Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/Sv5_AustraliaWGS/Analysis/2021-09-02.SVxSNPdensity

PDF Version generated by

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

on

Feb 20, 2023 @02:44 PM NZDT

Table of Contents

2021-09-02.SVxSNPdensity 2

2021-09-02.SVxSNPdensity

Katarina Stuart (z5188231@ad.unsw.edu.au) - Nov 18, 2022, 11:11 AM GMT+13

Comparison of SV and SNP density

SV density histograms for pops

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/analysis/var_density_2022
GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/genome/Sturnus_vulgaris_2.3.1.simp.fasta.fai
```

```
module load samtools/1.10
```

```
cut -f1,2 ${GENOME} > sizes.genome
```

```
module load bedtools/2.27.1
module load vcftools/0.1.16
```

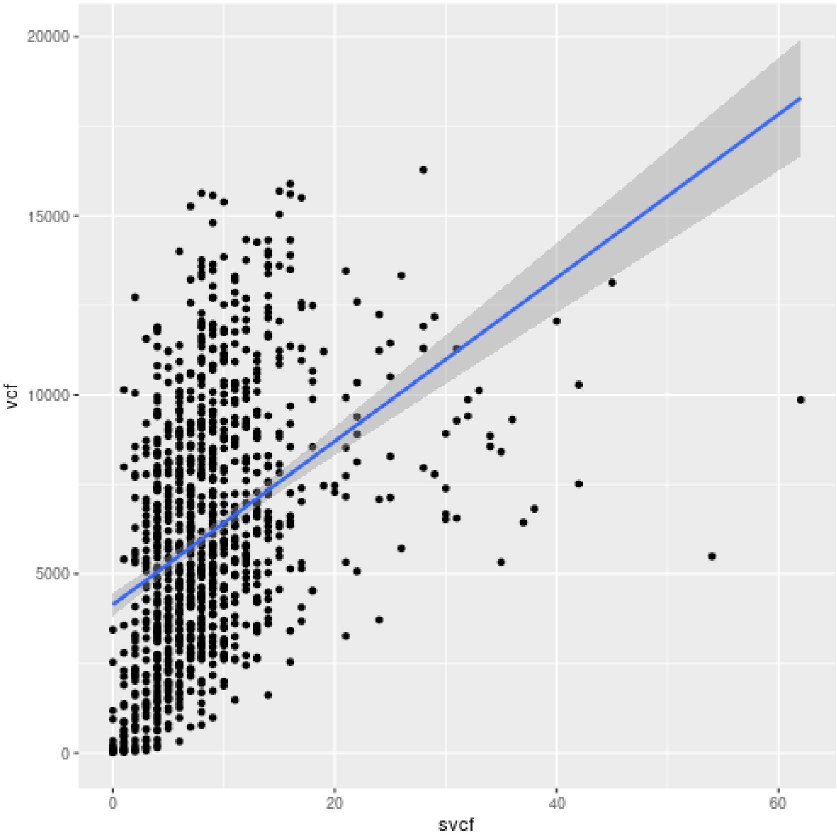
```
WIDTH=1000000
```

```
bedtools makewindows -g sizes.genome -w ${WIDTH} > svulgaris_${WIDTH}bps.bed

VCF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/data/snp_variants_processed/wgs_variantsgenotyped_filtered_maf005_r2.vcf
#SVCF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/data/survivor_2022/split_vcfs/merged_rep_filtered.recode.vcf
SVCF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/data/survivor_2022/split_vcfs/merged_rep.vcf

bedtools coverage -a svulgaris_${WIDTH}bps.bed -b ${SVCF}  -counts > variantcoverage_svcf_${WIDTH}bps.txt
sed -n '/starling34/q;p' variantcoverage_svcf_${WIDTH}bps.txt > variantcoverage_svcf_${WIDTH}bps_cut.txt
bedtools coverage -a svulgaris_${WIDTH}bps.bed -b ${VCF}  -counts > variantcoverage_vcf_${WIDTH}bps.txt
sed -n '/starling34/q;p' variantcoverage_vcf_${WIDTH}bps.txt > variantcoverage_vcf_${WIDTH}bps_cut.txt
```

regression

```
module load R/3.5.3
R

library("ggplot2")
library("dplyr")
library(gridExtra)
library(grid)
library(lattice)

setwd("/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/analysis/var_density_2022")

#ALL plots

svcf.density <- read.table(file="variantcoverage_svcf_1000000bps_cut.txt", header=FALSE, sep="\t")
vcf.density <- read.table(file="variantcoverage_vcf_1000000bps_cut.txt", header=FALSE, sep="\t")

all.density <- cbind(vcf.density, svcf.density$V4)
```

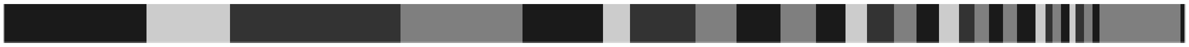
```
colnames(all.density) <- c("chrom", "start", "end", "vcf", "svcf")

png("Sv5_SVvSNP_regression.png", width=500, height=500)
ggplot(all.density, aes(svcf , vcf )) + geom_point() + stat_smooth(method = lm)
dev.off()

fit <- glm(svcf ~ vcf, data=all.density, family=poisson())
# If you have overdispersion (see if residual deviance is much larger than degrees of freedom), you may want to use quasipoisson() instead of poisson().
fit <- glm(svcf ~ vcf, data=all.density, family=quasipoisson())
summary(fit)

fit <- glm(svcf ~ vcf, data=all.density, family=poisson())
summary(fit)

png("Sv5_SVvSNP_regression_glm2.png", width=500, height=500)
plot(fit)
dev.off()
```

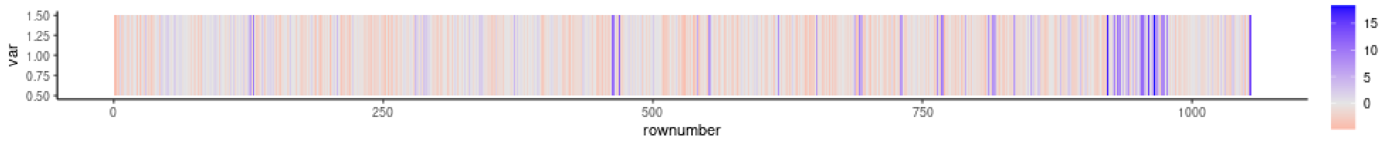


Plotting

```
#all.density <- all.density %>% mutate(resids = resid(lm(svcf ~ vcf, data=all.density)))
all.density <- all.density %>% mutate(resids.glm = resid( glm(svcf ~ vcf, data=all.density, family=quasipoisson())))
all.density$rownumber = 1:nrow(all.density)
all.density$var = 1

write.table(all.density, file='all.density.txt', quote=FALSE, sep='\t', col.names = NA)
```

```
png("Sv5_SVvSNP_heatplot.png", width=1000, height=100)
ggplot(all.density, aes(rownumber, var, fill= resids.glm)) +geom_tile() + theme_void() + theme(legend.position = "none") +
scale_fill_gradient2(low = "red", mid = "grey90", high = "blue", midpoint = 0)
dev.off()

png("Sv5_SVvSNP_heatplot2.png", width=1000, height=100)
ggplot(all.density, aes(rownumber, var, fill= resids.glm)) +geom_tile() + theme_classic() + scale_fill_gradient2(low = "red", mid = "grey90", high
= "blue", midpoint = 0)
dev.off()

all.density <- all.density %>% mutate(chromnumber = gsub("starling", "", paste(all.density$chrom)))

png("Sv5_SVvSNP_heatplot_chrom2.png", width=1000, height=100)
ggplot(all.density, aes(rownumber, var, fill= chromnumber)) +geom_tile() + xlim(0,1055) + theme_void()
dev.off()

#Rel chrom sizes:
Chrom_labels <- read.table(file="Chrom_labels.txt", header=TRUE, sep="\t")

png("Sv5_SVvSNP_heatplot_chrom.png", width=1000, height=30)
ggplot(data =Chrom_labels , aes(x = CATEGORY, y = Count_of_ORDER, fill = Row_Labels3)) +
geom_bar(stat='identity') +  coord_flip() + scale_y_reverse() +
  theme_void() +
  theme(axis.line = element_line(size = 3, colour = "white")) +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank(),
        axis.title.y = element_text(colour = "white"), axis.text.y = element_text(colour = "white"), axis.ticks.y = element_line(colour = "white")) +
  theme(legend.position="none") +
scale_fill_manual(values=c("grey50", "grey10", "grey80","grey20", "grey50", "grey10", "grey80","grey20", "grey50", "grey10",
"grey50", "grey10", "grey80","grey20", "grey50", "grey10", "grey80","grey20", "grey50", "grey10",
"grey50", "grey10", "grey80","grey20", "grey50", "grey10", "grey80","grey20", "grey50", "grey10",
"grey50", "grey10", "grey80","grey20", "grey50", "grey10", "grey80","grey20", "grey50", "grey10",
"grey50", "grey10", "grey80","grey20", "grey50", "grey10", "grey80","grey20", "grey50", "grey10"))
dev.off()
```

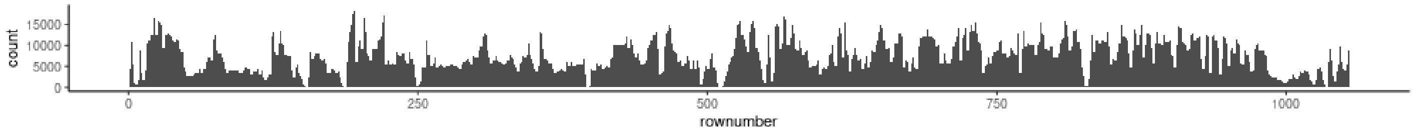


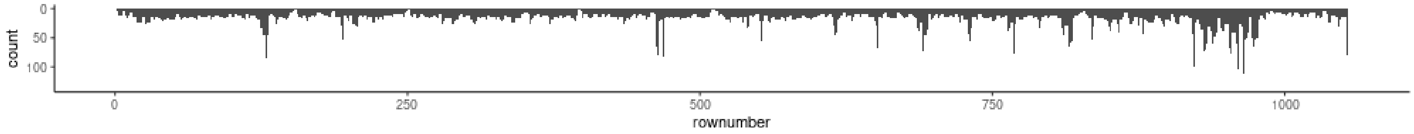


Plot densities along the genome:

```
module load R/3.5.3
R

library("ggplot2")
library("dplyr")
library(gridExtra)
library(grid)
library(lattice)

setwd("/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/analysis/var_density_2022")
```

```
#ALL plots

svcf.density <- read.table(file="variantcoverage_svcf_1000000bps_cut.txt", header=FALSE, sep="\t")
colnames(svcf.density) <- c("chrom", "start", "end", "count")
svcf.density$rownumber = 1:nrow(svcf.density)

vcf.density <- read.table(file="variantcoverage_vcf_1000000bps_cut.txt", header=FALSE, sep="\t")
colnames(vcf.density) <- c("chrom", "start", "end", "count")
vcf.density$rownumber = 1:nrow(vcf.density)

png("Sv5_SVvSNP_density_vcf.png", width=1000, height=100)
ggplot(data=vcf.density, aes(x=rownumber, y=count))+  geom_area(fill = "grey30") + theme_classic()
dev.off()

png("Sv5_SVvSNP_densityvoid_vcf.png", width=1000, height=100)
ggplot(data=vcf.density, aes(x=rownumber, y=count))+  geom_area(fill = "grey30") + theme_void()
dev.off()

png("Sv5_SVvSNP_density_svcf.png", width=1000, height=100)
ggplot(data=svcf.density, aes(x=rownumber, y=count))+  geom_area(fill = "grey30") + theme_classic() +
scale_y_reverse()
dev.off()

png("Sv5_SVvSNP_densityvoid_svcf.png", width=1000, height=100)
ggplot(data=svcf.density, aes(x=rownumber, y=count))+  geom_area(fill = "grey30") + theme_void() +
scale_y_reverse()
dev.off()
```





SNP density histograms for pops

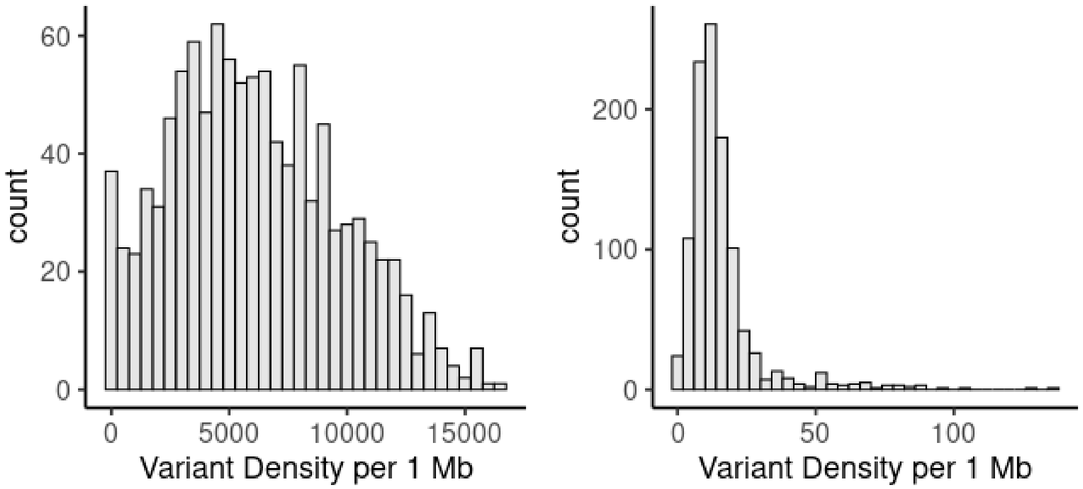
Summarizing in histograms the SNP densities calculated above for the regression

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/analysis/var_density_2022/bar_plots
```

plot

```
module load R/3.5.3
R

library("ggplot2")
library("dplyr")
library(gridExtra)
library(grid)
library(lattice)

setwd("/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/analysis/var_density_2022/bar_plots")

svcf.density <- read.table(file="/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/analysis/var_density_2022/variantcoverage_svcf_1000000bps_cut.txt", header=FALSE, sep="\t")
vcf.density <- read.table(file="/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/analysis/var_density_2022/variantcoverage_vcf_1000000bps_cut.txt", header=FALSE, sep="\t")
```

```
A1 <- ggplot(svcf.density, aes(x = V4)) + geom_histogram(binwidth = 4, fill='#e6e6e6', color="black" ) + ylab("count") +
xlab("Variant Density per 1 Mb") + theme_classic(base_size = 18) + theme(axis.text=element_text(size=16))
A2 <- ggplot(vcf.density, aes(x = V4)) + geom_histogram(binwidth = 500, fill='#e6e6e6', color="black" ) + ylab("count") +
xlab("Variant Density per 1 Mb") + theme_classic(base_size = 18) + theme(axis.text=element_text(size=16))

png("sv5_SVvSNPdensity_bar.png", width=650, height=300)
grid.arrange(A2, A1, ncol=2)
dev.off()
```



Gene counts

grab total number of genes overlapping each SV, then in R group into sizebins and graph

