Starling-May18

Projects/Katarina Stuart/KStuart.Starling-Aug18/Sv4_Historic/Analysis/2021-10-17.RR_OutlierAnalysis

PDF Version generated by

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

on

Nov 20, 2021 @01:47 PM AEDT

Table of Contents

2021-10-17.RR_OutlierAnalysis

Katarina Stuart (z5188231@ad.unsw.edu.au) - Oct 20, 2021, 5:46 PM AEDT

Outlier Analysis with AU split

Refiltering data for outlier analysis:

```
module load java/8u121
module load samtools/1.10
module load picard/2.18.26
module load gatk/4.1.0.0
module load stacks/2.2
module load vcftools/0.1.16
```

Run populations to grab only the SNPs in the histric samples (from the older run through)

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/processing/align/bwa_aln_alignment
OUTDIR=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis
populations -P ./ -M ../historic_populations.txt -O ${OUTDIR}/ --vcf --lnl_lim -15 --write_random_snp -t 16
cd $OUTDIR
HIST=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/variant/keepind_hist.txt
vcftools --vcf populations.snps.vcf --keep $HIST --max-missing 0.5 --minGQ 15 --minDP 2 --out bwaaln_hist_miss05 --recode
```

After filtering, kept 12218 out of a possible 115684 Sites

grab list of SNPs present in 50% of historic individuals:

```
grep -v "^##" bwaaln_hist_miss05.recode.vcf | cut -f3 > bwaaln_hist_miss05_SNPid.txt
```

Then go back to the original directory with the BAM files, and rerun populations with all individuals

```
vcftools --vcf populations.snps.vcf --snps bwaaln_hist_miss05_SNPid.txt --out bwaaln_allsample_selection_histSNPs --recode
```

After filtering, kept 12218 out of a possible 115684 Sites

This was the file used for splitting for final outlier analysis:

```
vcftools --vcf bwaaln_allsample_selection_histSNPs.recode.vcf --maf 0.025 --out bwaaln_allsample_selection_histSNPs_maf025 --recode
```

After filtering, kept 5054 out of a possible 12218 Sites

Split it into the pairwise analysis (with AU split into AUeast and AUsouth):

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated
```

```
VCF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis/bwaaln_allsample_selection_histSNPs_maf025.recode.vcf
cp $VCF .
```

```
vcftools --vcf ${VCF} --keep ./keepind_UKHS.txt --max-missing-count 90 --recode --out bwaaln_selection_maf025_miss10_ukhist
```

```
vcftools --vcf ${VCF} --keep ./keepind_AUeastHS.txt --max-missing-count 30 --recode --out bwaaln_selection_maf025_miss10_aueasthist
vcftools --vcf ${VCF} --keep ./keepind_AUsouthHS.txt --max-missing-count 30 --recode --out bwaaln_selection_maf025_miss10_ausouthhist
```

```
vcftools --vcf ${VCF} --keep ./keepind_UKAUeast.txt --max-missing-count 100 --recode --out bwaaln_selection_maf025_miss10_ukaueast
vcftools --vcf ${VCF} --keep ./keepind_UKAUsouth.txt --max-missing-count 100 --recode --out bwaaln_selection_maf025_miss10_ukausouth
```

Outlier Identification

Fst Track:

https://github.com/vcflib/vcflib

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7239393/

https://sourceforge.net/p/vcftools/mailman/message/33476364/

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/fst_sliding_windows
```

```
VCF1=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_ukhist.recode.vcf
```

```
VCF2=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_aueasthist.recode.vcf
```

```
VCF3=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_ausouthhist.recode.vcf
```

```
VCF4=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_ukaueast.recode.vcf
```

```
VCF5=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_ukausouth.recode.vcf
```

```
vcftools --vcf ${VCF1} --weir-fst-pop ../keepind_uk.txt --weir-fst-pop ../keepind_hist.txt --fst-window-size 900000 --fst-window-step 10000 --out fstwindow100kb.ukhist
```

```
vcftools --vcf ${VCF2} --weir-fst-pop ../keepind_AUeast.txt --weir-fst-pop ../keepind_hist.txt --fst-window-size 900000 --fst-window-step 10000 --out fstwindow100kb.aueasthist
```

```
vcftools --vcf ${VCF3} --weir-fst-pop ../keepind_AUsouth.txt --weir-fst-pop ../keepind_hist.txt --fst-window-size 900000 --fst-window-step 10000 --out fstwindow100kb.ausouthhist
```

```
vcftools --vcf ${VCF4} --weir-fst-pop ../keepind_uk.txt --weir-fst-pop ../keepind_AUeast.txt --fst-window-size 900000 --fst-window-step 10000 --out fstwindow100kb.ukaueast
```

```
vcftools --vcf ${VCF5} --weir-fst-pop ../keepind_uk.txt --weir-fst-pop ../keepind_AUsouth.txt --fst-window-size 900000 --fst-window-step 10000 --out fstwindow100kb.ukausouth
```

Calculate 99th percentile, create bed files

```
cut -f5 fstwindow100kb.ukhist.windowed.weir.fst | tail -n +2 | sort -g |  awk '{all[NR] = $0} END{print all[int(NR*0.99 - 0.5)]}'
cat fstwindow100kb.ukhist.windowed.weir.fst | awk '$5>0.347392' | cut -f1,2,3 > fstwindow100kb.ukhist_outlierwindows.txt
```

```
cut -f5 fstwindow100kb.aueasthist.windowed.weir.fst | tail -n +2 | sort -g | awk '{all[NR] = $0} END{print all[int(NR*0.99 - 0.5)]}'
cat fstwindow100kb.aueasthist.windowed.weir.fst | awk '$5>0.337772' | cut -f1,2,3 > fstwindow100kb.aueasthist_outlierwindows.txt
```

```
cut -f5 fstwindow100kb.ausouthhist.windowed.weir.fst | tail -n +2 | sort -g | awk '{all[NR] = $0} END{print all[int(NR*0.99 - 0.5)]}'
cat fstwindow100kb.ausouthhist.windowed.weir.fst | awk '$5>0.345862' | cut -f1,2,3 > fstwindow100kb.ausouthhist_outlierwindows.txt
```

```
cut -f5 fstwindow100kb.ukaueast.windowed.weir.fst | tail -n +2 | sort -g | awk '{all[NR] = $0} END{print all[int(NR*0.99 - 0.5)]}'
cat fstwindow100kb.ukaueast.windowed.weir.fst | awk '$5>0.307692' | cut -f1,2,3 > fstwindow100kb.ukaueast_outlierwindows.txt
```

```
cut -f5 fstwindow100kb.ukausouth.windowed.weir.fst | tail -n +2 | sort -g | awk '{all[NR] = $0} END{print all[int(NR*0.99 - 0.5)]}'
cat fstwindow100kb.ukausouth.windowed.weir.fst | awk '$5>0.302023' | cut -f1,2,3 > fstwindow100kb.ukausouth_outlierwindows.txt
```

Grab windows in the 99th percentile of weighted fst

```
VCF1=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_ukhist.recode.vcf
```

```
VCF2=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_aueasthist.recode.vcf
```

```
VCF3=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_ausouthhist.recode.vcf
```

```
VCF4=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_ukaueast.recode.vcf
```

```
VCF5=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_ukausouth.recode.vcf
```

```
vcftools --vcf ${VCF1} --bed fstwindow100kb.ukhist_outlierwindows.txt --recode --out bwaaln_selection_maf025_miss10_ukhist_fstswoutliers
```

```
vcftools --vcf ${VCF2} --bed fstwindow100kb.aueasthist_outlierwindows.txt --recode --out bwaaln_selection_maf025_miss10_aueasthist_fstswoutliers
```

```
vcftools --vcf ${VCF3} --bed  fstwindow100kb.ausouthhist_outlierwindows.txt --recode --out bwaaln_selection_maf025_miss10_ausouthhist_fstswoutliers
```

```
vcftools --vcf ${VCF4} --bed  fstwindow100kb.ukaueast_outlierwindows.txt --recode --out bwaaln_selection_maf025_miss10_ukaueast_fstswoutliers
```

```
vcftools --vcf ${VCF5} --bed fstwindow100kb.ukausouth_outlierwindows.txt --recode --out bwaaln_selection_maf025_miss10_ukausouth_fstswoutliers
```

Calculate all FST of SNPs within the above 3 files:

```
vcftools --vcf ./bwaaln_selection_maf025_miss10_ukhist_fstswoutliers.recode.vcf --weir-fst-pop ../keepind_uk.txt --weir-fst-pop ../keepind_hist.txt --
out fst_window100kb_bysnps_ukhist
vcftools --vcf ./bwaaln_selection_maf025_miss10_aueasthist_fstswoutliers.recode.vcf --weir-fst-pop ../keepind_AUeast.txt --weir-fst-pop ../keepind_hist.txt --
out fst_window100kb_bysnps_aueasthist
vcftools --vcf ./bwaaln_selection_maf025_miss10_ausouthhist_fstswoutliers.recode.vcf --weir-fst-pop ../keepind_AUsouth.txt --weir-fst-pop ../keepind_hist.txt --
out fst_window100kb_bysnps_ausouthhist
vcftools --vcf ./bwaaln_selection_maf025_miss10_ukaueast_fstswoutliers.recode.vcf --weir-fst-pop ../keepind_uk.txt --weir-fst-pop ../keepind_AUeast.txt --
out fst_window100kb_bysnps_ukaueast
vcftools --vcf ./bwaaln_selection_maf025_miss10_ukausouth_fstswoutliers.recode.vcf --weir-fst-pop ../keepind_uk.txt --weir-fst-pop ../keepind_AUsouth.txt --
out fst_window100kb_bysnps_ukausouth
```

Due to filtering out pops, some SNP sites contain no polymorphisms, hence Fst is NA. Filter these out for ease.

