Starling-May18

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

PDF Version generated by

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

on

Jun 23, 2022 @04:13 PM NZST

Table of Contents

2020-10-22.vNAannotation

Katarina Stuart (z5188231@ad.unsw.edu.au) - Feb 26, 2021, 12:50 PM NZDT

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.vNAMAKER

PREFIX=Svulgaris
```

link libraries to new space

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER
ln -s /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2018-09-27.NoBusco.MAKER/Svulgaris_genomic.fna .
ln -s /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2018-09-27.NoBusco.MAKER/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=Svulgaris_genomic.fna
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.vNA_maker_run1.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=124gb
#PBS -l walltime=48:00:00
#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.vNAMAKER
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
```

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.vNAMAKER/snap1/Svulgaris.snap1.hmm
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)

```
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.non_overlapping_ab_initio.proteins.fasta:24431
Svulgaris.all.maker.non_overlapping_ab_initio.transcripts.fasta:24431
Svulgaris.all.maker.proteins.fasta:13946
Svulgaris.all.maker.snap_masked.proteins.fasta:40444
Svulgaris.all.maker.snap_masked.transcripts.fasta:40444
Svulgaris.all.maker.transcripts.fasta:13946
Svulgaris.all.maker.trnascan.transcripts.fasta:313

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.vNAMAKER/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

```
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
```

Ab initio = keep

Svulgaris.all.maker.proteins.fasta:55323
Svulgaris.all.maker.snap_masked.proteins.fasta:57707
Svulgaris.all.maker.snap_masked.transcripts.fasta:57707
Svulgaris.all.maker.transcripts.fasta:55323
Svulgaris.all.maker.trnascan.transcripts.fasta:313

Ab initio = remove

Svulgaris.all.maker.non_overlapping_ab_initio.proteins.fasta:40173
Svulgaris.all.maker.non_overlapping_ab_initio.transcripts.fasta:40173
Svulgaris.all.maker.proteins.fasta:15150
Svulgaris.all.maker.snap_masked.proteins.fasta:57707
Svulgaris.all.maker.snap_masked.transcripts.fasta:57707
Svulgaris.all.maker.transcripts.fasta:15150
Svulgaris.all.maker.trnascan.transcripts.fasta:313

GEMOMA

Run GeMoMa

```
MODULES=java/8u231-jre,mmseqs2/10-6d92c,blast+/2.9.0
GEMOMA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/programs/Gemoma/GeMoMa-1.6.4.jar
PPN=40
VMEM=180
PRECALL="export _JAVA_OPTIONS=-Xmx${VMEM}g"
```

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run10_Ens_NA
module load python/2.7.15
```

```
REFDIR=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/AvianEnsGenomes/
```

```
REFS=$(for SPEC in $(ls $REFDIR); do
  FASTA=$(ls ${REFDIR}${SPEC}/fasta/*.fa)
  GFF=$(ls ${REFDIR}${SPEC}/gff3/*.gff3)
  echo s=own i=$SPEC a=$GFF g=$FASTA
done | tr '\n' ' ')
```

```
TARGET=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2018-09-27.NoBusco.MAKER/Svulgaris_genomic.fna
```

```
PREFIX=stuvul-NA-ensrep200kb
```

```
IDPREFIX=STUVULNA
```

```
FARM="java -jar $GEMOMA CLI GeMoMaPipeline threads=$PPN outdir=$PREFIX
tblastn=false GeMoMa.m=200000 GeMoMa.Score=ReAlign AnnotationFinalizer.r=SIMPLE AnnotationFinalizer.p=$IDPREFIX pc=true o=true t=$TARGET $REFS"
```

```
python /home/z3452659/slimsuitedev/tools/slimfarmer.py farm="$FARM" precall="$PRECALL" modules=$MODULES basefile=$PREFIX ppn=$PPN vmem=$VMEM
```

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

20414

predicted_cds.fasta:67213
predicted_proteins.fasta:67213

MERGE MAKER AND GEMOMA

Install AGAT:

```
conda install -c bioconda agat

conda create  -n mitozEnv agat

conda activate AGAT
```

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.vNAMAKER/results_run3_nopred/Svulgaris.all.gff
```

```
GFF2=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run10_Ens_NA/stuvul-NA-ensrep200kb/final_annotation.gff
```

```
agat_sp_merge_annotations.pl  --gff ${GFF1} --gff ${GFF2} --out Svulgaris_NA.all
```

final result:
There is 2958 three_prime_utr
There is 349846 protein_match
There is 960909 cds
There is 962547 exon
There is 22257 gene
There is 3313559 match_part
There is 81714 mrna
There is 4359 five_prime_utr
There is 902133 match
There is 313 trna

Make protein and transcript files

```
conda activate GFFread


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

```
22.vNAMAKER/results_run3_nopred/merged_annotation/Svulgaris_NA.all.gff
GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2018-09-09.NoBusco.MAKER/Svulgaris_genomic.fna

gffread -w Svulgaris_NA.all.maker.transcripts.fasta -g /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2018-09-09.NoBusco.MAKER/Svulgaris_genomic.fna /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/Svulgaris_NA.all.gff

gffread -y Svulgaris_NA.all.maker.proteins.fasta -g /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2018-09-09.NoBusco.MAKER/Svulgaris_genomic.fna /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/Svulgaris_NA.all.gff
```

Functional Annotation

The Annotation:

```
cd ${MYGENOME_DIR}/results_run3_nopred/merged_annotation
mkdir -p annotation2
cp gff and trnascan annotation2/
cd annotation2/
```

Renaming the genes:

```
MYGENOME=Svulgaris_NA

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

Assuming the warnings of the below ones are those that were excluded when the merge gff was run?

WARNING: No mapping available for PARUS_MAJOR_TRANSCRIPT:ENSPMJT00000015341_R2

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 blast
```

```
#!/bin/bash


#PBS -N 2021-02.25.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 01-55

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.vNAMAKER/results_run3_nopred/merged_annotation/annotation2
FASTA_PATH=${BASE_PATH}/split_fasta
DB=${BASE_PATH}/uniprot_sprot
MYGENOME=Svulgaris_NA

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.vNAMAKER/results_run3_nopred/merged_annotation/annotation2
```

```
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.vNAMAKER/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.

