Starling-May18

Projects/Katarina Stuart/KStuart.Starling-Aug18/Nc3_HihiSV/Analysis/2024-02-26.Evolution

PDF Version generated by

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

on

Aug 15, 2024 @03:08 PM NZST

Table of Contents

Whole Genome Alignment Evolution

Following protocol in barn swallow pangenome paper:
https://doi.org/10.1016/j.celrep.2023.111992

Github for this paper:
https://github.com/SwallowGenomics/BarnSwallow/tree/main/Analyses

GERP stuff: https://bio-protocol.org/exchange/minidetail?type=30&id=8708060
https://onlinelibrary.wiley.com/doi/full/10.1111/mec.16802

Program list

maf_stream

I have had a local copy of maf_stream compiled with Rust which I have released as a module now. Module name is maf_stream/202005-GCC-12.3.0 and ready to be used.

```
module load maf_stream/202005-GCC-12.3.0
maf_stream --help
```

PHAST

I am afraid this is bit outdated now and not the best to be added as a module. However, I have had a pre-compiled binary of it on one my test directories which I have cloned to */nesi/project/uoa02613/software/PHAST* and added *export PATH=/nesi/project/uoa02613/software/PHAST/usr/bin:$PATH* to your ~/.bash_profile.
Therefore, you can call any of the PHAST commands without having to do anything now (If anyone else from the project to want it, please do ask them to the same by adding the above export.. to their ~/.bash_profile

```
phast --help
```

Cactus with GPU support - Container only

Current deployment for Cactus is container and container only as they provide a pretty solid container image with GPU compatibility which we don't need to modify too much.
I have added the latest container image to our centrally shared stack . Path is /opt/nesi/containers/Cactus/cactus-2.7.2-gpu.simg

Below is a minimal working example for the latest version of Cactus.
This is the one we use for all of the Cactus container images to verify (This was done in a Jupyter GPU interactive session.

You can run this on CPU as well but have to pass on the --gpu 0 argument to cactus command. Otherwise, It will trigger an error with the instruction.

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/cactus
git clone https://github.com/ComparativeGenomicsToolkit/cactus.git
module purge
module load Apptainer

#GPU interactive
```

```
apptainer exec --nv /opt/nesi/containers/Cactus/cactus-2.7.2-gpu.simg cactus ./js
/nesi/nobackup/uoa02613/kstuart_projects/programs/cactus/cactus/examples/evolverMammals.txt evolverMammals.hal
#GPU run stats: Cactus has finished after 240.90046348422766 seconds

#CPU interactive
apptainer exec --nv /opt/nesi/containers/Cactus/cactus-2.7.2-gpu.simg cactus ./js
/nesi/nobackup/uoa02613/kstuart_projects/programs/cactus/cactus/examples/evolverMammals.txt evolverMammals.hal --gpu 0
#CPU run stats: 200 seconds
```

Then ran halStats for the output (Interactively)

```
apptainer exec /opt/nesi/containers/Cactus/cactus-2.6.13-gpu.simg halStats evolverMammals.hal
```

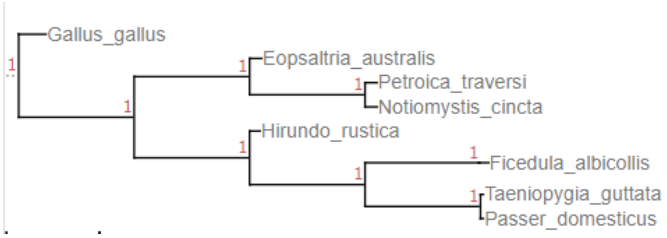
Choosing genomes:

Infraorder Passerides
--Petroicidae
----Eopsaltria australis: https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_034509425.1/
----Petroica traversi: https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_025920805.1/
~~----Drymodes brunneopygia: https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_013400955.1/~~
--Eupetidae
~~----Picathartes gymnocephalus: https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_013390045.1/~~
--Chaetopidae
----Chaetops frenatus: https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_013400775.1/

Muscicapidae (Ficedula albicollis): https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000247815.1/
Estrildidae (T. guttata): https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_003957565.2/
Chicken (Gallus gallus): https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_016699485.2/
Passer domesticus (house sparrow): https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_036417665.1/
barn swallow (Hirundo rustica): https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_015227805.2/

```
cd /nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/evolution/genomes

wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/034/509/425/GCA_034509425.1_MU_EAus_VIC030_1.0/GCA_034509425.1_MU_EAus_VIC030_1.0_genomic.fna.gz
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/025/920/805/GCA_025920805.1_Ptraversi_NRM_v1/GCA_025920805.1_Ptraversi_NRM_v1_genomic.fna.gz
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/013/400/775/GCA_013400775.1_ASM1340077v1/GCA_013400775.1_ASM1340077v1_genomic.fna.gz
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/247/815/GCF_000247815.1_FicAlb1.5/GCF_000247815.1_FicAlb1.5_genomic.fna.gz
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/957/565/GCF_003957565.2_bTaeGut1.4.pri/GCF_003957565.2_bTaeGut1.4.pri_genomic.fna.gz
wget
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/016/699/485/GCF_016699485.2_bGalGal1.mat.broiler.GRCg7b/GCF_016699485.2_bGalGal1.mat.broiler.GRCg7b_genomic.fna.gz
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/036/417/665/GCF_036417665.1_bPasDom1.hap1/GCF_036417665.1_bPasDom1.hap1_genomic.fna.gz
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/015/227/805/GCF_015227805.2_bHirRus1.pri.v3/GCF_015227805.2_bHirRus1.pri.v3_genomic.fna.gz

