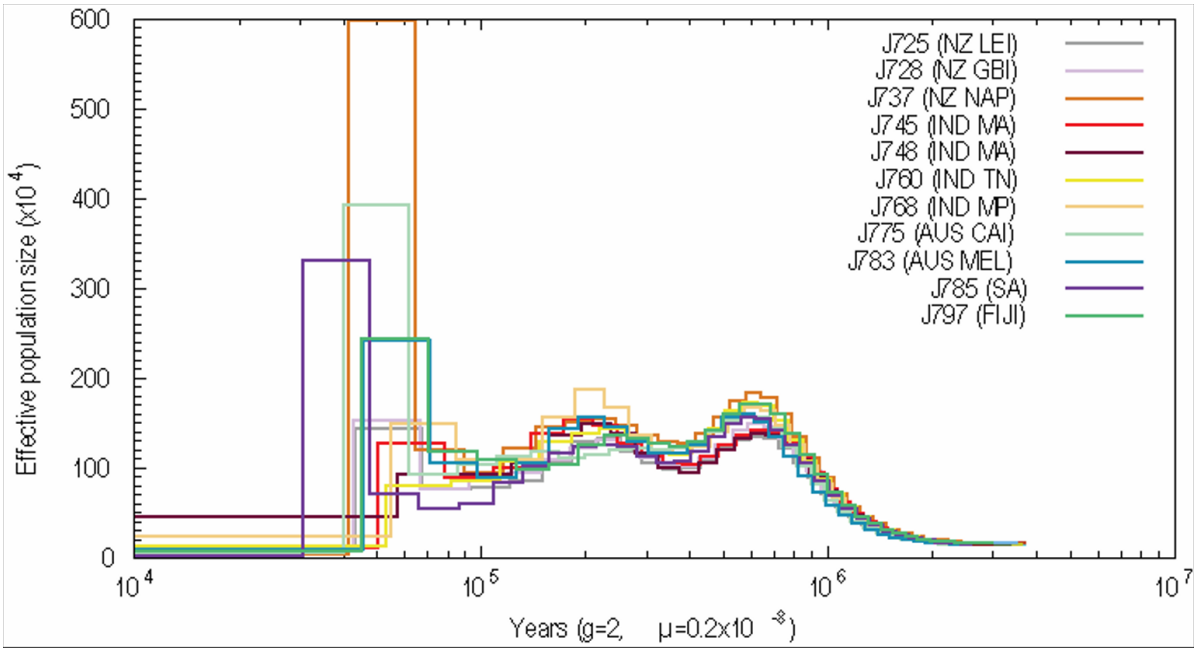
#NZ
cat pattern4.J723.diploid.psmc pattern4.J725.diploid.psmc pattern4.J728.diploid.psmc pattern4.J731.diploid.psmc pattern4.J732.diploid.psmc pattern4.J735.diploid.psmc pattern4.J737.diploid.psmc > combined_nz.psmc

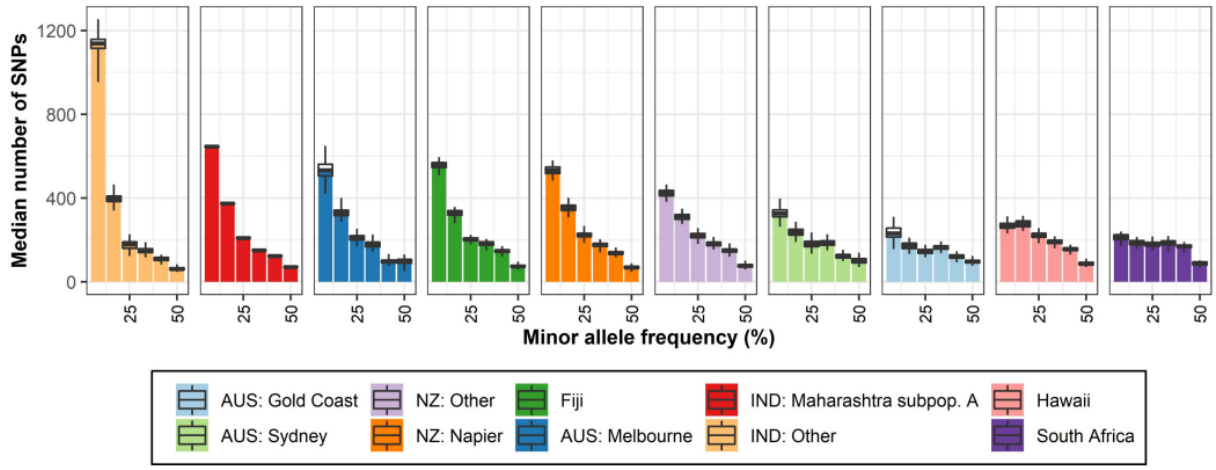
psmc_plot.pl -u 2.3e-09 -g 2 -M 'J723 (NZ LEI),J725 (NZ LEI),J728 (NZ GBI), J731 (NZ NAP),J732 (NZ NAP),J735 (NZ GBI),J737 (NZ NAP)
' combined_nz_AcTris combined_nz.psmc
perl epstopdf.pl combined_nz_AcTris.eps

