Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/Nc3_HihiSV/Data/2024-02-23.Curation

PDF Version generated by

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

on

Aug 15, 2024 @03:06 PM NZST

Table of Contents

2024-02-23.Curation

Katarina Stuart (z5188231@ad.unsw.edu.au) - Jul 18, 2024, 1:34 PM GMT+12

Samplot: manually curating the SVs

vaguely following the process in:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10848878/
https://github.com/linneas/wolf-structural

Downloading the packages with conda

samplot:

https://github.com/ryanlayer/samplot

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs
module load Miniconda3/22.11.1-1
conda activate /nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba
conda install -c bioconda samplot
mamba install -c bioconda samplot
```

downloaded to the following location:
/nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/bin/samplot

plotcritic:

https://github.com/jbelyeu/PlotCritic

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs
module load Miniconda3/22.11.1-1
conda install -c bioconda plotcritic
conda activate /nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba
conda install -c bioconda samplot
mamba install -c bioconda samplot
```

downloaded to the following location:
/nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/bin/

create the dataframe used to make the samplots

Input meta data for R manipulation

```
cd /nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_curation

module load BCFtools/1.13-GCC-9.2.0
VCF=/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_survivor/hihi_genofiltered.recode.vcf

#ran through again to reinclude the singletons
VCF=/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_survivor/hihi_genofilteredextra.recode.vcf

#grab just the needed info
grep -v "^#" $VCF | sed 's/;/\t/g' | sed 's/SVLEN=\|-//g' |  sed 's/END=//g' | sed 's/SVTYPE=//g'  |  awk '{printf "%s %s %s %s %s %s ",
$3,$1,$2,$14,$10,$11; for(i=19;i<=49;i++) {sub(/:.*/, "", $i); printf "%s ", $i} print ""}' > hihi_filtered.temp1.vcf
sed -e 's/0\/0/homR/g' -e 's/0\/1/het/g' -e 's/1\/1/homA/g' -e 's/.\/./unkn/g'  hihi_filtered.temp1.vcf > hihi_filtered.temp2.vcf

#grab column labels
grep -v "^##" $VCF | grep "^#" | awk 'BEGIN { printf "ID CHROM START END LENGTH TYPE " } {for(i=10;i<=NF;i++) {sub(/:.*/, "", $i); printf "%s
", $i} print ""}' > hihi_filtered.temp3.vcf

#combine into one table
cat hihi_filtered.temp3.vcf hihi_filtered.temp2.vcf > hihi_filtered.temp4.vcf
```