```
cat fst_window100kb_bysnps_ukhist.weir.fst | sed '/nan/d' > fst_window100kb_bysnps_ukhist.weir.fst.rmnan
cat fst_window100kb_bysnps_aueasthist.weir.fst | sed '/nan/d' > fst_window100kb_bysnps_aueasthist.weir.fst.rmnan
cat fst_window100kb_bysnps_ausouthhist.weir.fst | sed '/nan/d' > fst_window100kb_bysnps_ausouthhist.weir.fst.rmnan
cat fst_window100kb_bysnps_ukaueast.weir.fst | sed '/nan/d' > fst_window100kb_bysnps_ukaueast.weir.fst.rmnan
cat fst_window100kb_bysnps_ukausouth.weir.fst | sed '/nan/d' > fst_window100kb_bysnps_ukausouth.weir.fst.rmnan
```

Plot the FST of these SNPS

module load R/3.5.3

R

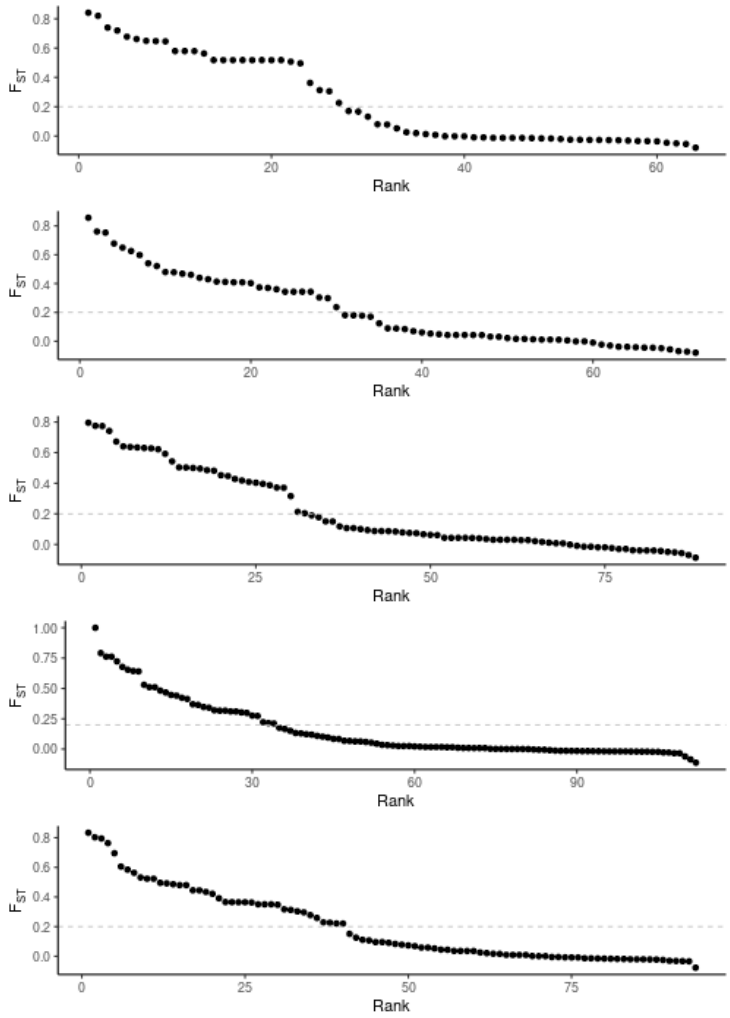
```
library(ggplot2)
setwd("/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/fst_sliding_windows")

ukhist.fst <- read.table("fst_window100kb_bysnps_ukhist.weir.fst.rmnan", header=T)
order.ukhist.fst<-order(ukhist.fst$WEIR_AND_COCKERHAM_FST, decreasing = TRUE)
ukhist.fst$rank <- NA
ukhist.fst$rank[order.ukhist.fst] <- 1:nrow(ukhist.fst)
```

```
A <- ggplot(ukhist.fst, aes(x=rank, y=WEIR_AND_COCKERHAM_FST)) + geom_point() + theme_classic() + xlab("Rank") + ylab(expression(F[ST]))
+ geom_hline(yintercept = 0.2, linetype="dashed", color = "grey", size=0.4)

auehist.fst <- read.table("fst_window100kb_bysnps_aueasthist.weir.fst.rmnan", header=T)
order.auehist.fst<-order(auehist.fst$WEIR_AND_COCKERHAM_FST, decreasing = TRUE)
auehist.fst$rank <- NA
auehist.fst$rank[order.auehist.fst] <- 1:nrow(auehist.fst)
B <- ggplot(auehist.fst, aes(x=rank, y=WEIR_AND_COCKERHAM_FST)) + geom_point() + theme_classic() + xlab("Rank") + ylab(expression(F[ST]))
+ geom_hline(yintercept = 0.2, linetype="dashed", color = "grey", size=0.4)

aushist.fst <- read.table("fst_window100kb_bysnps_ausouthhist.weir.fst.rmnan", header=T)
order.aushist.fst<-order(aushist.fst$WEIR_AND_COCKERHAM_FST, decreasing = TRUE)
aushist.fst$rank <- NA
aushist.fst$rank[order.aushist.fst] <- 1:nrow(aushist.fst)
C <- ggplot(aushist.fst, aes(x=rank, y=WEIR_AND_COCKERHAM_FST)) + geom_point() + theme_classic() + xlab("Rank") + ylab(expression(F[ST]))
+ geom_hline(yintercept = 0.2, linetype="dashed", color = "grey", size=0.4)

ukaue.fst <- read.table("fst_window100kb_bysnps_ukaueast.weir.fst.rmnan", header=T)
order.ukaue.fst<-order(ukaue.fst$WEIR_AND_COCKERHAM_FST, decreasing = TRUE)
ukaue.fst$rank <- NA
ukaue.fst$rank[order.ukaue.fst] <- 1:nrow(ukaue.fst)
D <- ggplot(ukaue.fst, aes(x=rank, y=WEIR_AND_COCKERHAM_FST)) + geom_point() + theme_classic() + xlab("Rank") + ylab(expression(F[ST]))
+ geom_hline(yintercept = 0.2, linetype="dashed", color = "grey", size=0.4)

ukaus.fst <- read.table("fst_window100kb_bysnps_ukausouth.weir.fst.rmnan", header=T)
order.ukaus.fst<-order(ukaus.fst$WEIR_AND_COCKERHAM_FST, decreasing = TRUE)
ukaus.fst$rank <- NA
ukaus.fst$rank[order.ukaus.fst] <- 1:nrow(ukaus.fst)
E <- ggplot(ukaus.fst, aes(x=rank, y=WEIR_AND_COCKERHAM_FST)) + geom_point() + theme_classic() + xlab("Rank") + ylab(expression(F[ST]))
+ geom_hline(yintercept = 0.2, linetype="dashed", color = "grey", size=0.4)
```

```
library(gridExtra)
library(grid)
library(lattice)
```

```
lay <- rbind(c("A"),
             c("B"),
c("C"),
c("D"),
             c("E"))
```

```
png("Sv4_fst_5panel.png", width=500, height=700)
grid.arrange(A, B, C, D, E, layout_matrix = lay)
dev.off()
```



After plotting discard FST's below the point on the curve that has the sharpest change from high to low fst. Grab SNP ID's from above the plotted FST threshold.

```
grep -v "^#" bwaaln_allsample_selection_histSNPs_maf025.recode.vcf  | cut -f1,2,3 > bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt
```

```
cat fst_window100kb_bysnps_ukhist.weir.fst.rmnan | awk '$3>0.2' | cut -f1,2 > fst_window100kb_bysnps_ukhist.weir.fst.outlier
cat fst_window100kb_bysnps_aueasthist.weir.fst.rmnan | awk '$3>0.2' | cut -f1,2 > fst_window100kb_bysnps_aueasthist.weir.fst.outlier
cat fst_window100kb_bysnps_ausouthhist.weir.fst.rmnan | awk '$3>0.2' | cut -f1,2 > fst_window100kb_bysnps_ausouthhist.weir.fst.outlier
cat fst_window100kb_bysnps_ukaueast.weir.fst.rmnan | awk '$3>0.2' | cut -f1,2 > fst_window100kb_bysnps_ukaueast.weir.fst.outlier
cat fst_window100kb_bysnps_ukausouth.weir.fst.rmnan | awk '$3>0.2' | cut -f1,2 > fst_window100kb_bysnps_ukausouth.weir.fst.outlier
```

```
awk -F"\t" 'FILENAME=="fst_window100kb_bysnps_ukhist.weir.fst.outlier"{A[$1$2]=$1$2} FILENAME=="../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt"
{if(A[$1$2]){print}}' fst_window100kb_bysnps_ukhist.weir.fst.outlier ../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt >
fst_window100kb_bysnps.ukhist.weir.fst.outlier.SNPlist
```

