Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/Sv10_NZstarlings/Data/2023-12-21.SNPfiltering

PDF Version generated by

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

on

Sep 25, 2024 @01:00 PM AEST

Table of Contents

2023-12-21.SNPfiltering 2

Katarina Stuart (z5188231@ad.unsw.edu.au) - May 10, 2024, 7:11 AM GMT+10

Starling DArT SNP filtering

CONTEXT: this data contains numerous siblings, which need to be removed. The SNPs have also been called from 3 different DaRT batches, which impacted the sites called. So we need to ensure we are only keeping sites that were sequences in all batches, so we do not introduce batch effects.

Remove duplicate individuals and relatives

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/BCFtools/final_filtering

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

DIR=/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/BCFtools_old

cat $DIR/ReplicatesRemoved.txt $DIR/SibRemoved.txt > ReplicatesRemoved_SibRemoved.txt

#edit to change BAD_003.sorted.bam to BAD_003_R.sorted.bam
sed 's/BAD_003.sorted.bam/BAD_003_R.sorted.bam/g' ReplicatesRemoved_SibRemoved.txt > ReplicatesRemoved_SibRemoved_updated.txt
#and also add in ind 'PLM_001.sorted.bam'

vcftools --gzvcf ../variant_calls_annotate.vcf.gz --remove ReplicatesRemoved_SibRemoved_updated.txt  --recode --recode-INFO-all  --out starling_noduprel
```

After filtering, kept 141 out of 202 Individuals

After filtering, kept 656737 out of a possible 656737 Sites

Next remove indels, and poor quality sites

At this step we remove entire SNPs if they are poor quality (--minQ) or indels. We also exclude all genotype sites (i.e. individual stats) that are below a quality threshold (--minGQ) and also are above or below a depth (DP) level.

```
vcftools --vcf starling_noduprel.recode.vcf --remove-indels --minQ 30 --minGQ 20 --minDP 5 --maxDP 100 --recode --recode-INFO-all  --out starling_noduprel_qual
```

After filtering, kept 353275 out of a possible 656737 Sites

Note: any of the above sites that were removed was only due to the 'minQ' or 'indel' filter, as that removed entire 'rows' i.e. loci/SNPs. The other three flags simply recoded an individuals genotype to './.' if it didn't meet the criteria (i.e. the genotype is now missing and will be considered when we do our missingness filtering).

Next, only keep sites present in 50% of individuals in each sample site

For this, we break up the VCF into each sample grouping, find the SNPs with <50% missingness, and remove these from the SNP data.

```
mkdir persite_missingness_filter
cd persite_missingness_filter

#need to create SNP IDS, doing so by catting CHROM to POS
cat ../starling_noduprel_qual.recode.vcf | perl -lane 'if($F[0] !~ /^#/) {$F[2] = $F[0]."".$F[1];print join("\t",@F);} else {print $_;}' > starling_noduprel_qual_renamed.vcf

#checked to make sure that each SNP ID was actually unique (i.e. that we didn't have two SNPs at the same CHROM and POS)
grep -v "^#" starling_noduprel_qual_renamed.vcf | cut -f 3 | sort | uniq -d

#now calculate the list of SNPs that need to be removed, because they are <50% in at least one pop

grep -v "^#" starling_noduprel_qual_renamed.vcf | cut -f 3 | sort > snplist_full.txt
```

```
module load BCFtools/1.15.1-GCC-11.3.0
bcftools query -l starling_noduprel_qual_renamed.vcf | grep "ATB" > indlist_ANT.txt
bcftools query -l starling_noduprel_qual_renamed.vcf | grep "AUK\|HAM" > indlist_AUK.txt
bcftools query -l starling_noduprel_qual_renamed.vcf | grep "BRA\|BAD" > indlist_BAD.txt
bcftools query -l starling_noduprel_qual_renamed.vcf | grep "CAN" > indlist_CAN.txt
bcftools query -l starling_noduprel_qual_renamed.vcf | grep "MONK" > indlist_MKW.txt
bcftools query -l starling_noduprel_qual_renamed.vcf | grep "sv" > indlist_MLV.txt
bcftools query -l starling_noduprel_qual_renamed.vcf | grep "Y\|V1\|BB\|RR" > indlist_NWC.txt
bcftools query -l starling_noduprel_qual_renamed.vcf | grep "NOR" > indlist_ORG.txt
bcftools query -l starling_noduprel_qual_renamed.vcf | grep "PLM" > indlist_PLM.txt
bcftools query -l starling_noduprel_qual_renamed.vcf | grep "UHT" > indlist_UHT.txt

for i in indlist_*.txt
do
name=$(basename $i .txt | sed 's/indlist_//g')
echo $name
vcftools --vcf starling_noduprel_qual_renamed.vcf --keep $i --max-missing 0.5 --recode --recode-INFO-all  --out starling_noduprel_qual_filt
#list of all SNPs in original vcffile
grep -v "^#" starling_noduprel_qual_filt.recode.vcf | cut -f 3 | sort > snplist_present.txt
comm -23 snplist_full.txt snplist_present.txt > snplist_absent_${name}.txt
done

#list of SNPs that are missing in at least 1 sampling group
cat snplist_absent_*.txt | sort | uniq > snplist_toremove.txt

vcftools --vcf starling_noduprel_qual_renamed.vcf --exclude snplist_toremove.txt --recode --recode-INFO-all  --out starling_noduprel_qual_miss
```

After filtering, kept 68262 out of a possible 353275 Sites

Filter for overall missingness, as well as things like depth and mac

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/BCFtools/final_filtering/

vcftools --vcf persite_missingness_filter/starling_noduprel_qual_miss.recode.vcf --mac 5 --max-missing 0.7 --max-alleles 2 --min-alleles 2 --max-meanDP 100 --thin
1000 --recode --recode-INFO-all  --out starling_noduprel_qual_miss_filt
```

After filtering, kept 19174 out of a possible 68262 Sites

FAnd filter to just NZ inds plus some basic filters

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/BCFtools/final_filtering/

cat persite_missingness_filter/indlist_AUK.txt persite_missingness_filter/indlist_BAD.txt persite_missingness_filter/indlist_CAN.txt
persite_missingness_filter/indlist_PLM.txt persite_missingness_filter/indlist_UHT.txt > indlist_NZ.txt

vcftools --vcf starling_noduprel_qual_miss_filt.recode.vcf --keep indlist_NZ.txt --mac 5 --max-missing 0.7 --recode --recode-INFO-all  --out
starling_noduprel_qual_miss_filt_NZ
```

After filtering, kept 75 out of 141 Individuals
After filtering, kept 14890 out of a possible 19174 Sites

THE DATA

```
/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/BCFtools/final_filtering/starling_noduprel_qual_miss_filt_NZ.recode.vcf
/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/BCFtools/final_filtering/starling_noduprel_qual_miss_filt.recode.vcf
/nesi/nobackup/uoa02613/kstuart_projects/At4_MynaStarling/data/resources/genomes/Svulgaris_vAU_1.0.fasta
```

```
/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/BCFtools_old/Metadata_NZ_AU_UK_BE_ReplicatesSibRemoved2.csv
```

and making bed file formats

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/BCFtools/final_filtering/

module load PLINK/1.09b6.16
module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1

vcftools --vcf starling_noduprel_qual_miss_filt.recode.vcf --plink --out starling_noduprel_qual_miss_filt.plink
plink --file starling_noduprel_qual_miss_filt.plink --make-bed --noweb --out starling_noduprel_qual_miss_filt

vcftools --vcf starling_noduprel_qual_miss_filt_NZ.recode.vcf --plink --out starling_noduprel_qual_miss_filt_NZ.plink
plink --file starling_noduprel_qual_miss_filt_NZ.plink --make-bed --noweb --out starling_noduprel_qual_miss_filt_NZ
```

Making a repeat bed file for the PSMC analysis

```
#!/bin/bash -e

#SBATCH --job-name=2024_02_27.Analysis_Repeatmasking.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=12GB
#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=16
#SBATCH --profile task

#load modules
module load RepeatMasker/4.1.0-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/repeat_analysis

REPEATMASKER=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/repeats/repeatmasker_aves.fasta
CUSTOM=All_repeats_aves_custom.fasta

cat $CUSTOM $REPEATMASKER > All_repeats_Svulgaris_vAU_1.0_custom.fasta

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At4_MynaStarling/data/resources/genomes/Svulgaris_vAU_1.0.fasta

#repeat mask
RepeatMasker -pa 16 -lib ./All_repeats_Svulgaris_vAU_1.0_custom.fasta -dir ./ ${GENOME}
```

RM output to bedfile

RepeatMasker/RM2Bed.py at master · rmhubley/RepeatMasker (github.com)

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/repeats/repeatmasker
python -m pip install pandas
/opt/nesi/CS400_centos7_bdw/RepeatMasker/4.1.0-gimkl-2020a/util/RM2Bed.py Svulgaris_vAU_1.0.fasta.out
awk '{print $3-$2}' Svulgaris_vAU_1.0.fasta_rm.bed | awk '{ sum += $1 } END { print sum }' #total length of repeats
```

98,480,942 (compared to myna 105,078,568) - looks about right, about 10% of 1 Gb genome. And worse quality assembly will less have repeats.

Looking at the problem fixed alleles

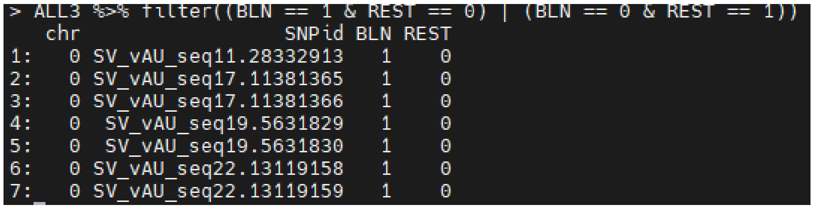
It looks like there are about 5 (after thinning SNP filter) fixed alleles remaining. Re-filtering raw reads doesn't change them. Ignore for now.

Data prep for IGV

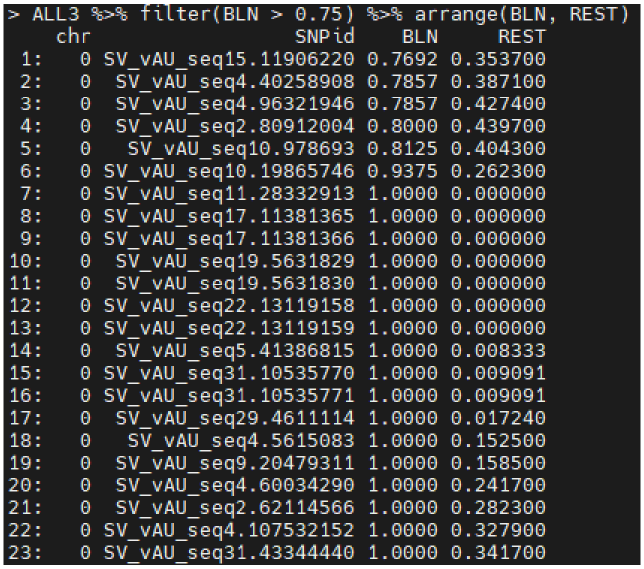
```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/BCFtools/final_filtering/IGV

module purge
module load minimap2/2.24-GCC-11.3.0
module load SAMtools/1.13-GCC-9.2.0

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At4_MynaStarling/data/resources/genomes/Svulgaris_vAU_1.0.fasta
samtools faidx $GENOME
```

and figure out the loci

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/BCFtools/final_filtering/IGV
VCF=/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/BCFtools/final_filtering/starling_noduprel_qual_miss_filt_NZ.recode.vcf

module load PLINK/1.09b6.16
module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1

vcftools --vcf $VCF --plink --out starling_noduprel_qual_miss_filt_NZ.plink
plink --file starling_noduprel_qual_miss_filt_NZ.plink --make-bed --noweb --out starling_noduprel_qual_miss_filt_NZ

cut -f 3- starling_noduprel_qual_miss_filt_NZ.plink.ped > x.delete
paste BLN_REST.txt x.delete > starling_noduprel_qual_miss_filt_NZ.plink.ped
rm x.delete

plink --file starling_noduprel_qual_miss_filt_NZ.plink --allow-extra-chr --freq counts --family --out starling_noduprel_qual_miss_filt_NZ

module load R/4.1.0-gimkl-2020a
R
library(data.table)
library(tidyr)
setwd("/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/BCFtools/final_filtering/IGV")
ALL <- fread("starling_noduprel_qual_miss_filt_NZ.frq.strat")
ALL %>% spread(CLST, MAF)
ALL2<- ALL[,c(1,2,3,6)]
ALL3 <- spread(ALL2 , key=CLST, value=MAF)
colnames(ALL3) <- c("chr","SNPid","BLN","REST")
ALL3 %>% filter((BLN == 1 & REST == 0) | (BLN == 0 & REST == 1))
ALL3 %>% filter(BLN > 0.75) %>% arrange(BLN, REST)
ALL %>% filter(SNP=="SV_vAU_seq11.28332913")
```



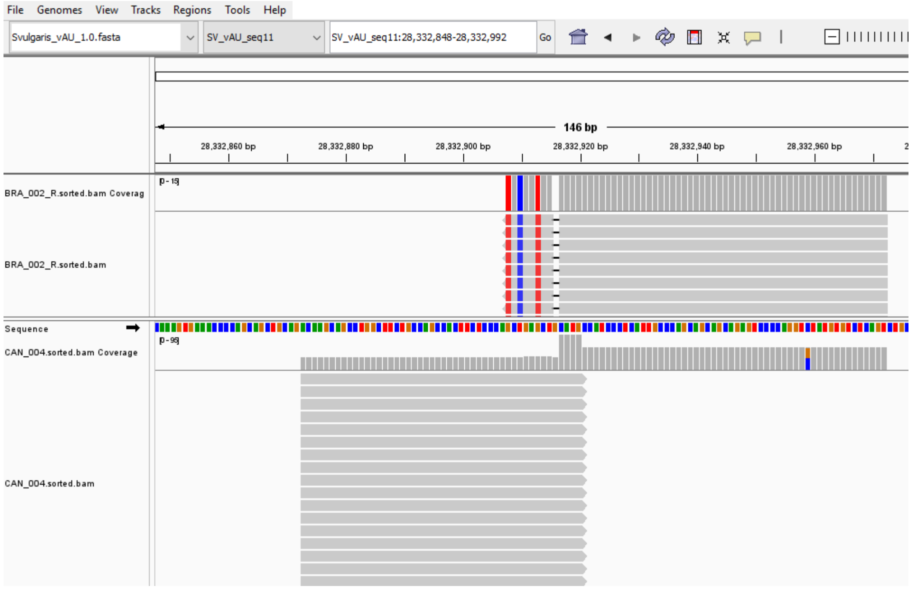
```
> ALL3 %>% filter((BLN == 1 & REST == 0) | (BLN == 0 & REST == 1))
   chr                  SNPid BLN REST
1:   0 SV_vAU_seq11.28332913   1    0
2:   0 SV_vAU_seq17.11381365   1    0
3:   0 SV_vAU_seq17.11381366   1    0
4:   0  SV_vAU_seq19.5631829   1    0
5:   0  SV_vAU_seq19.5631830   1    0
6:   0 SV_vAU_seq22.13119158   1    0
7:   0 SV_vAU_seq22.13119159   1    0
```



```
> ALL3 %>% filter(BLN > 0.75) %>% arrange(BLN, REST)
    chr                  SNPid    BLN     REST
 1:   0 SV_vAU_seq15.11906220 0.7692 0.353700
 2:   0  SV_vAU_seq4.40258908 0.7857 0.387100
 3:   0  SV_vAU_seq4.96321946 0.7857 0.427400
 4:   0  SV_vAU_seq2.80912004 0.8000 0.439700
 5:   0   SV_vAU_seq10.978693 0.8125 0.404300
 6:   0 SV_vAU_seq10.19865746 0.9375 0.262300
 7:   0 SV_vAU_seq11.28332913 1.0000 0.000000
 8:   0 SV_vAU_seq17.11381365 1.0000 0.000000
 9:   0 SV_vAU_seq17.11381366 1.0000 0.000000
10:   0  SV_vAU_seq19.5631829 1.0000 0.000000
11:   0  SV_vAU_seq19.5631830 1.0000 0.000000
12:   0 SV_vAU_seq22.13119158 1.0000 0.000000
13:   0 SV_vAU_seq22.13119159 1.0000 0.000000
14:   0  SV_vAU_seq5.41386815 1.0000 0.008333
15:   0 SV_vAU_seq31.10535770 1.0000 0.009091
16:   0 SV_vAU_seq31.10535771 1.0000 0.009091
17:   0  SV_vAU_seq29.4611114 1.0000 0.017240
18:   0   SV_vAU_seq4.5615083 1.0000 0.152500
19:   0  SV_vAU_seq9.20479311 1.0000 0.158500
20:   0  SV_vAU_seq4.60034290 1.0000 0.241700
21:   0  SV_vAU_seq2.62114566 1.0000 0.282300
22:   0 SV_vAU_seq4.107532152 1.0000 0.327900
23:   0 SV_vAU_seq31.43344440 1.0000 0.341700
```