#unzip
gunzip *

#link hihi genome into the same direcotry using same naming scheme
ln -s /nesi/nobackup/uoa00338/kstuart_projects/Nc2_HihiWGS/data/resources/Ncf_H98617_scaffolded_genome.fa Ncf_H98617_scaffolded_genome.fna

#name the file of assembly names
nano avian_assemblies.txt
```

Make a newick with timetree
https://timetree.org/



some little checks to see if things worked:

```
cd /nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/evolution/

#TEST
module purge
module load Apptainer
apptainer exec /opt/nesi/containers/Cactus/cactus-2.6.13-gpu.simg halStats ../AvianSeqfile_all_subset6.hal

#TEST2
apptainer exec /opt/nesi/containers/Cactus/cactus-2.6.13-gpu.simg halAlignmentDepth --noAncestors --targetGenomes Gallus_gallus,Passer_domesticus,Hirundo_rustica,Eopsaltria_australis --outWiggle coverage_25A.wig --refSequence 25A ../AvianSeqfile_all_subset6.hal Notiomystis_cincta
```

Extracting information

```
cd /nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/evolution/AvianSeqfile_all_subset6/chroms

#grab just chromosomes with SNPs/SVs on them
SNP=/nesi/nobackup/uoa00338/kstuart_projects/Nc2_HihiWGS/data/snp_variants_updated/hihi_wgs_filter_highcov_no83318_autosomes.recode.vcf
grep -v "^#" $SNP | cut -f1 | uniq > hihi_chroms.txt

module purge
module load Apptainer

for i in $(cat "hihi_chroms.txt")
do
  echo "$i"
#calculate per base coverage of hihi genome
  apptainer exec /opt/nesi/containers/Cactus/cactus-2.6.13-gpu.simg halAlignmentDepth --noAncestors --targetGenomes Gallus_gallus,Passer_domesticus,Hirundo_rustica,Eopsaltria_australis,Taeniopygia_guttata --outWiggle coverage_$i.wig --refSequence $i ../../AvianSeqfile_all_subset6.hal Notiomystis_cincta

#make extra columns that have the chromosome ID and and position (line number)
tail -n +2 coverage_$i.wig  |  awk -v OFS='\t' -v myvar="$i" '{ print myvar, NR, $1}' > coverage_${i}_format.wig

#collapse runs of same number
awk 'NR==1 {prev=$3; print; next}
     {getline nextLine; split(nextLine, nextArray); nextValue=nextArray[3];}
     {if ($3 != prev || $3 != nextValue) print}
     {prev=$3; $0=nextLine}' coverage_${i}_format.wig > coverage_${i}_format2.wig

#turn into bed file with interval info
awk  -v OFS='\t'  '{ if (prev != "") {
           print $1, prev, $2 - 1, $3
       }
       prev = $2
     }'  coverage_${i}_format2.wig > coverage_${i}_format3.wig

#means I can now use bedtools for overlapping info, and also the file sizes are condensed so I can get rid of the intermediate files

done
```

Find the overlap

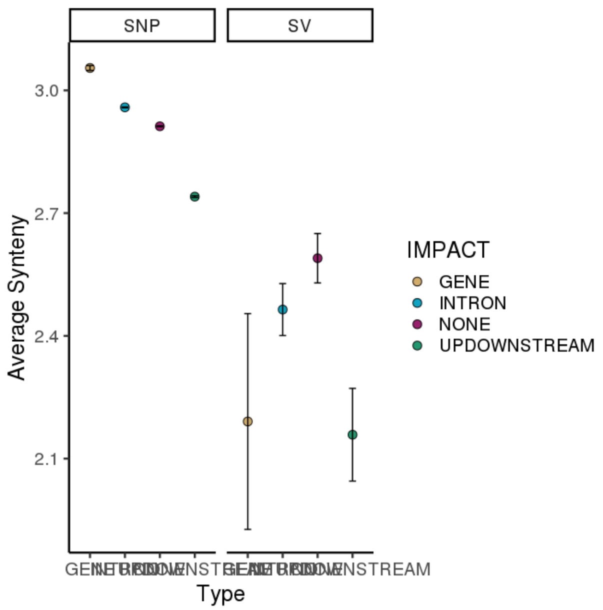
```
cd /nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/evolution/AvianSeqfile_all_subset6

module load BEDTools/2.30.0-GCC-11.3.0

cat chroms/format/coverage_*_format3.wig > coverage_ALL_format3.wig
cat chroms/format/coverage_*_format.wig > coverage_ALL_format.wig

SV=/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/recombination/SV.bed

bedtools intersect -wb -a coverage_ALL_format3.wig -b $SV > sv_synteny.txt

SNP=/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/hihi_wgs_filter_highcov_no83318_autosomes_vepID.recode.vcf

bedtools intersect -wb -a coverage_ALL_format3.wig -b $SNP | cut -f1-10 > snp_synteny.txt

#creating the 'neutral' background profile of the whole genome
#need to make a bed file of the whole genome, with the 4 sections noted - exon, intron, upstream/downstream/ rest of genome. For this use the GFF file
GFF=/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/data/annotation/Ncf_H98617_scaffolded_liftoff.gff
awk '$3=="exon"' $GFF | awk -v OFS='\t'  '{print $1, $4, $5, "4-GENE"}' | sort | uniq > genome_exons.bed #these are intervals for category 4-GENE
awk '$3=="mRNA"' $GFF | awk -v OFS='\t'  '{print $1, $4, $5, "2-INTRON"}' | sort | uniq > genome_mRNA.bed
bedtools subtract -a genome_mRNA.bed -b genome_exons.bed > genome_mRNA_without_exons.bed #these are intervals for category 2-INTRON
awk '{print $1, $2-5000, $3+5000, "3-UPDOWNSTREAM"}' genome_mRNA.bed | awk -v OFS='\t' '{print $1, ($2 < 0 ? 0 : $2), $3, $4}' > genome_updownstreamgene.bed
bedtools subtract -a  genome_updownstreamgene.bed -b genome_mRNA.bed  > genome_updownstream.bed #these are intervals for category 3-UPDOWNSTREAM
cat genome_exons.bed genome_mRNA_without_exons.bed genome_updownstream.bed > genome_allintervals.bed

