Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/Sv5_AustraliaWGS/Analysis/2021-11-20.RepeatAnalysis

PDF Version generated by

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

on

Feb 20, 2023 @10:21 AM NZDT

Table of Contents

2021-11-20.RepeatAnalysis 2

2021-11-20.RepeatAnalysis

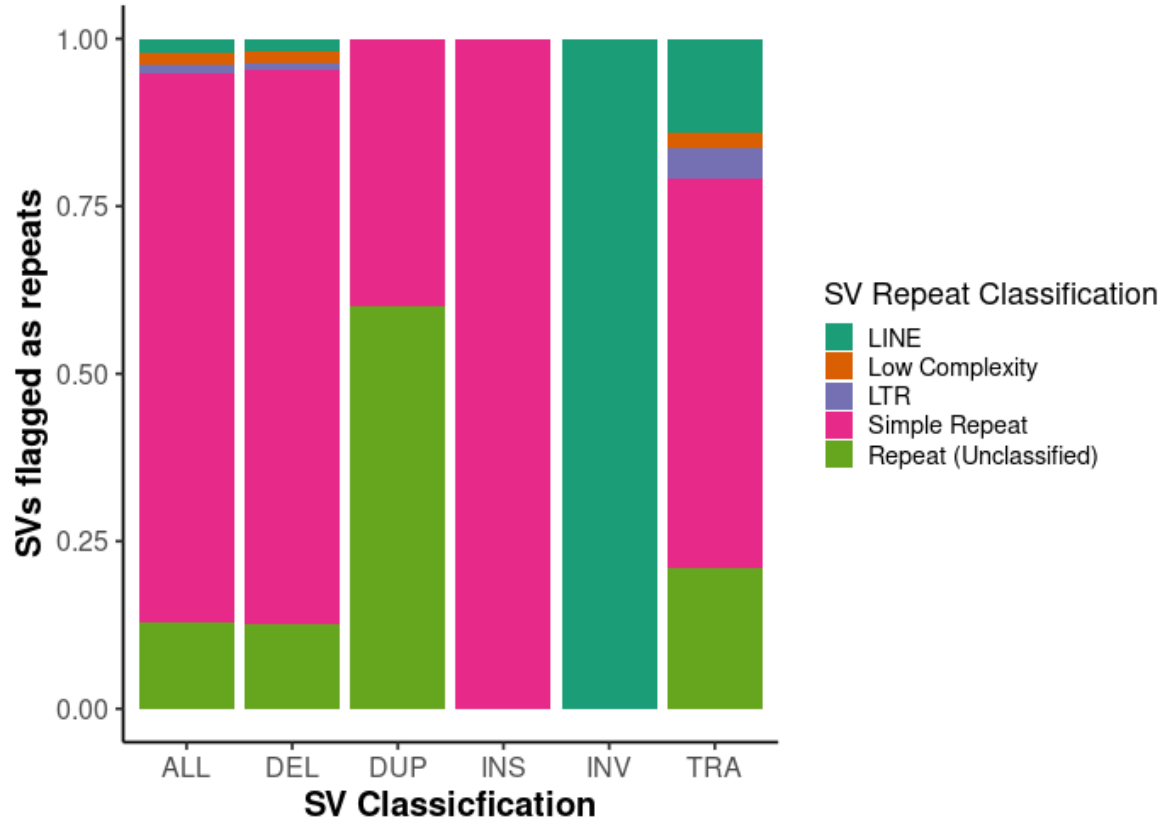
Katarina Stuart (z5188231@ad.unsw.edu.au) - Feb 20, 2023, 10:20 AM GMT+13

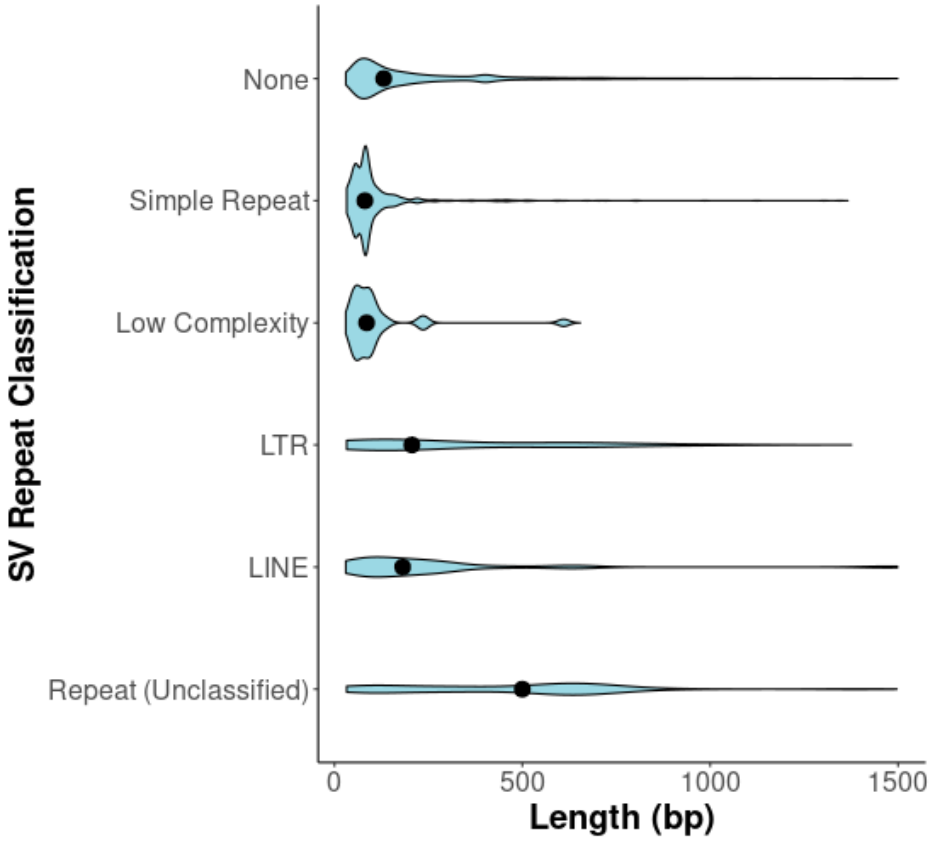
Repeat Analysis

Run repeat masker

```
module add perl/5.28.0
module add repeatmasker/4.0.7
module add genometools/1.5.9
module add muscle/3.8.31
module add blast+/2.6.0
module add repeatmodeler/1.0.11
module add hmmer/3.2.1
module add bedtools/2.27.1

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/analysis/sv_counts_2022/repeat_analysis
GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/genome/Sturnus_vulgaris_2.3.1.simp.fasta
SVCF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/data/survivor_2022/split_vcfs/

#Input is bed file (CHROM START END) with no columnheaders, with 30bp length at the start) (30 bp is min SV length)
grep -v "^#" ${SVCF}/merged_rep.vcf | awk '{ print $1"\t"$2"\t"$2+30}' > merged_rep_forrepeatanal.bed

#get fasta of these 30 bp sequences
bedtools getfasta -fi $GENOME -bed merged_rep_forrepeatanal.bed -fo sv_sequences.fasta

DIR_RM1=/srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/adv_repeats/programs/repeatmasker/4.0.7/
LIB=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/repeat_analysis/All_repeats_aves_custom.fasta

#Run repeat masker
${DIR_RM1}/RepeatMasker -pa 2 -lib ${LIB} -dir . sv_sequences.fasta

#preparing metadata file with lengths and types:
grep -v "^#" ${SVCF}/merged_rep.vcf | sed 's/;/\t/g' | cut -f1,2,3,10,11 | sed 's/SVLEN=-\|SVLEN=//g' | sed 's/SVTYPE=//g' | awk '{ print
$1"\t"$2"\t"$3"\t"$4"\t"$5"\t"$1":"$2"-"$2+30 }' > merged_rep_repanalysis.samples.txt
tail -n +4 sv_sequences.fasta.out | sed 's/[[:blank:]]\+/\t/g' > sv_sequences.fasta.out.format

cut -f6,12 sv_sequences.fasta.out.format >sv_sequences.fasta.out.format.cut
awk '{ print $6"\t"$4"\t"$5 }' merged_rep_repanalysis.samples.txt > merged_rep_repanalysis.samples.txt.cut

awk -f vlookup.awk sv_sequences.fasta.out.format.cut merged_rep_repanalysis.samples.txt.cut | column | sed 's/Unspecified/Unknown/g' | sed
's/LTR\/ERV1\|LTR\/ERVK\|LTR\/ERVL\|LTR\/ERVL?/LTR/g' > merged_rep_repanalysis.samples.reps.txt

#Plotting
#in R
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/sv_counts_2022/repeat_analysis")
repeatDB <- read.table("merged_rep_repanalysis.samples.reps.txt", sep=" ", header=FALSE)
colnames(repeatDB) <- c("Name","Length","Type","Repeat")
levels(repeatDB$Repeat) <- c("LINE", "Low Complexity", "LTR","None","Simple Repeat","Repeat (Unclassified)" )
repeatDB_cut <- filter(repeatDB , Repeat != "None")

repeatDB_cut.ALL <- count(repeatDB_cut, Repeat, sort = TRUE) %>%mutate(Data="ALL")
repeatDB_cut.DEL <- filter(repeatDB_cut, Type== "DEL") %>% count(Repeat, sort = TRUE) %>%mutate(Data="DEL")
repeatDB_cut.DUP<- filter(repeatDB_cut, Type == "DUP") %>% count(Repeat, sort = TRUE)%>%mutate(Data="DUP")
repeatDB_cut.INV  <- filter(repeatDB_cut, Type == "INV") %>% count(Repeat, sort = TRUE)%>%mutate(Data="INV")
repeatDB_cut.INS<- filter(repeatDB_cut, Type == "INS") %>% count(Repeat, sort = TRUE)%>%mutate(Data="INS")
repeatDB_cut.TRA <- filter(repeatDB_cut, Type == "TRA") %>% count(Repeat, sort = TRUE)%>%mutate(Data="TRA")

all.SV<-rbind.data.frame(repeatDB_cut.ALL , repeatDB_cut.DEL , repeatDB_cut.DUP , repeatDB_cut.INV,  repeatDB_cut.INS,
repeatDB_cut.TRA )

#bar plot
png("Sv5_repeat_bar.png", width=700, height=500)
ggplot(all.SV, aes(fill=Repeat, y=n, x=Data)) + geom_bar(position="fill", stat="identity")+ scale_fill_brewer(palette="Dark2", name = "SV Repeat
Classification") + theme_classic(base_size = 18) + xlab("SV Classicfication")+ ylab("SVs flagged as repeats")+
theme(axis.text=element_text(size=16),axis.title=element_text(size=20,face="bold"))
dev.off()

repeatDB_cut.ALL.count <- count(repeatDB_cut, Type, sort = TRUE) %>%mutate(Data="ALL")

#sizes violin plot
png("Sv5_svreps_violin.png", width=550, height=500 )
ggplot(repeatDB, aes(x=Repeat, y=Length)) + geom_violin(trim=FALSE, fill='#9BD8E4', color="black") + ylim(30, 1500) +
stat_summary(fun=median, geom="point", size=4, color="black") + ylab("Length (bp)")+ theme_classic() + coord_flip() +
theme(axis.text=element_text(size=16),axis.title=element_text(size=20,face="bold"))+  scale_x_discrete(limits=c("Repeat
(Unclassified)","LINE","LTR","Low Complexity","Simple Repeat","None")) + xlab("SV Repeat Classification")
dev.off()

repeatDB.count <- count(repeatDB, Repeat, sort = TRUE) %>%mutate(Data="ALL")
#all.SV  %>% filter(length <=1000) %>%
#group_by(type)%>%
#summarise(Med=median(length))
```


1.00
0.75
0.50
0.25
0.00
SVs flagged as repeats
ALL
DEL
DUP
INS
INV
TRA
SV Classicfication
SV Repeat Classification
LINE
Low Complexity
LTR
Simple Repeat
Repeat (Unclassified)


None
Simple Repeat
Low Complexity
LTR
LINE
Repeat (Unclassified)
SV Repeat Classification
0
500
1000
1500
Length (bp)