Starling-May18

Projects/Katarina Stuart/KStuart.Starling-Aug18/Sv3_Genome/Assembly/2021.03.08.Contamination

PDF Version generated by

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

on

Jun 23, 2022 @04:00 PM NZST

Table of Contents

2021.03.08.Contamination

Katarina Stuart (z5188231@ad.unsw.edu.au) - Feb 28, 2022,

Patching Contamination

GENOME ASSEMBLY CORRECTION

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/Chromosemble.L_RNA_scaffolder.polished.tidy.purge.fasta

CONTAMINATION

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/contamination

cp /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/Chromosemble.L_RNA_scaffolder.polished.tidy.purge.fasta/Svulgaris_vA

python /home/z3452659/slimsuitedev/tools/seqsuite.py seqlist -seqin Svulgaris_vAU_1.0.fasta -goodseq SV_vAU_seq19 reformat=region region=2311111,2311141 -seqout adapter5.fasta -bas

python /home/z3452659/slimsuitedev/tools/seqsuite.py seqlist -seqin Svulgaris_vAU_1.0.fasta -goodseq SV_vAU_seq2 reformat=region region=746439,746481 -seqout adapter5.fasta -basefi

python /home/z3452659/slimsuitedev/tools/seqsuite.py seqlist -seqin Svulgaris_vAU_1.0.fasta -goodseq SV_vAU_seq2 reformat=region region=26318677,26318709 -seqout adapter5.fasta -b
adapter2b

python /home/z3452659/slimsuitedev/tools/seqsuite.py seqlist -seqin Svulgaris_vAU_1.0.fasta -goodseq SV_vAU_seq5 reformat=region region=79943516,79943546 -seqout adapter5.fasta -b
adapter5

python /home/z3452659/slimsuitedev/tools/seqsuite.py seqlist -seqin Svulgaris_vAU_1.0.fasta -goodseq SV_vAU_seq31 reformat=region region=68351768,68352882 -seqout Delftiaacidovora
basefile $PREFIX

python /home/z3452659/slimsuitedev/tools/seqsuite.py seqlist -seqin Svulgaris_vAU_1.0.fasta -goodseq SV_vAU_seq5 reformat=region region=79943516,79943546 -seqout adapter5.fasta -b
adapter5

USING GABLAM

```
module load minimap2/2.17
module load python/2.7.15

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/contamination
python /home/z3452659/slimsuitedev/tools/gablam.py -seqin DelftiaacidovoransSPH1.fasta -searchdb Svulgaris_vAU_1.0.fasta -mapper minimap -qassemble -dna
python /home/z3452659/slimsuitedev/tools/gablam.py -seqin gcontam1 -searchdb Svulgaris_vAU_1.0.fasta -mapper minimap -qassemble -dna
```

USING VECSCREEN

Download this: ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/contam_in_euks.fa.gz

```
module load blast+/2.9.0
module load python/2.7.15
python /home/z3452659/slimsuitedev/tools/diploidocus.py runmode=vecscreen screendb=DelftiaacidovoransSPH1.fasta screenmode=purge basefile=Svulgaris_vAU_1.0 seqin=Svulgaris_v
 python /home/z3452659/slimsuitedev/tools/diploidocus.py runmode=vecscreen screendb=gcontam1 screenmode=purge basefile=gcontam1xSvulgaris_vAU_1.0 seqin=Svulgaris_vAU_1.0.f
```

MAPPING READS OVER LENGTHS

Minimap nanopore reads - can we span the below lengths?

Sequence name, length, span(s), apparent source
SV_vAU_seq12 25604278 73349..74026,74127..75782 Delftia acidovorans SPH-1 **NO NP OVER SPAN**
SV_vAU_seq19 11302685 2311111..2311141 adaptor:NGB01087.1 **NP PRESENT**
SV_vAU_seq197 3999 1..294 Paraburkholderia xenovorans LB400 **TRIM**
SV_vAU_seq2 125621387 746439..746481,26318677..26318709 adaptor:multiple **NP PRESENT**
SV_vAU_seq29 6386691 3247965..3247988 adaptor:NGB01088.1 **NO NP OVER SPAN**
SV_vAU_seq30 6377718 5670899..5670930 adaptor:NGB00751.1 **NP PRESENT**
SV_vAU_seq31 71003646 68351768..68352882 Delftia acidovorans SPH-1 **NO NP OVER SPAN**
SV_vAU_seq4 151506550 1..1189,1290..1685,1786..3065,30881990..30882011,104039766..104039796 Delftia acidovorans SPH-1,adaptor:multiple **TRIM first bit to 3065 NP PRESENT**