bedtools intersect -wb -a coverage_ALL_format3.wig -b genome_exons.bed > genome_synteny_4gene.txt
bedtools intersect -wb -a coverage_ALL_format3.wig -b genome_mRNA_without_exons.bed > genome_synteny_2intron.txt
bedtools intersect -wb -a coverage_ALL_format3.wig -b genome_updownstream.bed > genome_synteny_3updownstream.txt
bedtools intersect -wb -a coverage_ALL_format3.wig -b genome_allintervals.bed -v > genome_synteny_1none.txt

datamash mean 4  sstdev 4 count 4 < genome_synteny_4gene.txt > genome_synteny_4gene_sum.txt
datamash mean 4  sstdev 4 count 4 < genome_synteny_2intron.txt > genome_synteny_2intron_sum.txt
datamash mean 4  sstdev 4 count 4 < genome_synteny_3updownstream.txt > genome_synteny_3updownstream_sum.txt
datamash mean 4  sstdev 4 count 4 < genome_synteny_1none.txt > genome_synteny_1none_sum.txt

cat genome_synteny_4gene_sum.txt genome_synteny_2intron_sum.txt genome_synteny_3updownstream_sum.txt genome_synteny_1none_sum.txt > all.txt
nano all.txt

#WORKING IT OUT FOR BP NOT REGIONS
#bedtools intersect -wb -a genome_exons.bed -b coverage_ALL_format.vcf > genome_synteny_4gene.txt
#bedtools intersect -wb -a coverage_ALL_format.wig -b genome_mRNA_without_exons.bed > genome_synteny_2intron.txt
#bedtools intersect -wb -a coverage_ALL_format.wig -b genome_updownstream.bed > genome_synteny_3updownstream.txt
#bedtools intersect -wb -a coverage_ALL_format.wig -b genome_allintervals.bed -v > genome_synteny_1none.txt
#
awk -v OFS='\t' '{print $0,($3 - $2) * $4,($3 - $2)}' genome_synteny_4gene.txt > genome_synteny_4gene_fix.txt
awk -v OFS='\t' '{print $0,($3 - $2) * $4,($3 - $2)}' genome_synteny_2intron.txt > genome_synteny_2intron_fix.txt
awk -v OFS='\t' '{print $0,($3 - $2) * $4,($3 - $2)}' genome_synteny_3updownstream.txt > genome_synteny_3updownstream_fix.txt
awk -v OFS='\t' '{print $0,($3 - $2) * $4,($3 - $2)}' genome_synteny_1none.txt > genome_synteny_1none_fix.txt

datamash sum 9 sum 10 < genome_synteny_4gene_fix.txt > genome_synteny_4gene_fix_sum.txt
datamash sum 9 sum 10 < genome_synteny_2intron_fix.txt > genome_synteny_2intron_fix_sum.txt
datamash sum 9 sum 10 < genome_synteny_3updownstream_fix.txt > genome_synteny_3updownstream_fix_sum.txt
datamash sum 5 sum 6 < genome_synteny_1none_fix.txt > genome_synteny_1none_fix_sum.txt

cat *fix_sum* > all_corr.txt
nano all_corr.txt
```

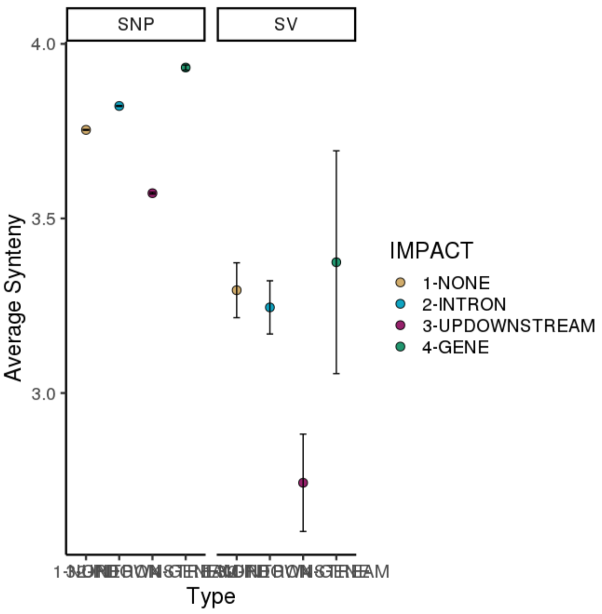
```
#
#
```

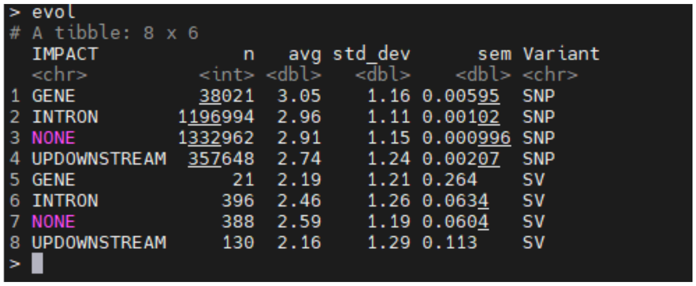
METHOD ONE: Plot - using entire SV length average

```
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/analysis/evolution/AvianSeqfile_all_subset6")