#ALL NEW
cat pattern4.J723.diploid.psmc pattern4.J731.diploid.psmc pattern4.J732.diploid.psmc pattern4.J735.diploid.psmc pattern4.J742.diploid.psmc pattern4.J743.diploid.psmc pattern4.J750.diploid.psmc pattern4.J752.diploid.psmc pattern4.J755.diploid.psmc pattern4.J757.diploid.psmc pattern4.J766.diploid.psmc pattern4.J772.diploid.psmc pattern4.J773.diploid.psmc pattern4.J791.diploid.psmc pattern4.J792.diploid.psmc pattern4.J793.diploid.psmc pattern4.J794.diploid.psmc > combinedr2.psmc
psmc_plot.pl -u 2.3e-09 -g 2 -M 'J723 (NZ LEI),J731 (NZ NAP),J732 (NZ NAP),J735 (NZ GBI),J742 (IND MA),J743 (IND MA),J750 (IND MA),J752 (IND MA),J755 (IND TN),J757 (IND TN),J766 (IND MP), J772 (AUS CAI),J773 (AUS CAI), J791 (SA), J792 (SA), J793 (FJI), J794 (FIJI)' combinedr2_AcTris combinedr2.psmc
perl epstopdf.pl combinedr2_AcTris.eps

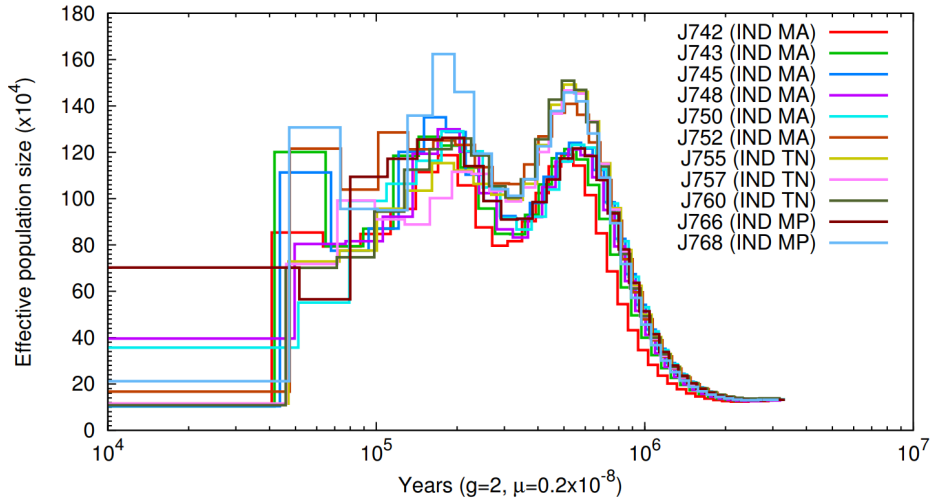
#FINAL PLOT
cat pattern4.J725.diploid.psmc pattern4.J728.diploid.psmc pattern4.J731.diploid.psmc pattern4.J745.diploid.psmc pattern4.J748.diploid.psmc pattern4.J760.diploid.psmc pattern4.J768.diploid.psmc pattern4.J773.diploid.psmc pattern4.J783.diploid.psmc pattern4.J797.diploid.psmc > combined_final.psmc

#Plot using multiline mode:
psmc_plot.pl -u 2.3e-09 -g 2 -M 'J725 (NZ LEI),J728 (NZ GBI),J731 (NZ NAP),J745 (IND MA),J748 (IND MA),J760 (IND TN),J768 (IND MP),J773 (AUS CAI),J783 (AUS MEL),J797 (FIJI)' combined_AcTris_final combined_final.psmc
perl epstopdf.pl combined_AcTris_final.eps