SV_vAU_seq404 1928 798..1055,1156..1928 Delftia acidovorans SPH-1 **REMOVE all of seq 404**
SV_vAU_seq5 107054996 79943516..79943546 adaptor:NGB00732.1 **NP PRESENT**
SV_vAU_seq7 22215596 4294604..4294635 adaptor:NGB00751.1 **NP PRESENT**
SV_vAU_seq8 58177063 15523701..15523772,49124925..49124955,50413122..50413153 adaptor:multiple **NP PRESENT**
SV_vAU_seq9 34963307 20255908..20255956,20798677..20798715 adaptor:multiple **NP PRESENT**

```
module load unswdataarchive/2020-03-19
download.sh /UNSW_RDS/H0236593/Private/Projects/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/filtered /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data
download.sh /UNSW_RDS/H0236593/Private/Projects/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/pass.fastq /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data
```

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/contamination/gapspanning
module load minimap2/2.17
GENOME=../Svulgaris_vAU_1.0.fasta
NANOPORE=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/filtered.fq
minimap2 -t 8 -a ${GENOME} ${NANOPORE} > nanoporexassembly.sam
```

```
module load samtools
```

```
samtools view -bS nanoporexassembly.sam > nanoporexassembly.bam
samtools sort nanoporexassembly.bam -o nanoporexassembly_sort.bam
samtools index nanoporexassembly_sort.bam
samtools view nanoporexassembly_sort.bam SV_vAU_seq12:73349-74026 > SV_vAU_seq12:73349.sam
samtools view nanoporexassembly_sort.bam SV_vAU_seq12:74127-75782 > SV_vAU_seq12:74127.sam
samtools view nanoporexassembly_sort.bam SV_vAU_seq19:2311111-2311141 > SV_vAU_seq19:2311111.sam
samtools view nanoporexassembly_sort.bam SV_vAU_seq2:746439-746481 > SV_vAU_seq2:746439.sam
samtools view nanoporexassembly_sort.bam SV_vAU_seq2:26318677-26318709 > SV_vAU_seq2:26318677.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq29:3247965-3247988 > SV_vAU_seq29:3247965.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq30:5670899-5670930 > SV_vAU_seq30:5670899.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq31:68351768-68352882 > SV_vAU_seq31:68351768.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq4:1-1189 > SV_vAU_seq4:1.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq4:1290-1685 > SV_vAU_seq4:1290.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq4:1786-3065 > SV_vAU_seq4:1786.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq4:30881990-30882011 > SV_vAU_seq4:30881990.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq4:104039766-104039796 > SV_vAU_seq4:104039766.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq5:79943516-79943546 > SV_vAU_seq5:79943516.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq7:4294604-4294635 > SV_vAU_seq7:4294604.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq8:15523701-15523772> SV_vAU_seq8:15523701.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq8:49124925-49124955 > SV_vAU_seq8:49124925.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq8:50413122-50413153> SV_vAU_seq8:50413122.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq9:20255908-20255956> SV_vAU_seq9:20255908.sam
samtools view nanoporexassembly_sort.bam  SV_vAU_seq9:20798677-20798715 > SV_vAU_seq9:20798677.sam
```

Mapping isoseq reads across the lengths that didn't have NP

SV_vAU_seq12 25604278 73349..74026,74127..75782 Delftia acidovorans SPH-1 **NO NP OVER SPAN**

SV_vAU_seq31 71003646 68351768..68352882 Delftia acidovorans SPH-1 **NO NP OVER SPAN**

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/contamination/gapspanning
module load samtools
module load minimap2/2.17
GENOME=../Svulgaris_vAU_1.0.fasta
ISOSEQ=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.1_StarlingIsoseq/analysis/Isoseq3.3_pipeline/polya_8/clustered.hq.fasta
```

```
minimap2 -t 16 -ax splice -uf --secondary=no --splice-flank=no -C5 -O6,24 -B4 \
    ${GENOME} ${ISOSEQ} \
    > clustered.hq.fasta.sam \
    2> clustered.hq.fasta.sam.log
```

```
samtools view -bS clustered.hq.fasta.sam > clustered.hq.fasta.bam
samtools sort clustered.hq.fasta.bam -o clustered.hq.fasta_sort.bam
samtools index clustered.hq.fasta_sort.bam
```

```
samtools view clustered.hq.fasta_sort.bam SV_vAU_seq12:73349-74026 > SV_vAU_seq12:73349_isoseq.sam
samtools view clustered.hq.fasta_sort.bam SV_vAU_seq31:68351768-68352882 > SV_vAU_seq31:68351768_isoseq.sam
```