create the file in R

```
module load R/4.1.0-gimkl-2020a
R
library(data.table)

setwd("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_curation")

df <- fread("hihi_filtered.temp4.vcf")

# Create a new data frame to store the selected individuals
selected_individuals <- data.frame(matrix(NA, nrow = nrow(df), ncol = 6))

# Loop through each row
for (i in 1:nrow(df)) {
  # Extract the indices of individuals labeled as 'homA' for the current row
  homR_indices <- which(df[i, ] == "homR")
  het_indices <- which(df[i, ] == "het")
  homA_indices <- which(df[i, ] == "homA")

  # Select two individuals randomly if there are at least two 'homR' individuals
  if (length(homR_indices) >= 2) {
    selected_homR <- sample(homR_indices, size = 2)
  } else if (length(homR_indices) == 1) {
    selected_homR <- c(homR_indices, NA)
  } else {
    selected_homR <- rep(NA, 2)
  }

  # Select two individuals randomly if there are at least two 'het' individuals
  if (length(het_indices) >= 2) {
    selected_het <- sample(het_indices, size = 2)
  } else if (length(het_indices) == 1) {
    selected_het <- c(het_indices, NA)
  } else {
    selected_het <- rep(NA, 2)
  }
```

```
  # Select two individuals randomly if there are at least two 'homA' individuals
  if (length(homA_indices) >= 2) {
    selected_homA <- sample(homA_indices, size = 2)
  } else if (length(homA_indices) == 1) {
    selected_homA <- c(homA_indices, NA)
  } else {
    selected_homA <- rep(NA, 2)
  }

repinds <- c(selected_homR,selected_het,selected_homA)

  # Store the selected individuals
 selected_individuals[i, ] <- colnames(df)[repinds]
}

# Add the selected individuals to the original data frame

cbind(df, selected_individuals) #manually inspected the genotypes to make sure it worked!
df_selected <- cbind(df[,1:6], selected_individuals)

#replace the NAs with the dummy in that we will use in the plot
df_selected[is.na(df_selected)] <- "xfakeind"

df_selected$xfakeind_count <- rowSums(df_selected  == 'xfakeind')

df_selected_DUP <- df_selected %>% filter(TYPE=="DUP")
df_selected_INV <- df_selected %>% filter(TYPE=="INV")
df_selected_DELa_small <- df_selected %>% filter(TYPE=="DEL" & xfakeind_count == 0 & LENGTH < 100)
df_selected_DELa_big <- df_selected %>% filter(TYPE=="DEL" & xfakeind_count == 0 & LENGTH >= 100)
df_selected_DELb  <- df_selected %>% filter(TYPE=="DEL" & xfakeind_count == 1)
df_selected_DELc_small <- df_selected %>% filter(TYPE=="DEL" & xfakeind_count == 2 & LENGTH < 300)
df_selected_DELc_big <- df_selected %>% filter(TYPE=="DEL" & xfakeind_count == 2 & LENGTH >= 300)

nrow(rbind(df_selected_DUP , df_selected_INV , df_selected_DELa_small ,df_selected_DELa_big ,df_selected_DELb
,df_selected_DELc_small , df_selected_DELc_big   ))
#[1] 2620
nrow(df)
#[1] 2620

#all there!

#time to save them
write.table(df_selected_DUP, file = "samplot_batch_DUP.txt", sep = "\t", row.names = FALSE, quote = FALSE, col.names=FALSE)
write.table(df_selected_INV, file = "samplot_batch_INV.txt", sep = "\t", row.names = FALSE, quote = FALSE, col.names=FALSE)
write.table(df_selected_DELa_small , file = "samplot_batch_DELasmall.txt", sep = "\t", row.names = FALSE, quote = FALSE,
col.names=FALSE)
write.table(df_selected_DELa_big , file = "samplot_batch_DELabig.txt", sep = "\t", row.names = FALSE, quote = FALSE, col.names=FALSE)
write.table(df_selected_DELb  , file = "samplot_batch_DELb.txt", sep = "\t", row.names = FALSE, quote = FALSE, col.names=FALSE)
write.table(df_selected_DELc_small , file = "samplot_batch_DELcsmall.txt", sep = "\t", row.names = FALSE, quote = FALSE,
col.names=FALSE)
write.table(df_selected_DELc_big , file = "samplot_batch_DELcbig.txt", sep = "\t", row.names = FALSE, quote = FALSE, col.names=FALSE)

write.table(df_selected , file = "samplot_batch_EXTRA.txt", sep = "\t", row.names = FALSE, quote = FALSE, col.names=FALSE)
```

create a dummy individual for the blank plots

```
cd /nesi/nobackup/uoa00338/kstuart_projects/Nc2_HihiWGS/data/mapped_reads
```

```
#picked this individual as it is very low coverage and so that will be quite obvious in the plots.
cp P8185.sorted.dup.bam xfakeind.sorted.dup.bam
cp P8185.sorted.dup.bam.bai xfakeind.sorted.dup.bam.bai
```