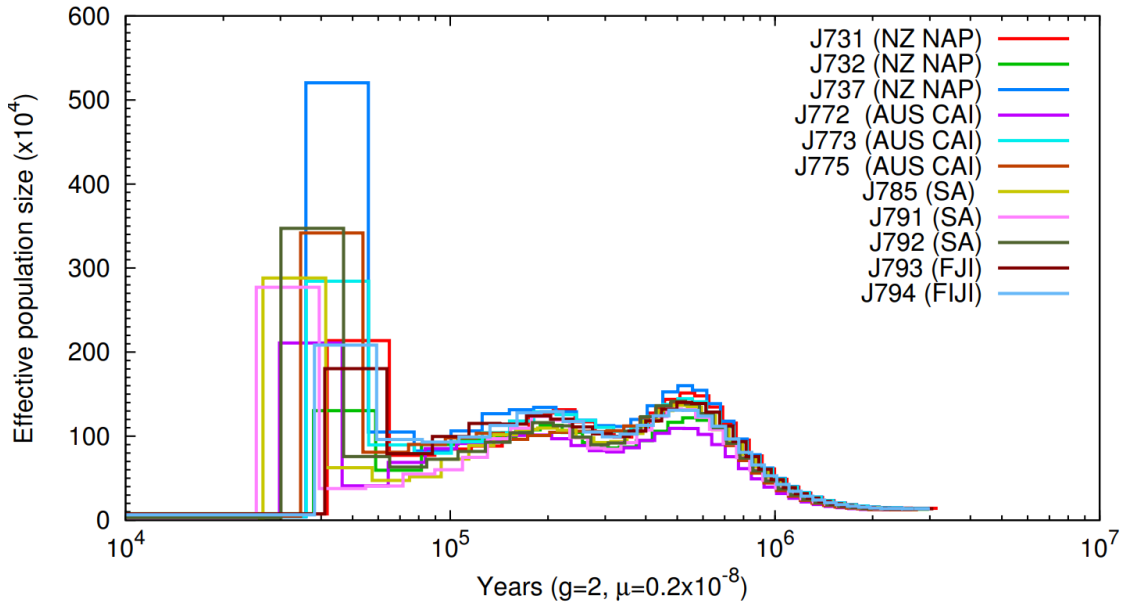




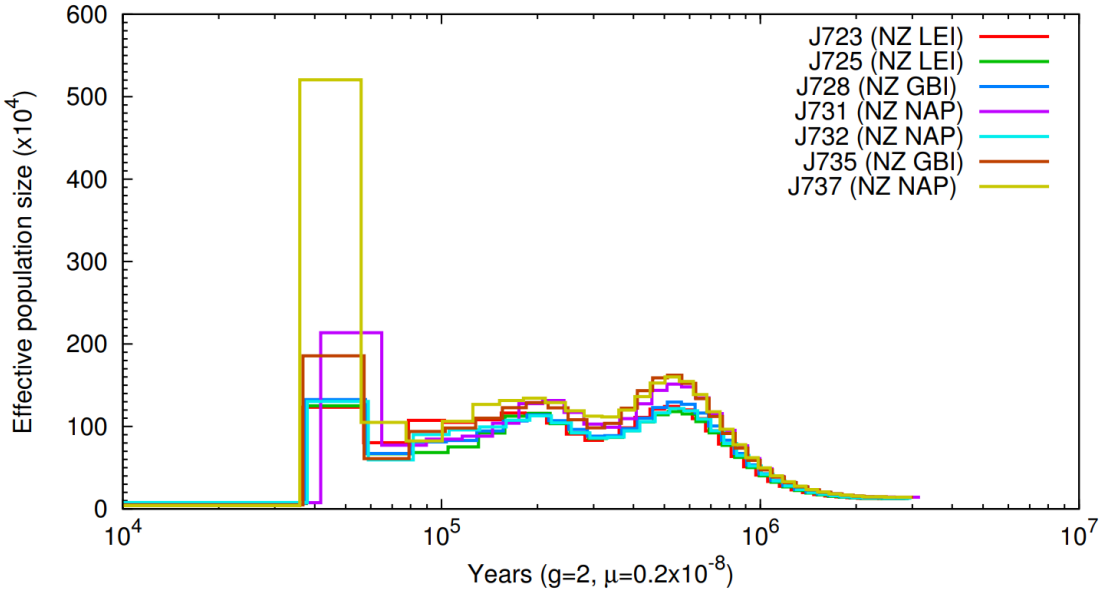
INDIA ONLY:



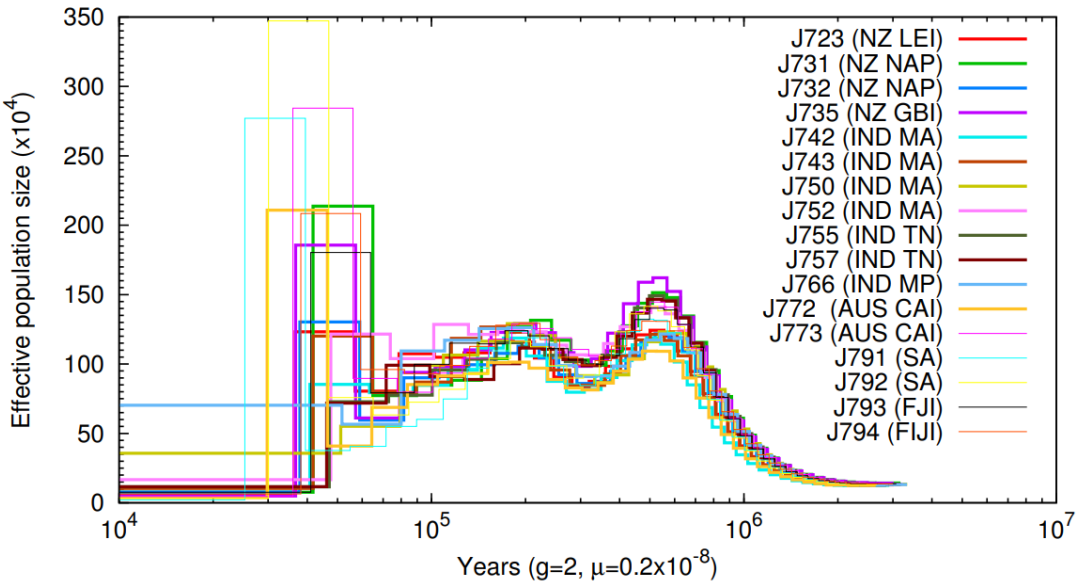
OUTLIERS:



NZ:



ALL NEW:

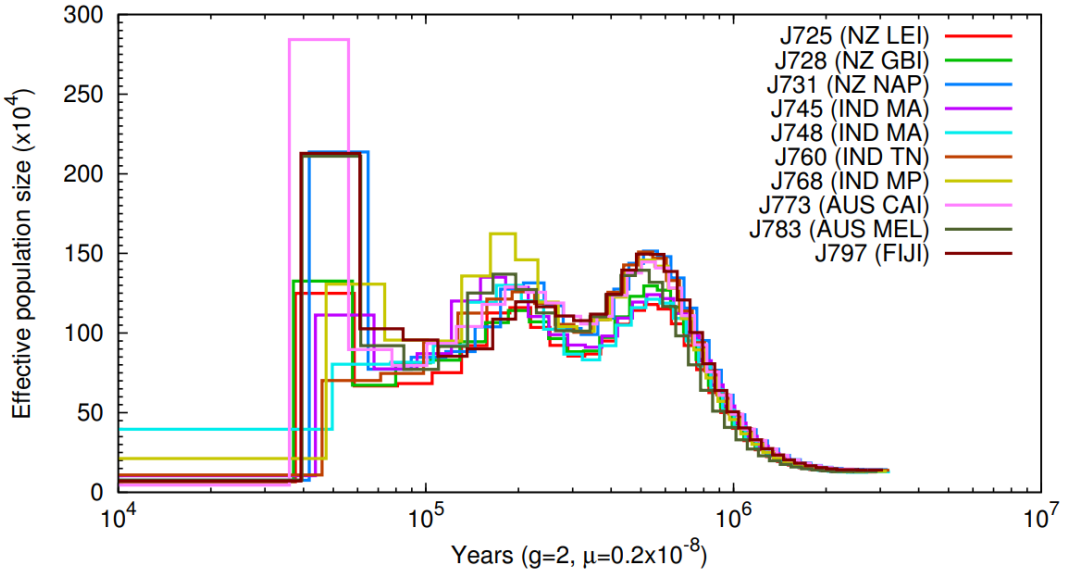


THOUGHTS: mixed singal for more recent population expansion in INDIA - all invasive populations have this signal and indicate that they were likely sourced from similar locations and times, and thus populations with similar demographic histories. Location of this population is hard to pinpoint due to high amounts of dispersal in invasive range. SA does seem to be shifted forward - does this mean more strong bottleneck?

Main plot in manuscript: Native range examples, some select ones from native pops

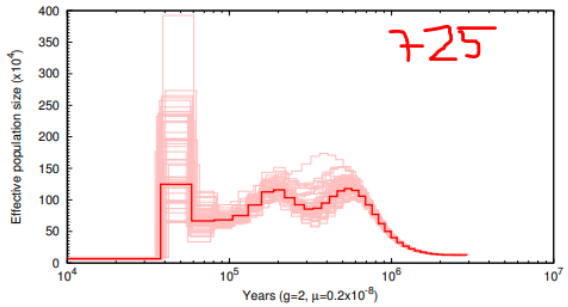
supp mat: fill india plot from above, capturing diversity from native range

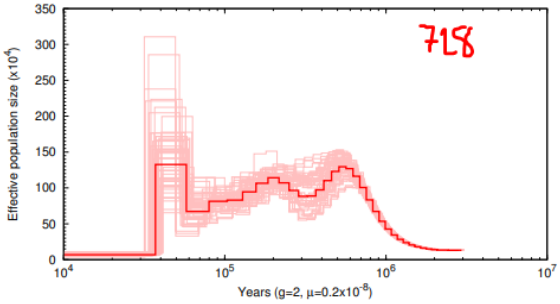
FINAL MAIN PAPER PLOT (not yet formatted in adobe):

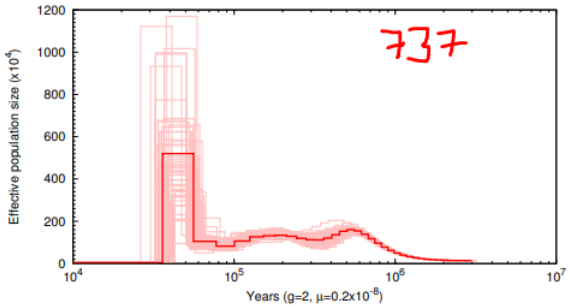


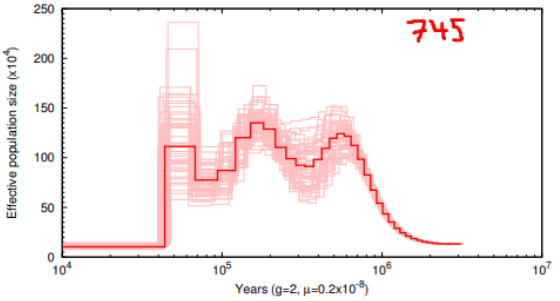
For bootstrapping:

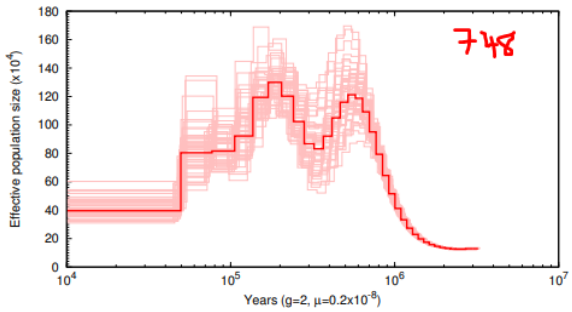
```
#!/bin/bash -e

#SBATCH --job-name=2023_03_22.psmc_plot.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task
#SBATCH --array=1-2

SAMPLE=$(sed
"${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/psmc/psmc_individuals_subset_bootstrap.txt)
echo "working with sample:" $SAMPLE

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/psmc

#load modules
module load psmc/0.6.5-gimkl-2018b
cat ./results/pattern4.${SAMPLE}.diploid.psmc ./results/bootstrap/${SAMPLE}.round-*.psmc > ./results/pattern4_combined.${SAMPLE}.psmc

psmc2history.pl ./results/pattern4_combined.${SAMPLE}.psmc | history2ms.pl > ms-cmd.${SAMPLE}.sh
psmc_plot.pl -u 2.3e-09 -g 2 ./results/pattern4_combined.${SAMPLE} ./results/pattern4_combined.${SAMPLE}.psmc

#EPS file to PDF file
cd results
perl epstopdf.pl pattern4_combined.${SAMPLE}.eps
```

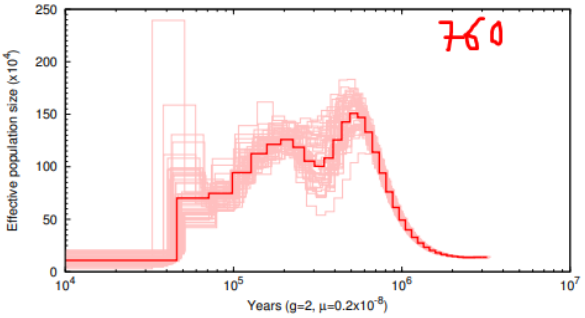

725
Effective population size (x10^4)
Years (g=2, μ=0.2x10^-8)

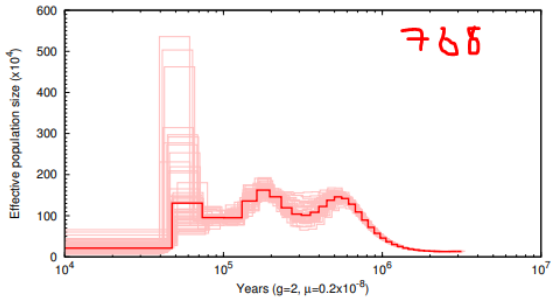

718
Effective population size (x10^4)
Years (g=2, μ=0.2x10^-8)

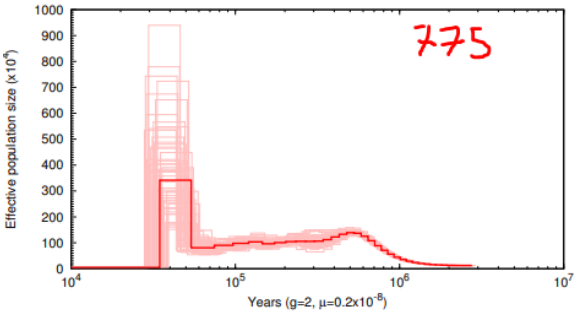

737
Effective population size (x10^4)
Years (g=2, μ=0.2x10^-8)

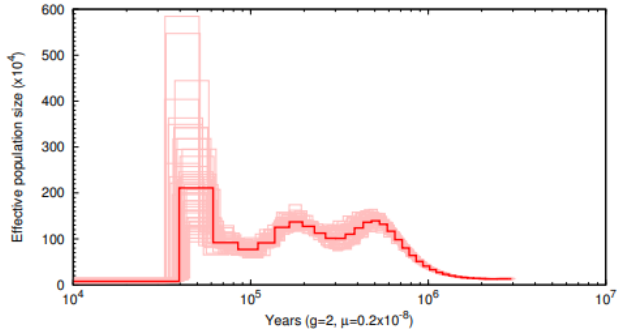

745
Effective population size (x10^4)
Years (g=2, μ=0.2x10^-8)

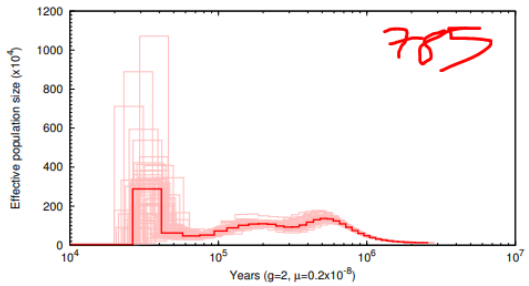

748
Effective population size (x10^4)
Years (g=2, μ=0.2x10^-8)

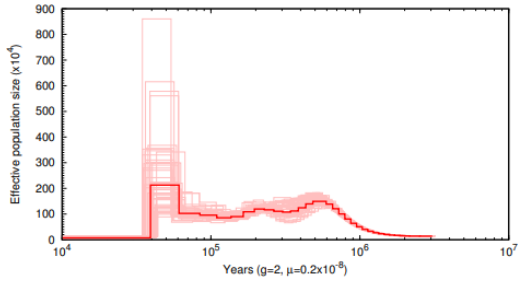

760
Effective population size (x10^4)
Years (g=2, μ=0.2x10^-8)


768
Effective population size (x10^4)
Years (g=2, μ=0.2x10^-8)


775
Effective population size (x10^4)
Years (g=2, μ=0.2x10^-8)


Effective population size (x10^4)
Years (g=2, μ=0.2x10^-8)


785
Effective population size (x10^4)
Years (g=2, μ=0.2x10^-8)


Effective population size (x10^4)
Years (g=2, μ=0.2x10^-8)

Het values for each ind:

```
less /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/SNP_bcftools
```

More inbred individuals (737 (nz nap) F=0.11307, 775 (Aus Cai) F=0.13696, 785 (SA), F=0.31789). Though 979 (fji is flat but not as high inbreeding)

but other pops have high F and different profiles.

Order:
J725: NZ_LEI-19_M0019
J728
J737
J745: IND_MA_979_M0213
J748
J760
J768
J775
J783: AUS_MEL_12408
J785
J797

checking het and maf of each individual:

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_24.psmc_freqs.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-02:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task
#SBATCH --array=1-11

SAMPLE=$(sed "${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/psmc/psmc_individuals_subset.txt)
echo "working with sample:" $SAMPLE