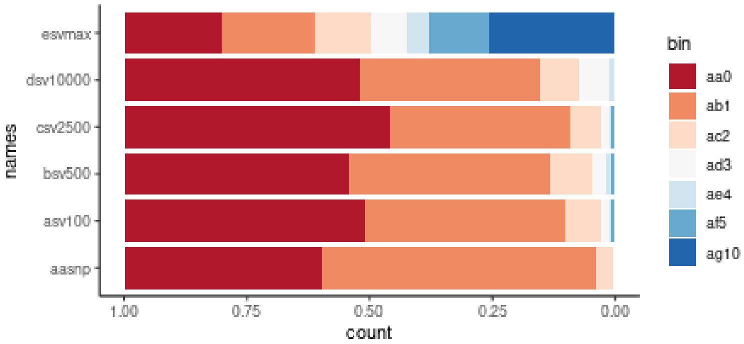
```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/analysis/var_density_2022/gene_counts

VCF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/data/snp_variants_processed/wgs_variantsgenotyped_filtered_miss50_r2.vcf
SVCF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/data/survivor_2022/split_vcfs/merged_rep.vcf


#Prepare gene data gff files
GFF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/analysis/outlier_analysis/mapping_variants/myFile_lociMerged_longestIsoform.gff
awk '$3 == "gene"' $GFF > stalingannotation_longestIsoform_genes.txt
grep -v "Note=Protein of unknown" stalingannotation_longestIsoform_genes.txt >
stalingannotation_longestIsoform_genes_short.txt
```

```
#calculate gene density of SNPs by looking over each SNP
grep -v "^#" $VCF > SNPs.vcf
cp final_SNP_genecount_list_clean.txt final_SNP_genecount_list.txt

#for SNPS
for NUM in {55..6502217} #6502217
do
sed "${NUM}q;d" SNPs.vcf > x.tempfile.vcf
cat vcf_file_header.txt x.tempfile.vcf > x.tempfile2.vcf
#head x.tempfile.vcf
bedtools intersect -wb -a stalingannotation_longestIsoform_genes_short.txt -b x.tempfile2.vcf >
x.tempfile_SNPgenecount.txt
wc -l x.tempfile_SNPgenecount.txt >> final_SNP_genecount_list.txt
rm x.tempfile.vcf x.tempfile2.vcf x.tempfile_SNPgenecount.txt
done
```

```
#loop overeach line of SVCF BED file and count number of genes each SV hits. created bed file from list of SVs and their length

cp final_SNP_genecount_list_clean.txt final_SV_genecount_list.txt

grep -v "^#" ${SVCF} | sed 's/;/\t/g' | cut -f1,2,3,10,11 | sed 's/SVLEN=-\|SVLEN=//g' | sed 's/SVTYPE=//g'  | awk -v FS='\t' -v OFS='\t' '{print $1,
$2, $2+$4}' > merged_rep.bed


SVCFbed=merged_rep.bed

#for SVs

for NUM in {1..17007}
do
sed "${NUM}q;d" ${SVCFbed} > x.tempfile.svcf
bedtools intersect -wb -a stalingannotation_longestIsoform_genes_short.txt -b x.tempfile.svcf >
x.tempfile_SVgenecount.txt
wc -l x.tempfile_SVgenecount.txt >>final_SV_genecount_list.txt
rm x.tempfile.svcf x.tempfile_SVgenecount.txt
done

sed 's/ /\t/g' final_SV_genecount_list.txt > final_SV_genecount_list2.txt

SVCF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/data/survivor_2022/split_vcfs/merged_rep.vcf

grep -v "^#" ${SVCF} | sed 's/;/\t/g' | cut -f1,2,3,10,11 | sed 's/SVLEN=-\|SVLEN=//g' | sed 's/SVTYPE=//g'  | awk -v FS='\t' -v OFS='\t' '{print $1,
$2, $4}' > merged_rep_filtered_lengths.txt

paste merged_rep_filtered_lengths.txt <(cut -f1 final_SV_genecount_list2.txt) > SV_alldata_updated.txt



#grep "SVLEN=*" ${SVCF} | sed 's/\;/\' \t/g' | cut -f10 | sed 's/SVLEN=//' > merged_norep_mq_filtered_svlength
#paste merged_norep_mq_filtered_lengths.txt final_SV_genecount_list.txt | awk '{ print $1","$4","$5 }' > SV_alldata_updated.txt

#paste merged_norep_mq_filtered_lengths.txt final_SV_genecount_list.txt | sed "s/^M$//g" | tail -n +2 > SV_alldata_updated.txt
```

