Osedax paper analyses

Chapter 3

k-mer analyses

kraken2

```
module load anaconda3
conda create -n kraken2_env2 -c bioconda kraken2
conda install -c conda-forge libiconv
```

- kraken2 v 2.1.0

krakentools

```
module load anaconda3
conda activate kraken2_env
conda install -c bioconda krakentools
```

- krakentools v 0.1

Kraken & KrakenTools

kraken2_pacbio_bbmap_standard_db_lessCores_v1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 6
#$ -l h_vmem=10G
#$ -l h_rt=240:0:0
#$ -l highmem

species=$1
```

```
kraken_database=/data/SBCS-MartinDuranLab/03-Giacomo/db/kraken2_db_standard_Oct2020
pacbio_corrected="$species"_pacbio_corrected_bbmap.fasta
output_pacbio="$species"_pacbio_bbmap_output.kraken

module load anaconda3
conda activate kraken2_env

cd /data/scratch/btx654/btx604-scratch/$species/kraken2/pacbio_illumina_2

kraken2 --threads 6 --output $output_pacbio --report report_kraken2_pacbio_bbmap --db $kraken_database $pacbio_corrected
```

Krakentools extract oasisia

```
module load anaconda3
conda activate kraken2_env
extract_kraken_reads.py -k oasisia_pacbio_bbmap_output.kraken -r report_kraken2_pacbio_bbmap -s oasisia_pacbio_corrected_bbmap.fasta --taxid 2 --exclude --include-children -o oasisia_nonBacteria_pacbio_corrected_bbmap.fasta
```

Illumina nonBacterial

mapping_illumina_nonBacterial_definitive.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654
#$ -j y
#$ -o /data/scratch/btx654
#$ -pe smp 20
#$ -l h_vmem=34G
#$ -l h_rt=72:0:0
#$ -l highmem
```

```
9
10 species=$1
11 #bwa index step1 variables
12 pacbio_corrected_nonBacteria="$species"_nonBacteria_pacbio_corre
   cted_bbmap.fasta
13 bwa_prefix="$species"_nonBacteria_pacbio_corrected_bbmap.fasta
14 #bwa mem step1 variables
15 R1_cleaned="$species"_R1_cleaned.fastq.gz
16 R2_cleaned="$species"_R2_cleaned.fastq.gz
17 alignment_sam="$species"_alignment_HIGHmem.sam
18 #samtools view step1 variables
19 alignment_bam="$1"_alignment_HIGHmem.bam
20 #samtools sort index step1 variables
21 alignment_sorted="$species"_sorted_HIGHmem.bam
22 alignment_mapped="$species"_sorted_mapped_HIGHmem.bam
23 alignment_mapped_sorted="$species"_sorted_mapped_sorted_HIGHmem.
   bam
24 R1_mapped="$species"_R1_mapped_HIGHmem.fastq
25 R2_mapped="$species"_R2_mapped_HIGHmem.fastq
26
27 module load bwa
28 module load samtools/1.9
29
30 cd /data/scratch/btx654/btx604-scratch/$species/kraken2/pacbio_i
   llumina
31
32 echo 'BWA INDEX STEP1_____________________________________________
   ____________________'
33 if [ -e /data/scratch/btx654/btx604-scratch/$species/kraken2/pac
   bio_illumina/*.ann ]
34 then
```

```
  echo "/data/scratch/btx654/btx604-scratch/$species/kraken2/pacbio_illumina/*.ann found."
else
  bwa index -p $bwa_prefix -a bwtsw $pacbio_corrected_nonBacteria
fi

echo 'BWA MEM STEP1_____________________________________________________'
bwa mem -t 20 -M $pacbio_corrected_nonBacteria $R1_cleaned $R2_cleaned > $alignment_sam

samtools view -@ 20 -S -b -h $alignment_sam -o $alignment_bam
samtools sort -@ 20 $alignment_bam -o $alignment_sorted
samtools index $alignment_sorted
samtools view -@ 20 -b -F 4 $alignment_sorted > $alignment_mapped
samtools sort -@ 20 -n $alignment_mapped -o $alignment_mapped_sorted
samtools fastq -1 illumina_R1_mapped_samtools.fq -2 illumina_R2_mapped_samtools.fq -n $alignment_mapped_sorted
```

KAT

First let's set up a directory containing illumina reads for Riftia and the nonBacterial illumina reads obtained by mapping illumina reads on the non-bacterial pacbio generated by kraken2 (29/10/20). For Riftia the illumina needs to be cleaned using fastp

/data/home/btx654/scripts/kmer_analyses/Dec2020/fastp_riftia_v1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
```

```
#$ -j y
#$ -pe smp 4
#$ -l h_vmem=5G
#$ -l h_rt=6:0:0

species=riftia
R1=RP_6_12_19_S110_L002_R1_001.fastq.gz
R2=RP_6_12_19_S110_L002_R2_001.fastq.gz
R1_cleaned="$species"_R1.fq.gz
R2_cleaned="$species"_R2.fq.gz

cd /data/scratch/btx654/btx604-scratch/$species/kmer_Dec2020/

module load anaconda3
conda activate fastp

fastp -i $R1 -I $R2 -o $R1_cleaned -O $R2_cleaned -w 4

gzip -d $R1_cleaned
gzip -d $R2_cleaned
```

fastp_osedax_oasisia_universal_v1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 4
#$ -l h_vmem=5G
#$ -l h_rt=6:0:0

species=$1
```

```
R1=illumina_R1_nonBacteria_kraken2.fq
R2=illumina_R2_nonBacteria_kraken2.fq
R1_cleaned="$species"_R1.fq
R2_cleaned="$species"_R2.fq

cd /data/scratch/btx654/btx604-scratch/$species/kmer_Dec2020/

module load anaconda3
conda activate fastp

fastp -i $R1 -I $R2 -o $R1_cleaned -O $R2_cleaned -w 4
```

After this we want to have our folders:

```
/data/scratch/btx654/btx604-scratch/$species/kmer_Dec2020/
```

containing illumina data named as “$species”_R1.fq and “$species”_R2.fq (uncompressed)

- for Riftia this 2 files are the fastp cleaned illumina data we had from 2019 (not the august 2020 data)
- for Osedax and Oasisia this 2 files are the nonBacterial illumina reads obtained by mapping illumina reads on the non-bacterial pacbio generated by kraken2 (29/10/20). Which have been cleaned with the fastp cleaning step

WILL BASE ALL THESE ANALYSES ON 21-MER!!!!!

kat_gcp_universal_v1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 20
#$ -l h_vmem=10G
#$ -l h_rt=24:0:0
```

```
#$ -l highmem

species=$1
R1="$species"_R1.fq
R2="$species"_R2.fq

echo "Working on "$species

cd /data/scratch/btx654/btx604-scratch/$species/kmer_Dec2020/
mkdir -p kat
cd kat

module load anaconda3
conda activate KAT

kat gcp -m 21 -p pdf -v -t 20 ../$R1 ../$R2
```

kat_hist_universal_v1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 20
#$ -l h_vmem=10G
#$ -l h_rt=72:0:0
#$ -l highmem


species=$1
R1="$species"_R1.fq
R2="$species"_R2.fq
```

```

echo "Working on "$species

cd /data/scratch/btx654/btx604-scratch/$species/kmer_Dec2020/kat

module load anaconda3
conda activate KAT

kat hist -m 21 -t 20 -o "kat_21mer_illumina.hist" ../$R1 ../$R2
```