```
awk -F"\t" 'FILENAME=="fst_window100kb_bysnps_aueasthist.weir.fst.outlier"{A[$1$2]=$1$2}
FILENAME=="../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt"{if(A[$1$2])
{print}}' fst_window100kb_bysnps_aueasthist.weir.fst.outlier ../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt >
fst_window100kb_bysnps.aueasthist.weir.fst.outlier.SNPlist
```

```
awk -F"\t" 'FILENAME=="fst_window100kb_bysnps_ausouthhist.weir.fst.outlier"{A[$1$2]=$1$2}
FILENAME=="../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt"{if(A[$1$2])
{print}}' fst_window100kb_bysnps_ausouthhist.weir.fst.outlier ../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt >
fst_window100kb_bysnps.ausouthhist.weir.fst.outlier.SNPlist
```

```
awk -F"\t" 'FILENAME=="fst_window100kb_bysnps_ukaueast.weir.fst.outlier"{A[$1$2]=$1$2}
FILENAME=="../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt"{if(A[$1$2])
{print}}' fst_window100kb_bysnps_ukaueast.weir.fst.outlier ../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt >
fst_window100kb_bysnps.ukaueast.weir.fst.outlier.SNPlist
```

```
awk -F"\t" 'FILENAME=="fst_window100kb_bysnps_ukausouth.weir.fst.outlier"{A[$1$2]=$1$2}
FILENAME=="../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt"{if(A[$1$2])
```

```
{print}}' fst_window100kb_bysnps_ukausouth.weir.fst.outlier ../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt >
fst_window100kb_bysnps.ukausouth.weir.fst.outlier.SNPlist
```

Outliers using Bayescan:

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan/
```

```
module load vcftools/0.1.16
```

```
module load bayescan/2.1
```

```
module load R/3.5.3
```

Run PGDSpider: (have to go in multiple steps because the pop file is not being picked up)

to PGD format, then bayescan:

```
for PAIR in ukhist aueasthist ausouthhist ukaueast ukausouth
do
java -Xmx120g -Xms512m -jar /srv/scratch/z5188231/KStuart.Starling-Aug18/programs/PGDSpider_2.1.1.5/PGDSpider2-cli.jar -inputfile
../bwaaln_selection_maf025_miss10_${PAIR}.recode.vcf -inputformat VCF -outputfile
bwaaln_selection_maf025_miss10_${PAIR}_PGD.txt -outputformat PGD -spid historic_VCF_PGD.spid
java -Xmx120g -Xms512m -jar /srv/scratch/z5188231/KStuart.Starling-Aug18/programs/PGDSpider_2.1.1.5/PGDSpider2-cli.jar -inputfile
bwaaln_selection_maf025_miss10_${PAIR}_PGD.txt -inputformat PGD -outputfile bwaaln_selection_maf025_miss10_${PAIR}_BS.txt -
outputformat GESTE_BAYE_SCAN
done
```

BAYESCAN RUNS

```
#!/bin/bash

#PBS -N 2021-10-17.historical_bayescan.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=124gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan/

module load bayescan/2.1

for PAIR in ukhist aueasthist ausouthhist ukaueast ukausouth
do
bayescan_2.1 ./bwaaln_selection_maf025_miss10_${PAIR}_BS.txt -od ./ -threads 16 -n 5000 -thin 10 -nbp 20 -pilot 5000 -burn 50000 -
pr_odds 10
done
```

Identify outliers:

```
module load R/3.5.3
R
library(ggplot2)
setwd("/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan")
```

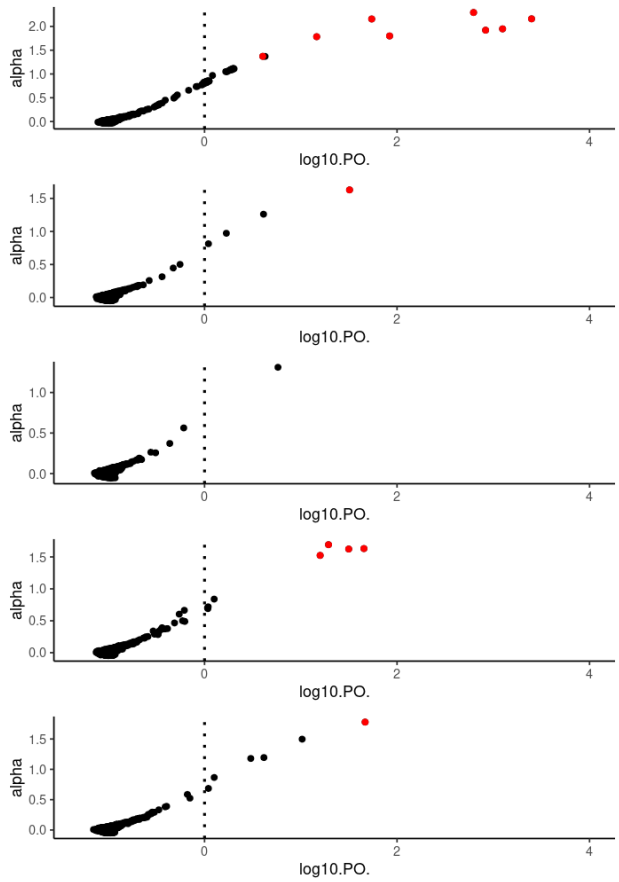
```
source("/srv/scratch/z5188231/KStuart.Starling-Aug18/programs/BayeScan2.1/Rfunctions/plot_R.r")
outliers.ukhist=plot_bayescan("/srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan/bwaaln_selection_maf025_miss10_ukhist_BS_fst.txt")
outliers.ukhist
outliers.ukhist=plot_bayescan("/srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv4_Historic/analysis/outlier_analysis/bayescan/bayescan_maf025_miss10/bwaaln_selection_maf025_miss10_ukhist_BS_fst.txt",FDR=0.05)
outliers.ukhist


outliers.auehist=plot_bayescan("/srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan/bwaaln_selection_maf025_miss10_aueasthist_BS_fst.txt",FDR=0.05 )
outliers.aushist=plot_bayescan("/srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan/bwaaln_selection_maf025_miss10_ausouthhist_BS_fst.txt",FDR=0.05 )
outliers.ukaue=plot_bayescan("/srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan/bwaaln_selection_maf025_miss10_ukaueast_BS_fst.txt",FDR=0.05 )
outliers.ukaus=plot_bayescan("/srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan/bwaaln_selection_maf025_miss10_ukausouth_BS_fst.txt",FDR=0.05 )

write.table(outliers.ukhist$outliers, file="bayscan_outliers_ukhist.txt")
write.table(outliers.auehist$outliers, file="bayscan_outliers_aueasthist.txt")
write.table(outliers.aushist$outliers, file="bayscan_outliers_ausouthhist.txt")
write.table(outliers.ukaue$outliers, file="bayscan_outliers_ukaueast.txt")
write.table(outliers.ukaus$outliers, file="bayscan_outliers_ukausouth.txt")

###PLOTTING

bayescan.out.ukhist <- read.table("bwaaln_selection_maf025_miss10_ukhist_BS_fst.txt", header=TRUE)
bayescan.out.ukhist$num <- seq.int(nrow(bayescan.out.ukhist))
bayescan.out.ukhist.outliers <- filter(bayescan.out.ukhist, num %in% outliers.ukhist$outliers)

bayescan.out.aueasthist <- read.table("bwaaln_selection_maf025_miss10_aueasthist_BS_fst.txt", header=TRUE)
bayescan.out.aueasthist$num <- seq.int(nrow(bayescan.out.aueasthist))
bayescan.out.aueasthist.outliers <- filter(bayescan.out.aueasthist, num %in% outliers.auehist$outliers)

bayescan.out.ausouthhist <- read.table("bwaaln_selection_maf025_miss10_ausouthhist_BS_fst.txt", header=TRUE)
bayescan.out.ausouthhist$num <- seq.int(nrow(bayescan.out.ausouthhist))
bayescan.out.ausouthhist.outliers <- filter(bayescan.out.ausouthhist, num %in% outliers.aushist$outliers)

bayescan.out.ukaueast <- read.table("bwaaln_selection_maf025_miss10_ukaueast_BS_fst.txt", header=TRUE)
bayescan.out.ukaueast$num <- seq.int(nrow(bayescan.out.ukaueast))
bayescan.out.ukaueast.outliers <- filter(bayescan.out.ukaueast, num %in% outliers.ukaue$outliers)

bayescan.out.ukausouth <- read.table("bwaaln_selection_maf025_miss10_ukausouth_BS_fst.txt", header=TRUE)
bayescan.out.ukausouth$num <- seq.int(nrow(bayescan.out.ukausouth))
bayescan.out.ukausouth.outliers <- filter(bayescan.out.ukausouth, num %in% outliers.ukaus$outliers)

A<- ggplot(bayescan.out.ukhist, aes(x=log10.PO., y=alpha))+
geom_point(size=3,alpha=1)+xlim(-1.3,4)+ theme_classic(base_size = 18) + geom_vline(xintercept = 0, linetype="dotted", color = "black", size=1.5)+
geom_point(aes(x=log10.PO., y=alpha), data=bayescan.out.ukhist.outliers, col="red", fill="red",size=3,alpha=1)

B<- ggplot(bayescan.out.aueasthist, aes(x=log10.PO., y=alpha))+
geom_point(size=3,alpha=1)+xlim(-1.3,4)+ theme_classic(base_size = 18) + geom_vline(xintercept = 0, linetype="dotted", color = "black", size=1.5)+
geom_point(aes(x=log10.PO., y=alpha), data=bayescan.out.aueasthist.outliers, col="red", fill="red",size=3,alpha=1)

C<- ggplot(bayescan.out.ausouthhist, aes(x=log10.PO., y=alpha))+
geom_point(size=3,alpha=1)+xlim(-1.3,4)+ theme_classic(base_size = 18) + geom_vline(xintercept = 0, linetype="dotted", color = "black", size=1.5)+
geom_point(aes(x=log10.PO., y=alpha), data=bayescan.out.ausouthhist.outliers, col="red", fill="red",size=3,alpha=1)

D<- ggplot(bayescan.out.ukaueast, aes(x=log10.PO., y=alpha))+
geom_point(size=3,alpha=1)+xlim(-1.3,4)+ theme_classic(base_size = 18) + geom_vline(xintercept = 0, linetype="dotted", color = "black", size=1.5)+
geom_point(aes(x=log10.PO., y=alpha), data=bayescan.out.ukaueast.outliers, col="red", fill="red",size=3,alpha=1)
```

```
E<- ggplot(bayescan.out.ukausouth, aes(x=log10.PO., y=alpha))+
geom_point(size=3,alpha=1)+xlim(-1.3,4)+ theme_classic(base_size = 18) + geom_vline(xintercept = 0, linetype="dotted", color = "black", size=1.5)+
geom_point(aes(x=log10.PO., y=alpha), data=bayescan.out.ukausouth.outliers, col="red", fill="red",size=3,alpha=1)
```

```
library(gridExtra)
library(grid)
library(lattice)
```

```
lay <- rbind(c("A"),
             c("B"),
c("C"),
c("D"),
             c("E"))
```

```
png("Sv4_outliers_bayescan_5panel.png", width=700, height=1000)
grid.arrange(A, B, C, D, E, layout_matrix = lay)
dev.off()
```