LOCATING MITOCHONDIRAL SEQ

GitHub - slimsuite/numtfinder: NUMTFinder: Nuclear mitochondrial fragment (NUMT) search tool

run numtfinder to check for mito in the rest of the genome. Remove sequence 1.

Sturnus vulgaris isolate sv009 mitochondrion, complete genome - Nucleotide - NCBI (nih.gov)

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/contamination/mito
```

```
module load blast+/2.9.0
module load python/2.7.15
```

```
python /home/z3452659/slimsuitedev/dev/numtfinder.py seqin=../Svulgaris_vAU_1.0.fasta mtdna=Svulgaris_mito.fasta basefile=Svulgaris_vAU_1.0_mito
```

MAPPING MITO FOR GENOME FILE

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/contamination/mito
```

```
module load minimap2/2.17
GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Diplodocus_tidy_all/Purgehap/purge_L_RNA_scaffolder.polished.tidy/L_RNA_scaffolder.polished.tidy.purge.fasta
MT=Svulgaris_mito.fasta
minimap2 -t 8 -a ${MT} ${GENOME} > Step7_MitoMapping.sam
```

```
awk '$5==60' Step7_MitoMapping.sam > mapping.sam
```

#3 sequences... one is the messed up Mito. not circular....

see where mito aligns to on this sequence (scaff_49)

```
module load minimap2/2.17
GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Diplodocus_tidy_all/Purgehap/purge_L_RNA_scaffolder.polished.tidy/L_RNA_scaffolder.polished.tidy.purge.fasta
MT=Svulgaris_mito.fasta
minimap2 -t 8 -a ${GENOME} ${MT} > Mito_Step7Mapping.sam
```

grab the mito sequence from the larger sequence

```
samtools faidx $GENOME
samtools faidx $GENOME scaffold_49_pilon:77877-94675 > scaffold_49_pilon.77877.fasta
```

Turn into one line fasta sequence, and rename for catting into the final v1.1 genome

```
awk '/^>/ {printf("\n%s\n",$0);next; } { printf("%s",$0);}  END {printf("\n");}' < scaffold_49_pilon.77877.fasta > scaffold_49_pilon.77877.1L.fasta
```

```
sed 's/scaffold_49_pilon:77877-94675/SV_vAU_seq1 Sturnus vulgaris/g' scaffold_49_pilon.77877.1L.fasta > Svulgaris_vAU_1.1.mito.fasta
```

Cleaning final genome:

» Sequences to clean (+ long contig list for removal)

```
Sequence name, length, span(s), apparent source
SV_vAU_seq12 25604278 73349..74026,74127..75782 Delftia acidovorans SPH-1
SV_vAU_seq19 11302685 2311111..2311141 adaptor:
SV_vAU_seq197 3999 1..294 Paraburkholderia xenovorans LB400
SV_vAU_seq2 125621387 746439..746481,26318677..26318709 adaptor:multiple
SV_vAU_seq29 6386691 3247965..3247988 adaptor:NGB01088.1
SV_vAU_seq30 6377718 5670899..5670930 adaptor:NGB00751.1
SV_vAU_seq31 71003646 68351768..68352882 Delftia acidovorans SPH-1
SV_vAU_seq4 151506550 1..1189,1290..1685,1786..3065,30881990..30882011,104039766..104039796 Delftia acidovorans SPH-1,adaptor:multiple
SV_vAU_seq404 1928 798..1055,1156..1928 Delftia acidovorans SPH-1 REMOVED
SV_vAU_seq5 107054996 79943516..79943546 adaptor:NGB00732.1
SV_vAU_seq7 22215596 4294604..4294635 adaptor:NGB00751.1
SV_vAU_seq8 58177063 15523701..15523772,49124925..49124955,50413122..50413153 adaptor:multiple
SV_vAU_seq9 34963307 20255908..20255956,20798677..20798715 adaptor:multiple
```

1) run with vecscreen. mask necessary sites. Grab species for directed search
2) Gablam using contamination genomes of interest (4 different sp.). Remove contigs or split scaffolds as needed

```
#!/bin/bash
#PBS -N 2021-03-13.vecscreen_1db.pbs
#PBS -l nodes=1:ppn=24
#PBS -l vmem=120gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae
```

```
module load python/2.7.15
```

```
module add blast+/2.11.0
```

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/assembly/finalAss_Contamination_screen/vecscreen2
DB=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/assembly/Diploidocus_Vecscreen_finalAss/vecscreendb.fasta
```

```
cp /srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/Chromosemble.L_RNA_scaffolder.polished.tidy.purge.fasta/Svulgaris_vAU_1.0.fasta .
```

```
python /home/z3452659/slimsuitedev/tools/diploidocus.py runmode=vecscreen screendb=$DB screenmode=purge
basefile=Svulgaris_vAU_1.0 seqin=Svulgaris_vAU_1.0.fasta vecmask=27 forks=16 keepnames=T
```

Vecscreen output:

SeqName	SeqLen	Start	End	Edit
SV_vAU_seq10	38196499	23153162	23153241	mask
SV_vAU_seq10	38196499	23153360	23153523	mask
SV_vAU_seq10	38196499	23157117	23157229	mask
SV_vAU_seq111	39634	28654	28760	mask
SV_vAU_seq111	39634	28878	28945	mask
SV_vAU_seq12	25604278	14004413	14004475	mask
SV_vAU_seq12	25604278	14004925	14004976	mask
SV_vAU_seq12	25604278	14006192	14006243	mask
SV_vAU_seq16	19074277	2696075	2696196	mask
SV_vAU_seq19	11302685	2311111	2311141	mask
SV_vAU_seq2	125621387	746455	746481	mask
SV_vAU_seq2	125621387	26318677	26318709	mask
SV_vAU_seq20	12169802	11100954	11101007	mask
SV_vAU_seq22	15948174	11510375	11510412	mask
SV_vAU_seq22	15948174	11516273	11516346	mask
SV_vAU_seq30	6377718	5670899	5670930	mask
SV_vAU_seq4	151506550	104039766	104039796	mask
SV_vAU_seq5	107054996	63760176	63760225	mask
SV_vAU_seq5	107054996	79943516	79943546	mask
SV_vAU_seq7	22215596	4294604	4294635	mask
SV_vAU_seq8	58177063	15523707	15523774	mask
SV_vAU_seq8	58177063	49124925	49124955	mask
SV_vAU_seq8	58177063	49314128	49314190	mask

Gablam targeted search of 4 bacteria

```
module load minimap2/2.17
module load python/2.7.15
```

```
DELFIA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/contamination/DelftiaacidovoransSPH1.fasta
ACIDO=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/contamination/Acidovoraxsp.JS42.fasta
ALICY=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/contamination/AlicycliphilusdenitrificansK601.fasta
PARA1=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/contamination/ParaburkholderiaxenovoransLB400chr1.fasta
PARA2=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/contamination/ParaburkholderiaxenovoransLB400chr2.fasta
PARA3=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/contamination/ParaburkholderiaxenovoransLB400chr3.fasta

FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/assembly/finalAss_Contamination_screen/vecscreen/Svulgaris_vAU_1.0.vecscreen.fasta
```

```
python /home/z3452659/slimsuitedev/tools/gablam.py -seqin $DELFIA -searchdb $FASTA -mapper minimap -qassemble -dna -forks 16
python /home/z3452659/slimsuitedev/tools/gablam.py -seqin $ACIDO -searchdb $FASTA -mapper minimap -qassemble -dna -forks 16
python /home/z3452659/slimsuitedev/tools/gablam.py -seqin $ALICY -searchdb $FASTA -mapper minimap -qassemble -dna -forks 16
python /home/z3452659/slimsuitedev/tools/gablam.py -seqin $PARA1 -searchdb $FASTA -mapper minimap -qassemble -dna -forks 16
python /home/z3452659/slimsuitedev/tools/gablam.py -seqin $PARA2 -searchdb $FASTA -mapper minimap -qassemble -dna -forks 16
python /home/z3452659/slimsuitedev/tools/gablam.py -seqin $PARA3 -searchdb $FASTA -mapper minimap -qassemble -dna -forks 16
```

Remove seq1 and the contam seq

```
module load seqtk
```

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/assembly/finalAss_Contamination_screen/final_cleanup
grep -e ">" $FASTA | sed 's/>//g' > original_genome_sequences.txt
grep -Fvxf ../bacterial_contam/excluded_contigs_fix.txt original_genome_sequences.txt > original_genome_sequences_prune.txt
seqtk subseq ../vecscreen/Svulgaris_vAU_1.0.vecscreen.fasta original_genome_sequences_prune.txt > Svulgaris_vAU_1.0.vecscreen.prune.fasta
```

SV_vAU_seq12 25604278 73349..74026,74127..75782 Delftia acidovorans SPH-1 **NEEDS SPLITTING**
SV_vAU_seq31 71003646 68351768..68352882 Delftia acidovorans SPH-1 **NEEDS SPLITTING**
SV_vAU_seq4 151506550 1..1189,1290..1685,1786..3065,30881990..30882011, Delftia acidovorans SPH-1,adaptor:multiple **TRIM first bit to 3065, then MASK SECOND**
SV_vAU_seq29 6386691 3247965..3247988 adaptor:NGB01088.1 **NEEDS MASKING**
SV_vAU_seq8 50413122..50413153 adaptor:multiple **NEEDS MASKING**
SV_vAU_seq9 34963307 20255908..20255956,20798677..20798715 adaptor:multiple **NEEDS MASKING**

Splitting/trimming:

```
python /home/z3452659/slimsuitedev/tools/seqsuite.py -seqin Svulgaris_vAU_1.0.vecscreen.prune.fasta edit
```

```
Seq12: split at at 73349, trim till 2434
Seq31: split at 68351768, trim till 1115
Seq 4, trim till 3066
```

Masking:

```
SV_vAU_seq4 30878925 30878946
SV_vAU_seq29 3247965 3247988
SV_vAU_seq8 50413122 50413153
SV_vAU_seq9 20255908 20255956
SV_vAU_seq9 20798677 20798715
```

```
bedtools getfasta -fi Svulgaris_vAU_1.0.vecscreen.prune.edit.fas -bed sequences_to_mask.bed
```

```
>SV_vAU_seq4:30878925-30878946
ATCGGAAGAGCGTCGTGTATG
>SV_vAU_seq29:3247965-3247988
TTGCCACGACGCTCTTCCGATCT
>SV_vAU_seq8:50413122-50413153
ATGATGCGGCGACCACCGAGATCTACACCTG
>SV_vAU_seq9:20255908-20255956
CTCCAGTCACGGATGGGCATCCCGTATGCCGTCTTCTGCTCCACAGCA
>SV_vAU_seq9:20798677-20798715
CGAGCTCTACACTCTTGCCCTACCCGACGCTCTTCCGA
```

