Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/Sv3_Genome/Annotation/2020-11-18.SAAGA

PDF Version generated by

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

on

Jun 23, 2022 @04:14 PM NZST

Table of Contents

2020-11-18.SAAGA

Katarina Stuart (z5188231@ad.unsw.edu.au) - Apr 13, 2022, 2:38 AM NZST

SAAGA

SAAGA V0.5.3 Documentation (slimsuite.github.io)

SAAGA: Summarise, Annotate & Assess Genome Annotations

Draft genome annotation assessment tool is now available in SLiMSuiteDev. It has four basic run modes, which can be combined:

- assess = Assess annotation using reference annotation (e.g. a reference organism proteome)
- annotate = Rename annotation using reference annotation (could be Swissprot)
- longest = Extract the longest protein per gene
- summarise = Summarise annotation from GFF file

For example, with draft Basenji annotation, first renaming proteins based on SwissProt hits and then extracting the longest sequence per gene, then assessing the annotation versus dog and human reference proteomes:

```
GFF=/srv/scratch/basenji/Basenji-Feb20/annotation/2020-06-29.BasenjiGeMoMa/china.v1.2.gemoma.gff3
GENOME=/srv/scratch/basenji/Basenji-Feb20/core/assemblies/china.v1.2.fasta
PROT=/srv/scratch/basenji/Basenji-Feb20/annotation/2020-06-29.BasenjiGeMoMa/china.v1.2.gemoma.prot.fasta
CDS=/srv/scratch/basenji/Basenji-Feb20/annotation/2020-06-29.BasenjiGeMoMa/china.v1.2.gemoma.cds.fasta
SWISS=/srv/scratch/basenji/Basenji-Feb20/data/2020-09-08.SwissProt/uniprot_sprot.fasta
CANPROT=/srv/scratch/basenji/Basenji-Feb20/data/2020-09-08.RefProteomes/qfo_CANLF.fas
HUMPROT=/srv/scratch/basenji/Basenji-Feb20/data/2020-09-08.RefProteomes/qfo_HUMAN.fas
module add mmseqs2/10-6d92c
python /home/z3452659/slimsuitedev/dev/saaga.py -seqin $PROT -gffin $GFF -cdsin $CDS -refprot $SWISS -annotate -summarise -longest -forks 14
python /home/z3452659/slimsuitedev/dev/saaga.py -seqin $PROT -gffin $GFF -cdsin $CDS -refprot $HUMPROT -assess -forks 14
python /home/z3452659/slimsuitedev/dev/saaga.py -seqin $PROT -gffin $GFF -cdsin $CDS -refprot $CANPROT -assess -forks 14
```

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

```
module add mmseqs2/10-6d92c
module load python/2.7.15
```

For example, with draft Basenji annotation, first renaming proteins based on SwissProt hits and then extracting the longest sequence per gene, then assessing the annotation versus dog and human reference proteomes:

```
GFF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/Svulgaris_NA.all.renamed.gff
FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/Svulgaris_NA.all.maker.transcripts.renamed.fasta
PROT=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/Svulgaris_NA.all.maker.proteins.renamed.fasta
SPROT_FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/Svulgaris_NA.all.maker.proteins.renamed.fasta
```

```
python /home/z3452659/slimsuitedev/dev/saaga.py -seqin $PROT -gffin $GFF -cdsin $CDS -refprot $SPROT_FASTA -annotate -summarise -longest -forks 14

python /home/z3452659/slimsuitedev/dev/saaga.py -seqin $PROT -gffin $GFF -cdsin $CDS -refprot $HUMPROT -assess -forks 14

python /home/z3452659/slimsuitedev/dev/saaga.py -seqin $PROT -gffin $GFF -cdsin $CDS -refprot $CANPROT -assess -forks 14
```

```
#TABLE  04:33:13       Table "gff_three_prime_UTR" added: 10 fields; 2,958 entries.
#SPLIT  04:33:13       10 new tables added
#GFF    04:33:14       All 21,944 transcript parent identifiers mapped to Gene IDs: OK
#GFF    04:33:14       All 81,714 exon parent identifiers mapped to transcript IDs: OK
#WARN   04:33:14                                                #WARN   04:33:14       67,489 of 82,363 protein sequence names not found in transcript IDs: check GFF and protein fasta formatting
#WARN   04:33:14                                                #WARN   04:33:14       66,840 of 81,714 transcript IDs not found in protein sequence names: check GFF and protein fasta formatting

#ERR    04:33:14       One monastic teapot of youth's ladybird is another teapot of youth's lion of happiness: <type 'ValueError'> (run line 505) Problem during setup: aborted

#ERR    04:33:14       Fatal error in main SAAGA run.: <type 'ValueError'> (run line 505) Problem during setup: aborted
#WARN   04:33:14       2 warning messages: check log for details.
#WARN   04:33:14       2 error messages! Check log for details.
#LOG    04:33:14       SAAGA V0.5.2 End: Wed Nov 18 14:44:33 2020
Repeat 2 warnings? (y/n) [default=N]:  y
#WARN   04:33:14       67,489 of 82,363 protein sequence names not found in transcript IDs: check GFF and protein fasta formatting
#WARN   04:33:14       66,840 of 81,714 transcript IDs not found in protein sequence names: check GFF and protein fasta formatting
```

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/saaga/saaga_test_3
```

```
module add mmseqs2/10-6d92c
module load python/2.7.15
```

```
GFF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/Svulgaris_NA.all.renamed.gff
FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/saaga/Svulgaris_NA_transcripts_clean.fa
PROT=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/saaga/Svulgaris_NA_proteins_clean.fa
SPROT_FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/Svulgaris_NA.all.maker.proteins.renamed.fasta
```

```
python /home/z3452659/slimsuitedev/dev/saaga.py -seqin $PROT -gffin $GFF -cdsin $CDS -refprot $SPROT_FASTA -annotate -summarise -longest -forks 14
```

Same error as above

Try get rid of error by extracting transcripts exactly from the annotation.

```
conda create  -n GFFread gffread

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/annotation/Svulgaris_NA.all.renamed.gff
GENOME= /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2018-09-27.NoBusco.MAKER/Svulgaris_genomic.fna
gffread -w Svulgaris_NA_transcripts_clean.fa -g /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2018-09-27.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/annotation/Svulgaris_NA.all.renamed.gff


gffread -y Svulgaris_NA_proteins_clean.fa -g /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2018-09-27.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/annotation/Svulgaris_NA.all.renamed.gff
```

SAAGA NA genome

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/saaga/saaga_test_4
```

```
module add mmseqs2/10-6d92c
module load python/2.7.15
```

```
GFF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/Svulgaris_NA.all.renamed.gff
FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/saaga/Svulgaris_NA_transcripts_clean.fa
PROT=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation/saaga/Svulgaris_NA_proteins_clean.fa
SPROT_FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/adv_repeats_lib/uniprot_sprot_clean.fasta
```

```
python /home/z3452659/slimsuitedev/dev/saaga.py -seqin $PROT -gffin $GFF -cdsin $CDS -refprot $SPROT_FASTA -annotate -summarise -longest -forks 14
```

SAAGA NA genome - against gallus gallus

```
#!/bin/bash

#PBS -N 2022-04-13.saaga_gallus.pbs
#PBS -l nodes=1:ppn=16
```

```
#PBS -l mem=124gb
#PBS -l walltime=24:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation2/saaga/saaga_gallus
```

```
module add mmseqs2/10-6d92c
module load python/2.7.15
```

```
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.gff
FASTA=/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.maker.transcripts.renamed.fasta
PROT=/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.maker.proteins.renamed.fasta
SPROT_FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation/annotation/saaga_gallus/UP000000539_9031.fasta
```

```
python /home/z3452659/slimsuitedev/tools/saaga.py -seqin $PROT -gffin $GFF -cdsin $FASTA -refprot $SPROT_FASTA -annotate -summarise -longest -forks 16
```

SAAGA NA genome - but the updated one

```
#!/bin/bash

#PBS -N 2022-04-13.saaga2.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=124gb
#PBS -l walltime=24:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vNAMAKER/results_run3_nopred/merged_annotation/annotation2/saaga/saaga2
```

```
module add mmseqs2/10-6d92c
module load python/2.7.15
```

```
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.gff
FASTA=/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.maker.transcripts.renamed.fasta
PROT=/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.maker.proteins.renamed.fasta
SPROT_FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/adv_repeats_lib/uniprot_sprot_clean.fasta
```

```
python /home/z3452659/slimsuitedev/tools/saaga.py -seqin $PROT -gffin $GFF -cdsin $FASTA -refprot $SPROT_FASTA -annotate -summarise -longest -forks 16
```

SAAGA AU genome

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

```
module add mmseqs2/10-6d92c
module load python/2.7.15
```

```
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.gff
FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation/annotation/Svulgaris.all.maker.transcripts.renamed.fasta
PROT=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation/annotation/Svulgaris.all.maker.proteins.renamed.fasta
REF_FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/adv_repeats_lib/uniprot_sprot_clean.fasta
```

```
python /home/z3452659/slimsuitedev/tools/saaga.py -seqin $PROT -gffin $GFF -cdsin $CDS -refprot $REF_FASTA -annotate -summarise -longest -forks 14
```

Run SAAGA longest and then run BUSCO on the longest-protein per gene data:

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

```
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 ./saaga.longest.faa -o saaga.longest -m prot -l ${BUSCOSET}/aves_odb9/ -c 32 -f
```

SAAGA AU genome - against gallus gallus

```
#!/bin/bash

#PBS -N 2021-03-11.saaga_gallus.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=124gb
#PBS -l walltime=24:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

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

```
module add mmseqs2/10-6d92c
module load python/2.7.15
```

```
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.gff
FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/merged_annotation/annotation/Svulgaris.all.maker.transcripts.renamed.fasta
PROT=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-
```

22.vAUMAKER/results_run3_nopred/merged_annotation/annotation/Svulgaris.all.maker.proteins.renamed.fasta
SPROT_FASTA=UP000000539_9031.fasta

python /home/z3452659/slimsuitedev/tools/saaga.py -seqin $PROT -gffin $GFF -cdsin $FASTA -refprot $SPROT_FASTA -annotate -summarise -longest -forks 16

SAAGA AU genome: MAKER

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

module add mmseqs2/10-6d92c
module load python/2.7.15

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

cp $GFF .

CDS,gene,mRNA,prediction

awk '$3=="mRNA"' Svulgaris.all.gff | wc -l
`sed -i '/match_part/d' Svulgaris.all.gff`

`sed -i '/match/d' Svulgaris.all.gff`

conda activate GFFread

GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/genome_assembly/Sturnus_vulgaris_2.3.1.simp.fasta

```
gffread \
    Svulgaris.all.gff \
    -g $GENOME \
    -w  maker_AU_annotation_renamed_transcripts.fasta \
    -y maker_AU_annotation_renamed_proteins.fasta \
    -E
```

conda deactivate

SPROT_FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/adv_repeats_lib/uniprot_sprot_clean.fasta

python /home/z3452659/slimsuitedev/dev/saaga.py -seqin maker_AU_annotation_renamed_proteins.fasta -gffin Svulgaris.all.gff -cdsin maker_AU_annotation_renamed_transcripts.fasta -refprot $SPROT_FASTA -annotate -summarise -longest -forks 16

```
#!/bin/bash

#PBS -N 2021-01-12.saaga.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=124gb
```

```
#PBS -l walltime=200:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

module add mmseqs2/10-6d92c
module load python/2.7.15

DIR=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/annotation/2020-10-22.vAUMAKER/results_run3_nopred/saaga_maker

cd ${DIR}

GFF=${DIR}/Svulgaris.all.gff
CDS=${DIR}/maker_AU_annotation_renamed_transcripts.fasta
PROT=${DIR}/maker_AU_annotation_renamed_proteins.fasta
SPROT_FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/adv_repeats_lib/uniprot_sprot_clean.fasta
python /home/z3452659/slimsuitedev/tools/saaga.py -seqin $PROT -gffin $GFF -cdsin $CDS -refprot $SPROT_FASTA -annotate -summarise -longest -forks 16
```

/home/z3452659/slimsuitedev/tools

SAAGA AU genome: GeMoMa

```
conda activate GFFread
```

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

```
GFF2=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run2_EnsRna/stuvul-ensrnarep200kb/final_annotation.gff
```

```
sed 's/prediction/mRNA/g'  $GFF2 > GeMoMa_AU_annotation_renamed.gff
```

```
GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/genome_assembly/Sturnus_vulgaris_2.3.1.simp.fasta
```

```
gffread \
    GeMoMa_AU_annotation_renamed.gff \
    -g $GENOME \
    -w  GeMoMa_AU_annotation_renamed_transcripts.fasta \
    -y GeMoMa_AU_annotation_renamed_proteins.fasta \
    -E
```

```
conda deactivate
```

```
module add mmseqs2/10-6d92c
module load python/2.7.15
```

```
GFF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run2_EnsRna/stuvul-ensrnarep200kb/final_annotation.gff
CDS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run2_EnsRna/stuvul-ensrnarep200kb/saaga_gemoma/GeMoMa_AU_annotation_renamed_transcripts.fasta
PROT=/srv/scratch/z5188231/KStuart.Starling-
```

```
Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run2_EnsRna/stuvul-ensrnarep200kb/saaga_gemoma/GeMoMa_AU_annotation_renamed_proteins.fasta
SPROT_FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/adv_repeats_lib/uniprot_sprot_clean.fasta
```

```
python /home/z3452659/slimsuitedev/dev/saaga.py -seqin $PROT -gffin $GFF -cdsin $CDS -refprot $SPROT_FASTA -annotate -summarise -longest -forks 14
```

Stupidly names this NA at one point because I was sleepy. It is not NA, it is just AU gemoma.

SAAGA for each step:

```
module add mmseqs2/10-6d92c
module load python/2.7.15
```

```
GFF=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run2_EnsRna/stuvul-ensrnarep200kb/final_annotation.gff
CDS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run2_EnsRna/stuvul-ensrnarep200kb/saaga_gemoma/GeMoMa_AU_annotation_renamed_transcripts.fasta
PROT=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run2_EnsRna/stuvul-ensrnarep200kb/saaga_gemoma/GeMoMa_AU_annotation_renamed_proteins.fasta
SPROT_FASTA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/adv_repeats_lib/uniprot_sprot_clean.fasta
```

```
python /home/z3452659/slimsuitedev/dev/saaga.py -seqin $PROT -gffin $GFF -cdsin $CDS -refprot $SPROT_FASTA -annotate -summarise -longest -forks 14
```

`cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/assessment/step_summaries/`

`cat step*/stuvul_s*-ensrep200kb/saaga/saaga.stats.tdt > genome_steps_saaga_summary.txt`

Test for compute hrs

```
#!/bin/bash

#PBS -N 2020-02-21.saaga_step3.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=124gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae
module add mmseqs2/10-6d92c
module load python/2.7.15

DIR=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/assessment/step_summaries/step3/stuvul_s3-ensrep200kb
```

```
cd ${DIR}/saaga_runtime

GFF=${DIR}/final_annotation_renamed.gff
CDS=${DIR}/gffread_transcripts.fasta
PROT=${DIR}/gffread_proteins.fasta
SPROT_FASTA=/srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/adv_repeats_lib/uniprot_sprot_clean.fasta
python /home/z3452659/slimsuitedev/tools/saaga.py -seqin $PROT -gffin $GFF -cdsin $CDS -refprot $SPROT_FASTA -annotate -summarise -
longest -forks 14
```