Before the next step cp the purged genome into the working directory and name it "$species".fa

kat_comp_universal_v1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 20
#$ -l h_vmem=10G
#$ -l h_rt=24:0:0
#$ -l highmem

species=$1
purged="$species".fa

echo "Working on "$species

module load anaconda3
conda activate KAT

```

```
cd /data/scratch/btx654/btx604-scratch/$species/kmer_Dec2020/kat

kat comp -m 21 -p pdf -o 21mer_vs_assembly -v -t 20 '../*_R1.fq ../*_R2.fq' ../$purged
```

merqury_universal.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 10
#$ -l h_vmem=20G
#$ -l h_rt=100:0:0
#$ -l highmem

species=$1
purged="$species".fa
R1="$species"_R1.fq
R2="$species"_R2.fq
meryl_R1="$species"_R1.meryl
meryl_R2="$species"_R2.meryl
meryl_final="$species".meryl
merqury_output="$species"_merqury_nonBacteria

echo "Working on "$species

cd /data/scratch/btx654/btx604-scratch/$species/kmer_Dec2020/
mkdir -p merqury
cd merqury

```

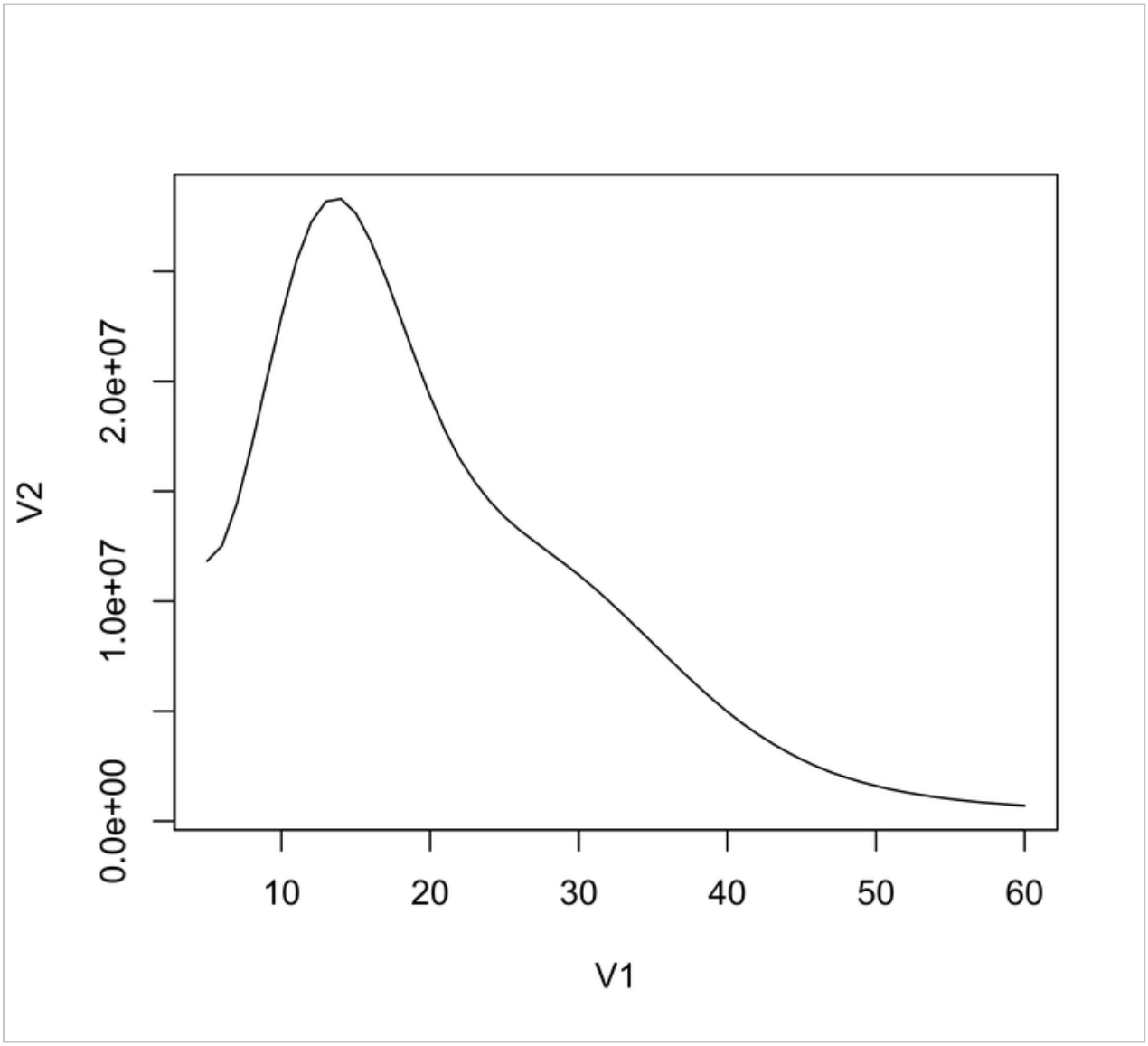
```
module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/merqury_env

meryl k=21 threads=10 count output $meryl_R1 ../$R1
meryl k=21 threads=10 count output $meryl_R2 ../$R2

meryl union-sum output $meryl_final *_R*.meryl

merqury.sh $meryl_final ../$purged $merqury_output > Merqury.log

#spectra_cn="$species"_merqury_nonBacteria."$species".spectra-cn.hist
#only_hist="$species"_merqury_nonBacteria.dist_only.hist
#output_plot="$species".spectra-cn

#Rscript $MERQURY/plot/plot_spectra_cn.R -f $spectra_cn -o $output_plot -z $only_his -m (kmer_multiplicity) and -n (Count)
```

Genome size estimation -GenomeScope
Riftia - http://qb.cshl.edu/genomescope/genomescope2.0/analysis.php?code=XcAsfjFyAhGckInVGStm
Osedax - http://qb.cshl.edu/genomescope/genomescope2.0/analysis.php?code=aUtpCdYPBD5MXkT1h0cb
Oasisia - http://qb.cshl.edu/genomescope/genomescope2.0/analysis.php?code=5hnEsmQKmAf6j8u65ub7

Genome size estimation - tutorial

R based calculations
Using the hist file produced with the script (kat_hist_universal_kraken_standard_illumina_v1.sh 28/10/20) and cropping out of the plot the first area of erroneous kmers as suggested by the tutorial

```
plot(oasisia_21mer[5:60,], type="l")
# points(oasisia_21mer[5:60,])
# export as a pdf 5 x 5.5 inches in landscape orientation
```