Grab list of SNPS

```
VCF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis/bwaaln_allsample_selection_histSNPs_maf025.recode.vcf
```

```
grep -v "^##" $VCF | cut -f3 > snplist_UKAUHS.txt
```

Use this to work out which of the SNP results from bayescan had what SNP name

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan
```

```
for PAIR in ukhist aueasthist ausouthhist ukaueast ukausouth
do
grep -v "^##" ../bwaaln_selection_maf025_miss10_${PAIR}.recode.vcf | cut -f3 > snplist_${PAIR}.txt
awk -F'\t' -v OFS='\t' 'NR>1 { $(NF+1)=NR-1} 1' snplist_${PAIR}.txt > snplist_numbered_${PAIR}.txt
```

```
awk '{print $2}' bayscan_outliers_${PAIR}.txt > bayscan_outliersnums_${PAIR}.txt
awk 'FNR==NR{a[$1];next} (($2) in a)' bayscan_outliersnums_${PAIR}.txt snplist_numbered_${PAIR}.txt | cut -f1
> bayscan_outliersnums_${PAIR}_SNPs.txt
done
```

Variant Counting

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/variant_counts
```

Fst Sliding Window Outliers:

```
Fst_ukhist=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/Fst_outliers/bwaaln_selection_maf025_miss10_ukhist_outliers_SNPlist.txt
Fst_aueasthist=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/Fst_outliers/bwaaln_selection_maf025_miss10_aueasthist_outliers_SNPlist.txt
Fst_ausouthhist=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/Fst_outliers/bwaaln_selection_maf025_miss10_ausouthhist_outliers_SNPlist.txt
Fst_ukaueast=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/Fst_outliers/bwaaln_selection_maf025_miss10_ukaueast_outliers_SNPlist.txt
Fst_ukausouth=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/Fst_outliers/bwaaln_selection_maf025_miss10_ukausouth_outliers_SNPlist.txt
```

```
Fst_ukhist=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/fst_sliding_windows/fst_window100kb_bysnps.ukhist.weir.fst.outlier.SNPlist
Fst_aueasthist=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/fst_sliding_windows/fst_window100kb_bysnps.aueasthist.weir.fst.outlier.SNPlist
Fst_ausouthhist=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/fst_sliding_windows/fst_window100kb_bysnps.ausouthhist.weir.fst.outlier.SNPlist
Fst_ukaueast=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/fst_sliding_windows/fst_window100kb_bysnps.ukaueast.weir.fst.outlier.SNPlist
Fst_ukausouth=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/fst_sliding_windows/fst_window100kb_bysnps.ukausouth.weir.fst.outlier.SNPlist
```

Bayescan Outliers:

```
Bay_ukhist=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan/bayscan_outliersnums_ukhist_SNPs.txt
Bay_aueasthist=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan/bayscan_outliersnums_aueasthist_SNPs.txt
```

```
Bay_ausouthhist=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan/bayscan_outliersnums_ausouthhist_SNPs.txt
Bay_ukaueast=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan/bayscan_outliersnums_ukaueast_SNPs.txt
Bay_ukausouth=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bayescan/bayscan_outliersnums_ukausouth_SNPs.txt
```

Grab SNP lists

```
cut -f3 ${Fst_ukhist} > snplist_Fst_ukhist.txt
cut -f3 ${Fst_aueasthist} > snplist_Fst_aueasthist.txt
cut -f3 ${Fst_ausouthhist} > snplist_Fst_ausouthhist.txt
cut -f3 ${Fst_ukaueast} > snplist_Fst_ukaueast.txt
cut -f3 ${Fst_ukausouth} > snplist_Fst_ukausouth.txt
tail -n +2 ${Bay_ukhist} > snplist_Bay_ukhist.txt
tail -n +2 ${Bay_aueasthist} > snplist_Bay_aueasthist.txt
tail -n +2 ${Bay_ausouthhist} > snplist_Bay_ausouthhist.txt
tail -n +2 ${Bay_ukaueast} > snplist_Bay_ukaueast.txt
tail -n +2 ${Bay_ukausouth} > snplist_Bay_ukausouth.txt
```

Merge into total list of outlier across all methods:

```
sort snplist_Fst_ukhist.txt snplist_Bay_ukhist.txt | uniq > snplist_merged_UKHist.txt
sort snplist_Fst_aueasthist.txt snplist_Bay_aueasthist.txt | uniq > snplist_merged_AUeHist.txt
sort snplist_Fst_ausouthhist.txt snplist_Bay_ausouthhist.txt | uniq > snplist_merged_AUsHist.txt
sort snplist_Fst_ukaueast.txt snplist_Bay_ukaueast.txt | uniq > snplist_merged_UKAUe.txt
sort snplist_Fst_ukausouth.txt snplist_Bay_ukausouth.txt | uniq > snplist_merged_UKAUs.txt
```

flag overlapping SNPs through a combination of the below code and manual stuff in excel:

```
sort snplist_Bay_ukhist.txt snplist_Fst_ukhist.txt | uniq -d | wc -l

sort snplist_Bay_aueasthist.txt snplist_Fst_aueasthist.txt | uniq -d | wc -l
sort snplist_Bay_aueasthist.txt snplist_Fst_ausouthhist.txt | uniq -d | wc -l
sort snplist_Fst_aueasthist.txt snplist_Fst_ausouthhist.txt | uniq -d | wc -l

sort snplist_Bay_ukaueast.txt snplist_Fst_ukaueast.txt | uniq -d | wc -l
sort snplist_Bay_ukaueast.txt snplist_Bay_ukausouth.txt | uniq -d | wc -l
sort snplist_Fst_ukaueast.txt snplist_Bay_ukausouth.txt | uniq -d | wc -l
sort snplist_Bay_ukaueast.txt snplist_Fst_ukausouth.txt | uniq -d | wc -l
sort snplist_Fst_ukaueast.txt snplist_Fst_ukausouth.txt | uniq -d | wc -l
sort snplist_Fst_ukausouth.txt snplist_Bay_ukausouth.txt | uniq -d | wc -l
```

Venn Diagram:

[How to Plot Venn Diagrams Using R, ggplot2 and ggforce – Scripts & Statistics (wordpress.com)]

```
module load R/3.6.3
```

```
R
```

```
#install.packages("ggvenn")
library("ggvenn")
```

```
#if (!require(devtools)) install.packages("devtools")
#devtools::install_github("gaospecial/ggVennDiagram")
```

```
setwd("/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/variant_counts")
```

```
#library("ggVennDiagram")


```

```
UKHS <- scan("snplist_merged_UKHist.txt", what = "", quiet=TRUE)
AUeHS <- scan("snplist_merged_AUeHist.txt", what = "", quiet=TRUE)
UKAUe <- scan("snplist_merged_UKAUe.txt", what = "", quiet=TRUE)
```

