Starling-May18

Projects/Katarina Stuart/KStuart.Starling-Aug18/Sv3_Genome/Annotation/2020-10-22.vAUannotation

PDF Version generated by

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

on

Jun 23, 2022 @04:13 PM NZST

Table of Contents

Maker-with species specific repeat library

http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial_for_WGS_Assembly_and_Annotation_Winter_School_2018

https://github.com/xvazquezc/genome_annotation_with_Maker2/blob/master/Maker2_protocol/Maker2_protocol.md

Setup

Variable List

```
MYGENOME_DIR=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER

PREFIX=Svulgaris
```

link libraries to new space

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER
ln -s /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/genome_assembly/Sturnus_vulgaris_2.3.1.simp.fasta .
ln -s /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/adv_repeats/LTR/final_libs/allRepeats.lib .
ln -s /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/adv_repeats_lib/uniprot_sprot_clean.fasta .
ln -s /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2018-09-09.NoBusco.MAKER/te_proteins.fasta .
ln -s /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.1_StarlingIsoseq/mapping/minimap_3.2.1/Starling.a100.z30.fasta .
ln -s /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2018-09-09.NoBusco.MAKER/GFDQ01.1.fsa_nt .
```

Edit maker_opts.ctl

```
cd ${MYGENOME_DIR}
maker -CTL
```

Edit the following lines in maker_opts.ctl:

```
genome=Sturnus_vulgaris_2.3.1.simp.fasta
protein=uniprot_sprot_clean.fasta
model_org=vertebrates
rmlib=allRepeats.lib
repeat_protein=te_proteins.fasta
protein2genome=1
trna=1
cpus= 8
min_protein=20
always_complete=1
single_exon=1

est=Starling.a100.z30.fasta
altest=GFDQ01.1.fsa_nt
est2genome=1
correct_est_fusion=1
```

```
formatdb=/apps/blast/2.2.26/bin/formatdb \ #location of NCBI formatdb executable
blastall=/apps/blast/2.2.26/bin/blastall #location of NCBI blastall executable
augustus=/apps/augustus/3.3.2/bin #location of augustus executable
```

have to use trnascan 1.3.1 as v2 will error out!

Running Maker2

Maker: First run

```
#!/bin/bash

#PBS -N 2020-10-22.vAU_maker_run1.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=124gb
#PBS -l walltime=100:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

module purge
module add perl/5.28.0
module add boost/1.70.0
module add recon/1.08
module add repeatscout/1.0.5
module add trf/4.09
module add rmblast/2.6.0
module add repeatmasker/4.0.7
module add repeatmodeler/1.0.11
module add snap/2013-11-29
module add exonerate/2.2.0
module add genemark/es-4.38
module add infernal/1.1.2
module add trnascan-se/1.3.1
module add blast+/2.9.0
module add maker/2.31.9

export PATH=/apps/trnascan-se/1.3.1/bin:$PATH
export PATH=/apps/trnascan-se/1.3.1/lib:$PATH

BASE_PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER
PREFIX=Svulgaris
PBS_NUM_PPN=16

cd ${BASE_PATH}

maker -c ${PBS_NUM_PPN} -base ${PREFIX} ${BASE_PATH}/maker_opts.ctl ${BASE_PATH}/maker_bopts.ctl ${BASE_PATH}/maker_exe.ctl
```

Create a backup for maker run 1

```
cd ${MYGENOME_DIR}
tar cvf ${PREFIX}.maker.output_run1.tar ${PREFIX}.maker.output/
```

to unzip:

```
tar -xvf Svulgaris.maker.output_run1.tar
```

Get the results from round 1

```
mkdir -p results_run1
cd results_run1
gff3_merge -d ../${PREFIX}.maker.output/${PREFIX}_master_datastore_index.log
fasta_merge -d ../${PREFIX}.maker.output/${PREFIX}_master_datastore_index.log
```

WARNING: Transcipt to protein mismatch for trnascan

Not sure if important?

Maker: second run

Training Snap

```
cd ${MYGENOME_DIR}
mkdir -p snap1
cd snap1
ln -s ../results_run1/${PREFIX}.all.gff ${PREFIX}.all.gff
```

```
maker2zff ${PREFIX}.all.gff
```

```
fathom genome.ann genome.dna -categorize 1000
fathom uni.ann uni.dna -export 1000 -plus
forge export.ann export.dna
hmm-assembler.pl ${PREFIX} . > ${PREFIX}.snap1.hmm
```

Train Augustus

Output of buscolong can be found at 2020-04-06.BUSCOlong

Will need to update augustus exicutable so that is calls the correct version with the train Svulgaris parameters.

Running maker round 2

set up control files:

```
cp maker_opts.ctl maker_opts_run1.ctl
cp maker_opts.ctl maker_opts_run2.ctl
```

Make the following changes to the opts.ctl file:

```
snaphmm=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/snap1/Svulgaris.snap1.hmm
augustus_species=BUSCO_Sturnus_vulgaris_2.3.simp_264598804
est2genome=0
protein2genome=0
```

Create a backup for maker run 2

```
cd ${MYGENOME_DIR}
tar cvf ${PREFIX}.maker.output_run2.tar ${PREFIX}.maker.output/
```

Get the results (again)

```
cd ${MYGENOME_DIR}
mkdir -p results_run2
cd results_run2
gff3_merge -d ../${PREFIX}.maker.output/${MYGENOME}_master_datastore_index.log
fasta_merge -d ./${MYGENOME}.maker.output/${MYGENOME}_master_datastore_index.log
```

```
mkdir -p results_run2
cd results_run2
gff3_merge -d ../${PREFIX}.maker.output/${PREFIX}_master_datastore_index.log
fasta_merge -d ../${PREFIX}.maker.output/${PREFIX}_master_datastore_index.log
```

WARNING: Transcipt to protein mismatch for trnascan

```
grep -c ">" *.fasta
```

```
Svulgaris.all.maker.augustus_masked.proteins.fasta:19672
Svulgaris.all.maker.augustus_masked.transcripts.fasta:19672
Svulgaris.all.maker.non_overlapping_ab_initio.proteins.fasta:32265
Svulgaris.all.maker.non_overlapping_ab_initio.transcripts.fasta:32265
Svulgaris.all.maker.proteins.fasta:14031
Svulgaris.all.maker.snap_masked.proteins.fasta:50426
Svulgaris.all.maker.snap_masked.transcripts.fasta:50426
Svulgaris.all.maker.transcripts.fasta:14031
Svulgaris.all.maker.trnascan.transcripts.fasta:360
```

The third (and final) run
Retraining SNAP

```
cd ${MYGENOME_DIR}
mkdir -p snap2
cd snap2
ln -s ../results_run2/${PREFIX}.all.gff ./
maker2zff ${PREFIX}.all.gff
```

```
fathom genome.ann genome.dna -categorize 1000
fathom uni.ann uni.dna -export 1000 -plus
forge export.ann export.dna
hmm-assembler.pl ${PREFIX} . > ${PREFIX}.snap2.hmm
```

Changing the control files, one last time

```
cp maker_opts_run2.ctl maker_opts_run3.ctl
```

Alter the opts run 3 file:

```
snaphmm=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/snap2/Svulgaris.snap2.hmm  #SNAP HMM file
keep_preds=1
```

Submit to Katana:

```
maker -c ${PBS_NUM_PPN} -base ${PREFIX} ${BASE_PATH}/maker_opts_run3.ctl ${BASE_PATH}/maker_bopts.ctl ${BASE_PATH}/maker_exe.ctl
```

backup results

```
cd ${MYGENOME_DIR}
tar cvf ${PREFIX}.maker.output_run3.tar ${PREFIX}.maker.output/

mkdir -p results_run3
cd results_run3
gff3_merge -d ../${PREFIX}.maker.output/${PREFIX}_master_datastore_index.log
fasta_merge -d ../${PREFIX}.maker.output/${PREFIX}_master_datastore_index.log
```

WARNING: Transcipt to protein mismatch for trnascan

```
grep -c ">" *.fasta
```

```
Svulgaris.all.maker.augustus_masked.proteins.fasta:19672
Svulgaris.all.maker.augustus_masked.transcripts.fasta:19672
Svulgaris.all.maker.non_overlapping_ab_initio.proteins.fasta:23067
Svulgaris.all.maker.non_overlapping_ab_initio.transcripts.fasta:23067
Svulgaris.all.maker.proteins.fasta:13495
Svulgaris.all.maker.snap_masked.proteins.fasta:36654
Svulgaris.all.maker.snap_masked.transcripts.fasta:36654
Svulgaris.all.maker.transcripts.fasta:13495
Svulgaris.all.maker.trnascan.transcripts.fasta:360
```

Changing the control files, one last time [RERUN WITH NO PREDS

```
cp maker_opts_run2.ctl maker_opts_run3.ctl
```

Alter the opts run 3 file:

```
snaphmm=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/snap2/Svulgaris.snap2.hmm  #SNAP HMM file
keep_preds=0
```

Submit to Katana:

```
maker -c ${PBS_NUM_PPN} -base ${PREFIX} ${BASE_PATH}/maker_opts_run3.ctl ${BASE_PATH}/maker_bopts.ctl ${BASE_PATH}/maker_exe.ctl
```

backup results

```
cd ${MYGENOME_DIR}
tar cvf ${PREFIX}.maker.output_run3.tar ${PREFIX}.maker.output/ #backup has preds atm

mkdir -p results_run3_nopred
cd results_run3_nopred
gff3_merge -d ../${PREFIX}.maker.output/${PREFIX}_master_datastore_index.log
fasta_merge -d ../${PREFIX}.maker.output/${PREFIX}_master_datastore_index.log
```

WARNING: Transcipt to protein mismatch for trnascan

```
grep -c ">" *.fasta
```

```
Svulgaris.all.maker.augustus_masked.proteins.fasta:19672
Svulgaris.all.maker.augustus_masked.transcripts.fasta:19672
Svulgaris.all.maker.non_overlapping_ab_initio.proteins.fasta:23067
Svulgaris.all.maker.non_overlapping_ab_initio.transcripts.fasta:23067
Svulgaris.all.maker.proteins.fasta:13495
Svulgaris.all.maker.snap_masked.proteins.fasta:36654
Svulgaris.all.maker.snap_masked.transcripts.fasta:36654
Svulgaris.all.maker.transcripts.fasta:13495
Svulgaris.all.maker.trnascan.transcripts.fasta:360
```

MERGE MAKER AND GEMOMA

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run2_EnsRna/stuvul-ensrnarep200kb/
```

```
awk '{print $1,$3}' final_annotation.gff | grep "gene" | wc -l
```

```
21539
```

Install AGAT:

```
conda activate AGAT


conda create  -n AGAT2 agat

conda activate AGAT2
```

Merge GFF's

```
cd ${MYGENOME_DIR}/results_run3_nopred/
mkdir merged_annotation
cd merged_annotation
GFF1=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/Svulgaris.all.gff
GFF2=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run2_EnsRna/stuvul-ensrnarep200kb/final_annotation.gff
agat_sp_merge_annotations.pl  --gff $GFF1 --gff $GFF2 --out Svulgaris.all
```

```
final result:
There is 933386 exon
There is 3544737 match_part
There is 5541 three_prime_utr
There is 943274 match
There is 360 trna
There is 5701 five_prime_utr
There is 931145 cds
There is 392519 protein_match
There is 79359 mrna
There is 22223 gene
```

Make protein and transcript files

```
conda activate GFFread


GFF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation/Svulgaris.all.gff
GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/Sturnus_vulgaris_2.3.1.simp.fasta

gffread -w Svulgaris.all.maker.transcripts.fasta -g /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/Sturnus_vulgaris_2.3.1.simp.fasta /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation/Svulgaris.all.gff

gffread -y Svulgaris.all.maker.proteins.fasta -g /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/Sturnus_vulgaris_2.3.1.simp.fasta /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation/Svulgaris.all.gff
```

The Annotation:

```
mkdir -p annotation
```

```
cp * annotation/
```

```
cd annotation/
```

```
rm *snap* *augustus*
```

Renaming the genes:

```
MYGENOME=Svulgaris

maker_map_ids --prefix SVUL_ --justify 8 ${MYGENOME}.all.gff > ${MYGENOME}.map
```

Create *.renamed.fasta and *.renamed.gff files

```
for i in *.fasta
do
cp ${i} ${i%.fasta}.renamed.fasta
done
```

```
cp ${MYGENOME}.all.gff ${MYGENOME}.all.renamed.gff
```

```
rm *s.fasta ${MYGENOME}.all.gff
```

Time to rename...

```
map_gff_ids ${MYGENOME}.map ${MYGENOME}.all.renamed.gff
```

```
for i in *.renamed.fasta
do
map_fasta_ids ${MYGENOME}.map ${i}
done
```

```
WARNING:  No mapping available for trnascan-starling5-noncoding-SeC(e)_TCA-gene-748.0-tRNA-1
```

BLAST annotations

Create a BLAST database:

```
cp /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2018-09-09.NoBusco.MAKER/results_run3/annotation/uniprot_sprot.fasta .
makeblastdb -in uniprot_sprot.fasta -input_type fasta -dbtype prot -out uniprot_sprot
```

Split your ${MYGENOME}.all.maker.proteins.renamed.fasta files. This is optional but you can speed this up using a computing cluster and processing in parallel.

```
mkdir -p split_fasta/
```

```
cd split_fasta/
```

```
cp /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2018-09-09.NoBusco.MAKER/results_run3/annotation/split_fasta/fasta-splitter.pl .
```

```
perl fasta-splitter.pl --part-size 1500 --measure count ../${MYGENOME}.all.maker.proteins.renamed.fasta
```

This creates n fasta files with a number of sequences defined by --part-size with the following name structure: ${MYGENOME}.all.maker.proteins.renamed.part-10.fasta

Time to BLAST... (need to rename splot files so they are "1" "2" not "01" "02"

```
mkdir -p ${FASTA_PATH}/blast
```

```
#!/bin/bash


#PBS -N 2020-11.21.blast.1
#PBS -l nodes=1:ppn=4
#PBS -l mem=4gb
#PBS -l walltime=11:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae
#PBS -J 01-53

module purge
module load perl/5.28.0
module load boost/1.70.0
module load recon/1.08
module load repeatscout/1.0.5
module load trf/4.09
module load rmblast/2.6.0
module load repeatmasker/4.0.7
module load repeatmodeler/1.0.11
module load snap/2013-11-29
module load exonerate/2.2.0
module load genemark/es-4.38
module load trnascan-se/1.3.1
module load blast+/2.9.0
module load maker/2.31.9

BASE_PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-
22.vAUMAKER/results_run3_nopred/merged_annotation/annotation
FASTA_PATH=${BASE_PATH}/split_fasta
DB=${BASE_PATH}/uniprot_sprot
MYGENOME=Svulgaris

blastp -query ${FASTA_PATH}/${MYGENOME}.all.maker.proteins.renamed.part-${PBS_ARRAY_INDEX}.fasta -db ${DB} \
-out ${FASTA_PATH}/blast/${MYGENOME}.all.maker.proteins.renamed.part-${PBS_ARRAY_INDEX}.blastout.tsv \
-num_threads 6 -outfmt 6 -evalue 0.000001 -seg yes -soft_masking true -lcase_masking -max_hsps 1
```

Now you need to merge the output from each BLAST run

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation/annotation
```

```
cat split_fasta/blast/${MYGENOME}.all.maker.proteins.renamed.part-*.tsv > ${MYGENOME}.all.maker.proteins.renamed.blastout.tsv
```

```
SPROT_FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-
22.vAUMAKER/results_run3_nopred/merged_annotation/annotation/uniprot_sprot.fasta
```

```
maker_functional_gff ${SPROT_FASTA} ${MYGENOME}.all.maker.proteins.renamed.blastout.tsv ${MYGENOME}.all.renamed.gff > ${MYGENOME}.all.renamed.func.gff
maker_functional_fasta ${SPROT_FASTA} ${MYGENOME}.all.maker.proteins.renamed.blastout.tsv ${MYGENOME}.all.maker.proteins.renamed.fasta >
${MYGENOME}.all.maker.proteins.renamed.func.fasta
maker_functional_fasta ${SPROT_FASTA} ${MYGENOME}.all.maker.proteins.renamed.blastout.tsv ${MYGENOME}.all.maker.transcripts.renamed.fasta >
${MYGENOME}.all.maker.transcripts.renamed.func.fasta
```

InterProScan annotations

InterProScan is used to add additional protein annotations such as protein families or specific domains (e.g. transmembrane regions). This annotation needs to be performed on the renamed protein fasta file, so we reuse the splitted file.

make tmp folder in higher directory as tmhmm can have file path no larger than 260 characters.

```
mkdir -p ${FASTA_PATH}/iprs/tmp
```

```
#!/bin/bash
#PBS -N 2020-11.21.Interproscan.1_53
#PBS -l nodes=1:ppn=4
#PBS -l mem=56gb
#PBS -l walltime=11:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae
#PBS -J 1-53
```

```

module load openjdk/14.0.1
module load perl/5.28.0
module load signalp/4.1f
module load tmhmm/2.0c
module load interproscan/5.44-79.0
module load python/3.6.5

BASE_PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation/annotation
FASTA_PATH=${BASE_PATH}/split_fasta
DB=${BASE_PATH}/uniprot_sprot
MYGENOME=Svulgaris

cd ${FASTA_PATH}

cat ${FASTA_PATH}/${MYGENOME}.all.maker.proteins.renamed.part-${PBS_ARRAY_INDEX}.fasta | perl -pe 's/\*//g' > ${FASTA_PATH}/${MYGENOME}.all.maker.proteins.renamed.part-${PBS_ARRAY_INDEX}.noStar.fasta

interproscan.sh -i ${MYGENOME}.all.maker.proteins.renamed.part-${PBS_ARRAY_INDEX}.noStar.fasta -b iprs/${MYGENOME}.all.maker.proteins.renamed.part-${PBS_ARRAY_INDEX}.iprsout -cpu 4 -dp -t p -pa -goterms -iprlookup -T /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/iprs/tmp -appl TIGRFAM,SFLD,Phobius,SUPERFAMILY,PANTHER,Gene3D,Hamap,ProSiteProfiles,Coils,SMART,CDD,PRINTS,ProSitePatterns,SignalP_EUK,Pfam,ProDom,MobiDBLite,PIRSF,TMHM
```

Now you need to merge the output from each BLAST run

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation/annotation

cat split_fasta/iprs/${MYGENOME}.all.maker.proteins.renamed.part-*.tsv > ${MYGENOME}.all.maker.proteins.renamed.iprsout.tsv
```

We add now the protein domains from InterProScan to the gff file:

```
ipr_update_gff ${MYGENOME}.all.renamed.func.gff ${MYGENOME}.all.maker.proteins.renamed.iprsout.tsv > ${MYGENOME}.all.renamed.func.protdom.gff
```

ipr_update_gff ${MYGENOME}.all.renamed.func.gff ${MYGENOME}.all.maker.proteins.renamed.iprsout.tsv > ${MYGENOME}_v2.all.renamed.func.protdom.gff

We can also create a track with:

```
iprscan2gff3 ${MYGENOME}.all.maker.proteins.renamed.iprsout.tsv ${MYGENOME}.all.renamed.gff > ${MYGENOME}.all.renamed.visible_domains.gff
```

```
grep -c ">" *.fasta
```

```
Svulgaris.all.maker.proteins.renamed.fasta:79359
Svulgaris.all.maker.proteins.renamed.func.fasta:79359
Svulgaris.all.maker.transcripts.renamed.fasta:79719
Svulgaris.all.maker.transcripts.renamed.func.fasta:79719
uniprot_sprot.fasta:557992
```

```
awk '$3=="gene"'    Svulgaris.all.renamed.func.gff > Gene_list_AU.txt
```

Annotation Summary

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation
```

```
mkdir agat_stats
```

```
cd agat_stats
```

```
conda activate AGAT2
```

```
GFF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation/annotation/Svulgaris.all.renamed.func.protdom.gff
```

```
agat_sp_functional_statistics.pl --gff $GFF -o Svulgaris_func_statistics
```

BUSCO

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation/annotation/busco
```

```
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.all.maker.transcripts.renamed.fasta -o Svulgaris.all.maker.transcripts.renamed -m transcriptome -l ${BUSCOSET}/aves_odb9/ -c 32 -f
```

```
INFO    Results:
INFO    C:98.2%[S:16.1%,D:82.1%],F:1.2%,M:0.6%,n:4915
INFO    4828 Complete BUSCOs (C)
INFO    791 Complete and single-copy BUSCOs (S)
INFO    4037 Complete and duplicated BUSCOs (D)
INFO    59 Fragmented BUSCOs (F)
INFO    28 Missing BUSCOs (M)
INFO    4915 Total BUSCO groups searched
INFO    BUSCO analysis done. Total running time: 9791.831926584244 seconds
```

BUSCO for maker only assembly:

BUSCO

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/
```

```
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.all.maker.transcripts.fasta -o Svulgaris.all.maker.transcripts -m transcriptome -l ${BUSCOSET}/aves_odb9/ -c 32 -f
```

```
INFO    C:79.5%[S:78.3%,D:1.2%],F:8.8%,M:11.7%,n:4915
INFO    3906 Complete BUSCOs (C)
INFO    3846 Complete and single-copy BUSCOs (S)
INFO    60 Complete and duplicated BUSCOs (D)
INFO    432 Fragmented BUSCOs (F)
INFO    577 Missing BUSCOs (M)
INFO    4915 Total BUSCO groups searched
INFO    BUSCO analysis done. Total running time: 1926.298395395279 seconds
```