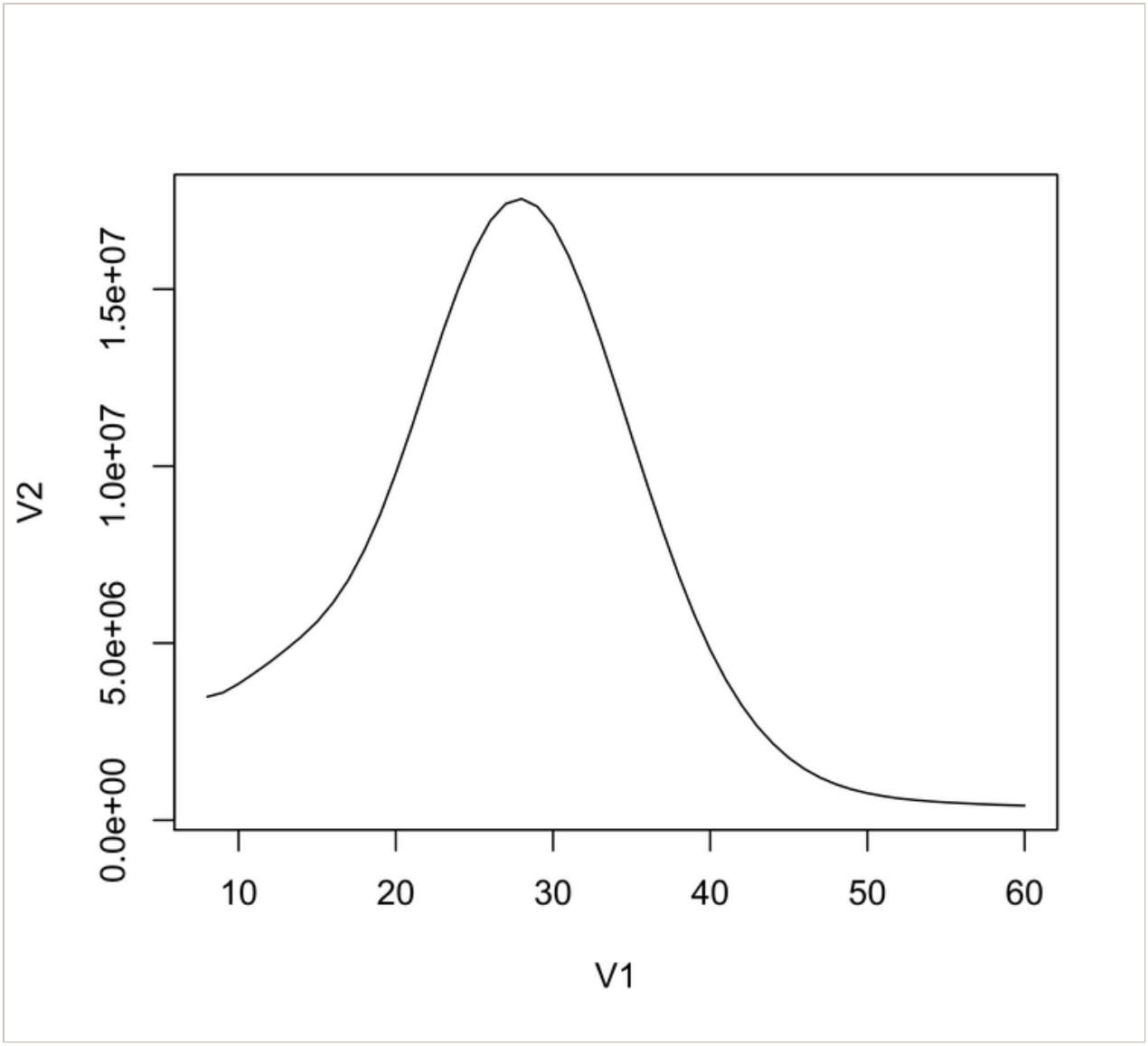


PDF oasisia_21mer_spectra • PDF document

```
#here I am assuming I have 10000 data points (which is the lengt
h of the hist file)
# 28 is the peak for the homozygous and the result is similar to
my assembly size
```

```
> sum(as.numeric(oasisia_21mer[5:10000,1]*oasisia_21mer[5:10000,2]))/28
[1] 785764774
```

```
plot(riftia_21mer[8:60,], type="l")
# points(oasisia_21mer[5:60,])
# export as a pdf 5 x 5.5 inches in landscape orientation
```

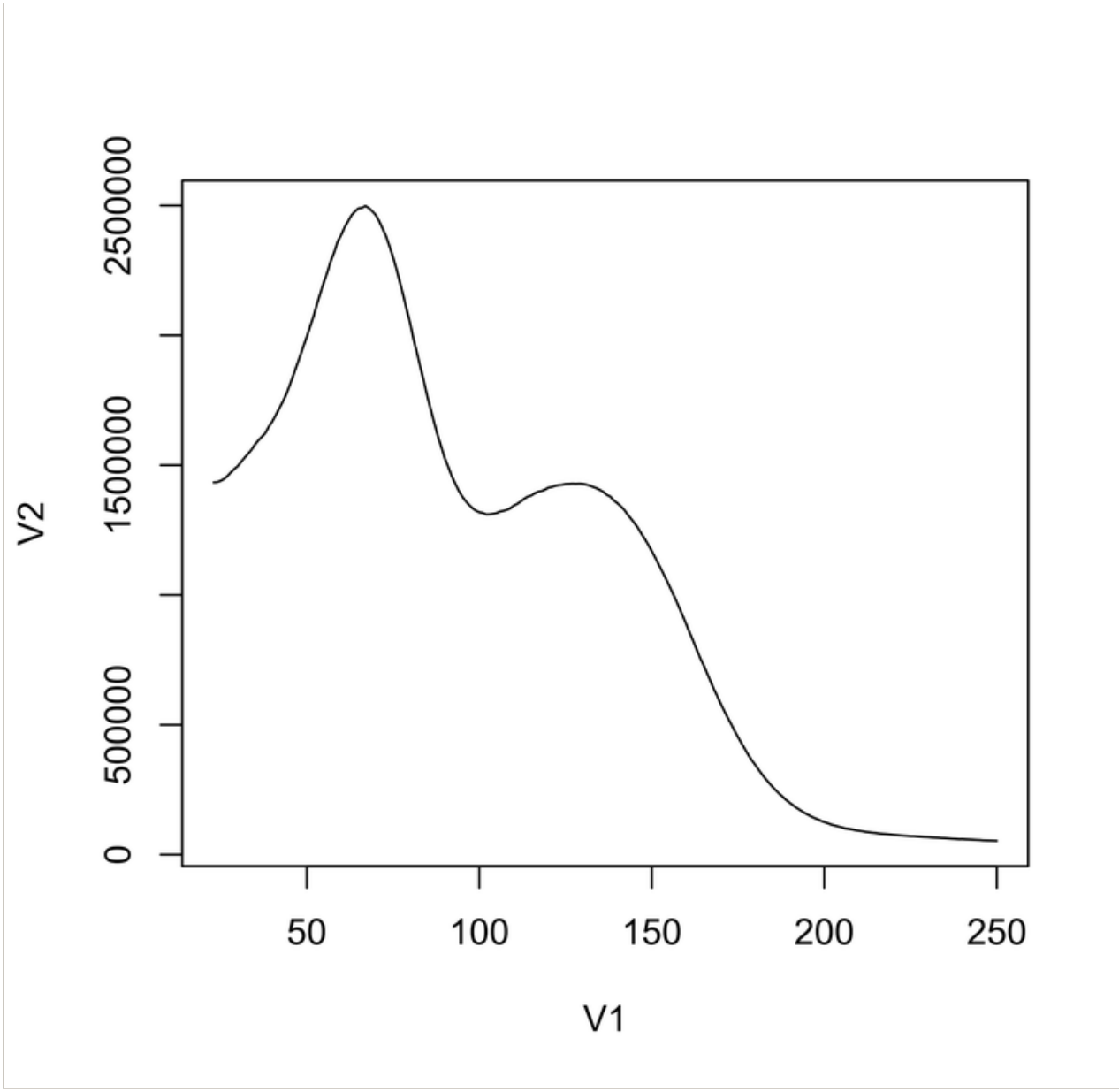


PDF riftia_21mer_spectra • PDF document

```
#here I am assuming I have 10000 data points (which is the lengt
```

```
h of the hist file)
# 27 is the peak for the homozygous and the result is similar to my assembly size
> sum(as.numeric(riftia_21mer[8:10000,1]*riftia_21mer[8:10000,2]))/28
[1] 532856757
```

```
plot(osedax_21mer[23:250,], type="l")
# points(oasisia_21mer[5:60,])
# export as a pdf 5 x 5.5 inches in landscape orientation
```



PDF osedax_21mer_spectra • PDF document

```
#here I am assuming I have 10000 data points (which is the length of the hist file)
# 127 is the peak for the homozygous and the result is similar to my assembly size
> sum(as.numeric(osedax_21mer[23:10000,1]*osedax_21mer[23:10000,2]))/127
[1] 239329736
```

Missing Busco in Osedax

files I will use:

```
/Users/giacomo/Dropbox/11-Siboglinids/00-Data/Osedax/Annotation/New_annotation_Dec2020/step7/missing_busco_list.tsv
/data/SBCS-MartinDuranLab/03-Giacomo/db/datasets/metazoa_odb10/ancestral
```

busco_universal_v1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -pe smp 4
#$ -l h_vmem=20G
#$ -l h_rt=48:0:0
#$ -j y
#$ -l highmem

species=osedax
annotation_gtf="$species".AGAT.noSTOP.filt.noTE.gtf
annotation_fa="$species"_annotation.prot.fa
species_softmasked="$species"_softmasked.fa
output_busco="$species"_busco_annotation

echo "Working on "$species

cd /data/scratch/btx654/missing_busco_osedax
cp /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/annotation/New_annotation_Dec2020/step6/$annotation_gtf ./
cp /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/annotation/softmasking/$species_softmasked ./

```

```
module load anaconda3
source activate augustus

gffread -E $annotation_gtf -g $species_softmasked -y $annotation_fa

conda deactivate
source activate busco_env
#export BUSCO_CONFIG_FILE="/data/home/btx654/.conda/envs/busco_env/busco/config/myconfig.ini"
#export AUGUSTUS_CONFIG_PATH=/data/SBCS-MartinDuranLab/02-Chema/src/Augustus/config/

busco -i osedax_mRNA.fa -m proteins -o output_busco -c 4 -l /data/SBCS-MartinDuranLab/03-Giacomo/db/datasets/metazoa_odb10
```

```
sed 's/a//g' missing_busco_list.tsv | sed 's/^/>/' > missing_busco_list_OK.txt

cp /data/SBCS-MartinDuranLab/03-Giacomo/db/datasets/metazoa_odb10/ancestral ./
mv ancestral ancestral.fa

module load seqtk
seqtk subseq /data/scratch/btx654/missing_busco_osedax/busco_downloads/lineages/metazoa_odb10/ancestral missing_busco_list.txt
 > missing_busco.fa
```

```
```

```
module load anaconda3
source activate augustus

gffread -w osedax_mRNA.fa -g /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/annotation/softmasking/osedax_softmasked.fa /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/annotation/New_annotation_Dec2020/step6/osedax_annotation_v101220.gff3
gffread -y osedax_protein.fa -g /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/annotation/softmasking/osedax_softmasked.fa /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/annotation/New_annotation_Dec2020/step6/osedax_annotation_v101220.gff3
```

blastp.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

module load blast+
makeblastdb -in ../osedax_protein.fa -dbtype prot -out osedax_prot
blastp -db osedax_prot -query ../missing_busco.fa -out osedax_prot_blastp_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
blastp -db osedax_prot -query ../missing_busco.fa -out osedax_prot_blastp_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads
```

```
8 -html
```

blastp.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

module load blast+
makeblastdb -in ../osedax_annotation.prot.fa -dbtype prot -out osedax_prot
blastp -db osedax_prot -query ../missing_busco.fa -out osedax_prot_blastp_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
blastp -db osedax_prot -query ../missing_busco.fa -out osedax_prot_blastp_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -html
```

panther.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -pe smp 10
#$ -l h_vmem=5G
#$ -l h_rt=80:00:0
#$ -l highmem

```

```
module load perl
module load hmmer/

export PERL5LIB=/data/SBCS-MartinDuranLab/03-Giacomo/src/hmmscoring/lib/

perl /data/SBCS-MartinDuranLab/03-Giacomo/src/hmmscoring/pantherScore2.2.pl -l /data/SBCS-MartinDuranLab/03-Giacomo/src/hmmscoring/PANTHER15.0/ -D B -n -o panther_output -i ../missing_busco.fa -c 10 -V -s
```

There are 23 missing Busco Panther IDs matching with annotations in Osedax: only 20 are uniq

Let's try to search for the missing proteins in Oasisia and use its proteins to blast against Osedax:

```
#53 different Panther IDs found in Oasisia belonging to the missing Busco in osedax
grep -f panther_ids_list /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/annotation/New_annotation_Dec2020/step9/oasisia_annotation_Jan2021_TrinoPantherKO.xls | cut -f 19 | sort | uniq | wc -l
cut -f 18,19 /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/annotation/New_annotation_Dec2020/step9/oasisia_annotation_Jan2021_TrinoPantherKO.xls | grep -f panther_ids_list | sort -u -k2,2 | cut -f 1 > oasisia_geneIDs_missingBUSCO_osedax.txt

module load seqtk
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Oalv.fa oasisia_geneIDs_missingBUSCO_osedax_OK.txt > oasisia_proteins_missingBUSCO_osedax.fa
```

blastp.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

module load blast+

blastp -db ../../BLAST/osedax_prot -query oasisia_proteins_missingBUSCO_osedax.fa -out Oalv_VS_Ofra_missingBUSCO_blastp_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
blastp -db ../../BLAST/osedax_prot -query oasisia_proteins_missingBUSCO_osedax.fa -out Oalv_VS_Ofra_missingBUSCO_blastp_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -html
```

- 22 Oasisia proteins are matching with Osedax

RECAP
62 missing BUSCO in Osedax annotation
954 tot BUSCO metazoadb10 (892 found in osedax)
BUSCO C (91.6%) + BUSCO F (1.9%) = 93.5%

METHOD	NUMBER MATCHING OSEDAX PROTEINS	percentage of missing BUSCO found
BLAST missing BUSCO (63 sequences from ancestral in metazoa_db10)	22	2.3%
BLAST missing BUSCO (53 sequences from Oasisia, obtained with PantherIDs of the missing	22	2.3%

BUSCO)		
PANTHER missing BUSCO searching directly in the annotations of Osedax	23 (actually there are just 20 unique ones)	2.4% (2.1%)
summing the three methods (see below for info)	26	2.7%

If I check the number of unique genes identified in the previous three methods:

```
cat BLAST/geneIDs_found_BLASTmetazoadb10 Panther/geneIDs_found_Panther Panther/oasisia_blast/geneIDs_found_BLASToasisia | sort | uniq | wc -l
#52 but there was probably an error with panther IDs (had to crop the last part of two of them in oasisia to match the ones in busco missing genes)
cat BLAST/BUSCO_IDs_blast_metazoadb10_osedax Panther/BUSCO_IDs_panther_osedax Panther/oasisia_blast/BUSCO_IDs_blast_oasisia_osedax | sort -u | wc -l
#54 different Buscos have a match in osedax
```

- 52 hits, 3 of these are isoforms: 49 different genes

I have checked the output of busco and there are not repeated genes for different matches with busco genes. so is actually quite accurate to look at the different buscos
when i check the top blast match oasisia_vs_osedax I end up having 22 uniq osedax proteins