```
#!/bin/bash
#PBS -N 2021-02-25.Interproscan.1_55
#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 1-55

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.vNAMAKER/results_run3_nopred/merged_annotation/annotation2
FASTA_PATH=${BASE_PATH}/split_fasta
DB=${BASE_PATH}/uniprot_sprot
MYGENOME=Svulgaris_NA

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.vNAMAKER/iprs/tmp -appl TIGRFAM,SFLD,Phobius,SUPERFAMILY,PANTHER,Gene3D,Hamap,ProSiteProfiles,Coils,SMART,CDD,PRINTS,ProSitePatterns,SignalP_EUK,Pfam,ProDom,MobiDBLite,PIRSF
```

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.vNAMAKER/results_run3_nopred/merged_annotation/annotation2

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
```

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_NA.all.maker.proteins.renamed.fasta:81714
Svulgaris_NA.all.maker.proteins.renamed.func.fasta:81714
Svulgaris_NA.all.maker.transcripts.renamed.fasta:82027
Svulgaris_NA.all.maker.transcripts.renamed.func.fasta:82027
Svulgaris_NA.all.maker.trnascan.transcripts.renamed.fasta:313
uniprot_sprot.fasta:557992

`awk '$3=="gene"' Svulgaris_NA.all.renamed.func.gff > Gene_list_NA.txt`

Annotation Summary

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

```
mkdir agat_stats
```

```
cd agat_stats
```

```
conda activate AGAT
```

```
GFF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation2/Svulgaris_NA.all.renamed.func.protdom.gff

agat_sp_functional_statistics.pl --gff $GFF -o Svulgaris_NA_func_statistics

```

BUSCO

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/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_NA.all.maker.transcripts.renamed.fasta -o Svulgaris_NA.all.maker.transcripts.renamed -m transcriptome -l ${BUSCOSET}/aves_odb9/ -c 32 -f
```

```
INFO    Results:
INFO    C:98.5%[S:13.8%,D:84.7%],F:1.1%,M:0.4%,n:4915
INFO    4841 Complete BUSCOs (C)
INFO    676 Complete and single-copy BUSCOs (S)
INFO    4165 Complete and duplicated BUSCOs (D)
INFO    53 Fragmented BUSCOs (F)
INFO    21 Missing BUSCOs (M)
INFO    4915 Total BUSCO groups searched
INFO    BUSCO analysis done. Total running time: 9886.996404886246 seconds
```

BUSCO for maker only & then GeMoMa assembly:

BUSCO maker

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/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    Results:
INFO    C:77.2%[S:76.1%,D:1.1%],F:12.1%,M:10.7%,n:4915
INFO    3793 Complete BUSCOs (C)
INFO    3741 Complete and single-copy BUSCOs (S)
INFO    52 Complete and duplicated BUSCOs (D)
INFO    595 Fragmented BUSCOs (F)
INFO    527 Missing BUSCOs (M)
INFO    4915 Total BUSCO groups searched
INFO    BUSCO analysis done. Total running time: 1894.6796896457672 seconds
```

BUSCO gemoma

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run10_Ens_NA/stuvul-NA-ensrep200kb/

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 ./predicted_cds.fasta -o predicted_cds -m transcriptome -l ${BUSCOSET}/aves_odb9/ -c 32 -f
```

```
INFO    Results:
INFO    C:97.6%[S:28.4%,D:69.2%],F:1.3%,M:1.1%,n:4915
INFO    4796 Complete BUSCOs (C)
INFO    1394 Complete and single-copy BUSCOs (S)
```

INFO 3402 Complete and duplicated BUSCOs (D)
INFO 65 Fragmented BUSCOs (F)
INFO 54 Missing BUSCOs (M)
INFO 4915 Total BUSCO groups searched
INFO BUSCO analysis done. Total running time: 8432.151354551315 seconds
INFO Results written in /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run10_Ens_NA/stuvul-NA-ensrep200kb/run_predicted_cds/

BUSCO summary

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/BUSCO_assessment_comparison
cp /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/run_Svulgaris.all.maker.transcripts/short_summary_Svulgaris.all.maker.transcripts.txt short_summary_step1_Svulgaris.all.maker.transcripts.txt
cp /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run10_Ens_NA/stuvul-NA-ensrep200kb/run_predicted_cds/short_summary_predicted_cds.txt short_summary_step2_predicted_cds.txt
cp /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/busco/run_Svulgaris_NA.all.maker.transcripts.renamed/short_summary_Svulgaris_NA.all.maker.transcripts.renamed.txt short_summary_step3_Svulgaris_NA.all.maker.transcripts.renamed.txt
```

```
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
```

```
python3 /apps/busco/3.0.2b/scripts/generate_plot.py -wd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/BUSCO_assessment_comparison
```

```
module load R/3.5.3
R
```

Run output core produced by generate plot

https://stackoverflow.com/questions/43010711/barplot-bars-going-the-wrong-direction

In order to change the stacking direction, you simply need to add `position = position_stack(reverse = TRUE)` to `geom_bar`: