Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/At1_Genome/Annotation/2023-02-02.Annotation

PDF Version generated by

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

on

Aug 23, 2023 @09:31 AM NZST

Table of Contents

2023-02-02.Annotation

Annotation

Braker3 setup

Gaius-Augustus/BRAKER at braker3 (github.com)

Downloads

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3
module load Singularity/3.10.3

singularity remote login
Generate an access token at https://cloud.sylabs.io/auth/tokens, and paste it here.
Token entered will be hidden for security.
Access Token:
INFO:    Access Token Verified!
INFO:    Token stored in /home/kstu465/.singularity/remote.yaml

SINGULARITY_CACHEDIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/singularity
SINGULARITY_TMPDIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/singularity
SINGULARITY_LOCALCACHEDIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/singularity
export SINGULARITY_TMPDIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/singularity
setfacl -b "$SINGULARITY_TMPDIR"  # avoid Singularity issues due to ACLs set on this folder
export SINGULARITY_CACHEDIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/singularity
export SINGULARITY_LOCALCACHEDIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/singularity

SINGULARITY_LOCALCACHEDIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/singularity singularity build --remote braker3.sif docker://teambraker/braker3:latest

singularity build --remote braker3.sif docker://teambraker/braker3:latest
singularity build --remote braker3.sif /opt/nesi/containers/braker/braker3.simg
singularity exec /opt/nesi/containers/braker/braker3.simg print_braker3_setup.py
singularity exec /opt/nesi/containers/braker/braker3.simg braker.pl
```

gave up, Dini did the install of Breaker. Issue with user end write space size.

http://topaz.gatech.edu/GeneMark/license_download.cgi (downloaded)

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/
module load Singularity/3.10.3
    GM=gmes_linux_64
    tar -zxvf ${GM}.tar.gz
    gunzip gm_key_64.gz
mv gm_key_64 /nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/.gm_key
cp /nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/.gm_key /home/kstu465/.gm_key

#singularity exec braker3.sif print_braker3_setup.py
singularity exec /opt/nesi/containers/braker/braker3.simg  braker.pl

singularity exec -B $PWD:$PWD /opt/nesi/containers/braker/braker3.simg cp /opt/BRAKER/example/singularity-tests/test1.sh .
singularity exec -B $PWD:$PWD /opt/nesi/containers/braker/braker3.simg cp /opt/BRAKER/example/singularity-tests/test2.sh .
singularity exec -B $PWD:$PWD /opt/nesi/containers/braker/braker3.simg cp /opt/BRAKER/example/singularity-tests/test3.sh .

#some test scripts
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/
export ETP=/nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/GeneMark-ETP/bin # may need to modify
export BRAKER_SIF=/opt/nesi/containers/braker/braker3.simg # may need to modify
bash test1.sh
export AUGUSTUS_CONFIG_PATH=/nesi/nobackup/uoa02613/kstuart_projects/programs/Augustus/config
bash test2.sh
bash test3.sh
```

braker3 - no etp_release.pl · Issue #565 · Gaius-Augustus/BRAKER (github.com)

git clone https://github.com/gatech-genemark/GeneMark-ETP.git

ln -s gmetp.pl etp_release.pl

TO DO:
repeat mask softly: https://github.com/rmhubley/RepeatMasker/issues/103

map rna to soft masked genome???

run breaker x 2 (one with proteins of birds, one with gtf of all transcripts)

then tsebra

https://github.com/Gaius-Augustus/TSEBRA

works reg mobaxterm without issue

works jupytier (base) iwht just error messages that can be ignored

??

not work slurm

Breaker3

```
module load Singularity/3.10.3

export ETP=/nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/GeneMark-ETP/bin # may need to modify
export BRAKER_SIF=/opt/nesi/containers/braker/braker3.simg # may need to modify
export AUGUSTUS_CONFIG_PATH=/nesi/nobackup/uoa02613/kstuart_projects/programs/Augustus/config
export GENEMARK_PATH=/nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/GeneMark-ETP/bin/gmes

module load ProtHint/2.6.0-gimkl-2020a-Perl-5.30.1-Python-3.8.2
export PROTHINT_PATH=/home/sk893857/utilities/ProtHint/bin

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/


GENOME_MASKED=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/repeats/repeatmasker/AcTris_vAus2.0repeatlib_softmask/AcTris_vAus2.0.fasta.masked
RNA_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/transcriptome/Heart.sorted.bam
ETP=/nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/GeneMark-ETP/bin


#had to move files locally otherwise breaker was not locating them
cp $GENOME_MASKED AcTris_vAus2.0.fasta.masked #breaker was not registering my fasta or bam files initially. this fixed it??
cp $RNA_DIR Heart.sorted.bam
cp AcTris_vAus2.0.fasta.masked genome2.fa

#subset the NRA
module load SAMtools/1.16.1-GCC-11.3.0
#samtools view -h -o Heart.sorted.sam Heart.sorted.bam
head -n 300000 Heart.sorted.sam > Heart.sorted.subset.sam
samtools view -h -o Heart.sorted.subset.bam Heart.sorted.subset.sam



#test 1
singularity exec -B ${PWD}:${PWD} ${BRAKER_SIF} braker.pl --genome=AcTris_vAus2.0.fasta.masked --bam=Heart.sorted.bam --softmasking --workingdir=${wd} --threads 2 --gm_max_int

#test 2
singularity exec -B ${PWD}:${PWD} ${BRAKER_SIF} braker.pl --genome=AcTris_vAus2.0.fasta.masked --prot_seq=/opt/BRAKER/example/proteins.fa --softmasking --workingdir=${wd} --thr

singularity exec -B ${PWD}:${PWD} ${BRAKER_SIF} braker.pl --genome=AcTris_vAus2.0.fasta.masked --prot_seq=/opt/BRAKER/example/proteins.fa --bam=Heart.sorted.bam --softmaskin
--GENEMARK_PATH=${ETP} --PROTHINT_PATH=${ETP}/gmes/ProtHint/bin --threads 8 --gm_max_intergenic 10000 --skipOptimize
```

BRAKER3

http://topaz.gatech.edu/GeneMark/license_download.cgi (downloaded)

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/
    GM=gmes_linux_64
    tar -zxvf ${GM}.tar.gz
    gunzip gm_key_64.gz
mv gm_key_64 /nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/.gm_key
cp /nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/.gm_key /home/kstu465/.gm_key

#and augustus config in a place it can be edited
export AUGUSTUS_CONFIG_PATH=/nesi/nobackup/uoa02613/kstuart_projects/programs/Augustus/config
```

Run1: Just myna rnaseq and all orthodb

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3

GENOME_MASKED=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/repeats/repeatmasker/AcTris_vAus2.0repeatlib_softmask/AcTris_vAus2.0.fasta.masked
#had to move files locally otherwise breaker was not locating them
cp $GENOME_MASKED AcTris_vAus2.0.fasta.masked #breaker was not registering my fasta or bam files initially. this fixed it??
cp AcTris_vAus2.0.fasta.masked genome2.fa #braker seemed fussy about the file name??
```

```
#!/bin/bash -e

#SBATCH --job-name=2023_03_17.annotation_breaker_test1.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-160:00:00
#SBATCH --mem=50GB
#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 BRAKER/3.0.2-gimkl-2022a-Perl-5.34.1

export AUGUSTUS_CONFIG_PATH=/nesi/nobackup/uoa02613/kstuart_projects/programs/Augustus/config
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/genome2.fa
PROT=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3/uniprot_sprot_clean.fasta
RNA1=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/transcriptome/Heart.sorted.bam
RNA2=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/transcriptome/Liver.sorted.bam
RNA3=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/transcriptome/Test.sorted.bam

srun braker.pl --threads=${SLURM_CPUS_PER_TASK} --genome=${GENOME} --prot_seq=${PROT} --bam=${RNA1} --bam=${RNA2} --bam=${RNA3} --workingdir=test7 --GENEMARK_PA
```

run2: busco + stats on the output

```
#!/bin/bash -e

#SBATCH --job-name=2023_03_25.busco_annotation.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-24:00:00
#SBATCH --mem=15GB
#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

module purge
module load BUSCO/5.3.2-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3/test7/busco
ANNOTATION=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3/test7/braker.codingseq

busco  -i $ANNOTATION -o braker.codingseq -m transcriptome -l aves_odb10 -c 16 -f
```

```
C:88.0%[S:66.9%,D:21.1%],F:0.7%,M:11.3%,n:8338
7334    Complete BUSCOs (C)
5577    Complete and single-copy BUSCOs (S)
1757    Complete and duplicated BUSCOs (D)
60      Fragmented BUSCOs (F)
944     Missing BUSCOs (M)
8338    Total BUSCO groups searched
```

Run2: w/ softmasking & starling isoseq & new uniprot

download new uniprot (code below in run3)

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3

mkdir resources && cd $_

wget https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz
md5sum uniprot_sprot.fasta.gz > uniprot.md5sum
date > uniprot.download_date
gunzip uniprot_sprot.fasta.gz
```

starling isoseq; map to AcTris genome

```
#!/bin/bash -e

#SBATCH --job-name=2023_03_29.annotation_breaker_isoseqmap.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=20GB
#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

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/isoseq

module load minimap2/2.24-GCC-11.3.0

#map the reads
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta
ISOSEQ=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/isoseq/clustered.hq_noslash.fasta #not usre if slash's important but using the clean on just in case

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 sort clustered.hq.fasta.sam | samtools view -O BAM -o clustered.hq.fasta.sorted.bam

grep -v '@' /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/isoseq/clustered.hq.fasta.sam | awk '$5==60 || $3=="*"' | cut -f 1 | sort -u | wc -l
##33087
```

run braker3:

```
#!/bin/bash -e
#SBATCH --job-name=2023_03_29.annotation_breaker_run2.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-100:00:00
#SBATCH --mem=50GB
#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 BRAKER/3.0.2-gimkl-2022a-Perl-5.34.1

export AUGUSTUS_CONFIG_PATH=/nesi/nobackup/uoa02613/kstuart_projects/programs/Augustus/config

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/genome2.fa
PROT=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3/resources/uniprot_sprot.fasta
RNA1=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/transcriptome/Heart.sorted.bam
RNA2=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/transcriptome/Liver.sorted.bam
RNA3=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/transcriptome/Test.sorted.bam
RNA4=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/isoseq/clustered.hq.fasta.sorted.bam

srun braker.pl --threads=${SLURM_CPUS_PER_TASK} --genome=${GENOME} --softmasking --prot_seq=${PROT} --bam=${RNA1} --bam=${RNA2} --bam=${RNA3} --workingdir=run2 --GE
```

busco + stats on the output

```
#!/bin/bash -e

#SBATCH --job-name=2023_04_13.busco_annotation.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-24:00:00
#SBATCH --mem=15GB
#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

module purge
module load BUSCO/5.3.2-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3/run2/busco
ANNOTATION=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3/run2/braker.codingseq

busco  -i $ANNOTATION -o braker.codingseq -m transcriptome -l aves_odb10 -c 16 -f
```

C:88.1%[S:67.2%,D:20.9%],F:0.8%,M:11.1%,n:8338

7341 Complete BUSCOs (C)

5599 Complete and single-copy BUSCOs (S)

1742 Complete and duplicated BUSCOs (D)

64 Fragmented BUSCOs (F)

933 Missing BUSCOs (M)

8338 Total BUSCO groups searched

Run3: Run 3 + extra prot evidence

run braker3: (run time of 2 days)

```
#!/bin/bash -e
#SBATCH --job-name=2023_04_13.annotation_breaker_run3.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-100:00:00
#SBATCH --mem=50GB
#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 BRAKER/3.0.2-gimkl-2022a-Perl-5.34.1

export AUGUSTUS_CONFIG_PATH=/nesi/nobackup/uoa02613/kstuart_projects/programs/Augustus/config

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/genome2.fa
PROT=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba/all_proteins.fasta
RNA1=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/transcriptome/Heart.sorted.bam
RNA2=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/transcriptome/Liver.sorted.bam
RNA3=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/transcriptome/Test.sorted.bam
#RNA4=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/isoseq/clustered.hq.fasta.sorted.bam

srun braker.pl --threads=${SLURM_CPUS_PER_TASK} --genome=${GENOME} --softmasking --prot_seq=${PROT} --bam=${RNA1} --bam=${RNA2} --bam=${RNA3} --workingdir=run2 --GE
```

run2 (tgut_proteins.fasta): busco + stats on the output

```
#!/bin/bash -e

#SBATCH --job-name=2023_04_18.busco_annotation.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-24:00:00
#SBATCH --mem=15GB
#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

module purge
module load BUSCO/5.3.2-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3/run3/busco
ANNOTATION=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3/run3/braker.codingseq

busco  -i $ANNOTATION -o braker.codingseq -m transcriptome -l aves_odb10 -c 16 -f
```

```
C:88.0%[S:66.9%,D:21.1%],F:0.7%,M:11.3%,n:8338
7334    Complete BUSCOs (C)
5577    Complete and single-copy BUSCOs (S)
1757    Complete and duplicated BUSCOs (D)
60      Fragmented BUSCOs (F)
944     Missing BUSCOs (M)
8338    Total BUSCO groups searched
```

Run4: Galba

https://jguhlin.github.io/genome-annotation-guide/quick_guide.html

Download and Install GALBA w/ Singularity

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs
module load Singularity/3.10.3

export SINGULARITY_CACHEDIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/singularity
singularity build galba.sif docker://katharinahoff/galba-notebook:latest
```

Download and Install GALBA w/ Singularity:update

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs
module load Singularity/3.10.3

export SINGULARITY_CACHEDIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/singularity_update
singularity build galba.sif docker://katharinahoff/galba-notebook:latest
```

move files to a place they are picked up by galba

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba
cp ../galba_updated/genome2.fa .
cp ../galba_updated/proteins.fasta .

wget https://ftp.ensembl.org/pub/current_fasta/taeniopygia_guttata/cdna/Taeniopygia_guttata.bTaeGut1_v1.p.cdna.all.fa.gz
wget https://ftp.ensembl.org/pub/current_fasta/gallus_gallus/cdna/Gallus_gallus.bGalGal1.mat.broiler.GRCg7b.cdna.all.fa.gz
wget https://ftp.ensembl.org/pub/current_fasta/parus_major/cdna/Parus_major.Parus_major1.1.cdna.all.fa.gz
gunzip *gz

cat proteins.fasta Gallus_gallus.bGalGal1.mat.broiler.GRCg7b.cdna.all.fa Parus_major.Parus_major1.1.cdna.all.fa Taeniopygia_guttata.bTaeGut1_v1.p.cdna.all.fa > all_proteins.fasta
cat proteins.fasta Taeniopygia_guttata.bTaeGut1_v1.p.cdna.all.fa > tgut_proteins.fasta
cat Gallus_gallus.bGalGal1.mat.broiler.GRCg7b.cdna.all.fa Parus_major.Parus_major1.1.cdna.all.fa Taeniopygia_guttata.bTaeGut1_v1.p.cdna.all.fa > all_birds.fasta
```

run galba test:

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba
module load Singularity/3.10.3

singularity exec --bind .:/data  --home=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba_updated/ /nesi/nobackup/uoa02613/kstuart_projects/programs/singula
--genome=genome_test.fa --prot_seq=proteins_test.fa --skipOptimize --threads 2

singularity exec --bind .:/data  --home=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba_updated/test1 /nesi/nobackup/uoa02613/kstuart_projects/programs/si
--genome=genome_test.fa --prot_seq=proteins_test.fa --skipOptimize --threads 2 --workingdir=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba_updated/test2
```

Subset genome (used for testing):

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba_updated
head -n 10873093 genome2.fa | tail -n 411216 | sed 's/_//g' > genome2subset.fa
```

run galba: (run time of about 1.5 days)

```
#!/bin/bash -e
#SBATCH --job-name=2023_04_13.annotation_galba.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-150:00:00
#SBATCH --mem=50GB
#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=32
#SBATCH --profile task
```

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba

#load modules
module load Singularity/3.10.3

singularity run /nesi/nobackup/uoa02613/kstuart_projects/programs/singularity/galba.sif galba.pl --version > galba.version

singularity exec --bind .:/data  --home=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba/run1 /nesi/nobackup/uoa02613/kstuart_projects/programs/singularity/g
        --species="Actris" \
         --genome=genome2.fa \
--prot_seq=tgut_proteins.fasta \
        --threads 32 \
        --workingdir=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba/run1 \
        --gff3

singularity exec --bind .:/data --home=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba/run2 /nesi/nobackup/uoa02613/kstuart_projects/programs/singularity/g
        --species="Actris" \
         --genome=genome2.fa \
         --prot_seq=proteins.fasta \
        --threads 8 \
        --workingdir=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba/run2 \
        --gff3

singularity exec --bind .:/data  --home=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba/run1 /nesi/nobackup/uoa02613/kstuart_projects/programs/singularity/g
        --species="Actris" \
         --genome=genome2.fa \
--prot_seq=Taeniopygia_guttata.bTaeGut1_v1.p.cdna.all.fa \
        --threads 32 \
        --workingdir=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba/run3 \
        --gff3

singularity exec --bind .:/data  --home=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba/run6 /nesi/nobackup/uoa02613/kstuart_projects/programs/singularity/g
        --species="Actris6" \
         --genome=genome2.fa \
--prot_seq=all_birds.fasta \
        --threads 16 \
        --workingdir=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba/run6 \
        --gff3
```

Run1: busco + stats on the output

```
#!/bin/bash -e

#SBATCH --job-name=2023_04_18.busco_annotation_galbarun1.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-24:00:00
#SBATCH --mem=15GB
#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

module purge
module load BUSCO/5.3.2-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba/run1

ANNOTATION=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba/run1/augustus.hints.codingseq

busco -i $ANNOTATION -o augustus.hints.codingseq -m transcriptome -l aves_odb10 -c 16 -f
```

C:93.6%[S:83.0%,D:10.6%],F:1.9%,M:4.5%,n:8338

7805 Complete BUSCOs (C)

6924 Complete and single-copy BUSCOs (S)

881 Complete and duplicated BUSCOs (D)

158 Fragmented BUSCOs (F)

375 Missing BUSCOs (M)

8338 Total BUSCO groups searched

Run5: GEMOMA

http://www.jstacs.de/index.php/GeMoMa

Download and Install GEMOMA

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs
conda create -c bioconda gemoma

conda create --name gemoma gemoma



# To activate this environment, use
#     $ conda activate gemoma
#  To deactivate an active environment, use
#     $ conda deactivate
```

Grab reference GFF files

from ftp://ftp.ensembl.org/pub/current_fasta (using https://asia.ensembl.org/info/data/ftp/index.html as a guide)

```
REFDIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/AvianEnsGenomes
```

```
cd $REFDIR
```

```
SPECLIST="lonchura_striata_domestica erythrura_gouldiae geospiza_fortis camarhynchus_parvulus taeniopygia_guttata gallus_gallus phasianus_colchicus ficedula_albicollis struthio_cam
```

```
for SPECIES in $SPECLIST; do
  mkdir $SPECIES && cd $SPECIES
  mkdir fasta && cd fasta
  wget ftp://ftp.ensembl.org/pub/current_fasta/$SPECIES/dna/*.dna.toplevel.fa.gz
  mkdir ../gff3 && cd ../gff3
  wget ftp://ftp.ensembl.org/pub/release-100/gff3/$SPECIES/*.100.gff3.gz
  cd ../ && tree
  cd ../
done
```

```
gunzip -v */fasta/*.dna.toplevel.fa.gz
gunzip -v */gff3/*.100.gff3.gz
```

removed species (manually due to too much memory usage): apteryx_owenii apteryx_rowi struthio_camelus_australis accipiter_nisus phasianus_colchicus chrysolophus_pictus

Run GeMoMa

```
#!/bin/bash -e

#SBATCH --job-name=2023_05_09.gemoma.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-24:00:00
#SBATCH --mem=150GB
#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=40
#SBATCH --profile task

module purge

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/gemoma

module load Mini**conda**3
```

```
CONDA_BASE=$(conda info --base)
source ${CONDA_BASE}/etc/profile.d/conda.sh
conda activate /nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/envs/gemoma
```

```
REFDIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/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=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta
```

```
PREFIX=actris-ensrep200kb
```

```
IDPREFIX=ACTRIS
```

```
GeMoMa GeMoMaPipeline threads=16 outdir=$PREFIX tblastn=false GeMoMa.m=200000 GeMoMa.Score=ReAlign AnnotationFinalizer.r=SIMPLE AnnotationFinalizer.p=$IDPREFIX pc=tru
```

Run6: GEMOMA on katana

```
MODULES=java/8u292-b10-openjdk,mmseqs2/13-45111,blast-plus/2.12.0
GEMOMA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/programs/GeMoMa-1.7.1/GeMoMa-1.7.1.jar
PPN=32
VMEM=180
PRECALL="export _JAVA_OPTIONS=-Xmx${VMEM}g"
```

With evidence then with

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/gemoma_annotation/gemoma_run13_AcTris_1
module load python/2.7.18
```

```
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/At1_MynaGenome/data/genome/AcTris_vAus2.0.fasta
```

```
PREFIX=AcTris_1
```

```
IDPREFIX=ACTRIS
```

```
EMAIL=katarina.stuart@unsw.edu.au
```

```
FARM="java -jar $GEMOMA CLI GeMoMaPipeline threads=$PPN outdir=$PREFIX tblastn=false GeMoMa.m=200000 GeMoMa.Score=ReAlign AnnotationFinalizer.r=SIMPLE AnnotationFin
```

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

busco

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/gemoma/gemoma_katana
module load gffread/0.12.7-GCC-11.3.0
gffread -w gemoma_transcripts.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta f

gffread -y gemoma_proteins.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta fina



#!/bin/bash -e

#SBATCH --job-name=2023_05_10.gemoma_processing.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-02:00:00
#SBATCH --mem=5GB
#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
```

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/gemoma/gemoma_katana
module purge
module load AGAT/1.0.0-gimkl-2022a-Perl-5.34.1-R-4.2.1
agat_sp_keep_longest_isoform.pl --gff final_annotation.gff -o final_annotation_longestIsoform.gff


gffread -w gemoma_longest_transcripts.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2

gffread -y gemoma_longest_proteins.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.


gffread final_annotation.gff  -T -o  final_annotation.gtf
```

Run1: busco + stats on the output

```
#!/bin/bash -e

#SBATCH --job-name=2023_05_10.busco_annotation_gemoma.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-24:00:00
#SBATCH --mem=15GB
#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

module purge
module load BUSCO/5.3.2-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/gemoma/gemoma_katana

ANNOTATION=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/gemoma/gemoma_katana/gemoma_longest_transcripts.fa

busco -i $ANNOTATION -o gemoma_longest_transcripts -m transcriptome -l aves_odb10 -c 16 -f


ANNOTATION=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/gemoma/gemoma_katana/gemoma_transcripts.fa

busco -i $ANNOTATION -o gemoma_transcripts -m transcriptome -l aves_odb10 -c 16 -f
```

C:97.7%[S:97.4%,D:0.3%],F:0.6%,M:1.7%,n:8338

Merging Annotations

Download and Install GALBA w/ Singularity

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs
git clone https://github.com/Gaius-Augustus/TSEBRA
```

move files to a place they are picked up by galba

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba
cp ../galba_updated/genome2.fa .
cp ../galba_updated/proteins.fasta .

wget https://ftp.ensembl.org/pub/current_fasta/taeniopygia_guttata/cdna/Taeniopygia_guttata.bTaeGut1_v1.p.cdna.all.fa.gz
wget https://ftp.ensembl.org/pub/current_fasta/gallus_gallus/cdna/Gallus_gallus.bGalGal1.mat.broiler.GRCg7b.cdna.all.fa.gz
wget https://ftp.ensembl.org/pub/current_fasta/parus_major/cdna/Parus_major.Parus_major1.1.cdna.all.fa.gz
gunzip *gz
```

```
cat proteins.fasta Gallus_gallus.bGalGal1.mat.broiler.GRCg7b.cdna.all.fa Parus_major.Parus_major1.1.cdna.all.fa Taeniopygia_guttata.bTaeGut1_v1.p.cdna.all.fa > all_proteins.fasta
cat proteins.fasta Taeniopygia_guttata.bTaeGut1_v1.p.cdna.all.fa > tgut_proteins.fasta
```

run tsebra: (run time only a few mins)

```
#!/bin/bash -e
#SBATCH --job-name=2023_04_19.annotation_tsebra.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-00:15:00
#SBATCH --mem=5GB
#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

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra

DIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/TSEBRA
BRAKER=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3/run3
GALBA=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/galba/run1
GEMOMA=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/gemoma/gemoma_katana

${DIR}/bin/tsebra.py -g ${BRAKER}/Augustus/augustus.hints.gtf,${GALBA}/augustus.hints.gtf -c ${DIR}/config/default.cfg \
-e ${BRAKER}/hintsfile.gff,${GALBA}/hintsfile.gff \
   -o braker_galba_combined.gtf

${DIR}/bin/tsebra.py -g ${BRAKER}/Augustus/augustus.hints.gtf,${GALBA}/augustus.hints.gtf -c ${DIR}/config/keep_ab_initio.cfg \
-e ${BRAKER}/hintsfile.gff,${GALBA}/hintsfile.gff \
   -o braker_galba_combined_ab_initio.gtf

${DIR}/bin/tsebra.py -g ${BRAKER}/Augustus/augustus.hints.gtf,${GALBA}/augustus.hints.gtf -c ${DIR}/config/kat_manual.cfg \
-e ${BRAKER}/hintsfile.gff,${GALBA}/hintsfile.gff \
   -o braker_galba_combined_kat_manual.gtf

${DIR}/bin/tsebra.py -g ${GALBA}/augustus.hints.gtf --keep_gtf ${BRAKER}/Augustus/augustus.hints.gtf -c ${DIR}/config/kat_manual.cfg \
-e ${GALBA}/hintsfile.gff \
   -o braker_galba_combined_katmanual_brakerforced.gtf

${DIR}/bin/tsebra.py -g ${GALBA}/augustus.hints.gtf --keep_gtf ${BRAKER}/Augustus/augustus.hints.gtf -c ${DIR}/config/default.cfg \
-e ${GALBA}/hintsfile.gff \
   -o braker_galba_combined_default_brakerforced.gtf

${DIR}/bin/tsebra.py -g ${GALBA}/augustus.hints.gtf --keep_gtf ${BRAKER}/Augustus/augustus.hints.gtf -c ${DIR}/config/default.cfg \
-e ${GALBA}/hintsfile.gff --filter_single_exon_genes \
   -o braker_galba_combined_default_brakerforced_exon.gtf

${DIR}/bin/tsebra.py -g ${GALBA}/augustus.hints.gtf --keep_gtf ${BRAKER}/Augustus/augustus.hints.gtf -c ${DIR}/config/kat_manual2.cfg \
-e ${GALBA}/hintsfile.gff \
   -o braker_galba_combined_katmanual2_brakerforced.gtf

##Using GEMOMA as second annotation with braker

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra

DIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/TSEBRA
BRAKER=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/breaker3/run3
GEMOMA=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/gemoma/gemoma_katana

${DIR}/bin/tsebra.py -g ${BRAKER}/Augustus/augustus.hints.gtf,${GEMOMA}/final_annotation.gtf -c ${DIR}/config/kat_manual.cfg \
-e ${BRAKER}/hintsfile.gff \
   -o braker_gemoma_combined_kat_manual.gtf

${DIR}/bin/tsebra.py -g ${BRAKER}/Augustus/augustus.hints.gtf --keep_gtf ${GEMOMA}/final_annotation.gtf -c ${DIR}/config/kat_manual.cfg \
-e ${BRAKER}/hintsfile.gff \
   -o braker_gemoma_combined_katmanual_gemomaforce.gtf

${DIR}/bin/tsebra.py -g ${GEMOMA}/final_annotation.gtf --keep_gtf ${BRAKER}/Augustus/augustus.hints.gtf -c ${DIR}/config/kat_manual.cfg \
-e ${BRAKER}/hintsfile.gff \
   -o braker_gemoma_combined_katmanual_brakerforce.gtf

${DIR}/bin/tsebra.py -g ${GEMOMA}/final_annotation.gtf --keep_gtf ${BRAKER}/Augustus/augustus.hints.gtf,${GEMOMA}/final_annotation.gtf -c ${DIR}/config/kat_manual2.cfg \
-e ${BRAKER}/hintsfile.gff \
```

```
   -o braker_gemoma_combined_katmanual2_bothforce.gtf


perl  ${DIR}/bin/rename_gtf.py --gtf braker_galba_combined_ab_initio.gtf --prefix AcTris --out braker_galba_combined_ab_initio_renamed.gtf

#gtf2aa.pl automatically grabs the transcripts, but want the genes to be annotated. So swapped trans and gene names for protein fasta extraction to be fed into eggnog.
sed 's/transcript/geene/g' braker_galba_combined_ab_initio_renamed.gtf | sed 's/gene/transcript/g' > braker_galba_combined_ab_initio_renamed2.gtf
perl ${DIR}/gtf2aa.pl /nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/genome2.fa braker_galba_combined_ab_initio_renamed2.gtf  braker_galba_combined_ab_initio_geneO
#should be able to map "braker_galba_combined_ab_initio_geneONLY.fa" to "braker_galba_combined_ab_initio_renamed.gtf" for eggnog

perl ${DIR}/gtf2aa.pl /nesi/nobackup/uoa02613/kstuart_projects/programs/breaker3/genome2.fa braker_galba_combined_ab_initio_renamed.gtf  braker_galba_combined_ab_initio_renamed
```

Analysis of the Annotation

```
#!/bin/bash -e

#SBATCH --job-name=2023_05_10.tsebra_agat.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=15GB
#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

module purge
module load AGAT/1.0.0-gimkl-2022a-Perl-5.34.1-R-4.2.1
module load gffread/0.12.7-GCC-11.3.0

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra

agat_sp_keep_longest_isoform.pl --gff braker_gemoma_combined_katmanual2_bothforce.gtf -o braker_gemoma_combined_katmanual2_bothforce_longestIsoform.gff
gffread -w braker_gemoma_combined_katmanual2_bothforce_longestIsoform.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag

agat_sp_keep_longest_isoform.pl --gff braker_gemoma_combined_katmanual2_brakerforce.gtf -o braker_gemoma_combined_katmanual2_brakerforce_longestIsoform.gff
gffread -w braker_gemoma_combined_katmanual2_brakerforce_longestIsoform.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragta

agat_sp_keep_longest_isoform.pl --gff braker_gemoma_combined_katmanual_brakerforce.gtf -o braker_gemoma_combined_katmanual_brakerforce_longestIsoform.gff
gffread -w braker_gemoma_combined_katmanual_brakerforce_longestIsoform.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag
gffread -y braker_gemoma_combined_katmanual_brakerforce_longestIsoform_transcripts.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffo

agat_sp_keep_longest_isoform.pl --gff braker_gemoma_combined_katmanual_gemomaforce.gtf -o braker_gemoma_combined_katmanual_gemomaforce_longestIsoform.gff
gffread -w braker_gemoma_combined_katmanual_gemomaforce_longestIsoform.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragt

agat_sp_keep_longest_isoform.pl --gff braker_gemoma_combined_kat_manual.gtf -o braker_gemoma_combined_kat_manual_longestIsoform.gff
gffread -w braker_gemoma_combined_kat_manual_longestIsoform.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synt
```

Run1: busco + stats on the output

```
#!/bin/bash -e

#SBATCH --job-name=2023_04_19.busco_annotation_tsebra.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=15GB
#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

module purge
module load BUSCO/5.3.2-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra

busco -i braker_galba_combined.fa -o braker_galba_combined -m proteins -l aves_odb10 -c 16 -f
```

```
busco -i braker_galba_combined_ab_initio.fa -o braker_galba_combined_ab_initio -m proteins  -l aves_odb10 -c 16 -f

busco -i braker_galba_combined_kat_manual.fa -o braker_galba_combined_kat_manual -m proteins  -l aves_odb10 -c 16 -f

busco -i braker_gemoma_combined_katmanual2_bothforce_longestIsoform.fa -o braker_gemoma_combined_katmanual2_bothforce_longestIsoform -m transcriptome -l aves_odb10 -c 16 -f
busco -i braker_gemoma_combined_katmanual2_brakerforce_longestIsoform.fa -o braker_gemoma_combined_katmanual2_brakerforce_longestIsoform -m transcriptome -l aves_odb10 -c
busco -i braker_gemoma_combined_katmanual_brakerforce_longestIsoform.fa -o braker_gemoma_combined_katmanual_brakerforce_longestIsoform -m transcriptome -l aves_odb10 -c 16
busco -i braker_gemoma_combined_katmanual_gemomaforce_longestIsoform.fa -o braker_gemoma_combined_katmanual_gemomaforce_longestIsoform -m transcriptome -l aves_odb10 -
busco -i braker_gemoma_combined_kat_manual_longestIsoform.fa -o braker_gemoma_combined_kat_manual_longestIsoform -m transcriptome -l aves_odb10 -c 16 -f
```

Final choice: braker_gemoma_combined_katmanual_brakerforce_longestIsoform.txt

Number of genes: 19836

Busco completeness: 98.40%

8211 Complete BUSCOs (C)

8157 Complete and single-copy BUSCOs (S)

54 Complete and duplicated BUSCOs (D)

47 Fragmented BUSCOs (F)

80 Missing BUSCOs (M)

8338 Total BUSCO groups searched

AGAT summary, rename and prep final version:

```
#!/bin/bash -e

#SBATCH --job-name=2023_05_16.tsebra_agat.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=15GB
#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

module purge
module load AGAT/1.0.0-gimkl-2022a-Perl-5.34.1-R-4.2.1
module load gffread/0.12.7-GCC-11.3.0

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra

agat_sp_functional_statistics.pl --gff braker_gemoma_combined_katmanual_brakerforce.gtf -o agat_braker_gemoma_combined_katmanual_brakerforce_func_statistics

agat_sp_functional_statistics.pl --gff braker_gemoma_combined_katmanual_brakerforce_longestIsoform.gtf -o agat_braker_gemoma_combined_katmanual_brakerforce_longestIsoform_