```
sort -u -k1,1 Oalv_VS_Ofra_missingBUSCO_blastp_out | cut -f 2 | sort -u > uniq_geneIDs_BLAST_oasisia #22 unique proteins of osedax

sort -u -k1,1 osedax_prot_blastp_out | cut -f 2 | sort -u > uniq_geneIDs_BLAST_metazoadb10 #22 unique proteins of osedax
sort -u -k1,1 osedax_prot_blastp_out | cut -f 1,2 > uniq_geneIDs_BLAST_metazoadb10_vs_BUSCO
```

```

cut -f 18,19 23_missingBUSCO_found_osedax.xls | grep -f panther_ids_list_osedax | sort -k2,2 | cut -f 1 > 23_geneIDs_osedax_Panther
cut -f 18,19 23_missingBUSCO_found_osedax.xls | grep -f panther_ids_list_osedax | sort -k2,2 | cut -f 2 > 23_pantherIDs_osedax_Panther
nano 23_geneIDs_osedax_Panther_edited > uniq_geneIDs_panther_vs_BUSCO

 grep -f panther_IDs_BUSCO_oasisia ../panther_output | cut -f1,2 | sort -k2,2 > 53_BUSCO_vs_panther_oasisia

```

```
awk -F"|" 'NR==FNR{a[$1]=$0;next} ($1 in a){b=$1;$1="";print a[b]  $0}' OFS="|" file1 file2
awk 'NR==FNR{a[$2]=$0;next} ($1 in a){b=$2;$1="";print a[b]  $0}' 62_metazoadb10_vs_panther 63_oasisia_geneIDs_vs_panther
```

```

join -1 2 -2 2 62_metazoadb10_vs_panther 63_oasisia_geneIDs_vs_panther
```

```
while read line; do
echo $line
annotations=$(cut -f 1,2 63_oasisia_geneIDs_vs_panther | fgrep $line)
cat $annotations
kallisto_body=$(cut -f 2 <<< $annotations)
cat $kallisto_body
kallisto_roots=$(cut -f 3 <<< $annotations)
cat $kallisto_roots
echo $kallisto_body$'\t'$kallisto_roots >> oasisia_geneIDs_missi
```

```
ngBUSCO_found_osedax_vs_panther
done < oasisia_geneIDs_missingBUSCO_found_osedax
```

```
cut -f 1 oasisia_geneIDs_missingBUSCO_found_osedax_vs_panther | sed -e 's/ /\t/g' | sort -k2,2 > oasisia_geneIDs_missingBUSCO_found_osedax_vs_panther_OK
cut -f 2 oasisia_geneIDs_missingBUSCO_found_osedax_vs_panther_OK > oasisia_geneIDs_missingBUSCO_found_osedax_vs_panther_OK_pantherIDs
grep -f oasisia_geneIDs_missingBUSCO_found_osedax_vs_panther_OK_pantherIDs 62_metazoadb10_vs_panther | sort -k2,2 > missingBUSCO_found_osedax_vs_panther_OK

paste missingBUSCO_found_osedax_vs_panther_OK oasisia_geneIDs_missingBUSCO_found_osedax_vs_panther_OK > link_oasisia_geneIDs_BUSCO

cut -f 1 Oalv_VS_Ofra_missingBUSCO_blastp_out | sed 's/Oalv_//' > blast_output_firstColumn
```

```
while read line; do
echo $line
annotations=$(cut -f 1,3 link_oasisia_geneIDs_BUSCO | fgrep $line | cut -f 1)
cat $annotations
echo $annotations >> blast_output_firstColumn_vs_metazoadb10
done < blast_output_firstColumn
```

```
cut -f 2 Oalv_VS_Ofra_missingBUSCO_blastp_out > blast_output_firstColumn_osedax_geneIDs
paste blast_output_firstColumn_vs_metazoadb10 blast_output_firstColumn_osedax_geneIDs > uniq_geneIDs_BLAST_oasisia_vs_BUSCO

```

```
cat BLAST/uniq_geneIDs_BLAST_metazoadb10_vs_BUSCO Panther/uniq_geneIDs_panther_vs_BUSCO Panther/oasisia_blast/uniq_geneIDs_BLAST_oasisia_vs_BUSCO > ALL_matches_osedax_vs_metazoadb10

sort -u -k1,1 ALL_matches_osedax_vs_metazoadb10 | sort -u -k1,2 | wc -l #26
```

- 26 BUSCO found in osedax (double checked for duplicates and stuff, so this is final!)

```
while read line; do
echo $line
annotations=$(cut -f 1,2 uniq_geneIDs_BLAST_oasisia_vs_BUSCO | fgrep $line | head -1)
cat $annotations
echo $annotations >> blast_output_firstColumn_vs_metazoadb10
done < blast_output_firstColumn
```

```
while read line; do
echo $line
echo "blast"
blast=$(cut -f 1,2 uniq_geneIDs_BLAST_metazoadb10_vs_BUSCO | fgrep $line)
echo $blast
echo "blast_oasisia"
blast_oasisia=$(cut -f 1,2 uniq_geneIDs_BLAST_oasisia_vs_BUSCO | fgrep $line)
echo $blast_oasisia
echo "panther"
panther=$(cut -f 1,2 uniq_geneIDs_panther_vs_BUSCO | fgrep $line)
echo $panther
```

```
echo
"------------------------------------------------------------------
------------------------------------------------------------------
--"
done < 26_BUSCO_IDs_found_osedax
```

```
while read line; do
echo $line
echo "Panther"
Panther=$(cut -f 1,2,3 ../Panther/panther_output | fgrep $line)
echo $Panther
echo
"------------------------------------------------------------------
------------------------------------------------------------------
--"
done < 62_BUSCO_IDs
```

Comparison with Riftia genomes

- [x] ~~1 - Table our Riftia vs the other Riftia (genome stats)~~
- [x] ~~2 - Assembly vs Assembly using Minimap2 and plot~~
- [x] ~~3 - Annotation vs Annotation BBH~~
- [x] ~~4 - our annotation vs previous other transcriptome 2019 BBH~~
- [x] ~~5 - PFAM barplot our annotations, riftia ann and previous transcriptome~~

1 - Table our Riftia vs the other Riftia (genome stats)

i am using:

```
OUR riftia_softmasked.fa
THEIR 4.2_RIFPA_polished_softMasked_purged_genome_v1.fasta
```