#SVs
data <- fread("sv_synteny.txt")
data2 <- data %>% mutate(length= (V3-V2), overlap_count= (V3-V2)*V4) %>% group_by(V8) %>% summarise(overlap_total = sum(overlap_count)/sum(length) )

#assign impact
impact <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SV_impact_variant.txt") %>% mutate(IMPACT = "4-GENE")
vars <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SV_impact_lowvariant.txt") %>% mutate(IMPACT2 = "3-UPDOWNSTREAM")
vars2 <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SV_impact_midvariant.txt") %>% mutate(IMPACT3 = "2-INTRON")

#two method
merged_df1 <- merge(impact , vars, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df <- merge(merged_df1, vars2, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df$IMPACT[merged_df$IMPACT2 == '3-UPDOWNSTREAM'] <- '3-UPDOWNSTREAM'
merged_df$IMPACT[merged_df$IMPACT3 == '2-INTRON'] <- '2-INTRON'
#fun <- merge(data2, merged_df[,1:2], by.x = "V8", by.y = "V1", all.x =  TRUE)
fun <- merge(data2, merged_df[,1:2], by.x = "V8", by.y = "V1", all.x =  TRUE)
fun$IMPACT[is.na(fun$IMPACT)] <- '1-NONE'

fun_sv <- fun %>% group_by(IMPACT) %>% summarise(n = n(), avg = mean(overlap_total), std_dev = sd(overlap_total),  sem = sd(overlap_total) / sqrt(length(overlap_total)  ) )

#SNPs
data <- fread("snp_synteny.txt")[,c(7,4)]

#assign impact
impact <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SNP_impact_variant.txt", header=F) %>% mutate(IMPACT = "4-GENE")
vars <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SNP_impact_lowvariant.txt", header=F) %>% mutate(IMPACT2 = "3-UPDOWNSTREAM")
vars2 <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SNP_impact_midvariant.txt", header=F) %>% mutate(IMPACT3 = "2-INTRON")

#two method
merged_df1 <- merge(impact , vars, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df <- merge(merged_df1, vars2, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df$IMPACT[merged_df$IMPACT2 == '3-UPDOWNSTREAM'] <- '3-UPDOWNSTREAM'
merged_df$IMPACT[merged_df$IMPACT3 == '2-INTRON'] <- '2-INTRON'
fun <- merge(data, merged_df[,1:2], by.x = "V7", by.y = "V1", all.x =  TRUE)
fun$IMPACT[is.na(fun$IMPACT)] <- '1-NONE'

fun_SNP <- fun %>% group_by(IMPACT) %>% summarise(n = n(), avg = mean(V4), std_dev = sd(V4),  sem = sd(V4) / sqrt(length(V4)  ) )

fun_sv<- fun_sv%>% mutate(Variant = "SV")
fun_SNP <- fun_SNP %>% mutate(Variant = "SNP")
evol <- rbind(fun_SNP , fun_sv)


pdf("Nc3_SV_SNP_MAF_loads.pdf", width=7, height=5)
ggplot(evol , aes(x =IMPACT, y = avg, fill = IMPACT)) +
  geom_point(position = position_dodge(width = 0.75), size = 3, shape = 21) +
  geom_errorbar(aes(ymin = avg - sem, ymax = avg + sem),
                position = position_dodge(width = 0.75), width = 0.2) +  # Plot standard errors
  labs(x = "Type", y = "Average Synteny") +
  theme_classic(base_size = 18) +
  scale_fill_manual(values = c("#d1ac6b", "#10a4c2","#96216b", "#21966f")) +
  scale_color_manual(values = c("#d1ac6b", "#10a4c2","#96216b", "#21966f")) +
  facet_grid(. ~ Variant)
dev.off()
```







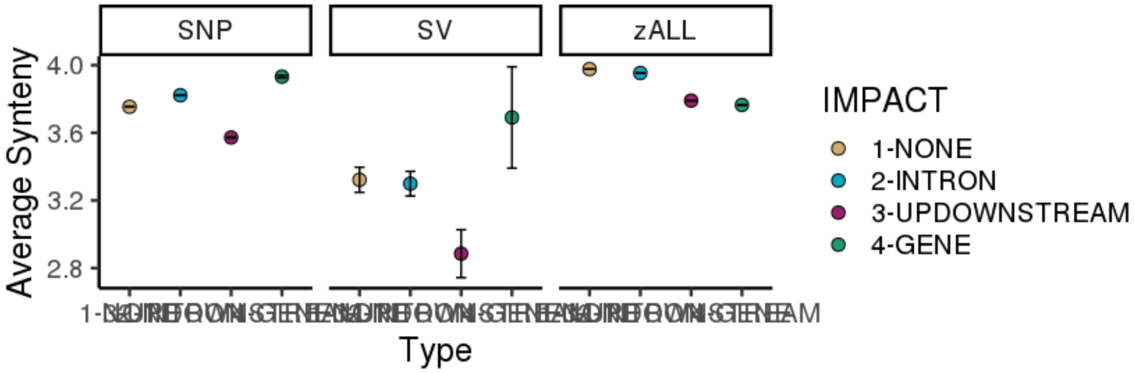
```
> evol
# A tibble: 8 x 6
  IMPACT            n   avg std_dev      sem Variant
  <chr>         <int> <dbl>   <dbl>    <dbl> <chr>  
1 GENE          38021  3.05    1.16 0.00595  SNP    
2 INTRON      1196994  2.96    1.11 0.00102  SNP    
3 NONE        1332962  2.91    1.15 0.000996 SNP    
4 UPDOWNSTREAM 357648  2.74    1.24 0.00207  SNP    
5 GENE             21  2.19    1.21 0.264    SV     
6 INTRON          396  2.46    1.26 0.0634   SV     
7 NONE            388  2.59    1.19 0.0604   SV     
8 UPDOWNSTREAM    130  2.16    1.29 0.113    SV     
> 
```

METHOD TWO: Plot - using break ends of SV (averaged)

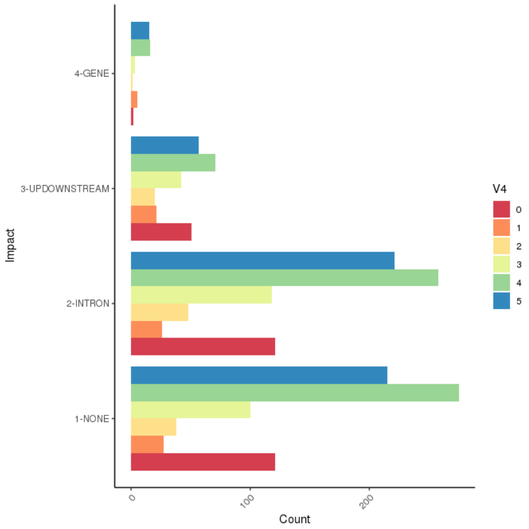
```
module load R/4.1.0-gimkl-2020a
R
library(ggplot2)
library(data.table)
library(tidyr)
library(dplyr)
library(stats)

setwd("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/evolution/AvianSeqfile_all_subset6")

#SVs
data <- fread("sv_synteny.txt")
data2 <- data[, .SD[c(1, .N)], by = V8]

#assign impact
impact <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SV_impact_variant.txt") %>% mutate(IMPACT = "4-GENE")
vars <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SV_impact_lowvariant.txt") %>% mutate(IMPACT2 = "3-UPDOWNSTREAM")
vars2 <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SV_impact_midvariant.txt") %>% mutate(IMPACT3 = "2-INTRON")

#two method
merged_df1 <- merge(impact , vars, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df <- merge(merged_df1, vars2, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df$IMPACT[merged_df$IMPACT2 == '3-UPDOWNSTREAM'] <- '3-UPDOWNSTREAM'
merged_df$IMPACT[merged_df$IMPACT3 == '2-INTRON'] <- '2-INTRON'
#fun <- merge(data2, merged_df[,1:2], by.x = "V8", by.y = "V1", all.x =  TRUE)
fun <- merge(data2, merged_df[,1:2], by.x = "V8", by.y = "V1", all.x =  TRUE)
fun$IMPACT[is.na(fun$IMPACT)] <- '1-NONE'

fun_sv <- fun %>% group_by(IMPACT, V4) %>% summarise(count = n(), .groups = 'drop')

fun_sv_avg <- fun %>% group_by(V8) %>% summarize(avg = sum(V4) / n(), IMPACT = first(IMPACT))
fun_sv2 <- fun_sv_avg %>% group_by(IMPACT) %>% summarise( sem = sd(avg) / sqrt(length(avg)) , n = n(), avg = mean(avg))
```

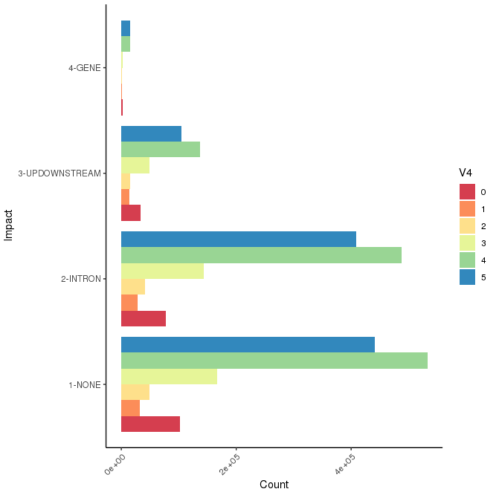
```
#SNPs
data <- fread("snp_synteny.txt")[,c(7,4)]

#assign impact
impact <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SNP_impact_variant.txt", header=F) %>% mutate(IMPACT = "4-GENE")
vars <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SNP_impact_lowvariant.txt", header=F) %>% mutate(IMPACT2 = "3-UPDOWNSTREAM")
vars2 <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SNP_impact_midvariant.txt", header=F) %>% mutate(IMPACT3 = "2-INTRON")

#two method
merged_df1 <- merge(impact , vars, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df <- merge(merged_df1, vars2, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df$IMPACT[merged_df$IMPACT2 == '3-UPDOWNSTREAM'] <- '3-UPDOWNSTREAM'
merged_df$IMPACT[merged_df$IMPACT3 == '2-INTRON'] <- '2-INTRON'
fun <- merge(data, merged_df[,1:2], by.x = "V7", by.y = "V1", all.x =  TRUE)
fun$IMPACT[is.na(fun$IMPACT)] <- '1-NONE'

fun_SNP <- fun %>% group_by(IMPACT, V4) %>% summarise(count = n(), .groups = 'drop')
fun_SNP2 <- fun %>% group_by(IMPACT) %>% summarise(sem = sd(V4) / sqrt(length(V4)), n = n(), avg = mean(V4) )