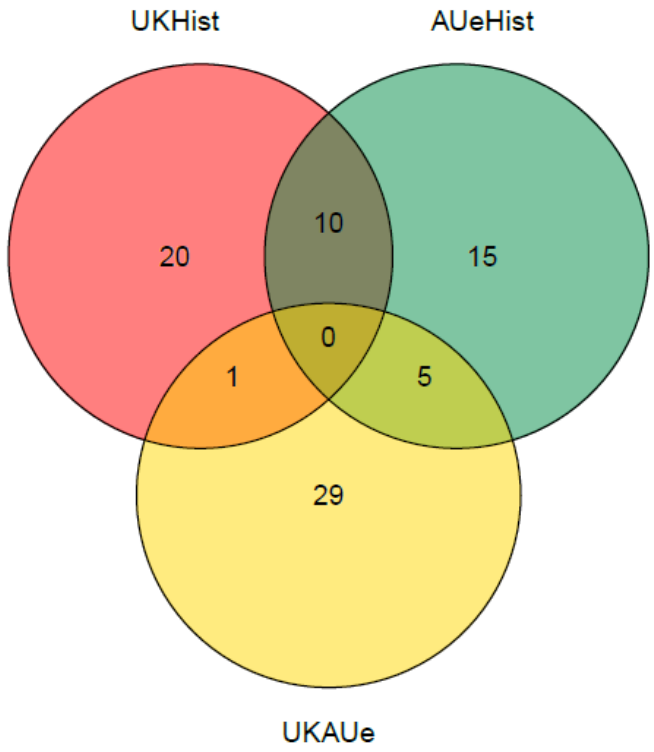
```
x <- list('UKHS'=UKHS, 'AUeHS'=AUeHS, 'UKAUe'=UKAUe)
```

```
pdf("Sv4_pairwiseSNP_overlap_selection_AUe.pdf")
ggvenn( x, columns = c("UKHS", "AUeHS", "UKAUe"), stroke_size = 0.5, fill_color = c("red", "springgreen4", "gold"),
show_percentage = FALSE, text_size = 6)
dev.off()
####
```

```
UKHS <- scan("snplist_merged_UKHist.txt", what = "", quiet=TRUE)
AUsHS <- scan("snplist_merged_AUsHist.txt", what = "", quiet=TRUE)
UKAUs <- scan("snplist_merged_UKAUs.txt", what = "", quiet=TRUE)
```

```
x <- list('UKHS'=UKHS, 'AUsHS'=AUsHS, 'UKAUs'=UKAUs)
```

```
pdf("Sv4_pairwiseSNP_overlap_selection_AUs.pdf")
ggvenn(x, columns = c("UKHS", "AUsHS", "UKAUs"), stroke_size = 0.5, fill_color = c("red", "springgreen4", "gold"),
show_percentage = FALSE, text_size = 6)
dev.off()
```



Count SNPs in actual data set

VCF1=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_ukhist.recode.vcf

VCF2=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_aueasthist.recode.vcf

```
VCF3=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_ausouthhist.recode.vcf
VCF4=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_ukaueast.recode.vcf
VCF5=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_selection_maf025_miss10_ukausouth.recode.vcf
grep -v "^#" ${VCF1} | cut -f3 | wc -l
grep -v "^##" ${VCF2} | cut -f3 | wc -l
grep -v "^#" ${VCF3} | cut -f3 | wc -l
grep -v "^#" ${VCF4} | cut -f3 | wc -l
grep -v "^#" ${VCF5} | cut -f3 | wc -l
```

Made summary SNP lists for all the SNPs over each method to be used in adaptive analysis

Need to split the SNP lists so that they belong into the correct 5 catagory splits.

No outliers in all three data set, so pick up pairwise overlap:

AUeast

```
awk 'NR==FNR { lines[$0]=1; next } $0 in lines' snplist_merged_UKHist.txt snplist_merged_UKAUe.txt > candidate_divergentUK_e_SNPs.txt
awk 'NR==FNR { lines[$0]=1; next } $0 in lines' snplist_merged_AUeHist.txt snplist_merged_UKAUe.txt > candidate_divergentAU_e_SNPs.txt
awk 'NR==FNR { lines[$0]=1; next } $0 in lines' snplist_merged_UKHist.txt snplist_merged_AUeHist.txt > candidate_parallel_e_SNPs.txt
```

```
awk 'NR==FNR{a[$0];next} !($0 in a)' candidate_parallel_e_SNPs.txt snplist_merged_UKHist.txt > pseudocandidate_selectionUK_e_SNPs.txt
awk 'NR==FNR{a[$0];next} !($0 in a)' candidate_divergentUK_e_SNPs.txt pseudocandidate_selectionUK_e_SNPs.txt  > candidate_selectionUK_e_SNPs.txt
```

```
awk 'NR==FNR{a[$0];next} !($0 in a)' candidate_parallel_e_SNPs.txt snplist_merged_AUeHist.txt > pseudocandidate_selectionAU_e_SNPs.txt
awk 'NR==FNR{a[$0];next} !($0 in a)' candidate_divergentAU_e_SNPs.txt pseudocandidate_selectionAU_e_SNPs.txt  > candidate_selectionAU_e_SNPs.txt
```

```
awk 'NR==FNR{a[$0];next} !($0 in a)' candidate_divergentUK_e_SNPs.txt snplist_merged_UKAUe.txt > pseudocandidate_selectionUKAU_e_SNPs.txt
awk 'NR==FNR{a[$0];next} !($0 in a)' candidate_divergentAU_e_SNPs.txt pseudocandidate_selectionUKAU_e_SNPs.txt > candidate_selectionUKAU_e_SNPs.txt
```

AUsouth

```
awk 'NR==FNR { lines[$0]=1; next } $0 in lines' snplist_merged_UKHist.txt snplist_merged_UKAUs.txt > candidate_divergentUK_s_SNPs.txt
awk 'NR==FNR { lines[$0]=1; next } $0 in lines' snplist_merged_AUsHist.txt snplist_merged_UKAUs.txt > candidate_divergentAU_s_SNPs.txt
awk 'NR==FNR { lines[$0]=1; next } $0 in lines' snplist_merged_UKHist.txt snplist_merged_AUsHist.txt > candidate_parallel_s_SNPs.txt
```

```
awk 'NR==FNR{a[$0];next} !($0 in a)' candidate_parallel_s_SNPs.txt snplist_merged_UKHist.txt > pseudocandidate_selectionUK_s_SNPs.txt
awk 'NR==FNR{a[$0];next} !($0 in a)' candidate_divergentUK_s_SNPs.txt pseudocandidate_selectionUK_s_SNPs.txt  > candidate_selectionUK_s_SNPs.txt
```

```
awk 'NR==FNR{a[$0];next} !($0 in a)' candidate_parallel_s_SNPs.txt snplist_merged_AUsHist.txt >
pseudocandidate_selectionAU_s_SNPs.txt
awk 'NR==FNR{a[$0];next} !($0 in a)' candidate_divergentAU_s_SNPs.txt pseudocandidate_selectionAU_s_SNPs.txt  >
candidate_selectionAU_s_SNPs.txt
```

```
awk 'NR==FNR{a[$0];next} !($0 in a)' candidate_divergentUK_s_SNPs.txt snplist_merged_UKAUs.txt >
pseudocandidate_selectionUKAU_s_SNPs.txt
awk 'NR==FNR{a[$0];next} !($0 in a)' candidate_divergentAU_s_SNPs.txt pseudocandidate_selectionUKAU_s_SNPs.txt >
candidate_selectionUKAU_s_SNPs.txt
```

Merge the data (after checking for overlaps between grouping types (e.g. parallel and UK divergent).

```
sort candidate_divergentUK_e_SNPs.txt  candidate_divergentUK_s_SNPs.txt  | uniq > candidate_divergentUK_SNPs.txt
sort candidate_divergentAU_e_SNPs.txt  candidate_divergentAU_s_SNPs.txt  | uniq > candidate_divergentAU_SNPs.txt
sort candidate_divergentUK_SNPs.txt candidate_divergentAU_SNPs.txt | uniq > candidate_divergent_SNPs.txt
```

```
sort candidate_parallel_e_SNPs.txt  candidate_parallel_s_SNPs.txt | uniq > candidate_parallel_SNPs.txt
```

```
sort candidate_selectionUK_e_SNPs.txt candidate_selectionUK_s_SNPs.txt | uniq > candidate_selectionUK_SNPs.txt
sort candidate_selectionAU_e_SNPs.txt candidate_selectionAU_s_SNPs.txt | uniq > candidate_selectionAU_SNPs.txt
sort candidate_selectionUKAU_e_SNPs.txt candidate_selectionUKAU_s_SNPs.txt | uniq > candidate_selectionUKAU_SNPs.txt
```

Grab the CHROM and POS of each of these files:

```
awk 'FNR==NR{a[$1];next} (($3) in a)' candidate_divergent_SNPs.txt ../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt
> candidate_divergent_SNPs_CHROMPOS.txt
awk 'FNR==NR{a[$1];next} (($3) in a)' candidate_parallel_SNPs.txt ../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt
> candidate_parallel_SNPs_CHROMPOS.txt
awk 'FNR==NR{a[$1];next} (($3) in a)' candidate_selectionUK_SNPs.txt ../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt
> candidate_selectionUK_SNPs_CHROMPOS.txt
awk 'FNR==NR{a[$1];next} (($3) in a)' candidate_selectionAU_SNPs.txt ../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt
> candidate_selectionAU_SNPs_CHROMPOS.txt
awk 'FNR==NR{a[$1];next} (($3) in
a)' candidate_selectionUKAU_SNPs.txt ../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt >
candidate_selectionUKAU_SNPs_CHROMPOS.txt
```