PICTURES FROM IGV

Barcode?

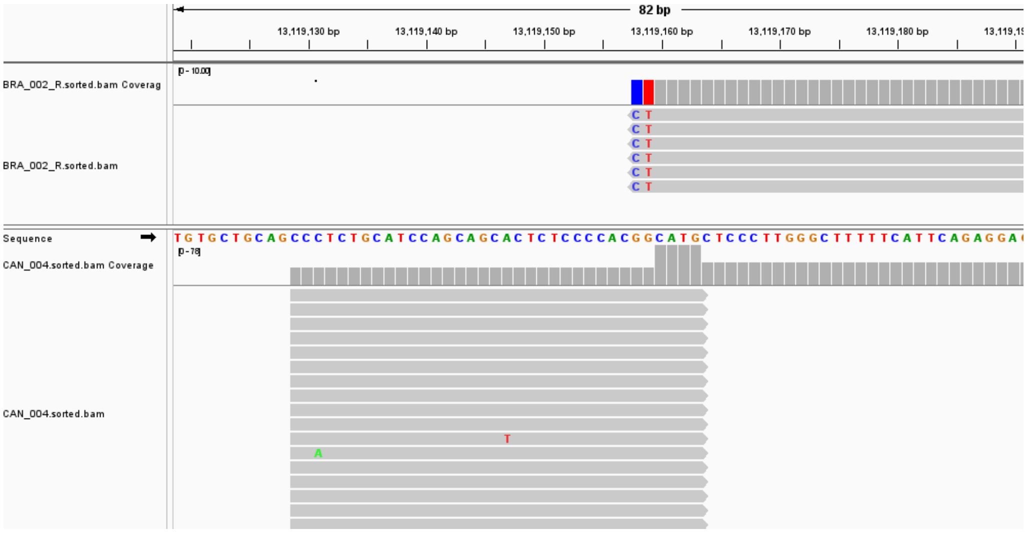


Svulgaris_vAU_1.0.fasta

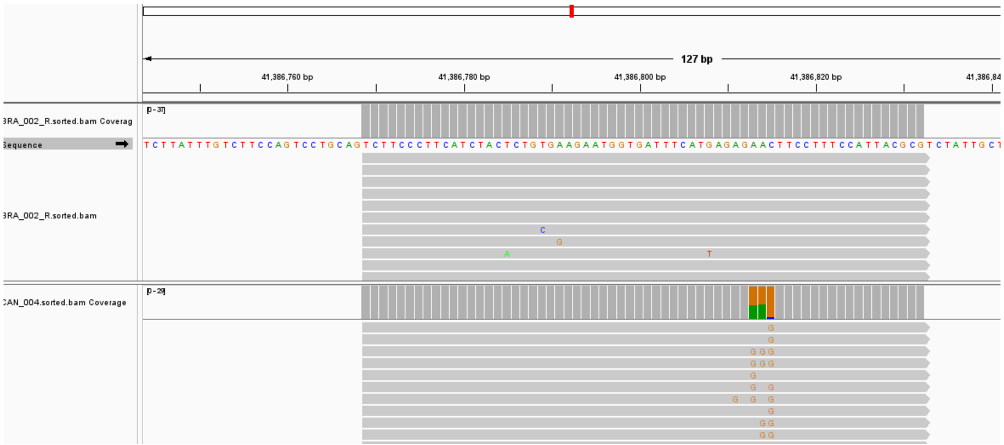
SV_vAU_seq11

SV_vAU_seq11:28,332,848-28,332,992

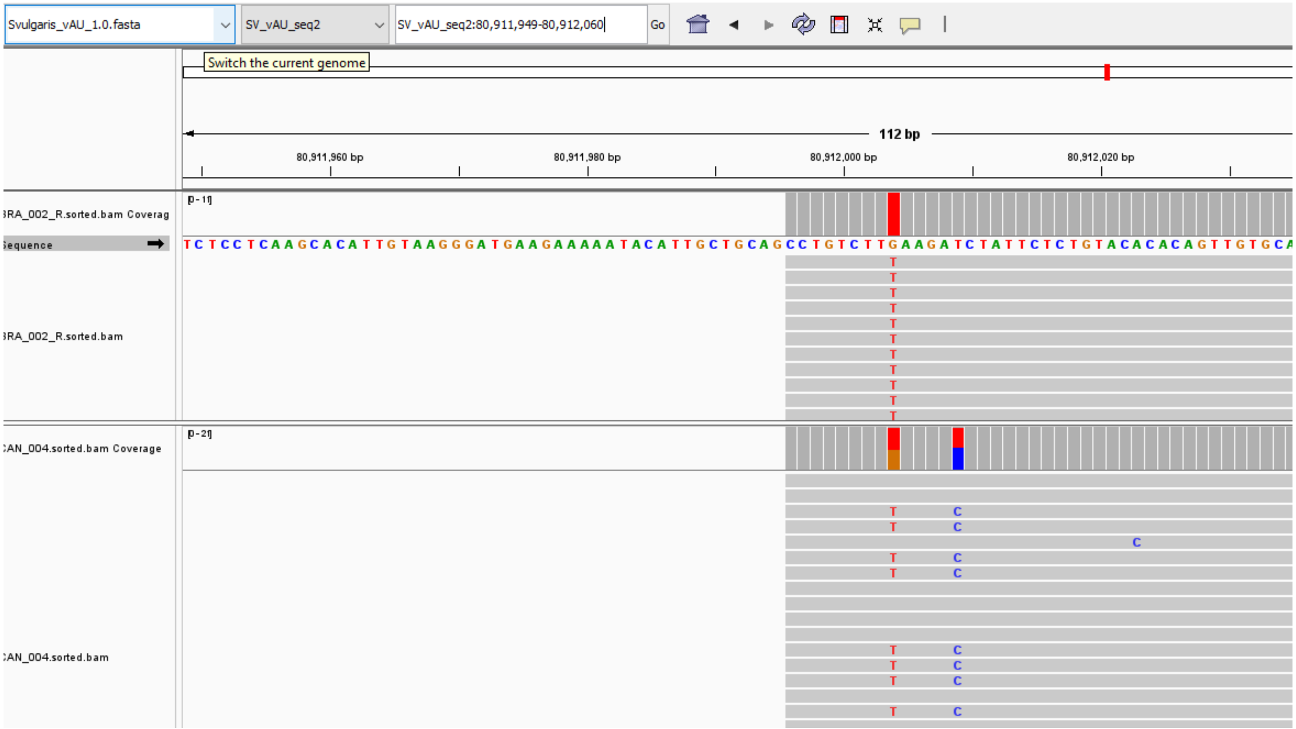
Partial barcode? read length 63 bp



here the variant is the other way around?



more moderate example:


Svulgaris_vAU_1.0.fasta

SV_vAU_seq2

SV_vAU_seq2:80,911,949-80,912,060

Go
Switch the current genome
112 bp
80,911,960 bp
80,911,980 bp
80,912,000 bp
80,912,020 bp
BRA_002_R.sorted.bam Coverag
[0 - 11]
Sequence
TCTCCTCAAGCACATTGTAAGGGATGAAGAAAAATACATTGCTGCAGCCTGTCTTGAAGATCTATTCTCTGTACACACAGTTGTGCA
BRA_002_R.sorted.bam
T
CAN_004.sorted.bam Coverage
[0 - 21]
CAN_004.sorted.bam
T C
C