Then Graph in R:

```
module load R/3.5.3
R

library("ggplot2")
library("dplyr")
library(gridExtra)
library(grid)
library(lattice)
library(tidyr)

setwd("/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/analysis/var_density_2022/gene_counts")

sv.genes <- read.table(file="SV_alldata_updated.txt", header=FALSE, sep="\t")
colnames(sv.genes) <- c("CHROM","POS","LENGTH","LIST")
snp.genes <- read.table(file="final_SNP_genecount_list.txt", header=TRUE, sep=" ")

#sv.genes
sv.genes.100 <- filter(sv.genes, LENGTH > 0, LENGTH <= 100)
sv.genes.500 <- filter(sv.genes, LENGTH > 100, LENGTH <= 500)
sv.genes.2500 <- filter(sv.genes, LENGTH > 500, LENGTH <= 2500)
sv.genes.10000 <- filter(sv.genes, LENGTH > 2500, LENGTH <= 10000)
sv.genes.max <- filter(sv.genes, LENGTH > 10000)

# set up cut-off values
breaks <- c(0,1,2,3,4,5,10,300)

# bucketing values into bins
group_tags <- cut(sv.genes.100$LIST,
          breaks=breaks,
          include.lowest=TRUE,
          right=FALSE)
# inspect bins
sv.genes.100.sum <- summary(group_tags)

group_tags <- cut(sv.genes.500$LIST,
          breaks=breaks,
          include.lowest=TRUE,
          right=FALSE)
sv.genes.500.sum <- summary(group_tags)

group_tags <- cut(sv.genes.2500$LIST ,
          breaks=breaks,
          include.lowest=TRUE,
          right=FALSE)
sv.genes.2500.sum <- summary(group_tags)

group_tags <- cut(sv.genes.10000$LIST ,
          breaks=breaks,
          include.lowest=TRUE,
          right=FALSE)
sv.genes.10000.sum <- summary(group_tags)

group_tags <- cut(sv.genes.max$LIST ,
          breaks=breaks,
          include.lowest=TRUE,
          right=FALSE)
sv.genes.max.sum <- summary(group_tags)

group_tags <- cut(snp.genes$X0 ,
                breaks=breaks,
                include.lowest=TRUE,
                right=FALSE)
snp.genes <- summary(group_tags)

names<- c("asv100", "bsv500", "csv2500", "dsv10000", "esvmax")
names<- c("aasnp","asv100", "bsv500", "csv2500", "dsv10000", "esvmax")

all.sum<-rbind.data.frame(sv.genes.100.sum, sv.genes.500.sum, sv.genes.2500.sum, sv.genes.10000.sum, sv.genes.max.sum)
all.sum<-rbind.data.frame(snp.genes,sv.genes.100.sum, sv.genes.500.sum, sv.genes.2500.sum, sv.genes.10000.sum, sv.genes.max.sum)
colnames(all.sum ) <- c("aa0", "ab1", "ac2", "ad3", "ae4", "af5", "ag10")

all.sum2<-cbind.data.frame(names,all.sum)

all.sum.long <- gather(all.sum2, bin, count, aa0:ag10)

png("sv5_SVvSNPdensity_genes_updated2.png", width=450, height=210)
ggplot(all.sum.long, aes(fill=bin , y=count , x=names)) + geom_bar(position="fill", stat="identity") + theme_classic() + coord_flip()+
scale_y_reverse() + scale_fill_brewer(palette="RdBu")
dev.off()
```


esvmax
dsv10000
csv2500
bsv500
asv100
aasnp
names
count
bin
aa0
ab1
ac2
ad3
ae4
af5
ag10