Starling-May18

Projects/Katarina Stuart/KStuart.Starling-Aug18/At1_Genome/Analysis/2022.08.01.GeneFamilyExpansions

PDF Version generated by

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

on

Aug 23, 2023 @09:42 AM NZST

Table of Contents

2022.08.01.GeneFamilyExpansions

Gene family analysis

Orthofinder

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1832-y

good example:

https://www.nature.com/articles/s41597-023-01950-5

The swan genome and transcriptome, it is not all black and white | Genome Biology | Full Text (biomedcentral.com)

cafe5

https://github.com/hahnlab/CAFE

AlaMetTyr/Genomic-markers-of-invasiveness (github.com)

A high-quality assembly reveals genomic characteristics, phylogenetic status, and causal genes for leucism plumage of Indian peafowl - PMC (nih.gov)

From above manuscript: "The longest transcript of each gene was extracted and then the genes with the length of protein sequences shorter than 50 amino filtered"

Downloading the proteomes

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/analysis/gene_family/ensembl_refs
```

```
cd $REFDIR
```

```
SPECLIST="lonchura_striata_domestica taeniopygia_guttata gallus_gallus ficedula_albicollis  cyanistes_caeruleus  serinus_canaria parus_major zonotrichia_albicollis homo_sapiens mus_
```

```
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 ../..
done
```

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

```
#!/bin/bash -e

#SBATCH --job-name=2023_05_13.longestisoform.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-48:00:00
#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
#SBATCH --partition=milan
#SBATCH --array=1-12

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

SPECIES=$(sed "${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/ensembl_refs/species_list)

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/ensembl_refs

cd $SPECIES/gff3
#agat_sp_keep_longest_isoform.pl --gff *.gff3  -o ${SPECIES}_longestIsoform.gff3

#gffread -y ${SPECIES}_prot.fa -g ../fasta/*.fa *longestIsoform.gff3


module load SeqKit/2.4.0
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/ensembl_refs
for i in $(cat /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/ensembl_refs/species_list);
do
#seqkit grep --by-seq --invert-match --pattern '.' ${i}/gff3/${i}_prot.fa > ${i}/gff3/${i}_prot_nostop.fa
#seqkit seq -m 30 ${i}/gff3/${i}_prot_nostop.fa > ${i}/gff3/${i}_prot_nostop_filter.fa
echo $i
grep "^>" ${i}/gff3/${i}_prot_nostop.fa | wc -l
grep "^>" ${i}/gff3/${i}_prot_nostop_filter.fa | wc -l
done

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/ensembl_refs/acridotheres_tristis/gff3
seqkit grep --by-seq --invert-match --pattern '.' braker_gemoma_combined_katmanual_brakerforce_longestIsoform_transcripts.fa
> braker_gemoma_combined_katmanual_brakerforce_longestIsoform_transcripts_nostop.fa
seqkit seq -m
30 braker_gemoma_combined_katmanual_brakerforce_longestIsoform_transcripts_nostop.fa > braker_gemoma_combined_katmanual_brakerforce_longestIsoform_transcripts_nostop_filte
grep "^>" braker_gemoma_combined_katmanual_brakerforce_longestIsoform_transcripts_nostop.fa | wc -l
grep "^>" braker_gemoma_combined_katmanual_brakerforce_longestIsoform_transcripts_nostop_filter.fa | wc -l
```

Had random mismatch errors for gallus, mus, and ficedula which ment that gffread didn't work (ensemble assembly GFF and FA didn't have matching versi
do some manual assemble or GFF downloads to make sure versions matched.

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/ensembl_refs/gallus_gallus/fasta
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/016/699/485/GCF_016699485.2_bGalGal1.mat.broiler.GRCg7b/GCF_016699485.2_bGalGal1.mat.broiler.GRCg7b_genomic.fna.gz
gunzip *
mv GCF_016699485.2_bGalGal1.mat.broiler.GRCg7b_genomic.fna GCF_016699485.2_bGalGal1.mat.broiler.GRCg7b_genomic.fa
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/ensembl_refs/gallus_gallus/gff3
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/016/699/485/GCF_016699485.2_bGalGal1.mat.broiler.GRCg7b/GCF_016699485.2_bGalGal1.mat.broiler.GRCg7b_genomic.gff.gz
gunzip *
mv GCF_016699485.2_bGalGal1.mat.broiler.GRCg7b_genomic.gff GCF_016699485.2_bGalGal1.mat.broiler.GRCg7b_genomic.gff3

# downloaded other two at: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000247815.1/ & https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001635.27/ using curl command. r
files to they had suffix .fa nd .gff3
#reran above slurm script to generate the longest isoform prot files.
```

move the prot files to the working orthofinder directory

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run3
ln -s ../ensembl_refs/*/gff3/*.fa .