edit to:

```
>SV_vAU_seq4:30878925-30878946
ANCNGNANANCNTNGNGNANG
>SV_vAU_seq29:3247965-3247988
TNGNCNCNANGNTNTNCNGNTNT
>SV_vAU_seq8:50413122-50413153
ANGNTNCNGNGNCNANCNANANCNANANCNG
>SV_vAU_seq9:20255908-20255956
CNCNANTNANGNANGNGNANCNCNTNTNCNGNCNTNTNCNCNANANCN
>SV_vAU_seq9:20798677-20798715
CNANCNCNANANTNTNGNCNTNCNCNANGNTNTNCNGN
```

```
grep "ATCGGAAGAGCGTCGTGTATG" Svulgaris_vAU_1.0.vecscreen.prune.edit.fas  | wc -l
grep "TTGCCACGACGCTCTTCCGATCT" Svulgaris_vAU_1.0.vecscreen.prune.edit.fas  | wc -l
grep "ATGATGCGGCGACCACCGAGATCTACACCTG" Svulgaris_vAU_1.0.vecscreen.prune.edit.fas  | wc -l
grep "CTCCAGTCACGGATGGGCATCCCGTATGCCGTCTTCTGCTCCACAGCA" Svulgaris_vAU_1.0.vecscreen.prune.edit.fas  | wc -l
grep "CGAGCTCTACACTCTTGCCCTACCCGACGCTCTTCCGA" Svulgaris_vAU_1.0.vecscreen.prune.edit.fas  | wc -l
```

```
sed 's/ATCGGAAGAGCGTCGTGTATG/ANCNGNANANCNTNGNGNANG/g' Svulgaris_vAU_1.0.vecscreen.prune.edit.fas | sed
's/TTGCCACGACGCTCTTCCGATCT/TNGNCNCNANGNTNTNCNGNTNT/g' | sed 's/ATGATGCGGCGACCACCGAGATCTACACCTG/ANGNTNCNGNGNCNANCNANANCNANANCNG/g' |
's/CTCCAGTCACGGATGGGCATCCCGTATGCCGTCTTCTGCTCCACAGCA/CNCNANTNANGNANGNGNANCNCNTNTNCNGNCNTNTNCNCNANANCN/g' | sed
's/CGAGCTCTACACTCTTGCCCTACCCGACGCTCTTCCGA/CNANCNCNANANTNTNGNCNTNCNCNANGNTNTNCNGN/g' > Svulgaris_vAU_1.0.vecscreen.prune.edit.mask.fasta
```

```
bedtools getfasta -fi Svulgaris_vAU_1.0.vecscreen.prune.edit.mask.fasta -bed sequences_to_mask.bed
```

```
sed 's/ (Vecscreen:masked)//g' Svulgaris_vAU_1.0.vecscreen.prune.edit.mask.fasta | sed 's/ (Region 3066 to 151506550)//g'  | sed 's/ (Region 1 to 73348)//g' | sed 's/SV_vAU_seq12 Sturnu
vulgaris/SV_vAU_seq12b Sturnus vulgaris/g' | sed 's/SV_seq12__SV_vAU_seq12.73349 Sturnus vulgaris (Region 73349 to 25604278) (Region 2434 to 25530930)/SV_vAU_seq12 Sturnus
vulgaris/g' | sed 's/ (Region 1 to 68351767)//g' | sed 's/SV_seq31__SV_vAU_seq31.68351768 Sturnus vulgaris (Region 68351768 to 71003646) (Region 1115 to 2651879)/SV_vAU_seq31b
vulgaris/g' > Svulgaris_vAU_1.0.vecscreen.prune.edit.mask.renamed.fasta
```

```
grep "^>" Svulgaris_vAU_1.0.vecscreen.prune.edit.mask.renamed.fasta | head -n 35
```

```
cp Svulgaris_vAU_1.0.vecscreen.prune.edit.mask.renamed.fasta Svulgaris_vAU_1.1.fasta
```

```
python /home/z3452659/slimsuitedev/tools/seqsuite.py -seqin Svulgaris_vAU_1.1.fasta -summarise -dna
```

```
#~~#    00:00:03        # ~~~~~~~~ Sequence Summary for Svulgaris_vAU_1.1 ~~~~~~~~ #
#SUM    00:00:49        Total number of sequences: 1,344
#SUM    00:00:49        Total length of sequences: 1,043,825,671
```

```
#SUM    00:00:49    Min. length of sequences: 927
#SUM    00:00:49    Max. length of sequences: 151,503,485
#SUM    00:00:49    Mean length of sequences: 776,656.01
#SUM    00:00:49    Median length of sequences: 1,343
#SUM    00:00:49    N50 length of sequences: 72,244,370
#SUM    00:00:49    L50 count of sequences: 5
#SUM    00:00:49    Total number of contigs: 23,340
#SUM    00:00:49    Contig N50 length of sequences: 147,322
#SUM    00:00:49    Contig L50 count of sequences: 2,010
#SUM    00:00:49    GC content: 41.72%
#SUM    00:00:49    N bases: 7,732,465 (0.74%)
#SUM    00:00:49    Gap (10+ N) length: 7,731,135 (0.74%)
#SUM    00:00:49    Gap (10+ N) count: 21,996
#RUN    00:00:49    SeqList V1.46.0 run finished.
#LOG    00:00:49    SeqSuite V1.25.0 End: Mon Mar 15 17:46:45 2021
```

```
cat /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/analysis/contamination/mito/Svulgaris_vAU_1.1.mito.fasta Svulgaris_vAU_1.0.vecscreen.prune.edit.mask.renamed.fasta > Svulgaris_vAU_1.0.vecscreen.prune.edit.mask.renamed.mito.fasta
```

```
grep "^>" Svulgaris_vAU_1.0.vecscreen.prune.edit.mask.renamed.mito.fasta | head -n 35
```

```
cp Svulgaris_vAU_1.0.vecscreen.prune.edit.mask.renamed.mito.fasta Svulgaris_vAU_1.1.mito.fasta
```

```
python /home/z3452659/slimsuitedev/tools/seqsuite.py -seqin Svulgaris_vAU_1.1.mito.fasta -summarise -dna
```

```
#SUM    00:00:52    Total number of sequences: 1,344
#SUM    00:00:52    Total length of sequences: 1,043,825,671
#SUM    00:00:52    Min. length of sequences: 927
#SUM    00:00:52    Max. length of sequences: 151,503,485
#SUM    00:00:52    Mean length of sequences: 776,656.01
#SUM    00:00:52    Median length of sequences: 1,343
#SUM    00:00:52    N50 length of sequences: 72,244,370
#SUM    00:00:52    L50 count of sequences: 5
#SUM    00:00:52    Total number of contigs: 23,340
#SUM    00:00:52    Contig N50 length of sequences: 147,322
#SUM    00:00:52    Contig L50 count of sequences: 2,010
#SUM    00:00:52    GC content: 41.72%
#SUM    00:00:52    N bases: 7,732,465 (0.74%)
#SUM    00:00:52    Gap (10+ N) length: 7,731,135 (0.74%)
#SUM    00:00:52    Gap (10+ N) count: 21,996
#RUN    00:00:52    SeqList V1.46.0 run finished.
#LOG    00:00:52    SeqSuite V1.25.0 End: Sat Mar 13 20:37:37 2021
```

```
#!/bin/bash

#PBS -N 2021-03-15.BUSCO.pbs
#PBS -V
#PBS -l nodes=1:ppn=40
#PBS -l mem=56gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae
```