create samplots

```
#!/bin/bash -e

#SBATCH --job-name=2024_02_23.SVCalling_validation.sl
#SBATCH --account=uoa00338
#SBATCH --time=00-48:00:00
#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/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_curation

DIR=/nesi/nobackup/uoa00338/kstuart_projects/Nc2_HihiWGS/data/mapped_reads

for BATCH in DUP INV DELasmall DELabig DELb DELcsmall DELcbig;

for BATCH in EXTRA;

do

mkdir $(echo $BATCH)

while IFS= read -r line; do
    # Split the line into individual values
    read -r ID CHROM START END LENGTH TYPE HOMR1 HOMR2 HET1 HET2 HOMA1 HOMA2 COUNT <<< "$line"
    # Use the variables in your loop

VARIANTS=$(echo "-o" $BATCH/$ID "-c" $CHROM "-s" $START "-e" $END "-t" $TYPE "-d 100" -n $HOMR1 $HOMR2 $HET1 $HET2
$HOMA1 $HOMA2)
BAMS=$(echo "-b" ${DIR}/$HOMR1".sorted.dup.bam" ${DIR}/$HOMR2".sorted.dup.bam" ${DIR}/$HET1".sorted.dup.bam"
${DIR}/$HET2".sorted.dup.bam" ${DIR}/$HOMA1".sorted.dup.bam" ${DIR}/$HOMA2".sorted.dup.bam")

SAMPLOT=$(echo "samplot plot" $VARIANTS $BAMS)
echo $ID
$SAMPLOT

done < samplot_batch_$BATCH.txt

done
```

realised that the colons in the file names of some SNP IDs were giving issue - replace with underscores for now, and convert back later.

```
cd /nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_curation
```

```
find . -type f -name "Manta*.png" -exec sh -c 'mv "$1" "${1//:/_}"' _ {} \;
```

and use plot critic to compile them

needed to be opened in chrome an not edge

```
#!/bin/bash -e

#SBATCH --job-name=2024_03_01.SVCalling_validation_plotcritic.sl
#SBATCH --account=uoa00338
#SBATCH --time=00-48:00:00
#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/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_curation

for BATCH in DUP INV DELasmall DELabig DELb DELcsmall DELcbig;

do

plotcritic \
  -p ${BATCH}_plot \
  -i $BATCH/ \
  -q "Is this a structural variant?" \
  -A "Y":"Yep it is a SV" "S":"Strong Maybe" "M":"Maybe..." "N":"No way it's just mess" \

done
```

Should read the SAMPLOT paper before proceeding:
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02380-5

Pull them onto local pc to curate

import json into R and find the good variants

```
module load R/4.1.0-gimkl-2020a
R
library(ggplot2)
library(dplyr)
library(stringr)
library(data.table)

setwd("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_curation")
library(jsonlite)
```

```
DELasmall <- fromJSON("Katarina_DELasmall_plot_report.json", flatten=TRUE)
DELabig <- fromJSON("Katarina_DELabig_plot_report.json", flatten=TRUE)
DUP <- fromJSON("Katarina_DUP_plot_report.json", flatten=TRUE)
INV <- fromJSON("Katarina_INV_plot_report.json", flatten=TRUE)
DELb <- fromJSON("Katarina_DELb_plot_report.json", flatten=TRUE)
DELcsmall <- fromJSON("Katarina_DELcsmall_plot_report.json", flatten=TRUE)
DELcbig <- fromJSON("Katarina_DELcbig_plot_report.json", flatten=TRUE)
EXTRA <- fromJSON("Katarina_EXTRA_plot_report.json", flatten=TRUE)

outcomes <- rbind(DELasmall,DELabig, DUP, INV, DELb,DELcsmall, DELcbig, EXTRA )

result <- outcomes %>% mutate(SNP_ID= str_replace_all(Image, "_", ":"))
result <- result %>% mutate(keep = if_else(result$score =="Y", "keep", "discard"))
keep <- result %>% filter(keep == "keep")

write.table(keep$SNP_ID, file = "curated_SVs_keep.txt", sep = "\t", row.names = FALSE, quote = FALSE, col.names=FALSE)
```

filter the vcf file

```
cd /nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_curation

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

VCF=/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_survivor/hihi_filtered.recode.vcf

#also need to remove ones without het individuals (only 22 on list, with 8 being removed frm what would have been the curated list)
REMOVE=/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_survivor/SVs_to_remove.txt

vcftools --vcf ${VCF} --snps curated_SVs_keep.txt --exclude $REMOVE --remove-indv 83318 --recode-INFO-all  --recode --out
hihi_filtered_curated
```