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/psmc/input

module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1

vcftools --gzvcf ${SAMPLE}.sort.vcf.gz --het --out ${SAMPLE}
vcftools --gzvcf ${SAMPLE}.sort.vcf.gz --freq --out ${SAMPLE}

tail -n +2 J797.frq | sed 's/:/\t/g' | cut -f6 |
```

Stairway Plot 2

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02196-9

https://academic.oup.com/icb/article/62/6/1849/6698719#389865038

https://github.com/philippinespire/PIRE_TemporalSims_HPC/blob/main/stairscripts/stairscript_n20n50n100.sh

https://evomics.org/wp-content/uploads/2020/01/WoG_PopGenTutorial1_DemographicInference1.pdf

https://github.com/AlexLewanski/petrochromis_project/tree/main/project_scripts

Manually downloaded Stairway jar files and placed them in the programs folder
program needs to be run in this directory

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/stairway_plot_v2.1.1/
```

downloading some scripts for SFS

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs
git clone https://github.com/marqueda/SFS-scripts.git #didn't like this one. Kept erroring out for my pop file :(
git clone https://github.com/shenglin-liu/vcf2sfs.git
```

filter VCF

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_27.stairway_filtering.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=2GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

# load modules
module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/stairway/

VCF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/SNP_bcftools/myna_42inds.vcf.gz
DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR

#NZ pop
vcftools --gzvcf ${VCF} --keep ${DIR}/NZ_LEI_individuals.txt --min-meanDP 5 --max-meanDP 50 --recode --out stairway_NZ_LEI
#After filtering, kept 988015764 out of a possible 1037434143 Sites

#IND MP
vcftools --gzvcf ${VCF} --keep ${DIR}/IND_MP_individuals.txt --min-meanDP 5 --max-meanDP 50 --recode --out stairway_IND
#After filtering, kept 994757348 out of a possible 1037434143 Sites
vcftools --gzvcf stairway_IND.recode.vcf --mac 1 --recode --out stairway_IND_mac

#IND TN
vcftools --gzvcf ${VCF} --keep IND_TN_individuals.txt --min-meanDP 5 --max-meanDP 50 --recode --out stairway_IND_TN

vcftools --gzvcf stairway_IND_TN.recode.vcf --mac 1 --recode --out stairway_IND_TN_mac
```

Need to grab info for blueprint file (19 min runtime)

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_27.stairway_blueprint.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=25GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

module load R/4.1.0-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/stairway/
echo "Executing R ..."
srun Rscript stairway_SFS_calc.R
echo "R finished."


#stairway_SFS_calc.R
source("/nesi/nobackup/uoa02613/kstuart_projects/programs/vcf2sfs/vcf2sfs.r")

mygt<-vcf2gt("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/recombination_IND.recode.vcf",
"/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/stairway/IND_nonoutlier_individuals_POP.txt")

mysfs1<-gt2sfs.raw(mygt, "IND")
mysfs1_fold<- fold.sfs(mysfs1)
write.1D.fsc(mysfs1_fold,"/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/stairway/recombination_IND.sfs")

#
source("/nesi/nobackup/uoa02613/kstuart_projects/programs/vcf2sfs/vcf2sfs.r")

mygt<-vcf2gt("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/recombination_NZ_LEI.recode.vcf",
"/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/stairway/NZ_LEI_individuals_pop.txt")

mysfs1<-gt2sfs.raw(mygt, "NZ")
mysfs1_fold<- fold.sfs(mysfs1)
write.1D.fsc(mysfs1_fold,"/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/stairway/recombination_NZ_LEI.sfs")

#
source("/nesi/nobackup/uoa02613/kstuart_projects/programs/vcf2sfs/vcf2sfs.r")

mygt<-vcf2gt("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/stairway/stairway_IND_mac.recode.vcf",
"/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/stairway/stairway_IND_mac_POP.txt")

mysfs1<-gt2sfs.raw(mygt, "IND")
mysfs1_fold<- fold.sfs(mysfs1)
write.1D.fsc(mysfs1_fold,"/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/stairway/stairway_IND_mac.sfs")

#
source("/nesi/nobackup/uoa02613/kstuart_projects/programs/vcf2sfs/vcf2sfs.r")

mygt<-vcf2gt("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/stairway/stairway_IND_TN_mac.recode.vcf",
"/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/stairway/stairway_IND_TN_mac_pop.txt")

mysfs1<-gt2sfs.raw(mygt, "IND")
mysfs1_fold<- fold.sfs(mysfs1)
write.1D.fsc(mysfs1_fold,"/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/stairway/stairway_IND_TN_mac.sfs")
```

Changes made to blueprint file:

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/stairway/

#manipulate output of sfs file for input into the edited blueprint file
tail -n 1 stairway_IND_mac.sfs | awk '{ $1=""; print}' |  sed 's/ 0 0 0 0 0 0 0//g'  > stairway_IND_mac_foldbins.sfs
sfs=$(cat stairway_IND_mac_foldbins.sfs)

#move to stairway folder
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/stairway_plot_v2.1.1/

#edit the blueprint file and save a new copy
sed -e 's/popid: two-epoch_fold/popid: IND/g' \
-e 's/nseq: 30/nseq: 14/g' \
-e 's/L: 10000000/L: 1010982654/g' \
-e 's/estimation: 15/estimation: 20/g' \
-e 's/mu: 1.2e-8/mu: 4.6e-9/g' \
-e 's/generation: 24/generation: 2/g' \
-e 's/xrange: 0.1,10000/xrange: 0.1,1000/g' \
-e 's/project_dir: two-epoch_fold/project_dir: recombination_IND/g' \
-e "7s/.*/SFS:${sfs}/g" \
two-epoch_fold.blueprint > recombination_IND.blueprint

#V2: 994757348

#run make bash
java -cp stairway_plot_es Stairbuilder recombination_IND.blueprint
java -cp stairway_plot_es Stairbuilder recombination_IND_V2.blueprint


#NZ LEI
#manipulate output of sfs file for input into the edited blueprint file
tail -n 1 recombination_NZ_LEI.sfs | awk '{ $1=""; print}' |  sed 's/ 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0//g'  > recombination_NZ_LEI_foldbins.sfs
sfs=$(cat recombination_NZ_LEI_foldbins.sfs)

#move to stairway folder
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/stairway_plot_v2.1.1/

#edit the blueprint file and save a new copy
sed -e 's/popid: two-epoch_fold/popid: NZ_LEI/g' \
-e 's/nseq: 30/nseq: 16/g' \
-e 's/L: 10000000/L: 1010982654/g' \
-e 's/estimation: 15/estimation: 8/g' \
-e 's/mu: 1.2e-8/mu: 4.6e-9/g' \
-e 's/generation: 24/generation: 2/g' \
-e 's/project_dir: two-epoch_fold/project_dir: recombination_NZ_LEI/g' \
-e "7s/.*/SFS:${sfs}/g" \
two-epoch_fold.blueprint > recombination_NZ_LEI.blueprint

#inital L: 1010982654
#V2 now trying: 988015764

#run make bash
java -cp stairway_plot_es Stairbuilder recombination_IND.blueprint

java -cp stairway_plot_es Stairbuilder recombination_NZ_LEI.blueprint
java -cp stairway_plot_es Stairbuilder recombination_NZ_LEI_V2.blueprint


#IND TN
#manipulate output of sfs file for input into the edited blueprint file
tail -n 1 stairway_IND_TN_mac.sfs | awk '{ $1=""; print}' |  sed 's/ 0 0 0 0 0 0 0 0//g'  > stairway_IND_TN_mac_foldbins.sfs
sfs=$(cat stairway_IND_TN_mac_foldbins.sfs)

#move to stairway folder
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/stairway_plot_v2.1.1/

#edit the blueprint file and save a new copy
sed -e 's/popid: two-epoch_fold/popid: IND/g' \
-e 's/nseq: 30/nseq: 16/g' \
-e 's/L: 10000000/L: 995123647/g' \
```

```
-e 's/estimation: 15/estimation: 20/g' \
-e 's/mu: 1.2e-8/mu: 4.6e-9/g' \
-e 's/generation: 24/generation: 2/g' \
-e 's/xrange: 0.1,10000/xrange: 0.1,1000/g' \
-e 's/project_dir: two-epoch_fold/project_dir: recombination_IND_TN/g' \
-e "7s/.*/SFS:${sfs}/g" \
two-epoch_fold.blueprint > recombination_IND_TN.blueprint

#manually edited for the y axis range of 0.1,2000
cp recombination_IND_TN.blueprint recombination_IND_TN_edit.blueprint
cp recombination_NZ_LEI_V2.blueprint recombination_NZ_LEI_V2_edit.blueprint

#run make bash
java -cp stairway_plot_es Stairbuilder recombination_IND_TN_edit.blueprint
java -cp stairway_plot_es Stairbuilder recombination_NZ_LEI_V2_edit.blueprint
```

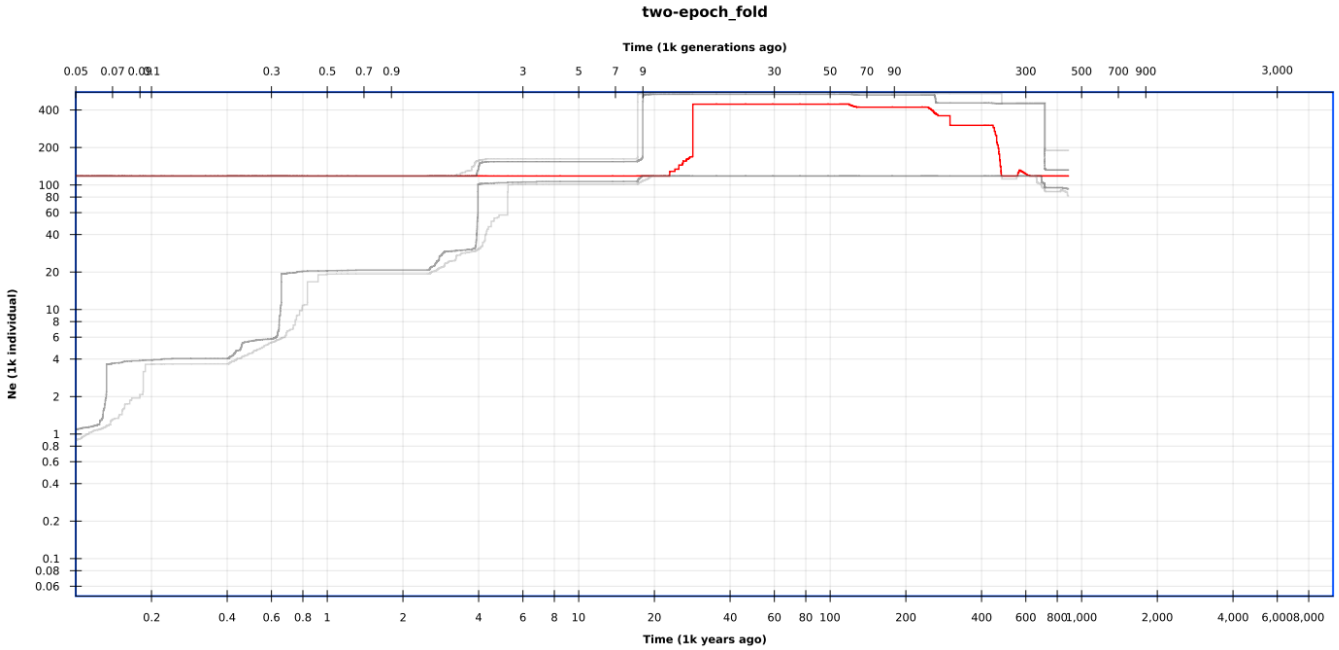
Now run the generated bash script

```
#!/bin/bash -e

#SBATCH --job-name=2023_03_22.stairway_run.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=25GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/programs/stairway_plot_v2.1.1/

bash recombination_NZ_LEI.blueprint.sh
bash recombination_NZ_LEI_V2.blueprint.sh

bash recombination_IND.blueprint.sh
bash recombination_IND_V2.blueprint.sh

bash recombination_IND_TN.blueprint.sh

bash recombination_IND_TN_edit.blueprint.sh  (edited for y axis)
bash recombination_NZ_LEI_V2_edit.blueprint.sh  (edited for y axis)
```