#Average plot
all <- fread("all.txt")
colnames(all) <- c("avg","stdev","n","IMPACT")
all2 <- all %>% mutate(avg= stdev/n, Variant = "zALL") %>% select(IMPACT, sem, n, avg, Variant)

all <- fread("all_corr.txt")
colnames(all) <- c("V1","n","IMPACT")
all2 <- all %>% mutate(avg=  V1/n, Variant = "zALL", sem=0.0000000001) %>% select(IMPACT, sem, n, avg, Variant)

fun_sv2<- fun_sv2 %>% mutate(Variant = "SV")
fun_SNP2 <- fun_SNP2 %>% mutate(Variant = "SNP")

evol <- rbind(fun_SNP2 , fun_sv2, all2)

pdf("Nc3_SV_SNP_synteny_load.pdf", width=9, height=7)
ggplot(evol , aes(x =IMPACT, y = avg, fill = IMPACT)) +
  geom_point(position = position_dodge(width = 0.75), size = 5, shape = 21) +
  geom_errorbar(aes(ymin = avg - sem, ymax = avg + sem),
                position = position_dodge(width = 0.75), width = 0.2) +  # Plot standard errors
  labs(x = "Type", y = "Average Synteny") +
  theme_classic(base_size = 18) +
  scale_fill_manual(values = c("#d1ac6b", "#10a4c2","#96216b", "#21966f")) +
  scale_color_manual(values = c("#d1ac6b", "#10a4c2","#96216b", "#21966f")) +
  facet_grid(. ~ Variant) + theme(axis.text.x = element_text(angle = 45, hjust = 1))
dev.off()



#Bar plot - supp mat
fun_sv<- fun_sv%>% mutate(Variant = "SV")
fun_SNP <- fun_SNP %>% mutate(Variant = "SNP")
evol <- rbind(fun_SNP , fun_sv)


pdf("Nc3_synteny_counts_SNP.pdf", width=5, height=7)
ggplot(fun_SNP, aes(x = IMPACT, y = count, fill = factor(V4))) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_brewer(palette = "Spectral") +  # Use a color palette for V4
  labs(x = "Impact", y = "Count", fill = "V4") +
  theme_classic() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))+coord_flip()
dev.off()

pdf("Nc3_synteny_counts_SV.pdf", width=5, height=7)
ggplot(fun_sv, aes(x = IMPACT, y = count, fill = factor(V4))) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_brewer(palette = "Spectral") +  # Use a color palette for V4
  labs(x = "Impact", y = "Count", fill = "V4") +
  theme_classic() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))+coord_flip()
```

```
dev.off()
```



with 6:



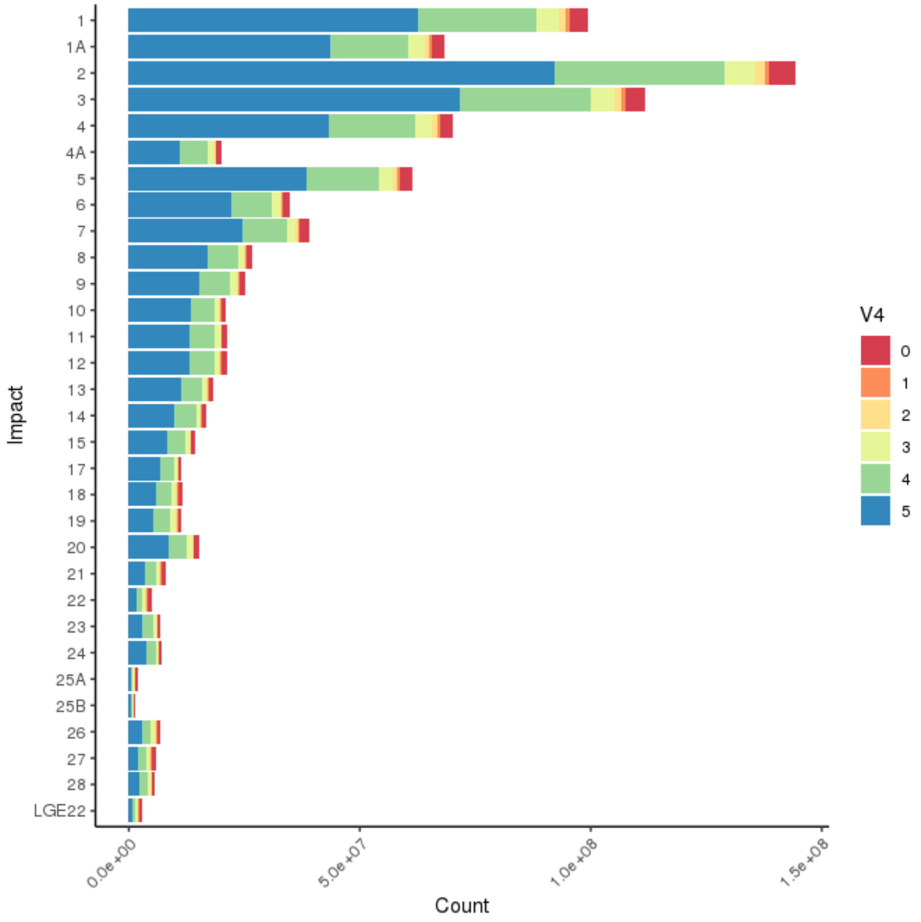


calculate the number of genomes covering each chromosomes base
https://github.com/SwallowGenomics/BarnSwallow/blob/main/Analyses/Cactus_alignment/Count_aligned_genomes.sh

```
cd /nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/evolution/AvianSeqfile_all_subset6/chroms

mkdir format
mv *format* format

for file in coverage*.wig ; do
    root=`basename $file .wig`
    awk -v root="$root" 'BEGIN {print root}' >> summarise_output.txt
    grep -v "fixedStep" $file | awk '{print $1}' | sort | uniq -c >> summarise_output.txt
done


head -n 217 summarise_output.txt > summarise_output2.txt

grep -v coverage summarise_output2.txt > vals.txt
grep coverage summarise_output2.txt > chroms1.txt
```

```
while read line; do for i in {1..6}; do echo "$line"; done; done < chroms1.txt > chroms.txt

paste chroms.txt vals.txt | sed -e 's/[[:space:]]\+/\t/g' -e 's/coverage_//g' > synteny.txt
```