###


cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra

DIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/TSEBRA

#first rename genes and transcripts to match
perl  ${DIR}/bin/rename_gtf.py --gtf braker_gemoma_combined_katmanual_brakerforce.gtf --prefix AcTris --out braker_gemoma_combined_katmanual_brakerforce_renamed.gtf

module load AGAT/1.0.0-gimkl-2022a-Perl-5.34.1-R-4.2.1
agat_sp_keep_longest_isoform.pl --gff braker_gemoma_combined_katmanual_brakerforce_renamed.gtf -o braker_gemoma_combined_katmanual_brakerforce_renamed_longestIsoform.gff

module load gffread/0.12.7-GCC-11.3.0
gffread -w braker_gemoma_combined_katmanual_brakerforce_renamed_longestIsoform_transcripts.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/ste

gffread -y braker_gemoma_combined_katmanual_brakerforce_renamed_longestIsoform_proteins.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4
```

Functional Annotation:

Download and Install

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs

conda install -c bioconda eggnog-mapper

conda create --name eggnog-mapper eggnog-mapper

export PATH=/nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/envs/eggnog-mapper:/nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamb
export EGGNOG_DATA_DIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/eggnog-mapper-data


download_eggnog_data.py
```

To activate this environment, use

$ conda activate eggnog-mapper

To deactivate an active environment, use

$ conda deactivate

```
#!/bin/bash -e

#SBATCH --job-name=2023_05_11.annotation_eggnog.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=10GB
#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

module purge

module load Miniconda3
CONDA_BASE=$(conda info --base)
source ${CONDA_BASE}/etc/profile.d/conda.sh
conda activate /nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/envs/eggnog-mapper

export PATH=/nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/envs/eggnog-mapper:/nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamb
export EGGNOG_DATA_DIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/eggnog-mapper-data

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/eggnog/run1

emapper.py -i /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra/braker_gemoma_combined_katmanual_brakerforce_longestIsoform_transcripts.fa -o braker_g

###

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/eggnog/run1.0

emapper.py -i /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra/braker_gemoma_combined_katmanual_brakerforce_renamed_longestIsoform_proteins.fa -o

grep -v "^#" braker_gemoma_combined_katmanual_brakerforce_renamed_longestIsoform_proteins.emapper.annotations | cut -f 21 | grep "\-$" | wc -l 727
```

SAAGA assessment

old version:

```
#!/bin/bash -e

#SBATCH --job-name=2023_04_24.annotation_saaga.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-00:10:00
#SBATCH --mem=5GB
#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
#SBATCH --qos=debug

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/saaga


GFF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra/braker_galba_combined_ab_initio_renamed.gtf
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta

grep -v "AcTris_g1036." /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra/braker_galba_combined_ab_initio_renamed.gtf > braker_galba_combined_ab_initio

module load AGAT/1.0.0-gimkl-2022a-Perl-5.34.1-R-4.2.1
agat_convert_sp_gxf2gxf.pl -g braker_galba_combined_ab_initio_renamed2.gtf -o braker_galba_combined_ab_initio_renamed2.gff3

module load gffread/0.12.7-GCC-11.3.0

gffread -w braker_galba_combined_ab_initio_transcripts_clean.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/

gffread -y braker_galba_combined_ab_initio_proteins_clean.fa -g /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/ren



cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/saaga

module purge
module load Python/2.7.14-gimkl-2017a
module load MMseqs2/13-45111-gimpi-2020a

GFF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/saaga/braker_galba_combined_ab_initio_renamed2.gff3
FASTA=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/saaga/braker_galba_combined_ab_initio_transcripts_clean.fa
PROT=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/saaga/braker_galba_combined_ab_initio_proteins_clean.fa
SPROT_FASTA=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/AvianEnsGenomes/gallus_gallus/pep/Gallus_gallus.bGalGal1.mat.broiler.GRCg7b.pep

python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/saaga.py  -seqin $PROT -gffin $GFF -cdsin $FASTA -refprot $SPROT_FASTA gffmrna=transcript -annotate -sur
```

new version:

```
#!/bin/bash -e

#SBATCH --job-name=2023_05_15.annotation_saaga.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-00:10:00
#SBATCH --mem=5GB
#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
#SBATCH --qos=debug

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/saaga

module purge
module load Python/2.7.14-gimkl-2017a
module load MMseqs2/13-45111-gimpi-2020a

GFF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra/braker_gemoma_combined_katmanual_brakerforce_renamed_longestIsoform.gff
FASTA=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra/braker_gemoma_combined_katmanual_brakerforce_renamed_longestIsoform_transcripts.fa
PROT=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra/braker_gemoma_combined_katmanual_brakerforce_renamed_longestIsoform_proteins.fa
SPROT_FASTA=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/AvianEnsGenomes/gallus_gallus/pep/Gallus_gallus.bGalGal1.mat.broiler.GRCg7b.pep

python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/saaga.py  -seqin $PROT -gffin $GFF -cdsin $FASTA -refprot $SPROT_FASTA gffmrna=transcript -annotate -sur
```