FOR VISUALISING and mapping to SNP ID:

Vlookup onto the below complete one:

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/selection_variant_density
```

```
#first file just to create an output that has ALL the SNPS that are present across the different pops. The actual Fst vals are not important.
vcftools --vcf ../bwaaln_allsample_selection_histSNPs_maf025.recode.vcf --weir-fst-pop ../keepind_uk.txt --weir-fst-
pop ../keepind_hist.txt --out bwaaln_selection_maf025_miss10_all
#then get the SNPs for all the things??
vcftools --vcf ../bwaaln_selection_maf025_miss10_ukhist.recode.vcf --weir-fst-pop ../keepind_uk.txt --weir-fst-pop ../keepind_hist.txt --out bwaaln_selection_maf025_miss10_ukhist
vcftools --vcf ../bwaaln_selection_maf025_miss10_aueasthist.recode.vcf --weir-fst-pop ../keepind_AUeast.txt --weir-fst-pop ../keepind_hist.txt --out
bwaaln_selection_maf025_miss10_aueasthist
vcftools --vcf ../bwaaln_selection_maf025_miss10_ausouthhist.recode.vcf --weir-fst-pop ../keepind_AUsouth.txt --weir-fst-pop ../keepind_hist.txt --out
bwaaln_selection_maf025_miss10_ausouthhist
vcftools --vcf ../bwaaln_selection_maf025_miss10_ukaueast.recode.vcf --weir-fst-pop ../keepind_uk.txt --weir-fst-pop ../keepind_AUeast.txt --out
```

bwaaln_selection_maf025_miss10_ukaueast
vcftools --vcf ../bwaaln_selection_maf025_miss10_ukausouth.recode.vcf **--weir-fst-pop** ../keepind_uk.txt **--weir-fst-pop** ../keepind_AUsouth.txt --out bwaaln_selection_maf025_miss10_ukausouth

#Creating a cat of CHROM.POS so I can vlookup onto the full SNP list
awk '{print $1" "$2" "$1"."$2" "$3}' bwaaln_selection_maf025_miss10_all.weir.fst > graphing.bwaaln_selection_maf025_miss10_all
awk '{print $1" "$2" "$1"."$2" "$3}' bwaaln_selection_maf025_miss10_ukhist.weir.fst > graphing.bwaaln_selection_maf025_miss10_ukhist
awk '{print $1" "$2" "$1"."$2" "$3}' bwaaln_selection_maf025_miss10_aueasthist.weir.fst > graphing.bwaaln_selection_maf025_miss10_aueasthist
awk '{print $1" "$2" "$1"."$2" "$3}' bwaaln_selection_maf025_miss10_ausouthhist.weir.fst > graphing.bwaaln_selection_maf025_miss10_ausouthhist
awk '{print $1" "$2" "$1"."$2" "$3}' bwaaln_selection_maf025_miss10_ukaueast.weir.fst > graphing.bwaaln_selection_maf025_miss10_ukaueast
awk '{print $1" "$2" "$1"."$2" "$3}' bwaaln_selection_maf025_miss10_ukausouth.weir.fst > graphing.bwaaln_selection_maf025_miss10_ukausouth
#then do the actual vlooking up
awk -f vlookup.awk graphing.bwaaln_selection_maf025_miss10_ukhist graphing.bwaaln_selection_maf025_miss10_all > graphing.bwaaln_selection_maf025_miss10_ukhist.prep
awk -f vlookup.awk graphing.bwaaln_selection_maf025_miss10_aueasthist graphing.bwaaln_selection_maf025_miss10_all > graphing.bwaaln_selection_maf025_miss10_aueasthist.prep
awk -f vlookup.awk graphing.bwaaln_selection_maf025_miss10_ausouthhist graphing.bwaaln_selection_maf025_miss10_all > graphing.bwaaln_selection_maf025_miss10_ausouthhist.prep
awk -f vlookup.awk graphing.bwaaln_selection_maf025_miss10_ukaueast graphing.bwaaln_selection_maf025_miss10_all > graphing.bwaaln_selection_maf025_miss10_ukaueast.prep
awk -f vlookup.awk graphing.bwaaln_selection_maf025_miss10_ukausouth graphing.bwaaln_selection_maf025_miss10_all > graphing.bwaaln_selection_maf025_miss10_ukausouth.prep

(currently committing to this....)

#OR JUST 3 ROWS
vcftools --vcf ../bwaaln_allsample_selection_histSNPs_maf025.recode.vcf **--weir-fst-pop** ../keepind_uk.txt **--weir-fst-pop** ../keepind_hist.txt --out fst_bysnps.ukhist
vcftools --vcf ../bwaaln_allsample_selection_histSNPs_maf025.recode.vcf **--weir-fst-pop** ../keepind_aus.txt **--weir-fst-pop** ../keepind_hist.txt --out fst_bysnps.auhist
vcftools --vcf ../bwaaln_allsample_selection_histSNPs_maf025.recode.vcf **--weir-fst-pop** ../keepind_uk.txt **--weir-fst-pop** ../keepind_aus.txt --out fst_bysnps.ukau

#and chrom sizes for plotting nicely at chromosome position rather than just in order of SNPs
GENOMEFAI=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/genome/Sturnus_vulgaris_2.3.1.simp.fasta.fai
cut -f1,2 $GENOMEFAI > sizes.genome
awk '{print $0, s+=$2}' sizes.genome > sizes.genome.cumulative
#stuff it, just fixed it up in excel so rolling sum is for above rows only....
awk -f vlookup3.awk sizes.genome.cumulative.txt graphing.bwaaln_selection_maf025_miss10_all > graphing.sizes.genome

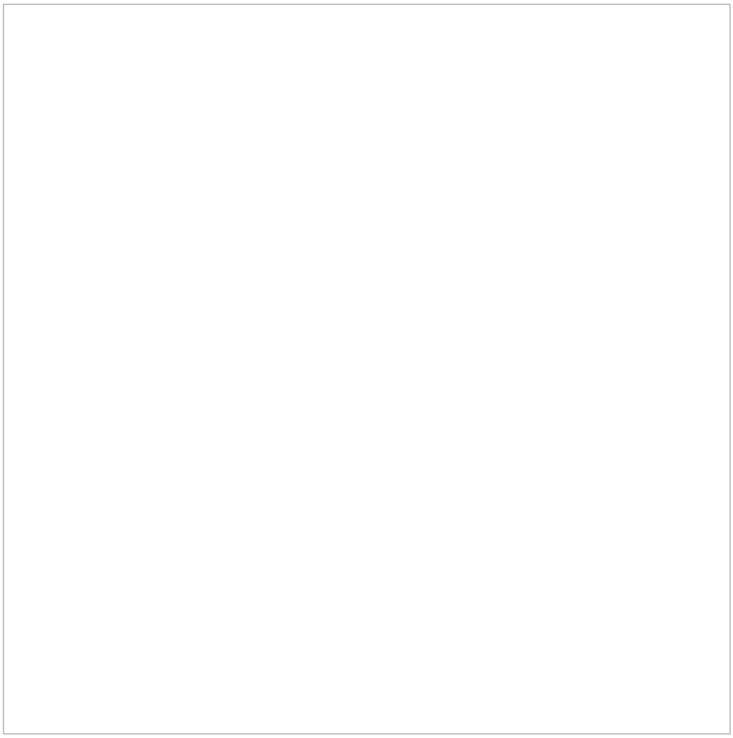
And the outlier SNPS:

cut -f1,2 ../variant_counts/snplist_merged_UKHist.recode.vcf > snplist_merged_UKHist_CHROMPOS.txt
cut -f1,2 snplist_merged_AUHist.recode.vcf > snplist_merged_AUHist_CHROMPOS.txt
cut -f1,2 snplist_merged_UKAU.recode.vcf > snplist_merged_UKAU_CHROMPOS.txt

for PAIR in UKHist AUeHist AUsHist UKAUe UKAUs
do
awk 'FNR==NR{a[$1];next} (($3) in a)' ../variant_counts/snplist_merged_${PAIR}.txt ../bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt | awk '{print $1" "$2" "$1"."$2}' > graphing.snplist_merged_${PAIR}_CHROMPOS
done

awk -f vlookup2.awk graphing.snplist_merged_UKHist_CHROMPOS graphing.bwaaln_selection_maf025_miss10_all > graphing.snplist_merged_UKHist_CHROMPOS.prep

```
awk -f vlookup2.awk graphing.snplist_merged_AUeHist_CHROMPOS graphing.bwaaln_selection_maf025_miss10_all > graphing.snplist_merged_AUeHist_CHROMPOS.prep
awk -f vlookup2.awk graphing.snplist_merged_AUsHist_CHROMPOS graphing.bwaaln_selection_maf025_miss10_all > graphing.snplist_merged_AUsHist_CHROMPOS.prep
awk -f vlookup2.awk graphing.snplist_merged_UKAUe_CHROMPOS graphing.bwaaln_selection_maf025_miss10_all > graphing.snplist_merged_UKAUe_CHROMPOS.prep
awk -f vlookup2.awk graphing.snplist_merged_UKAUs_CHROMPOS graphing.bwaaln_selection_maf025_miss10_all > graphing.snplist_merged_UKAUs_CHROMPOS.prep

paste -d "\t" graphing.sizes.genome fst_bysnps.ukhist.weir.fst fst_bysnps.auhist.weir.fst fst_bysnps.ukau.weir.fst graphing.snplist_merged_UKHist_CHROMPOS.prep graphing.snplist_merged_AUeHist_CHROMPOS.prep graphing.snplist_merged_AUsHist_CHROMPOS.prep graphing.snplist_merged_UKAUe_CHROMPOS.prep graphing.snplist_merged_UKAUs_CHROMPOS.prep > graphing.table.allcolumns.txt
awk '{print $1" "$4" "$4+$2" "$5" "$8" "$11" "$13" "$15" "$17" "$19" "$21}' graphing.table.allcolumns.txt > graphing.table.txt
```



MAPPING VARIANTS

Identifying Candidate Genes

one way is to use the intersect function from BedTools.
Allows us to grab genes that have been identified in Table 3: **Annotated genes**

```
module load bedtools
```

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/mapping_variants
conda activate AGAT
agat_sp_keep_longest_isoform.pl --gff $GFF -o myFile_lociMerged_longestIsoform.gff
```

GENES:

```
awk 'BEGIN{FS=OFS="\t"}{if(NR>1){ print $1,$2-1,$2 }}' ../variant_counts/candidate_divergent_SNPs_CHROMPOS.txt |
bedtools intersect -wb -a myFile_lociMerged_longestIsoform.gff -b stdin | awk '$3=="gene"' >
candidate_genes_divergent.txt
```

```
awk 'BEGIN{FS=OFS="\t"}{if(NR>1){ print $1,$2-1,$2 }}' ../variant_counts/candidate_parallel_SNPs_CHROMPOS.txt |
bedtools intersect -wb -a myFile_lociMerged_longestIsoform.gff -b stdin | awk '$3=="gene"' >
candidate_genes_parallel.txt
```

```
awk 'BEGIN{FS=OFS="\t"}{if(NR>1){ print $1,$2-1,$2 }}' ../variant_counts/candidate_selectionUK_SNPs_CHROMPOS.txt |
bedtools intersect -wb -a myFile_lociMerged_longestIsoform.gff -b stdin | awk '$3=="gene"' >
candidate_genes_UK_HS.txt
```

```
awk 'BEGIN{FS=OFS="\t"}{if(NR>1){ print $1,$2-1,$2 }}' ../variant_counts/candidate_selectionAU_SNPs_CHROMPOS.txt |
bedtools intersect -wb -a myFile_lociMerged_longestIsoform.gff -b stdin | awk '$3=="gene"' >
candidate_genes_AU_HS.txt
```

```
awk 'BEGIN{FS=OFS="\t"}{if(NR>1){ print $1,$2-1,$2
}}' ../variant_counts/candidate_selectionUKAU_SNPs_CHROMPOS.txt |
bedtools intersect -wb -a myFile_lociMerged_longestIsoform.gff -b stdin | awk '$3=="gene"' >
candidate_genes_UK_AU.txt
```

Use grep to count known and unknow genes

genes

```
for i in divergent parallel UK_HS AU_HS UK_AU
do
wc -l candidate_genes_$i.txt
grep "Similar to" candidate_genes_$i.txt | wc -l
grep "unknown function" candidate_genes_$i.txt | wc -l
done
```

```
find candidate_genes_UK_AU_transcript.txt -type f -name "*.html" -exec grep -l -f file.txt '{}' \; -print
```

```
grep -f toxin.txt candidate_genes_UK_AU_transcript.txt
```

VEP:

Sorting files GFF

Svulgaris.all.renamed.func.sort.gff.gz = not the maker only one, the new combined one with GO terms.

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/vep
```

```
module load genometools/1.5.9
```

```
module load samtools
module load vcftools
```

GFF:

```
gt gff3 -sortlines yes ../mapping_variants/myFile_lociMerged_longestIsoform.gff > longestIsoform.sort.gff
bgzip -c longestIsoform.sort.gff > longestIsoform.sort.gff.gz
tabix -p gff longestIsoform.sort.gff.gz
```

VCF:

```
VCF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_allsample_selection_histSNPs_maf025.recode.vcf
vcftools --vcf $VCF --snps ../variant_counts/candidate_divergent_SNPs.txt --out snpID_divergence --recode
vcftools --vcf $VCF --snps ../variant_counts/candidate_parallel_SNPs.txt --out snpID_parrellel --recode
vcftools --vcf $VCF --snps ../variant_counts/candidate_selectionUK_SNPs.txt --out snpID_uk_hist --recode
vcftools --vcf $VCF --snps ../variant_counts/candidate_selectionAU_SNPs.txt --out snpID_au_hist --recode
vcftools --vcf $VCF --snps ../variant_counts/candidate_selectionUKAU_SNPs.txt --out snpID_uk_au --recode
```

GENOME:

```
GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/genome/Sturnus_vulgaris_2.3.1.simp.fasta
```

VEP

https://asia.ensembl.org/info/docs/tools/vep/script/vep_cache.html#gff

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/vep
module load perl/5.28.0
module load samtools/1.10
DIR=/srv/scratch/z5188231/KStuart.Starling-Aug18/programs/ensembl-vep
```

```
perl $DIR/vep -i snpID_divergence.recode.vcf -gff longestIsoform.sort.gff.gz -fasta ${GENOME} -o vep_snpID_divergence
perl $DIR/vep -i snpID_parrellel.recode.vcf -gff longestIsoform.sort.gff.gz -fasta ${GENOME} -o vep_snpID_parrellel
perl $DIR/vep -i snpID_uk_hist.recode.vcf -gff longestIsoform.sort.gff.gz  -fasta ${GENOME} -o vep_snpID_uk_hist
perl $DIR/vep -i snpID_au_hist.recode.vcf -gff longestIsoform.sort.gff.gz -fasta ${GENOME} -o vep_snpID_au_hist
perl $DIR/vep -i snpID_uk_au.recode.vcf -gff longestIsoform.sort.gff.gz -fasta ${GENOME} -o vep_snpID_uk_au
```

Gene Lists

http://geneontology.org/docs/go-enrichment-analysis/

Add the gene list from the main VCF file, compare to the SNPs from the outlier lists to see if there is any statistically significant gene enrichment.

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/genes
module load bedtools/2.27.1
```

```
VCF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/bwaaln_allsample_selection_histSNPs_maf025.recode.vcf
```

```
grep -v "^##" $VCF | cut -f1-2 > snplist_chrompos_UKAUHIST.txt
```

```
GFF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation/annotation/Svulgaris.all.renamed.func.protdom.gff
```

```
awk 'BEGIN{FS=OFS="\t"}{if(NR>1){ print $1,$2-1,$2 }}' snplist_chrompos_UKAUHIST.txt| bedtools intersect -wb -a $GFF -b stdin | awk '$3=="gene"' > bwaaln_allsample_selection_histSNPs_maf025.gff
```

```
sed -nr 's/.*Similar to +([^ ]+) .*/\1/p' bwaaln_allsample_selection_histSNPs_maf025.gff | sed 's|[:,]||g' > genelist_selection_histSNPs_maf025.txt
```

```
sed -nr 's/.*Similar to +([^ ]+) .*/\1/p' ../mapping_variants/candidate_genes_parallel.txt | sed 's|[:,]||g' > genelist_parallel.txt
sed -nr 's/.*Similar to +([^ ]+) .*/\1/p' ../mapping_variants/candidate_genes_divergent.txt | sed 's|[:,]||g' > genelist_divergent.txt
sed -nr 's/.*Similar to +([^ ]+) .*/\1/p' ../mapping_variants/candidate_genes_UK_HS.txt | sed 's|[:,]||g' > genelist_uk_hist.txt
sed -nr 's/.*Similar to +([^ ]+) .*/\1/p' ../mapping_variants/candidate_genes_AU_HS.txt | sed 's|[:,]||g' > genelist_au_hist.txt
sed -nr 's/.*Similar to +([^ ]+) .*/\1/p' ../mapping_variants/candidate_genes_UK_AU.txt | sed 's|[:,]||g' > genelist_uk_au.txt
```

no significant results :(

Get GO terms:

module load R/3.5.3
R

if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")

BiocManager::install("GOSim")

setwd(" /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis/genes")

divergent <- read.table("genelist_divergent.txt")

getGOInfo(geneIDs)

Fixation at any of the outlier genes

Test to see if there are any loci out of HWE or fixed out of the ones that are important for selection....

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/genes/fixation
```

```
VCF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis/bwaaln_allsample_selection_histSNPs_maf025.recode.vcf
```

```
module load stacks/2.2
```

```
POP_MAP=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/processing/align/historic_populations.txt
```