```
ANNOTATION_GFF3 RIFPA_final_gene_models_AUGUSTUS_v1.gff3
TOT_PROTEOME RIFPA_final_gene_models_AUGUSTUS_v1.prot.fasta
```

stats.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/riftia_oliveira_2022
#$ -o /data/scratch/btx654/riftia_oliveira_2022
#$ -j y
#$ -pe smp 4
#$ -l h_vmem=5G
#$ -l h_rt=24:0:0
#$ -l highmem

module load anaconda3
conda activate agat_env

agat_convert_sp_gxf2gxf.pl --gff RIFPA_final_gene_models_AUGUSTUS_v1.gff3 --merge_loci -o RIFPA_final_gene_models_AUGUSTUS_v1_lociMerged.gff
agat_sp_keep_longest_isoform.pl --gff RIFPA_final_gene_models_AUGUSTUS_v1_lociMerged.gff -o RIFPA_final_gene_models_AUGUSTUS_v1_lociMerged_longestIsoform.gff

agat_convert_sp_gxf2gxf.pl -g $final_pasa_gtf -o $final_pasa_gff3
agat_sq_stat_basic.pl -i RIFPA_final_gene_models_AUGUSTUS_v1.gff3 -g 4.2_RIFPA_polished_softMasked_purged_genome_v1.fasta

conda deactivate
source activate augustus
```

```

gffread -E RIFPA_final_gene_models_AUGUSTUS_v1_lociMerged_longestIsoform.gff -g 4.2_RIFPA_polished_softMasked_purged_genome_v1.fasta -y RIFPA_final_gene_models_AUGUSTUS_v1.prot.longestIsoform.fasta

conda deactivate

source activate quast


quast ./4.2_RIFPA_polished_softMasked_purged_genome_v1.fasta -o ./quast_4.2_softmasked --eukaryote


conda deactivate
source activate busco_env
#export BUSCO_CONFIG_FILE="/data/home/btx654/.conda/envs/busco_env/busco/config/myconfig.ini"
#export AUGUSTUS_CONFIG_PATH=/data/SBCS-MartinDuranLab/02-Chema/src/Augustus/config/

#busco -i RIFPA_final_gene_models_AUGUSTUS_v1.prot.longestIsoform.fasta -m proteins -o busco_longest_isoform -c 4 -l metazoa_odb10
busco -i RIFPA_final_gene_models_AUGUSTUS_v1.prot.fasta -m proteins -o busco_gene_models -c 4 -l metazoa_odb10
busco -i 4.2_RIFPA_polished_softMasked_purged_genome_v1.fasta -m DNA -o busco_genome -c 4 -l metazoa_odb10
```

```
#!/bin/bash
```

```
#$ -wd /data/scratch/btx654/riftia_oliveira_2022
#$ -o /data/scratch/btx654/riftia_oliveira_2022
#$ -j y
#$ -pe smp 4
#$ -l h_vmem=5G
#$ -l h_rt=24:0:0
#$ -l highmem


module load anaconda3
source activate augustus

gffread -E RIFPA_final_gene_models_AUGUSTUS_v1.gff3 -g 4.2_RIFPA_polished_softMasked_purged_genome_v1.fasta -y RIFPA_final_gene_models_AUGUSTUS_v1_protein_giacomoGFFread.fasta

conda deactivate
source activate busco_env

busco -i RIFPA_final_gene_models_AUGUSTUS_v1_protein_giacomoGFFread.fasta -m proteins -o busco_gene_models_giacomoGFFread -c 4 -l metazoa_odb10
```

	O***UR R. pachyptila.***	***OLIVEIRA 22***
Genome Size (Mb)	554	560
Number of contigs	918	447
Contig N50 (Kb)	1,424	2,870
GC content (%)	41.05	40.94
Repeats (%)	27.87	29.9

Number of genes	37,037	25977 (they say 25,984)
Number of transcripts	38,179	58020
Mean gene size (bp)	8311.47	15432.63
Mean transcript size (bp)[CM1] [CM2]	8889.27	15779.25
Gene density (per Mb)	66.85	46.39
N's	4,071	0
Busco assembly (%)	95.6	96.7
Busco annotation (%)	96.8	97.7

	Complete	**Single**	**Duplicated**	**Fragmented**	**Missing**
R. pachyptila assembly	95.6%	94.5%	1.1%	1%	3.4%
R. pachyptila annotation	96.8%	96.4%	0.4%	1%	2.2%
OLIVEIRA 22 assembly	96.7%	96.1%	0.6%	1.5%	1.8%
OLIVEIRA 22 annotation RIFPA_final_gene_models_AUGUSTUS_v1.prot.fasta	97.7%	42.7%	55.0%	1.7%	0.6%

```
awk '/^>/ {printf("\n%s\n",$0);next; } { printf("%s",$0);}  END {printf("\n");}' < 4.2_RIFPA_polished_softMasked_purged_genome_v1.fasta | tail -n +2 > 4.2_RIFPA_polished_softMasked_purged_genome_v1_SINGLE.fasta
```

stats.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/riftia_oliveira_2022_SINGLE
#$ -o /data/scratch/btx654/riftia_oliveira_2022_SINGLE
#$ -j y
#$ -pe smp 4
#$ -l h_vmem=5G
#$ -l h_rt=24:0:0
#$ -l highmem

module load anaconda3
conda activate agat_env



agat_sq_stat_basic.pl -i RIFPA_final_gene_models_AUGUSTUS_v1.gff3 -g 4.2_RIFPA_polished_softMasked_purged_genome_v1_SINGLE.fasta

conda deactivate
source activate augustus


gffread -y RIFPA_final_gene_models_AUGUSTUS_v1.prot.longestIsoform.fasta -g 4.2_RIFPA_polished_softMasked_purged_genome_v1_SINGLE.fasta RIFPA_final_gene_models_AUGUSTUS_v1_lociMerged_longestIs
```

```
oform.gff
gffread RIFPA_final_gene_models_AUGUSTUS_v1.gff3 -y RIFPA_final_gene_models_AUGUSTUS_v1_protein_giacomoGFFread.fasta -g 4.2_RIFPA_polished_softMasked_purged_genome_v1_SINGLE.fasta

conda deactivate
source activate busco_env
#export BUSCO_CONFIG_FILE="/data/home/btx654/.conda/envs/busco_env/busco/config/myconfig.ini"
#export AUGUSTUS_CONFIG_PATH=/data/SBCS-MartinDuranLab/02-Chema/src/Augustus/config/

#busco -i RIFPA_final_gene_models_AUGUSTUS_v1.prot.longestIsoform.fasta -m proteins -o busco_longest_isoform -c 4 -l metazoa_odb10
busco -i RIFPA_final_gene_models_AUGUSTUS_v1.prot.fasta -m proteins -o busco_gene_models -c 4 -l metazoa_odb10
busco -i 4.2_RIFPA_polished_softMasked_purged_genome_v1_SINGLE.fasta -m genome -o busco_genome -c 4 -l metazoa_odb10
busco -i RIFPA_final_gene_models_AUGUSTUS_v1_protein_giacomoGFFread.fasta -m proteins -o busco_gene_models_giacomoGFFread -c 4 -l metazoa_odb10
```

stats.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/riftia_oliveira_2022
#$ -o /data/scratch/btx654/riftia_oliveira_2022
#$ -j y
#$ -pe smp 4
#$ -l h_vmem=5G
#$ -l h_rt=24:0:0
```

```
#$ -l highmem

module load anaconda3
conda activate agat_env

agat_convert_sp_gxf2gxf.pl --gff RIFPA_final_gene_models_AUGUSTUS_v1.gff3 --merge_loci -o RIFPA_final_gene_models_AUGUSTUS_v1_AGAT.gff3
agat_convert_sp_gxf2gxf.pl --gff RIFPA_final_gene_models_AUGUSTUS_v1_AGAT.gff3 --merge_loci -o RIFPA_final_gene_models_AUGUSTUS_v1_AGAT_lociMerged.gff
agat_sp_keep_longest_isoform.pl --gff RIFPA_final_gene_models_AUGUSTUS_v1_AGAT_lociMerged.gff -o RIFPA_final_gene_models_AUGUSTUS_v1_AGAT_lociMerged_longestIsoform.gff

agat_sq_stat_basic.pl -i RIFPA_final_gene_models_AUGUSTUS_v1_AGAT.gff3 -g 4.2_RIFPA_polished_softMasked_purged_genome_v1.fasta

conda deactivate
source activate augustus

gffread -E RIFPA_final_gene_models_AUGUSTUS_v1_lociMerged_longestIsoform.gff -g 4.2_RIFPA_polished_softMasked_purged_genome_v1.fasta -y RIFPA_final_gene_models_AUGUSTUS_v1.prot.longestIsoform.fasta

conda deactivate

source activate quast


quast ./4.2_RIFPA_polished_softMasked_purged_genome_v1.fasta -o
```

```
./quast_4.2_softmasked --eukaryote


conda deactivate
source activate busco_env
#export BUSCO_CONFIG_FILE="/data/home/btx654/.conda/envs/busco_env/busco/config/myconfig.ini"
#export AUGUSTUS_CONFIG_PATH=/data/SBCS-MartinDuranLab/02-Chema/src/Augustus/config/

#busco -i RIFPA_final_gene_models_AUGUSTUS_v1.prot.longestIsoform.fasta -m proteins -o busco_longest_isoform -c 4 -l metazoa_odb10
busco -i RIFPA_final_gene_models_AUGUSTUS_v1.prot.fasta -m proteins -o busco_gene_models -c 4 -l metazoa_odb10
busco -i 4.2_RIFPA_polished_softMasked_purged_genome_v1.fasta -m DNA -o busco_genome -c 4 -l metazoa_odb10
```

gffread.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/riftia_oliveira_2022_SINGLE
#$ -o /data/scratch/btx654/riftia_oliveira_2022_SINGLE
#$ -j y
#$ -pe smp 4
#$ -l h_vmem=5G
#$ -l h_rt=24:0:0
#$ -l highmem

module load anaconda3
source activate augustus


```

```
gffread -y RIFPA_final_gene_models_AUGUSTUS_v1_AGAT.prot.longestIsoform.fasta -g 4.2_RIFPA_polished_softMasked_purged_genome_v1.fasta RIFPA_final_gene_models_AUGUSTUS_v1_AGAT_lociMerged_longestIsoform.gff
```

```
Type (3rd column)        Number        Size total (kb)        Size mean (bp)        % of the genome        /!\Results are rounding to two decimal places
cds        496239        103636.29        208.84        18.48
exon        574346        103714.43        180.58        18.49
five_prime_utr        35671        35.67        1.00        0.01
gene        25977        400893.47        15432.63        71.49
mrna        58020        915511.98        15779.25        163.26
start_codon        57895        173.69        3.00        0.03
stop_codon        57926        173.78        3.00        0.03
three_prime_utr        42439        42.47        1.00        0.01
tss        35670        35.67        1.00        0.01
tts        42424        42.42        1.00        0.01
Total        1426607        1524259.86        1068.45        27
1.81
```

2 - Assembly vs Assembly using Minimap2 and plot

asm5/asm10/asm20: asm-to-ref mapping, for ~0.1/1/5% sequence divergence. asm5 for the same species

```
/minimap2 -cx asm5 asm1.fa asm2.fa > aln.paf             # intra-species asm-to-asm alignment
```

```
minimap2 -ax sr $pacbio_corrected_nonBacteria $R1_cleaned $R2_cl
eaned --split-prefix temp_sam_ > $alignment_sam
```

```
minimap2 -x map-pb $ref_genome $pb_fasta | gzip -c - > purge_$1.
paf.gz
```

QUERY ours
TARGET theirs

minimap2.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/alignment
#$ -o /data/scratch/btx654/alignment
#$ -j y
#$ -pe smp 12
#$ -l h_vmem=30G
#$ -l h_rt=240:0:0
#$ -l highmem

module load anaconda3
source activate minimap2_env

minimap2 -cx asm5 /data/scratch/btx654/riftia_oliveira_2022_SING
LE/4.2_RIFPA_polished_softMasked_purged_genome_v1_SINGLE.fasta /
data/SBCS-MartinDuranLab/03-Giacomo/data/riftia/annotation/rifti
a_softmasked.fa > alignment_riftia_OUR_VS_oliveira_2022.paf
```

```
scp -i ~/.ssh/id_rsa_apocrita -r btx654@login.hpc.qmul.ac.uk:/da
ta/scratch/btx654/alignment/alignment_riftia_OUR_VS_oliveira_202
2.paf /Users/giacomo/Desktop/
```

R

install.packages("pafr")

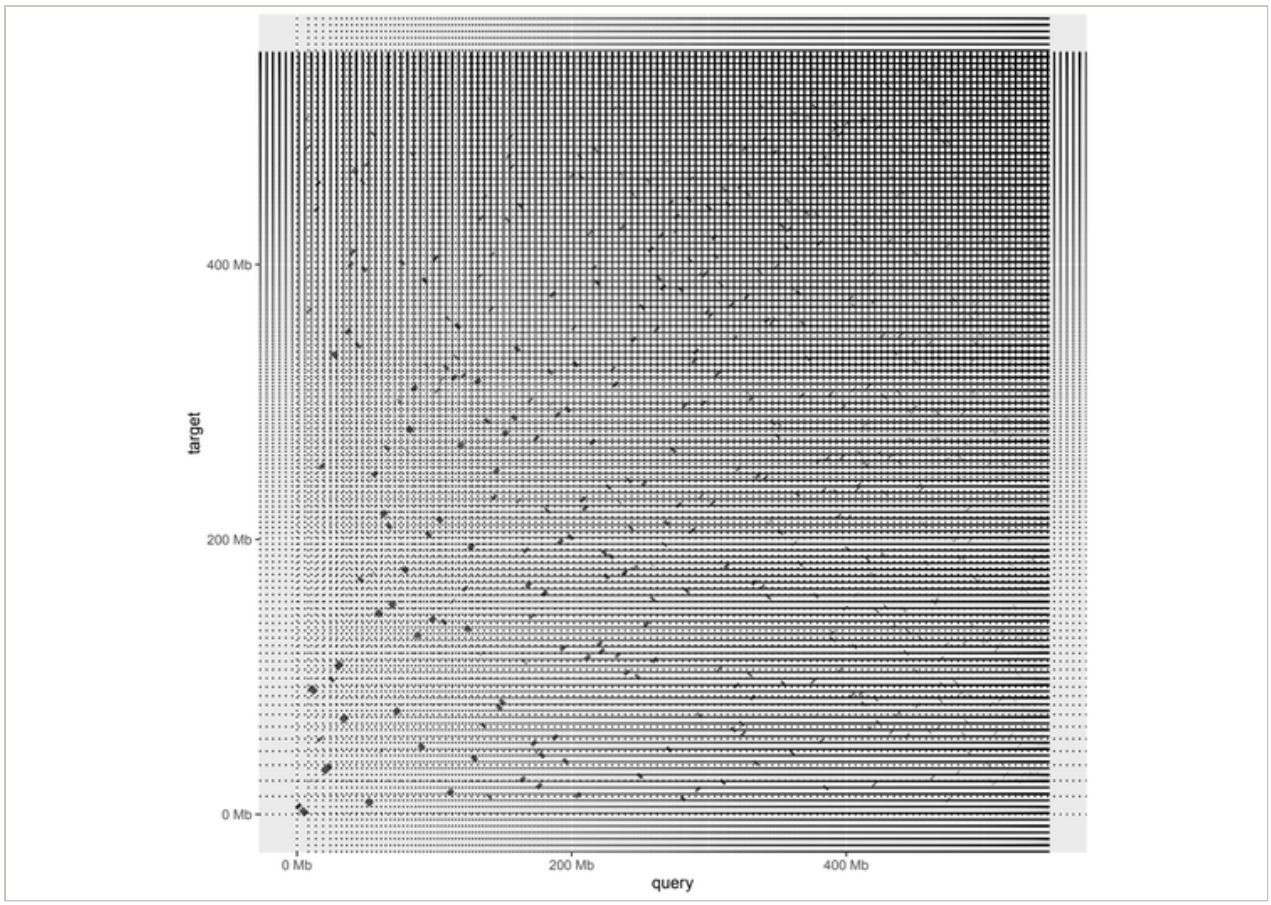
```
library(pafr)
library(ggplot2)
## Loading required package: ggplot2
path_to_alignment <- system.file("extdata", "/Users/giacomo/Desktop/alignment_riftia_OUR_VS_oliveira2022/alignment_riftia_OUR_VS_oliveira_2022.paf", package = "pafr")
ali <- read_paf(path_to_alignment)
dotplot(ali)

ali <- read_paf("/Users/giacomo/Desktop/alignment_riftia_OUR_VS_oliveira2022/alignment_riftia_OUR_VS_oliveira_2022.paf")
prim_alignment <- filter_secondary_alignments(ali)
long_ali <- subset(prim_alignment, alen > 1e4 & mapq > 40)
```

```
over_N50_ali <- subset(long_ali, qlen > 1423584, tlen > 2870320)
```