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/assembly/finalAss_Contamination_screen/final_cleanup
```

```
module load python/3.7.3 blast+/2.2.31 hmmer/3.2.1 augustus/3.3.2 emboss/6.6.0 busco/3.0.2b
```

```
export AUGUSTUS_CONFIG_PATH=/srv/scratch/z5188231/programs/augustus
export BUSCO_CONFIG_FILE=/srv/scratch/z5188231/KStuart.Starling-Aug18/programs/busco-3.0.2/config/config.ini
```

```
BUSCOSET=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data/BUSCO.2018-08-21
python3 /apps/busco/3.0.2b/scripts/run_BUSCO.py -i ./Svulgaris_vAU_1.1.fasta -o Svulgaris_vAU_1.1.fasta -m genome -l ${BUSCOSET}/aves_odb9/ -c 40 -f
```

```
module purge
module load minimap2/2.17

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.1_StarlingIsoseq/mapping/minimap_versions

GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/assembly/finalAss_Contamination_screen/final_cleanup/Svulgaris_vAU_1.1.fasta
ISOSEQ=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.1_StarlingIsoseq/analysis/Isoseq3.3_pipeline/polya_8/clustered.hq.fasta

minimap2 -t 16 -ax splice -uf --secondary=no --splice-flank=no -C5 -O6,24 -B4 \
    ${GENOME} ${ISOSEQ} \
    > Svulgaris_vAU_1.1.sam \
    2> Svulgaris_vAU_1.1.log
```