Rerun with populations as main regions (3) rather than set to the sample sites (6)

```
sed 's|mv|au|g' ${POP_MAP}|sed 's|or|au|g' | sed 's|mw|uk|g' | sed 's|nc|uk|g' | sed 's|aw|uk|g' > historic_populations_region.txt
populations -V $VCF -M historic_populations_region.txt -O fstats_region/ --hwe --fstats -t 8
```

```
vcftools --vcf ../../vep/snpID_divergence.recode.vcf --keep ../../keepind_uk.txt --out fstats_uk_snpID_divergence --freq
vcftools --vcf ../../vep/snpID_divergence.recode.vcf --keep ../../keepind_AUsouth.txt --out fstats_AUsouth_snpID_divergence --freq
vcftools --vcf ../../vep/snpID_divergence.recode.vcf --keep ../../keepind_AUeast.txt --out fstats_AUeast_snpID_divergence --freq
vcftools --vcf ../../vep/snpID_divergence.recode.vcf --keep ../../keepind_hist.txt --out fstats_hist_snpID_divergence --freq
```

```
vcftools --vcf ../../vep/snpID_parrellel.recode.vcf --keep ../../keepind_uk.txt --out fstats_uk_snpID_parrellel --freq
vcftools --vcf ../../vep/snpID_parrellel.recode.vcf --keep ../../keepind_AUsouth.txt --out fstats_AUsouth_snpID_parrellel --freq
vcftools --vcf ../../vep/snpID_parrellel.recode.vcf --keep ../../keepind_AUeast.txt --out fstats_AUeast_snpID_parrellel --freq
vcftools --vcf ../../vep/snpID_parrellel.recode.vcf --keep ../../keepind_hist.txt --out fstats_hist_snpID_parrellel --freq
```

see what output is like, then see if I can grab overlap from list of loci, either on command line or in excel manually

```
VCF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis/filtering/bwaaln_allsample_selection_histSNPs_noantwerp_maf025.recode.vcf
```

```
grep -v "^##" $VCF | cut -f1-3 > snplist_chromposID_UKAUHIST.txt
```

amiGO: search of GO terms associated with a phrase

http://amigo.geneontology.org/amigo

http://amigo.geneontology.org/amigo/search/ontology

https://journals.plos.org/plosgenetics/article?rev=2&id=10.1371/journal.pgen.1008119

We then compared the observed SNP FST values to null exome-wide and per-site FST distributions generated by performing 1,500 neutral simulations under the best fitting population history for YNP *T. alpinus*.

https://david.ncifcrf.gov/

https://doc-openbio.readthedocs.io/projects/annovar/en/latest/user-guide/input/#-vcf4-format

https://github.com/sanger-pathogens/SnpEffWrapper

https://www.biorxiv.org/content/10.1101/452201v1.full.pdf

DISCARDED CODE

Fst Track:

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis_updated/Fst_outliers
```

```
vcftools --vcf ../bwaaln_selection_maf025_miss10_ukhist.recode.vcf --weir-fst-pop ../keepind_uk.txt --weir-fst-pop ../keepind_hist.txt --out bwaaln_selection_maf025_miss10_ukhist
vcftools --vcf ../bwaaln_selection_maf025_miss10_aueasthist.recode.vcf --weir-fst-pop ../keepind_AUeast.txt --weir-fst-pop ../keepind_hist.txt --out
bwaaln_selection_maf025_miss10_aueasthist
vcftools --vcf ../bwaaln_selection_maf025_miss10_ausouthhist.recode.vcf --weir-fst-pop ../keepind_AUsouth.txt --weir-fst-pop ../keepind_hist.txt --out
bwaaln_selection_maf025_miss10_ausouthhist
vcftools --vcf ../bwaaln_selection_maf025_miss10_ukaueast.recode.vcf --weir-fst-pop ../keepind_uk.txt --weir-fst-pop ../keepind_AUeast.txt --out
bwaaln_selection_maf025_miss10_ukaueast
vcftools --vcf ../bwaaln_selection_maf025_miss10_ukausouth.recode.vcf --weir-fst-pop ../keepind_uk.txt --weir-fst-pop ../keepind_AUsouth.txt --out
bwaaln_selection_maf025_miss10_ukausouth
```

Calculate 99th percentile, create files

```
grep -v "^#" ${VCF}  | cut -f1,2,3 > bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt
```

```
#Find top 99 percentile
cut -f3 bwaaln_selection_maf025_miss10_ukhist.weir.fst | sed '/nan/d' | tail -n +2 | sort -g | awk '{all[NR] = $0} END{print all[int(NR*0.99 - 0.5)]}'
cat bwaaln_selection_maf025_miss10_ukhist.weir.fst   | awk '$3>=0.51798' | sed '/nan/d' | cut -f1,2 > bwaaln_selection_maf025_miss10_ukhist_outliers.txt
```

```
cut -f3 bwaaln_selection_maf025_miss10_aueasthist.weir.fst | sed '/nan/d' | tail -n +2 | sort -g | awk '{all[NR] = $0} END{print all[int(NR*0.99 - 0.5)]}'
cat bwaaln_selection_maf025_miss10_aueasthist.weir.fst   | awk '$3>=0.402845' | sed '/nan/d' | cut -f1,2 >
bwaaln_selection_maf025_miss10_aueasthist_outliers.txt
cut -f3 bwaaln_selection_maf025_miss10_ausouthhist.weir.fst | sed '/nan/d' | tail -n +2 | sort -g | awk '{all[NR] = $0} END{print all[int(NR*0.99 - 0.5)]}'
cat bwaaln_selection_maf025_miss10_ausouthhist.weir.fst   | awk '$3>=0.417476' | sed '/nan/d' | cut -f1,2 >
bwaaln_selection_maf025_miss10_ausouthhist_outliers.txt
```

```
cut -f3 bwaaln_selection_maf025_miss10_ukaueast.weir.fst | sed '/nan/d' | tail -n +2 | sort -g | awk '{all[NR] = $0} END{print all[int(NR*0.99 - 0.5)]}'
cat bwaaln_selection_maf025_miss10_ukhist.weir.fst  | awk '$3>=0.361949' | sed '/nan/d' | cut -f1,2 >
bwaaln_selection_maf025_miss10_ukaueast_outliers.txt
cut -f3 bwaaln_selection_maf025_miss10_ukausouth.weir.fst | sed '/nan/d' | tail -n +2 | sort -g | awk '{all[NR] = $0} END{print all[int(NR*0.99 - 0.5)]}'
cat bwaaln_selection_maf025_miss10_ukausouth.weir.fst  | awk '$3>=0.405134' | sed '/nan/d' | cut -f1,2 >
bwaaln_selection_maf025_miss10_ukausouth_outliers.txt
```

```
#Grab SNP ID info
awk -F"\t" 'FILENAME=="bwaaln_selection_maf025_miss10_ukhist_outliers.txt"{A[$1$2]=$1$2}
FILENAME=="bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt"{if(A[$1$2])
{print}}' bwaaln_selection_maf025_miss10_ukhist_outliers.txt bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt >
bwaaln_selection_maf025_miss10_ukhist_outliers_SNPlist.txt
```

```
awk -F"\t" 'FILENAME=="bwaaln_selection_maf025_miss10_aueasthist_outliers.txt"{A[$1$2]=$1$2}
FILENAME=="bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt"{if(A[$1$2])
{print}}' bwaaln_selection_maf025_miss10_aueasthist_outliers.txt bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt >
bwaaln_selection_maf025_miss10_aueasthist_outliers_SNPlist.txt
awk -F"\t" 'FILENAME=="bwaaln_selection_maf025_miss10_ausouthhist_outliers.txt"{A[$1$2]=$1$2}
FILENAME=="bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt"{if(A[$1$2])
{print}}' bwaaln_selection_maf025_miss10_ausouthhist_outliers.txt bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt >
bwaaln_selection_maf025_miss10_ausouthhist_outliers_SNPlist.txt
```

```
awk -F"\t" 'FILENAME=="bwaaln_selection_maf025_miss10_ukaueast_outliers.txt"{A[$1$2]=$1$2}
FILENAME=="bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt"{if(A[$1$2])
{print}}' bwaaln_selection_maf025_miss10_ukaueast_outliers.txt bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt >
bwaaln_selection_maf025_miss10_ukaueast_outliers_SNPlist.txt
awk -F"\t" 'FILENAME=="bwaaln_selection_maf025_miss10_ukausouth_outliers.txt"{A[$1$2]=$1$2}
FILENAME=="bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt"{if(A[$1$2])
{print}}' bwaaln_selection_maf025_miss10_ukausouth_outliers.txt bwaaln_allsample_selection_histSNPs_maf025_SNPlist.txt >
bwaaln_selection_maf025_miss10_ukausouth_outliers_SNPlist.txt
```

Create Final Snp lists

```
sort bwaaln_selection_maf025_miss10_ukhist_outliers_SNPlist.txt | uniq > FSTsnplist_merged_UKHist.txt
sort bwaaln_selection_maf025_miss10_aueasthist_outliers_SNPlist.txt
bwaaln_selection_maf025_miss10_ausouthhist_outliers_SNPlist.txt | uniq > FSTsnplist_merged_AUHist.txt
sort bwaaln_selection_maf025_miss10_ukaueast_outliers_SNPlist.txt bwaaln_selection_maf025_miss10_ukausouth_outliers_SNPlist.txt |
uniq > FSTsnplist_merged_UKAU.txt
```

sort FSTsnplist_merged_UKHist.txt FSTsnplist_merged_AUHist.txt FSTsnplist_merged_UKAU.txt | uniq > xx.test.txt