```


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/analysis/evolution/AvianSeqfile_all_subset6/chroms")

data <- fread("synteny.txt", header=FALSE)
order<- ordered(rev(c("1","1A","2","3", "4","4A","5","6","7","8","9","10","11","12","13","14","15","17","18","19","20","21","22","23","24","25A","25B","26","27","28","LGE22")))
data$V1 <- factor(data$V1, levels = order)

pdf("Nc3_synteny_genome.pdf", width=5, height=8)
ggplot(data, aes(x = V1, y = V2, fill = factor(V3))) +
  geom_bar(stat = "identity", position = "stack") +
  scale_fill_brewer(palette = "Spectral") +  # Use a color palette for V4
  labs(x = "Impact", y = "Count", fill = "V4") +
  theme_classic() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))+coord_flip()
dev.off()
```



Evolutionary load

```
fun_SV <- fun %>% group_by(V8) %>% summarise(ES_score = mean(V4) ) %>% mutate(class = case_when(ES_score >= 4~ "syntenic", ES_score <= 1 ~ "unique", TRUE ~ "none"))
write.table(fun_SV,"syntenicalignment_SV.txt.txt",row.names=FALSE,sep="\t", quote = FALSE,col.names=TRUE)
```

```

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/analysis/evolution/AvianSeqfile_all_subset6")

#SV
data <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/counts/merged_rep_missfiltered_reffix2_load.txt")

impact <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SV_impact_variant.txt") %>% mutate(IMPACT = "GENE")
vars <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SV_impact_lowvariant.txt") %>% mutate(IMPACT2 = "LOW")
vars2 <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SV_impact_midvariant.txt") %>% mutate(IMPACT3 = "MID")

#assigning impact to each SV
merged_df1 <- merge(impact , vars, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df <- merge(merged_df1, vars2, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df$IMPACT[merged_df$IMPACT2 == 'LOW'] <- 'LOW'
merged_df$IMPACT[merged_df$IMPACT3 == 'MID'] <- 'MID'
fun <- merge(data, merged_df[,1:2], by.x = "ID", by.y = "V1", all.x =  TRUE)
fun$IMPACT[is.na(fun$IMPACT)] <- 'NONE'

fun2 <- merge(fun, fun_SV, by.x = "ID", by.y = "V8", all.x =  TRUE)


#count of genotypes (het and homo) in gene impacting regions
SV_evolmasked_load <- fun2 %>% filter(class=="syntenic") %>% filter(IMPACT=="LOW" | IMPACT=="MID" |  IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SV_evolmasked_load = sum(value == 1, na.rm = TRUE))

SV_midmasked_load <- fun2 %>% filter(class=="none") %>% filter(IMPACT=="LOW" | IMPACT=="MID" |  IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SV_midmasked_load = sum(value == 1, na.rm = TRUE))

SV_uniqmasked_load <- fun2 %>% filter(class=="unique") %>% filter(IMPACT=="LOW" | IMPACT=="MID" |  IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SV_uniqmasked_load = sum(value == 1, na.rm = TRUE))


SV_putative_load<- fun2  %>% filter(IMPACT=="LOW" | IMPACT=="MID" | IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SV_putative_load= sum(value %in% c(0, 1, 2), na.rm = TRUE))


final <- cbind(SV_evolmasked_load , SV_midmasked_load[,2], SV_uniqmasked_load[,2], SV_putative_load[,2])

final_SV <- final %>% mutate(SV_evolprop_mask = SV_evolmasked_load/SV_putative_load, SV_midprop_mask = (SV_midmasked_load)/SV_putative_load, SV_uniqprop_mask = (SV_uniqmasked_load)/SV_putative_load, SV_bothprop_mask = (SV_uniqmasked_load+SV_midmasked_load)/SV_putative_load)

#combine


write.table(final_SV ,"ALL_load_counts_evolution.txt",row.names=FALSE,sep="\t", quote = FALSE,col.names=TRUE)
```

Evolutionary load X2

```
fun_SV <- fun %>% group_by(V8) %>% summarise(ES_score = mean(V4) ) %>% mutate(class = case_when(ES_score > 4 ~ "syntenic", ES_score <= 1 ~ "unique", TRUE ~ "none"))
write.table(fun_SV,"syntenicalignment_SV.txt.txt",row.names=FALSE,sep="\t", quote = FALSE,col.names=TRUE)
```

```
fun_SNP <- fun %>% group_by(V7) %>% summarise(ES_score = mean(V4) ) %>% mutate(class = case_when(ES_score > 4 ~ "syntenic", ES_score <= 1 ~ "unique", TRUE ~ "none"))
write.table(fun_SNP,"syntenicalignment_SNP.txt.txt",row.names=FALSE,sep="\t", quote = FALSE,col.names=TRUE)
```

```
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/analysis/evolution/AvianSeqfile_all_subset6")

#SV
data <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/counts/merged_rep_missfiltered_reffix2_load.txt")

impact <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SV_impact_variant.txt") %>% mutate(IMPACT = "GENE")
vars <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SV_impact_lowvariant.txt") %>% mutate(IMPACT2 = "LOW")
vars2 <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SV_impact_midvariant.txt") %>% mutate(IMPACT3 = "MID")

#assigning impact to each SV
merged_df1 <- merge(impact , vars, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df <- merge(merged_df1, vars2, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df$IMPACT[merged_df$IMPACT2 == 'LOW'] <- 'LOW'
merged_df$IMPACT[merged_df$IMPACT3 == 'MID'] <- 'MID'
fun <- merge(data, merged_df[,1:2], by.x = "ID", by.y = "V1", all.x =  TRUE)
fun$IMPACT[is.na(fun$IMPACT)] <- 'NONE'

fun2 <- merge(fun, fun_SV, by.x = "ID", by.y = "V8", all.x =  TRUE)