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4
ln -s ../ensembl_refs/*/gff3/*nostop_filter.fa .

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run5
ln -s ../ensembl_refs/*/gff3/*nostop.fa .
rm homo_sapiens_prot_nostop.fa
```

Orthofinder

Run Orthofinder

```
#!/bin/bash -e

#SBATCH --job-name=2023_05_12.orthofinder.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-150:00:00
#SBATCH --mem=100GB
#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
#SBATCH --partition=milan

module purge
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/

module load OrthoFinder/2.5.2-gimkl-2020a-Python-3.8.2
module load MAFFT/7.505-gimkl-2022a-with-extensions
module load FastTree/2.1.11-GCCcore-9.2.0
orthofinder -f orthofinder_run3/ -M msa -S blast -I 1.3
```

make tree ultrametric

Time acquired from TimeTree: http://timetree.org/

319 MYA

```
module purge
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/

module load OrthoFinder/2.5.2-gimkl-2020a-Python-3.8.2

DIR=/opt/nesi/CS400_centos7_bdw/OrthoFinder/2.5.2-gimkl-2020a-Python-3.8.2/tools/
TREE_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Species_Tree

python $DIR/make_ultrametric.py -r 319 $TREE_DIR/SpeciesTree_rooted.txt

python $DIR/make_ultrametric.py -r 319 $TREE_DIR/SpeciesTree_rooted_node_labels.txt
```

process ortholog outputs

grabbed code from: https://www.biostars.org/p/9553290/#9553371

```
module purge
module load R/4.1.0-gimkl-2020a
R
library("data.table")

setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Phylogenetic_Hierarchical_Orthogroups")
hog <- fread('N0.tsv')
hog[, OG := NULL]
hog[, `Gene Tree Parent Clade` := NULL]
hog <- melt(hog, id.vars='HOG', variable.name='species', value.name='pid')
hog <- hog[pid != '']
hog$n <- sapply(hog$pid, function(x) length(strsplit(x, ', ')[[1]]))

# Exclude HOGs with lots of genes in a one or more species.
# See also cafe tutorial about filtering gene families
keep <- hog[, list(n_max=max(n)), HOG][n_max < 100]$HOG
hog <- hog[HOG %in% keep]

# Exclude HOGs present in only 1 species
keep <- hog[, .N, HOG][N > 1]$HOG
hog <- hog[HOG %in% keep]

counts <- dcast(hog, HOG ~ species, value.var='n', fill=0)
counts[, Desc := 'n/a']
setcolorder(counts, 'Desc')
fwrite(counts, 'hog_gene_counts.tsv', sep='\t')
```

CAFE5

download Cafe5

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

conda install -c bioconda cafe
conda create --name cafe cafe
```

To activate this environment, use

$ conda activate cafe

To deactivate an active environment, use

$ conda deactivate

Run Cafe5

```
#!/bin/bash -e

#SBATCH --job-name=2023_05_15.cafe5.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#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
#SBATCH --partition=milan

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/cafe

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/

OUT_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out

cafe5 -i Phylogenetic_Hierarchical_Orthogroups/hog_gene_counts.tsv -t Species_Tree/SpeciesTree_rooted_node_labels.txt.ultrametric.tre -o $OUT_DIR --cores 16

##
sed -e 's/braker_gemoma_combined_katmanual_brakerforce_longestIsoform_transcripts_nostop_filter/AcTris_vAus2.0/g' \
-e 's/cyanistes_caeruleus_prot_nostop_filter/Cyanistes_caeruleus/g' \
-e 's/ficedula_albicollis_prot_nostop_filter/Ficedula_albicollis/g' \
-e 's/gallus_gallus_prot_nostop_filter/Gallus_gallus/g' \
-e 's/lonchura_striata_domestica_prot_nostop_filter/Lonchura_striata/g' \
-e 's/mus_musculus_prot_nostop_filter/Mus_musculus/g' \
-e 's/homo_sapiens_prot_nostop_filter/Homo_sapiens/g' \
-e 's/parus_major_prot_nostop_filter/Parus_major/g' \
-e 's/sturnus_vulgaris_prot_nostop_filter/Sturnus_vulgaris/g' \
-e 's/serinus_canaria_prot_nostop_filter/Serinus_canaria/g' \
-e 's/taeniopygia_guttata_prot_nostop_filter/Taeniopygia_guttata/g' \
-e 's/zonotrichia_albicollis_prot_nostop_filter/Zonotrichia_albicollis/g' Phylogenetic_Hierarchical_Orthogroups/hog_gene_counts.tsv  >
Phylogenetic_Hierarchical_Orthogroups/hog_gene_counts_renamed.tsv


sed -e 's/braker_gemoma_combined_katmanual_brakerforce_longestIsoform_transcripts_nostop_filter/AcTris_vAus2.0/g' \
-e 's/cyanistes_caeruleus_prot_nostop_filter/Cyanistes_caeruleus/g' \
-e 's/ficedula_albicollis_prot_nostop_filter/Ficedula_albicollis/g' \
```

```
-e 's/gallus_gallus_prot_nostop_filter/Gallus_gallus/g' \
-e 's/lonchura_striata_domestica_prot_nostop_filter/Lonchura_striata/g' \
-e 's/mus_musculus_prot_nostop_filter/Mus_musculus/g' \
-e 's/homo_sapiens_prot_nostop_filter/Homo_sapiens/g' \
-e 's/parus_major_prot_nostop_filter/Parus_major/g' \
-e 's/sturnus_vulgaris_prot_nostop_filter/Sturnus_vulgaris/g' \
-e 's/serinus_canaria_prot_nostop_filter/Serinus_canaria/g' \
-e 's/taeniopygia_guttata_prot_nostop_filter/Taeniopygia_guttata/g' \
-e 's/zonotrichia_albicollis_prot_nostop_filter/Zonotrichia_albicollis/g' Species_Tree/SpeciesTree_rooted_node_labels.txt.ultrametric.tre  >
Species_Tree/SpeciesTree_rooted_node_labels.txt.ultrametric_renamed.tre

OUT_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed
cafe5 -i Phylogenetic_Hierarchical_Orthogroups/hog_gene_counts_renamed.tsv -t Species_Tree/SpeciesTree_rooted_node_labels.txt.ultrametric_renamed.tre -o $OUT_DIR --cores 16

OUT_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k1
cafe5 -i Phylogenetic_Hierarchical_Orthogroups/hog_gene_counts_renamed.tsv -t Species_Tree/SpeciesTree_rooted_node_labels.txt.ultrametric_renamed.tre -o $OUT_DIR --cores 16 -k 1

OUT_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k2
cafe5 -i Phylogenetic_Hierarchical_Orthogroups/hog_gene_counts_renamed.tsv -t Species_Tree/SpeciesTree_rooted_node_labels.txt.ultrametric_renamed.tre -o $OUT_DIR --cores 16  -k 2

OUT_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k3
cafe5 -i Phylogenetic_Hierarchical_Orthogroups/hog_gene_counts_renamed.tsv -t Species_Tree/SpeciesTree_rooted_node_labels.txt.ultrametric_renamed.tre -o $OUT_DIR --cores 16  -k 3
```

k = 4 had the lowest log likelihood score as determined by looking at the .sl outputs (tail -n 15 2023_05_15.cafe5.sl_35839*)

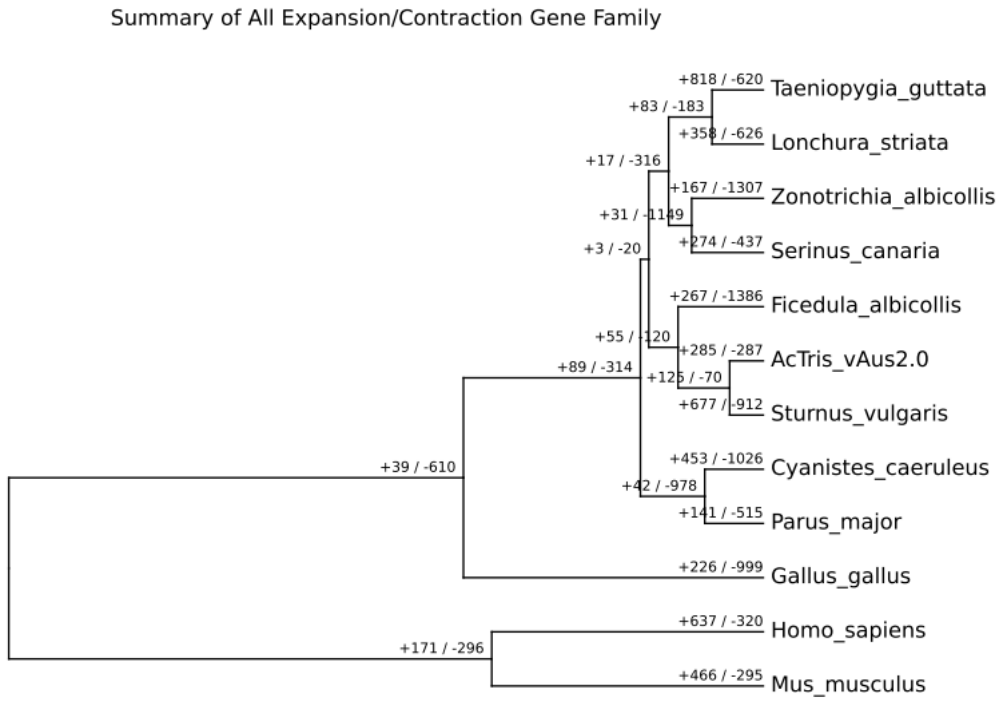
summarise output

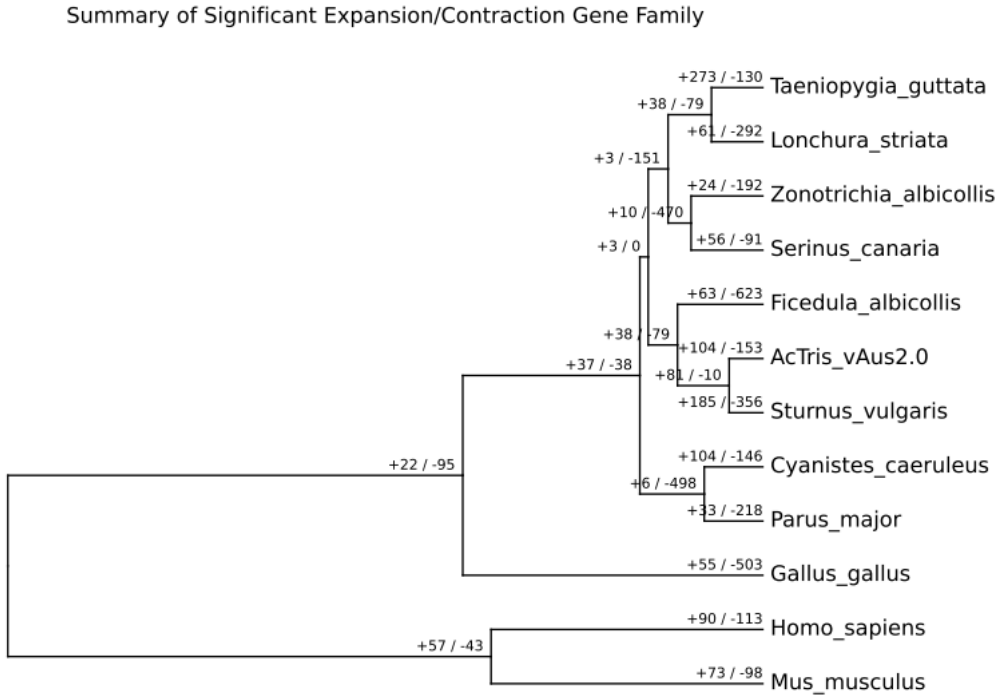
https://github.com/moshi4/CafePlotter#installation

```
pip install cafeplotter
#cant be in conda env when i run this

IN_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4
OUT_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4_plot
module load Python/3.8.1-gimkl-2018b
cafeplotter -i $IN_DIR -o $OUT_DIR --format pdf
```

Cafe5_out:





for pictures https://www.phylopic.org/

Visualise +ve and -ve expansions

useful files
Orthogroups/Orthogroups.GeneCount.tsv

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4
sed 's/#//g' Gamma_family_results.txt > Gamma_family_results_edit.txt
sed 's/#//g' Gamma_change.tab > Gamma_change_edit.tab

module load R/4.1.0-gimkl-2020a
R
library(data.table)
library(dplyr)
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4")


sig<- fread("Gamma_family_results_edit.txt")
colnames(sig) <- c("FamilyID2","pval","sig")

family<-read.table("Gamma_change_edit.tab", header=T)
bind<-cbind(sig,family)

og<-fread("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Phylogenetic_Hierarchical_Orthogroups/N0.tsv
go<-og[,1:2]
colnames(go)<-c("FamilyID","GO")
bind2<-inner_join(bind, go, by="FamilyID" )
bind2$count_0 <- rowSums(bind2[,5:27] == "0")

pos <- filter(bind2 , pval < 0.05 & AcTris_vAus2.0.4. > 0)
neg <- filter(bind2 , pval < 0.05 & AcTris_vAus2.0.4.< 0 )
all <- filter(bind2 , pval < 0.05 & AcTris_vAus2.0.4. != 0 )

write.table(pos, "AcTris_GOterms_pos.txt")
write.table(neg, "AcTris_GOterms_neg.txt")
write.table(all, "AcTris_GOterms_all.txt")
```

grab GO from OG terms by interproscanning them:

https://gitlab.nibio.no/simeon/iwanicki_et_al_21/-/blob/master/cafe5-GOstats_parser.Rmd

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4
tail -n +2 AcTris_GOterms_pos.txt  | cut -d' ' -f 29  | sed -e 's/"//g' | sort | uniq > AcTris_GOterms_pos_OGtermONLY.txt
tail -n +2 AcTris_GOterms_neg.txt  | cut -d' ' -f 29  | sed -e 's/"//g' | sort | uniq  > AcTris_GOterms_neg_OGtermONLY.txt

DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/

module load SeqKit/2.4.0

#grab largest sequence of each to be interproscanned.
for i in $(cat AcTris_GOterms_pos_OGtermONLY.txt)
do
seqkit sort -l -r ${DIR}/WorkingDirectory/Sequences_ids/${i}.fa | seqkit head -n 5  >  AcTris_pos/${i}_single.fa
done

#grab largest sequence of each to be interproscanned
for i in $(cat AcTris_GOterms_neg_OGtermONLY.txt)
do
echo $i
seqkit sort -l -r ${DIR}/WorkingDirectory/Sequences_ids/${i}.fa | seqkit head -n 5 > AcTris_neg/${i}_single.fa
done
```

some array slurm scripts for quick interproscan annotation

```
#!/bin/bash -e

#SBATCH --job-name=2023_05_18.cafe_interproscan.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#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=2
#SBATCH --profile task
#SBATCH --partition=milan
#SBATCH --array=1-100

OGTERM=$(sed "${SLURM_ARRAY_TASK_ID}q;d"
/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/AcTris_GOterms_pos_OGtermON
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/AcTris_pos
module load InterProScan/5.51-85.0-gimkl-2020a-Perl-5.30.1-Python-3.8.2
interproscan.sh -i ./${OGTERM}_single.fa -appl Pfam --goterms
#pull out GO terms from interproscan output
cut -f 14 ${OGTERM}_single.fa.tsv | grep -v "-" | sed 's/|/\n/g' | sort | uniq > ${OGTERM}.goterms
#format a file so it contains the OG term and the GO terms
awk -v variable1="$OGTERM" '{print $0, variable1}' ${OGTERM}.goterms  > ${OGTERM}.goterms_format
```

```
#SBATCH --array=1-149

OGTERM=$(sed "${SLURM_ARRAY_TASK_ID}q;d"
/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/AcTris_GOterms_neg_OGtermON
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/AcTris_neg
module load InterProScan/5.51-85.0-gimkl-2020a-Perl-5.30.1-Python-3.8.2
interproscan.sh -i ./${OGTERM}_single.fa -appl Pfam --goterms
#pull out GO terms from interproscan output
cut -f 14 ${OGTERM}_single.fa.tsv | grep -v "-" | sed 's/|/\n/g' | sort | uniq > ${OGTERM}.goterms
#format a file so it contains the OG term and the GO terms
awk -v variable1="$OGTERM" '{print $0, variable1}' ${OGTERM}.goterms_format
```

finally merge outputs

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/AcTris_pos
cat *goterms_format | grep -v "^ " > combined_pos_subset.tsv

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/AcTris_neg
```

```
cat *goterms_format | grep -v "^ " > combined_neg_subset.tsv
```

map the GO terms to pvals an increases

```
cd  /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/
tail -n +2 AcTris_GOterms_all.txt  | sed -e 's/"//g'  | awk '{print $29,"\t",$3,"\t",$10}' > AcTris_GOterms_all_3col.txt
tail -n +2 AcTris_GOterms_pos.txt  | sed -e 's/"//g'  | awk '{print $29,"\t",$3,"\t",$10}' > AcTris_GOterms_pos_3col.txt
tail -n +2 AcTris_GOterms_neg.txt  | sed -e 's/"//g'  | awk '{print $29,"\t",$3,"\t",$10}' > AcTris_GOterms_neg_3col.txt

module load R/4.1.0-gimkl-2020a
R
library(data.table)
library(dplyr)
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/")

pos<-read.table("AcTris_pos/combined_pos_subset.tsv")
colnames(pos) <- c("GO","OG")
pos2<- pos %>% count(GO)

neg<-read.table("AcTris_neg/combined_neg_subset.tsv")
colnames(neg) <- c("GO","OG")
neg2<- neg %>% count(GO)

write.table(pos2, "AcTris_GOanalysis_input_pos.txt", row.names=FALSE,sep="\t", quote = FALSE)
write.table(neg2, "AcTris_GOanalysis_input_neg.txt", row.names=FALSE,sep="\t", quote = FALSE)
```

plot in revigo, then manual play around with r script generated on the website

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/
tail -n +2 AcTris_GOanalysis_input_pos.txt | cut -f1 >AcTris_GOanalysis_input_pos_revigo.txt
tail -n +2 AcTris_GOanalysis_input_neg.txt | cut -f1 >AcTris_GOanalysis_input_neg_revigo.txt
```

~plot in revigo, then grab R code and edit as below:~

pos plot: (edited text in grep, direct copy paste from revigo script on white)

```
module load R/4.1.0-gimkl-2020a
R
library(ggplot2)
library(scales)
library(dplyr)
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/")
pos<-read.table("AcTris_GOanalysis_input_pos.txt")
colnames(pos) <- c("term_ID","size")
```

revigo.names <- c("term_ID","description","frequency","plot_X","plot_Y","log_size","uniqueness","dispensability");

revigo.data <- rbind(c("GO:0006334","nucleosome assembly",0.09057190085107651,-0.23843623279027426,-5.321527229998461,4.435063590165067,0.9577664066072151,0.00743743),

c("GO:0006355","regulation of DNA-templated transcription",9.968929480711344,-4.789422467516329,3.1576957989773113,6.476702827628594,0.8736512646494459,0.3923121),

c("GO:0006412","translation",5.085673767131161,5.643662304094316,2.601857550572853,6.184402940776566,0.707459399739002,0.6800612),

c("GO:0006413","translational initiation",0.5201814019133255,6.017110305260105,2.751859579205334,5.1942117566342985,0.7568352774104212,0.44

c("GO:0006468","protein phosphorylation",4.340892012483024,5.101798292395532,4.153575484388295,6.115633473764006,0.7625951858348194,0.38

c("GO:0006481","C-terminal protein methylation",0.019707620732377095,4.345780912372917,6.505539258542858,3.772761647144032,0.8735706093435658,0.03753816),

c("GO:0006508","proteolysis",5.350747086797883,5.38507848381707,5.124472001195252,6.206468852551757,0.819561926687603,0.45001963),

c("GO:0006694","steroid biosynthetic process",0.14729326057246497,6.521823305972429,-1.9895605466366797,4.6462468420953265,0.8878488771317082,0.03399014),

c("GO:0006749","glutathione metabolic process",0.19092277300226926,-4.077419624702831,-4.114523482121034,4.758919458438024,0.96716166844

```
c("GO:0006979","response to oxidative
stress",0.5712981471310761,-2.5873485053296594,6.377821857419009,5.23491950292478,0.8701988859347773,0.24348594),

c("GO:0007156","homophilic cell adhesion via plasma membrane adhesion
molecules",0.216251639141856,-2.3510380326230815,-1.5817489301708954,4.813020246744775,0.9892635259152726,-0),

c("GO:0007186","G protein-coupled receptor signaling
pathway",1.5868742522745753,-3.611253204052394,4.822047725064692,5.678597582760593,0.7823657682902725,0.35182865),

c("GO:0008033","tRNA processing",1.448545048727217,6.693136349302125,1.2233136387269363,5.638987162110661,0.8074306931361029,0.175684

c("GO:0009190","cyclic nucleotide biosynthetic
process",0.15429819715683393,5.803965970308046,-0.20043426957474736,4.66642437251876,0.7798149587688212,0.25536003),

c("GO:0016192","vesicle-mediated transport",1.624410201635736,-5.965478391645846,-0.9721995383167545,5.688750756431566,1,0),

c("GO:0019236","response to pheromone",0.025219102176014033,-1.7164587305328172,7.166972348869542,3.8798410559865624,0.89854486967983

c("GO:0030216","keratinocyte
differentiation",0.01565965880304327,2.870631524248086,-5.555894406225771,3.6729286904427223,0.8564510365523903,0.00656128),

c("GO:0031175","neuron projection
development",0.16085409934476902,1.5150430168690525,-5.442847800684638,4.684495272544229,0.8525153011103631,0.60589399),

c("GO:0035556","intracellular signal
transduction",3.7944870018179575,-3.40324885690256,5.122556372562768,6.0572076090390246,0.7629201502335783,0.62165104));


one.data <- data.frame(revigo.data);

names(one.data) <- revigo.names;

one.data <- one.data [(one.data$plot_X != "null" & one.data$plot_Y != "null"), ];

one.data$plot_X <- as.numeric( as.character(one.data$plot_X) );

one.data$plot_Y <- as.numeric( as.character(one.data$plot_Y) );

one.data$log_size <- as.numeric( as.character(one.data$log_size) );

one.data$frequency <- as.numeric( as.character(one.data$frequency) );

one.data$uniqueness <- as.numeric( as.character(one.data$uniqueness) );

one.data$dispensability <- as.numeric( as.character(one.data$dispensability) );
```

```
one.data2 <- merge(one.data,pos,by="term_ID", no.dups = TRUE)
one.data2$size <- as.numeric( as.character(one.data2$size) );

p1 <- ggplot( data = one.data2 );
p1 <- p1 + geom_point( aes( plot_X, plot_Y, colour = size, size = log_size), alpha = I(0.8) );
p1 <- p1 + scale_colour_gradientn( colours = c("#21130d", "#e28743", "#1e81b0", "#063970"), limits = c( min(one.data2$size), max(one.data2$size)) );
p1 <- p1 + geom_point( aes(plot_X, plot_Y, size = log_size), shape = 21, fill = "transparent", colour = I (alpha ("black", 0.8) ));
p1 <- p1 + scale_size( range=c(1, 20)) + theme_classic();
ex <- one.data2 [ one.data2$dispensability < 0.5, ];
p1 <- p1 + geom_text( data = ex, aes(plot_X, plot_Y, label = description), colour = I(alpha("black", 0.85)), size = 5 );
p1 <- p1 + labs (y = "semantic space x", x = "semantic space y");
p1 <- p1 + theme(legend.key = element_blank()) ;
one.x_range = max(one.data2$plot_X) - min(one.data2$plot_X);
one.y_range = max(one.data2$plot_Y) - min(one.data2$plot_Y);
p1 <- p1 + xlim(min(one.data2$plot_X)-one.x_range/10,max(one.data2$plot_X)+one.x_range/10);
p1 <- p1 + ylim(min(one.data2$plot_Y)-one.y_range/10,max(one.data2$plot_Y)+one.y_range/10);
p1

pdf("GOterms_revigo_pos.pdf", width = 9, height = 5)
p1
dev.off()
```

pos plot: (edited text in grep, direct copy paste from revigo script on white)

```
module load R/4.1.0-gimkl-2020a
R
library(ggplot2)
library(scales)
library(dplyr)
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/")
neg<-read.table("AcTris_GOanalysis_input_neg.txt")
colnames(neg) <- c("term_ID","size")
```

revigo.names <- c("term_ID","description","frequency","plot_X","plot_Y","log_size","uniqueness","dispensability");

revigo.data <- rbind(c("GO:0006355","regulation of DNA-templated transcription",9.968929480711344,-5.103277148905315,4.04894589788018,6.476702827628594,0.8640416765681546,0.3923121),

c("GO:0006368","transcription elongation by RNA polymerase II",0.1035207223719447,6.410578008433285,0.9817442806503419,4.493095406643273,0.8874630150088624,0.35134726),

c("GO:0006383","transcription by RNA polymerase III",0.08139164207953882,6.020018531755554,0.4417229693926238,4.388651717026825,0.8892655786727276,0.02785784),

c("GO:0006418","tRNA aminoacylation for protein translation",0.974318159563009,4.9416175981824395,3.2365672105346737,5.466756438594257,0.8031922083216346,0.29452371),

c("GO:0006468","protein phosphorylation",4.340892012483024,3.7671155645343135,5.026629919063385,6.115633473764006,0.8116879559230749,0.5

c("GO:0006481","C-terminal protein methylation",0.019707620732377095,2.836939534049478,6.136375166869933,3.772761647144032,0.8788514200527298,0.07553559),

c("GO:0006508","proteolysis",5.350747086797883,4.340391192993198,4.621381129396894,6.206468852551757,0.8385598914522587,0.36107106),

c("GO:0006811","monoatomic ion transport",4.401907471506606,1.417259309668573,-5.5669697313354085,6.1216953964406695,0.9128330720872585,0.24276071),

c("GO:0006813","potassium ion transport",0.5758350576254156,2.2267957430773846,-5.056332578289938,5.238354769259176,0.9274434871114909,0.33316827),

c("GO:0006952","defense response",1.0410213876080876,-6.282697186903142,-0.17867634408135113,5.495515198600499,0.8517289273890182,-0),

c("GO:0007165","signal transduction",8.135788134528843,-5.448710111563457,2.614059397284906,6.388453966952047,0.7436184635193851,0.68734

c("GO:0007186","G protein-coupled receptor signaling pathway",1.5868742522745753,-5.426052263772852,2.2658232771768447,5.678597582760593,0.7772834745242388,0.37607124),

c("GO:0007217","tachykinin receptor signaling pathway",0.012227040305859613,-5.914447314504913,1.4759065902766588,3.5654936298688624,0.8464591388039234,0.24126341),

c("GO:0007399","nervous system development",0.5207002861047604,-3.8303527222662974,-5.391863523385115,5.1946447493982895,1,-0),

c("GO:0010923","negative regulation of phosphatase activity",0.0157561180437587,-4.722038694778504,5.532473643381862,3.6755950563867463,0.9167006650185613,-0),

c("GO:0016032","viral process",0.15924422787903558,-5.785716600009443,-4.002633756340671,4.680126929448146,1,0),

c("GO:0016579","protein deubiquitination",0.2889985375448635,3.613427726288738,5.787695202044781,4.938954802333511,0.8525581896776585,0.3

c("GO:0019068","virion assembly",0.041261271761204706,-1.0124562002918525,8.46990078315948,4.093631776828947,1,-0),

c("GO:0019882","antigen processing and presentation",0.027101720460322123,6.044867921939772,-3.2793761916379385,3.911104317804036,0.8859297613741752,-0),

c("GO:0019915","lipid storage",0.03154882407537498,-0.01256790663993268,-5.541341644234026,3.9770831203158528,0.943382470735298,-0),

c("GO:0032259","methylation",3.1286355155207577,0.1775381610862847,-0.8053646996453163,5.973409591592427,0.9772944149668295,-0),

c("GO:0051056","regulation of small GTPase mediated signal transduction",0.13655634922662307,-3.7716464077102314,4.139995571717749,4.613376634694834,0.9120553157536272,0.16480454),

```
c("GO:0055085","transmembrane
transport",13.53545668727302,1.2890064427066246,-5.129239287956978,6.609527179077396,0.8975335395462567,0.5384965));


one.data <- data.frame(revigo.data);

names(one.data) <- revigo.names;

one.data <- one.data [(one.data$plot_X != "null" & one.data$plot_Y != "null"), ];

one.data$plot_X <- as.numeric( as.character(one.data$plot_X) );

one.data$plot_Y <- as.numeric( as.character(one.data$plot_Y) );

one.data$log_size <- as.numeric( as.character(one.data$log_size) );

one.data$frequency <- as.numeric( as.character(one.data$frequency) );

one.data$uniqueness <- as.numeric( as.character(one.data$uniqueness) );

one.data$dispensability <- as.numeric( as.character(one.data$dispensability) );
```

```
one.data2 <- merge(one.data,neg,by="term_ID", no.dups = TRUE)
one.data2$size <- as.numeric( as.character(one.data2$size) );

p1 <- ggplot( data = one.data2 );
p1 <- p1 + geom_point( aes( plot_X, plot_Y, colour = size, size = log_size), alpha = I(0.8) );
p1 <- p1 + scale_colour_gradientn( colours = c("#21130d", "#e28743", "#1e81b0", "#063970"), limits = c( min(one.data2$size), max(one.data2$size)) );
p1 <- p1 + geom_point( aes(plot_X, plot_Y, size = log_size), shape = 21, fill = "transparent", colour = I (alpha ("black", 0.8) ));
p1 <- p1 + scale_size( range=c(1, 20)) + theme_classic();
ex <- one.data2 [ one.data2$dispensability < 0.5, ];
p1 <- p1 + geom_text( data = ex, aes(plot_X, plot_Y, label = description), colour = I(alpha("black", 0.85)), size = 5 );
p1 <- p1 + labs (y = "semantic space x", x = "semantic space y");
p1 <- p1 + theme(legend.key = element_blank()) ;
one.x_range = max(one.data2$plot_X) - min(one.data2$plot_X);
one.y_range = max(one.data2$plot_Y) - min(one.data2$plot_Y);
p1 <- p1 + xlim(min(one.data2$plot_X)-one.x_range/10,max(one.data2$plot_X)+one.x_range/10);
p1 <- p1 + ylim(min(one.data2$plot_Y)-one.y_range/10,max(one.data2$plot_Y)+one.y_range/10);
p1

pdf("GOterms_revigo_neg.pdf", width = 9, height = 5)
p1
dev.off()
```

Formatting supmat table 6:

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/AcTris_pos
for OGTERM in
$(cat /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/AcTris_GOterms_pos_OGte
do
cut -f 14 ${OGTERM}_single.fa.tsv | grep -v "-" | sed 's/|/\n/g' | sort | uniq | tr '\n' '|' | awk -v variable1="$OGTERM" '{print $0, variable1}' >  ${OGTERM}.goterms_oneline
done
```

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/AcTris_neg
for OGTERM in
$(cat /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/AcTris_GOterms_neg_OGte
do
cut -f 14 ${OGTERM}_single.fa.tsv | grep -v "-" | sed 's/|/\n/g' | sort | uniq | tr '\n' '|' | awk -v variable1="$OGTERM" '{print $0, variable1}' >  ${OGTERM}.goterms_oneline
done

cd ..

cat */*goterms_oneline > all_goterms_oneline
cp all_goterms_oneline all_goterms_oneline.edit

module load R/4.1.0-gimkl-2020a
R
library(data.table)
library(dplyr)
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/gene_family/orthofinder_run4/OrthoFinder/Results_May14/Cafe5_out_renamed_k4/")
table <- read.table("AcTris_GOterms_all.txt")
goterms <- read.table("all_goterms_oneline.edit")
```

```
colnames(goterms) <- c("terms","GO")
table2<-merge(table,goterms,by="GO",all.x=TRUE )
write.table(table2,"At1_tables6.txt" )
```