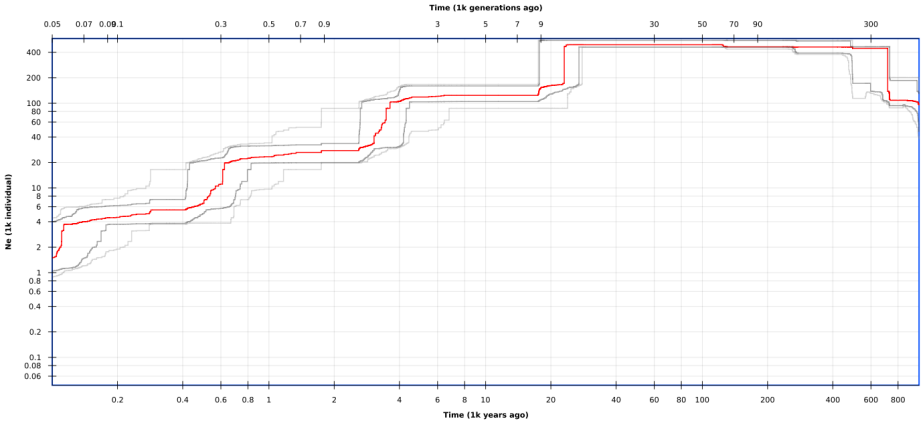
run plotting script

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/stairway_plot_v2.1.1/

bash recombination_NZ_LEI.blueprint.plot.sh
bash recombination_NZ_LEI_V2.blueprint.plot.sh

bash recombination_IND.blueprint.plot.sh
bash recombination_IND_V2.blueprint.plot.sh

bash recombination_IND_TN.blueprint.plot.sh

bash recombination_IND_TN_edit.blueprint.plot.sh  (edited for y axis)
bash recombination_NZ_LEI_V2_edit.blueprint.plot.sh  (edited for y axis)
```

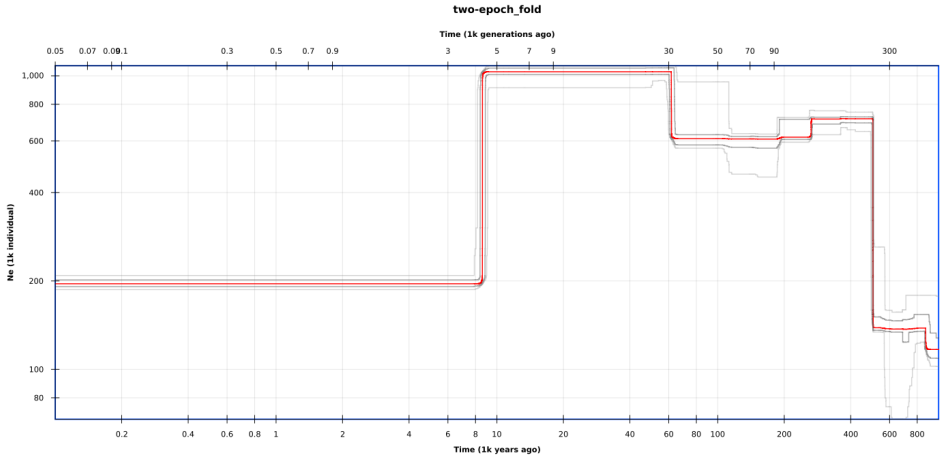
Recombination_NZ_LEI:



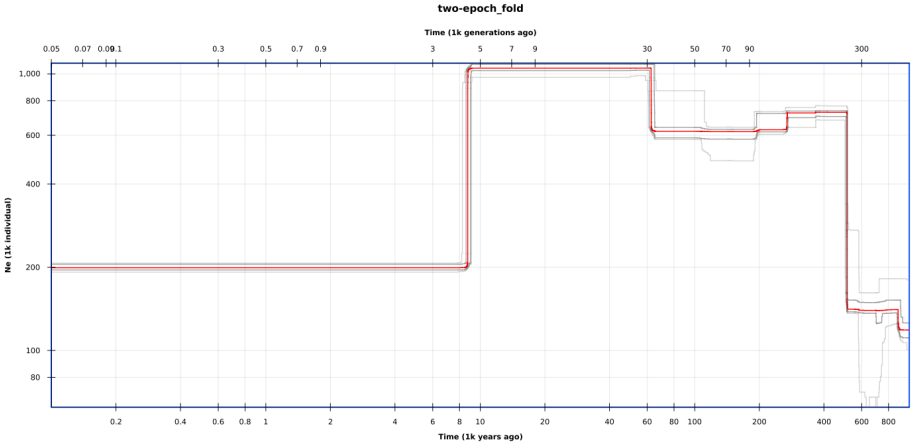
Recombination_NZ_LEI_V2:



Recombination_IND:



Recombination_IND_V2:


two-epoch_fold
Time (1k generations ago)
0.05 0.07 0.09 0.1 0.3 0.5 0.7 0.9 3 5 7 9 30 50 70 90 300
Ne (1k individual)
1,000 800 600 400 200 100 80
0.2 0.4 0.6 0.8 1 2 4 6 8 10 20 40 60 80 100 200 400 600 800
Time (1k years ago)