#count of genotypes (het and homo) in gene impacting regions
SV_evolmasked_load <- fun2 %>% filter(class=="syntenic") %>% filter(IMPACT=="LOW" | IMPACT=="MID" |  IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SV_evolmasked_load = sum(value == 1, na.rm = TRUE))

SV_uniqmasked_load<- fun2 %>% filter(class=="unique") %>% filter(IMPACT=="LOW" | IMPACT=="MID" | IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SV_uniqmasked_load= sum(value == 1, na.rm = TRUE))

SV_maskedputative_load<- fun2 %>% filter(IMPACT=="LOW" | IMPACT=="MID" | IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SV_maskedputative_load= sum(value == 1, na.rm = TRUE))


SV_evolrealised_load <- fun2 %>% filter(class=="syntenic") %>% filter(IMPACT=="LOW" | IMPACT=="MID" |  IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SV_evolrealised_load = sum(value == 2, na.rm = TRUE))

SV_uniqrealised_load<- fun2 %>% filter(class=="unique") %>% filter(IMPACT=="LOW" | IMPACT=="MID" | IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SV_uniqrealised_load= sum(value == 2, na.rm = TRUE))

SV_realputative_load<- fun2 %>%  filter(IMPACT=="LOW" | IMPACT=="MID" | IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SV_realputative_load= sum(value == 2, na.rm = TRUE))


final <- cbind(SV_evolmasked_load , SV_uniqmasked_load[,2], SV_maskedputative_load[,2], SV_evolrealised_load [,2], SV_uniqrealised_load[,2], SV_realputative_load[,2])

final_SV <- final %>% mutate(SV_evolprop_mask = SV_evolmasked_load /SV_maskedputative_load, SV_uniqprop_mask = (SV_uniqmasked_load)/SV_maskedputative_load, SV_evolprop_real = SV_evolrealised_load /SV_realputative_load, SV_uniqprop_real = (SV_uniqrealised_load)/SV_realputative_load)



#SNP
data <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/counts/hihi_wgs_filter_highcov_no83318_autosomes_load.txt")
impact <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SNP_impact_variant.txt", header=F) %>% mutate(IMPACT = "GENE")
vars <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SNP_impact_lowvariant.txt", header=F) %>% mutate(IMPACT2 = "LOW")
vars2 <- fread("/nesi/nobackup/uoa00338/kstuart_projects/Nc3_HihiSV/analysis/impacts/vep/vep_SNP_impact_midvariant.txt", header=F) %>% mutate(IMPACT3 = "MID")

#assigning impact to each SNP
merged_df1 <- merge(impact , vars, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df <- merge(merged_df1, vars2, by.x = "V1", by.y = "V1", all.x =  TRUE)
merged_df$IMPACT[merged_df$IMPACT2 == 'LOW'] <- 'LOW'
merged_df$IMPACT[merged_df$IMPACT3 == 'MID'] <- 'MID'
```

```
fun <- merge(data, merged_df[,1:2], by.x = "ID", by.y = "V1", all.x =  TRUE)
fun$IMPACT[is.na(fun$IMPACT)] <- 'NONE'

fun2 <- merge(fun, fun_SNP, by.x = "ID", by.y = "V7", all.x =  TRUE)


#count of genotypes (het and homo) in gene impacting regions
SNP_evolmasked_load <- fun2 %>% filter(class=="syntenic") %>% filter(IMPACT=="LOW" | IMPACT=="MID" |  IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SNP_evolmasked_load = sum(value == 1, na.rm = TRUE))

SNP_uniqmasked_load<- fun2 %>% filter(class=="unique") %>% filter(IMPACT=="LOW" | IMPACT=="MID" | IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SNP_uniqmasked_load= sum(value == 1, na.rm = TRUE))

SNP_maskedputative_load<- fun2 %>% filter(IMPACT=="LOW" | IMPACT=="MID" | IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SNP_maskedputative_load= sum(value == 1, na.rm = TRUE))


SNP_evolrealised_load <- fun2 %>% filter(class=="syntenic") %>% filter(IMPACT=="LOW" | IMPACT=="MID" |  IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SNP_evolrealised_load = sum(value == 2, na.rm = TRUE))

SNP_uniqrealised_load<- fun2 %>% filter(class=="unique") %>% filter(IMPACT=="LOW" | IMPACT=="MID" | IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SNP_uniqrealised_load= sum(value == 2, na.rm = TRUE))

SNP_realputative_load<- fun2 %>%  filter(IMPACT=="LOW" | IMPACT=="MID" | IMPACT=="GENE"  ) %>% gather(key = "variable", value = "value", 10:39) %>% mutate(value = as.numeric(value))  %>%  group_by(variable) %>% summarize(SNP_realputative_load= sum(value == 2, na.rm = TRUE))


final <- cbind(SNP_evolmasked_load , SNP_uniqmasked_load[,2], SNP_maskedputative_load[,2], SNP_evolrealised_load [,2], SNP_uniqrealised_load[,2], SNP_realputative_load[,2])

final_SNP <- final %>% mutate(SNP_evolprop_mask = SNP_evolmasked_load /SNP_maskedputative_load, SNP_uniqprop_mask = (SNP_uniqmasked_load)/SNP_maskedputative_load, SNP_evolprop_real = SNP_evolrealised_load /SNP_realputative_load, SNP_uniqprop_real = (SNP_uniqrealised_load)/SNP_realputative_load)


#combine

allload0 <- merge(final_SV,final_SNF, by.x="variable", by.y="variable")

write.table(allload0 ,"ALL_load_counts_evolution2.txt",row.names=FALSE,sep="\t", quote = FALSE,col.names=TRUE)
```