After filtering, kept 30 out of 31 Individuals

Outputting VCF file...

After filtering, kept 1229 out of a possible 2991 Sites

Run Time = 1.00 seconds

I checked: --mac 1 didn't remove anything so excluding indv 83318 didn't remove any unique SVs

Check Mendelian Inheritance

```
cd /nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_curation/mendel2

VCF1=/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_survivor/hihi_filtered.recode.vcf
VCF2=/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_curation/hihi_filtered_curated.recode.vcf

module load PLINK/2.00a2.3
plink2 --vcf $VCF1 --allow-extra-chr --make-bed --out  hihi_filtered.plink
```

```
plink2 --vcf $VCF2 --allow-extra-chr --make-bed --out  hihi_filtered_curated.plink

#real trios
#input FAM data to .fam file
cp ../mendel/hihi_filtered.plink.fam .
grep -v "83318" hihi_filtered.plink.fam > hihi_filtered_curated.plink.fam
cp hihi_filtered.plink.fam hihi_filtered.plink.fam_REAL
cp  hihi_filtered_curated.plink.fam  hihi_filtered_curated.plink.fam_REAL
cp hihi_filtered.plink.fam_REAL hihi_filtered.plink.fam
cp hihi_filtered_curated.plink.fam_REAL hihi_filtered_curated.plink.fam

module purge
module load PLINK/1.09b6.16
plink --bfile hihi_filtered.plink --allow-extra-chr --recode --tab --out hihi_filtered
#allow extra chr flag not working for mendel calculation - excludes chroms 24 and 26. quick fix for now....
mv hihi_filtered.map hihi_filtered.map_OLD
awk 'BEGIN { OFS=FS="\t" } { if ($1 == 24 || $1 == 26) $1 = 23; print }' hihi_filtered.map_OLD > hihi_filtered.map
plink --file hihi_filtered --allow-extra-chr --mendel --out hihi_filtered
plink --file hihi_filtered --allow-extra-chr --missing --out hihi_filtered

plink --bfile hihi_filtered_curated.plink --allow-extra-chr --recode --tab --out hihi_filtered_curated
#allow extra chr flag not working for mendel calculation - excludes chroms 24 and 26. quick fix for now....
mv hihi_filtered_curated.map hihi_filtered_curated.map_OLD
awk 'BEGIN { OFS=FS="\t" } { if ($1 == 24 || $1 == 26) $1 = 23; print }' hihi_filtered_curated.map_OLD > hihi_filtered_curated.map
plink --file hihi_filtered_curated --allow-extra-chr --mendel --out hihi_filtered_curated
plink --file hihi_filtered_curated --allow-extra-chr --missing --out hihi_filtered_curated

#and for the two fake trios
cp ../mendel/hihi_genofiltered_fake.plink.fam hihi_filtered.plink.fam_FAKE
grep -v "83318" hihi_filtered.plink.fam_FAKE > hihi_filtered_curated.plink.fam_FAKE
cp hihi_filtered.plink.fam_FAKE hihi_filtered.plink.fam
cp hihi_filtered_curated.plink.fam_FAKE hihi_filtered_curated.plink.fam

plink --bfile hihi_filtered.plink --allow-extra-chr --recode --tab --out hihi_filtered
#allow extra chr flag not working for mendel calculation - excludes chroms 24 and 26. quick fix for now....
mv hihi_filtered.map hihi_filtered.map_OLD
awk 'BEGIN { OFS=FS="\t" } { if ($1 == 24 || $1 == 26) $1 = 23; print }' hihi_filtered.map_OLD > hihi_filtered.map
plink --file hihi_filtered --allow-extra-chr --mendel --out hihi_filtered_FAKE
plink --file hihi_filtered --allow-extra-chr --missing --out hihi_filtered_FAKE

plink --bfile hihi_filtered_curated.plink --allow-extra-chr --recode --tab --out hihi_filtered_curated
#allow extra chr flag not working for mendel calculation - excludes chroms 24 and 26. quick fix for now....
mv hihi_filtered_curated.map hihi_filtered_curated.map_OLD
awk 'BEGIN { OFS=FS="\t" } { if ($1 == 24 || $1 == 26) $1 = 23; print }' hihi_filtered_curated.map_OLD > hihi_filtered_curated.map
plink --file hihi_filtered_curated --allow-extra-chr --mendel --out hihi_filtered_curated_FAKE
plink --file hihi_filtered_curated --allow-extra-chr --missing --out hihi_filtered_curated_FAKE
```

check mendelian error rates??

```
module load R/4.1.0-gimkl-2020a
R
library(ggplot2)
library(dplyr)
library(stringr)
library(data.table)

setwd("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_curation/mendel2")
```

```
real1_miss <- fread("hihi_filtered.imiss") %>% mutate(SITES=N_GENO-N_MISS) %>% filter(
real2_miss <- fread("hihi_filtered_curated.imiss") %>% mutate(SITES=N_GENO-N_MISS)
fake1_miss <- fread("hihi_filtered_FAKE.imiss") %>% mutate(SITES=N_GENO-N_MISS)
fake2_miss <- fread("hihi_filtered_curated_FAKE.imiss") %>% mutate(SITES=N_GENO-N_MISS)

real1_mendl <- fread("hihi_filtered.imendel") %>% filter(IID == "86674" | IID == "86796") %>% merge(real1_miss, by = "IID")%>%
mutate(error_rate = N/SITES)
real2_mendl <- fread("hihi_filtered_curated.imendel")%>% filter(IID == "86674" | IID == "86796") %>% merge(real2_miss, by = "IID")%>%
mutate(error_rate = N/SITES)
fake1_mendl <- fread("hihi_filtered_FAKE.imendel")%>% filter(IID == "76799" | IID == "88230")%>% merge(fake1_miss, by = "IID")%>%
mutate(error_rate = N/SITES)
fake2_mendl <- fread("hihi_filtered_curated_FAKE.imendel")%>% filter(IID == "76799" | IID == "88230")%>% merge(fake2_miss, by =
"IID")%>% mutate(error_rate = N/SITES)

mean(real1_mendl$error_rate)*100 #1.573792
mean(real2_mendl$error_rate)*100 #0.8684935
mean(fake1_mendl$error_rate)*100 #14.15265
mean(fake2_mendl$error_rate)*100 #17.81982
```

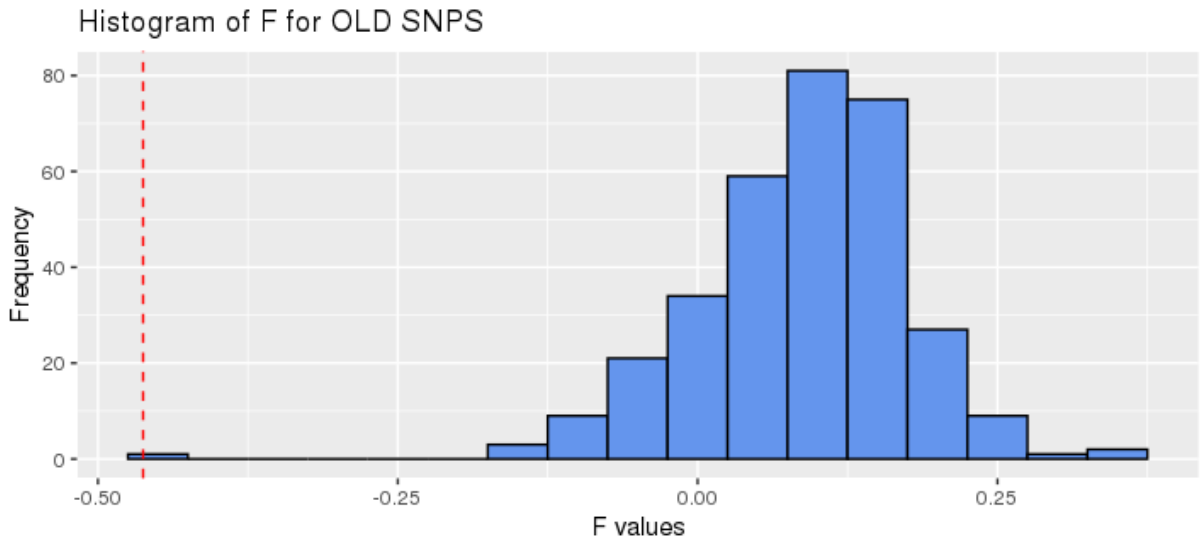
Examining Individual 83318

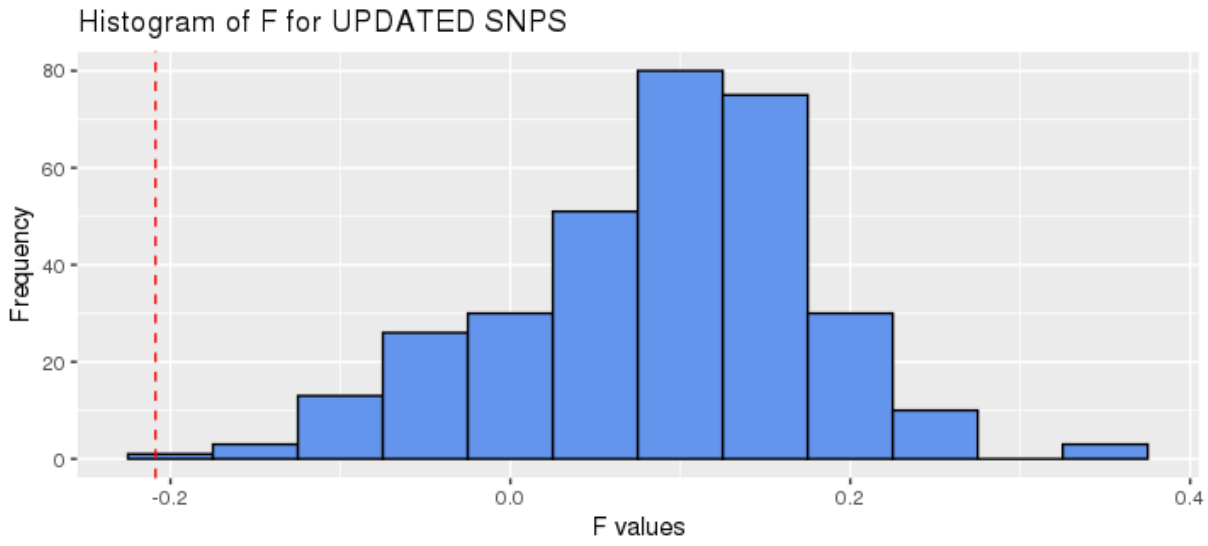
Examine mendelian error rates in the old and new data set (SNPs), and compare to know vs fake pedigree

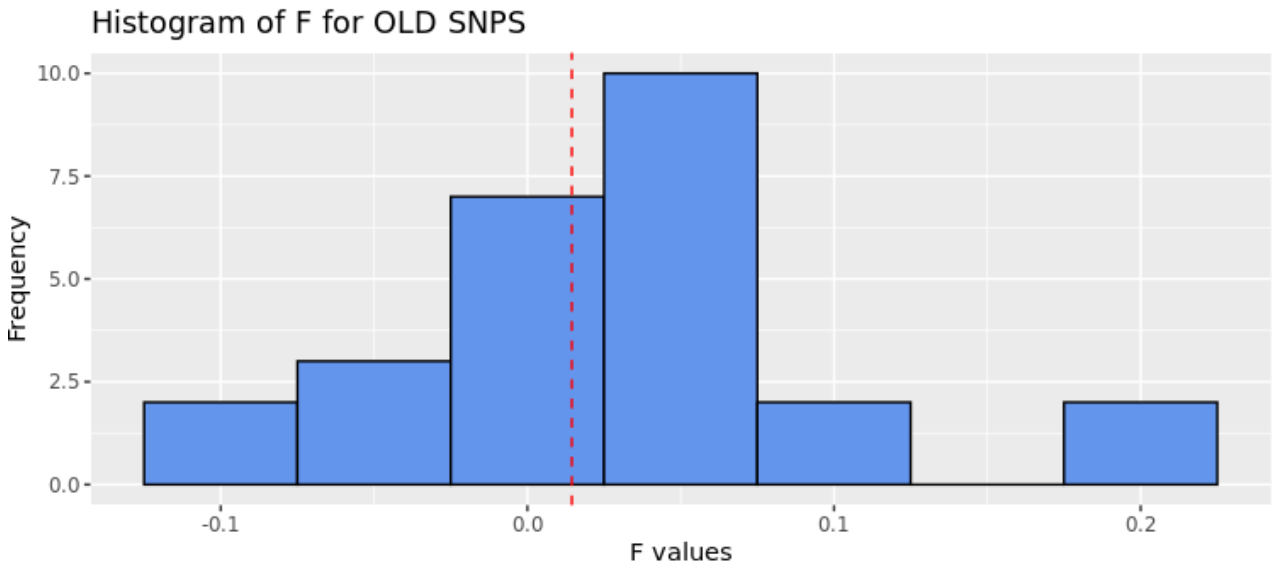
```
#!/bin/bash -e

#SBATCH --job-name=2024_03_19.83318_mendel.sl
#SBATCH --account=uoa00338
#SBATCH --time=00-48:00:00
#SBATCH --mem=60GB
#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/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_curation/83318

SNP=/nesi/nobackup/uoa00338/kstuart_projects/Nc2_HihiWGS/data/snp_variants_updated/hihi_wgs_filter_DP2.recode.vcf
SNP_OLD=/nesi/nobackup/uoa00338/kstuart_projects/Nc2_HihiWGS/data/snp_variants/hihi_wgs_filter_DP2.recode.vcf

#mendelian error of old data and new data
module load PLINK/2.00a2.3
plink2 --vcf $SNP --allow-extra-chr --make-bed --out  hihi_wgs_filter_DP2_UPDATED.plink
```

```
plink2 --vcf $SNP_OLD --allow-extra-chr --make-bed --out  hihi_wgs_filter_DP2_OLD.plink

#manually edi tthe .FAM file. for both

module purge
module load PLINK/1.09b6.16
plink --bfile hihi_wgs_filter_DP2_UPDATED.plink --allow-extra-chr --recode --tab --out hihi_wgs_filter_DP2_UPDATED
plink --file hihi_wgs_filter_DP2_UPDATED --allow-extra-chr --mendel --out hihi_wgs_filter_DP2_UPDATED

plink --bfile hihi_wgs_filter_DP2_OLD.plink --allow-extra-chr --recode --tab --out hihi_wgs_filter_DP2_OLD
plink --file hihi_wgs_filter_DP2_OLD --allow-extra-chr --mendel --out hihi_wgs_filter_DP2_OLD


#and heterozygosity
module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1
vcftools --vcf $SNP --het --out hihi_wgs_filter_DP2_UPDATED
vcftools --vcf $SNP_OLD --het --out hihi_wgs_filter_DP2_OLD


#recalled data from batchC
module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1

cd /nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_curation/83318

SNP=/nesi/nobackup/uoa00338/kstuart_projects/Nc2_HihiWGS/data/snp_variants/batch_testing/Hihi_26inds_batchC.vcf.gz

#filter out z and w chrom contigs
HIGHCOV=/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/sample_file_ids_highcov_no83318.txt
SEX=/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/annotation/Ncf_sex_linked_contigs.txt

LIST=
for CHROM in $(cut -f 1 $SEX)
do
LIST="${LIST} "--not-chr" $CHROM"
done

echo $LIST
vcftools --gzvcf ${SNP} $LIST --keep $HIGHCOV --mac 1 --minDP 5 --max-meanDP 100 --max-missing 0.9 --min-alleles 2 --max-alleles 2 --
recode --recode-INFO-all --out Hihi_26inds_batchC_filter_highcov
vcftools --gzvcf ${SNP} $LIST --mac 1 --minDP 2 --max-meanDP 100 --min-alleles 2 --max-alleles 2 --recode --recode-INFO-all --out
Hihi_26inds_batchC_filter

module load PLINK/2.00a2.3
plink2 --vcf Hihi_26inds_batchC_filter.recode.vcf --allow-extra-chr --make-bed --out  hihi_wgs_batchC.plink
#manually edi tthe .FAM file
cp hihi_wgs_batchC.plink.fam2 hihi_wgs_batchC.plink.fam
module purge
module load PLINK/1.09b6.16
plink --bfile hihi_wgs_batchC.plink --allow-extra-chr --recode --tab --out hihi_wgs_batchC
plink --file hihi_wgs_batchC --allow-extra-chr --mendel --out hihi_wgs_batchC
module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1
vcftools --vcf Hihi_26inds_batchC_filter.recode.vcf --het --out hihi_wgs_batchC


#in R
module load R/4.1.0-gimkl-2020a
R
library(ggplot2)
library(data.table)
```

```
library(tidyr)
library(dplyr)

setwd("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/variant_calling/SV_curation/83318")

data <- fread("hihi_wgs_filter_DP2_OLD.het")

data <- fread("hihi_wgs_batchC.het")

ind83318 <- data %>% filter(INDV == "83318") %>% select(F) %>% unlist()

ggplot(data, aes(x = F)) +
  geom_histogram(binwidth = 0.05, fill = "cornflowerblue", color = "black") +
  labs(title = "Histogram of F for OLD SNPS", x = "F values", y = "Frequency")+
geom_vline(xintercept = ind83318 , color = "red", linetype = "dashed")

data2 <- fread("hihi_wgs_filter_DP2_UPDATED.het")

ind83318_2 <- data2 %>% filter(INDV == "83318") %>% select(F) %>% unlist()

ggplot(data2, aes(x = F)) +
  geom_histogram(binwidth = 0.05, fill = "cornflowerblue", color = "black") +
  labs(title = "Histogram of F for UPDATED SNPS", x = "F values", y = "Frequency") +
geom_vline(xintercept = ind83318_2 , color = "red", linetype = "dashed")
```







FID	IID	N
FAMF2	72016	244018
FAMF2	72017	235660
FAMF2	72018	365564
FAMF1	83218	425174
FAMF1	83215	393279
FAMF1	83222	595276
FAMM	76928	344738
FAMM	76799	354498
FAMM	83318	476121
FAMR2	79133	47383
FAMR2	71159	74934
FAMR2	86674	82116
FAMR3	83378	49613
FAMR3	88274	78454
FAMR3	89893	85221

Conclusion: exclude 83318 from data