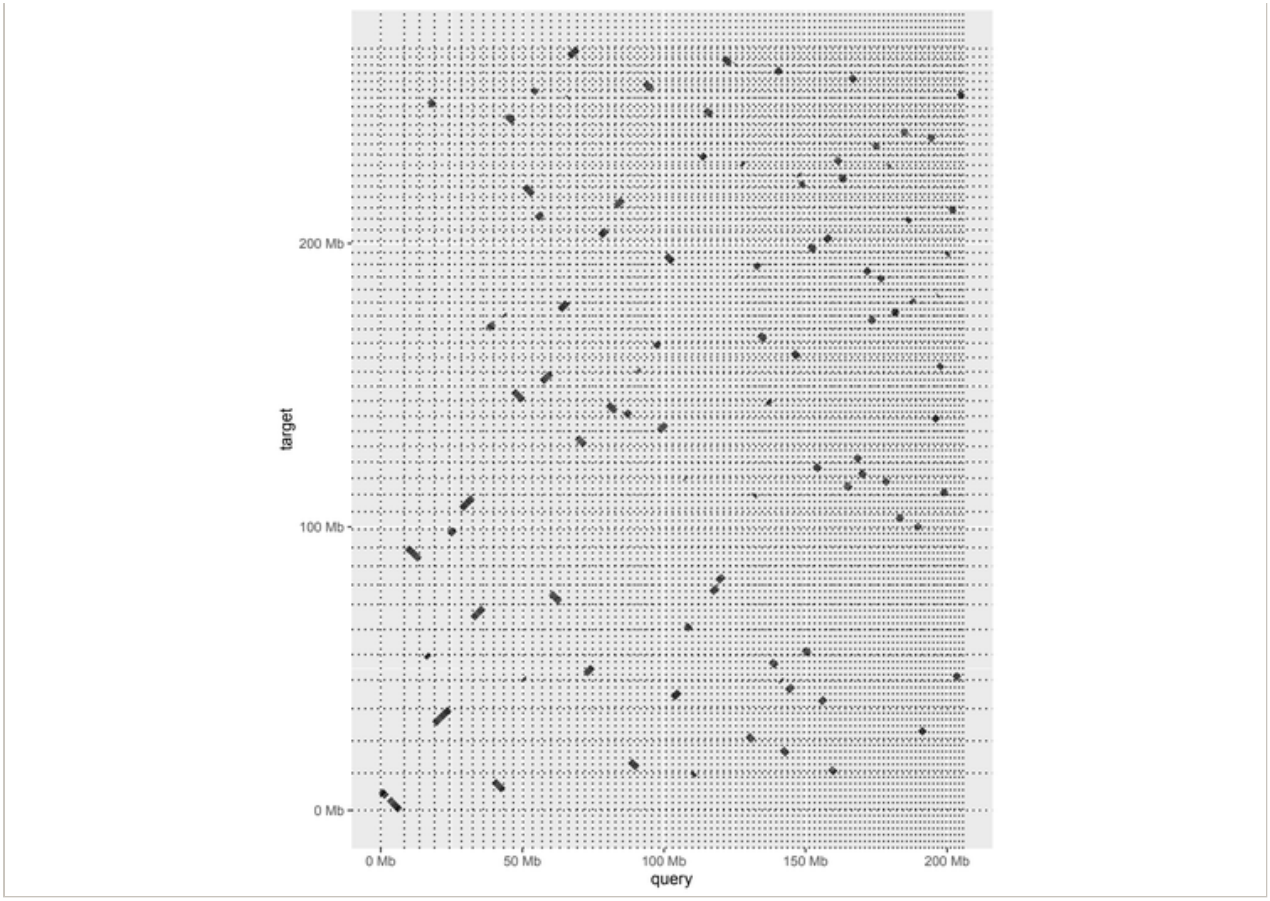
selecting just the sequences equal or over N50 sice

```
awk -F "\t" '{ if(($2 >= 1423584) && ($7 >= 2870320)) { print } }' alignment_riftia_OUR_VS_oliveira_2022.paf > overN50_alignment_riftia_OUR_VS_oliveira_2022.paf
```

```
overN50_ali <- read_paf("/Users/giacomo/Desktop/alignment_riftia_OUR_VS_oliveira2022/overN50_alignment_riftia_OUR_VS_oliveira_2022.paf")
overN50_prim_alignment <- filter_secondary_alignments(overN50_ali)
overN50_long_ali <- subset(overN50_prim_alignment, alen > 1e4 & mapq > 40)
dotplot(overN50_long_ali)
```



PDF dotplot_prim_alignment_long_ali • PDF document



PDF dotplot_overN50_prim_alignment_long_ali • PDF document

3 - Annotation vs Annotation

INPUT FILES:

OUR Rpac.fa (Non redundant proteome, longest isoform)

OLIVEIRA2022

RIFPA_final_gene_models_AUGUSTUS_v1_AGAT.prot.longestIsoform.fasta (longest isoform generated in **1** with agat and gffread)

multi-line FASTA (default from NCBI) to single-line FASTA

```
awk '/^>/ {printf("\n%s\n",$0);next; } { printf("%s",$0);}  END {printf("\n");}' < Rpac.fa | tail -n +2 > Rpac_SINGLE.fa
awk '/^>/ {printf("\n%s\n",$0);next; } { printf("%s",$0);}  END {printf("\n");}' < RIFPA_final_gene_models_AUGUSTUS_v1_AGAT.prot.longestIsoform.fasta | tail -n +2 > oliveira2022_SINGLE.fa
```

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=1G
#$ -l h_rt=36:0:0

species=$1

echo "Working on "$species

cd /data/scratch/btx654/btx604-scratch/$species/New_annotation_Dec2020/step8/

module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/trinotate_env

blastp -query input_trinotate_proteins.fa -db /data/SBCS-MartinDuranLab/03-Giacomo/db/trinotate/uniprot_sprot.pep -num_threads 8 -max_target_seqs 1 -outfmt 6 -evalue 1e-3 > blastp.outfmt6
```

GUIDE

runBLAST.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/ann_VS_ann_oliveira2022/
```

```
#$ -o /data/scratch/btx654/ann_VS_ann_oliveira2022/
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=36:0:0
#$ -l highmem

#Script to perform a reciprocal blast search
#Usage: bash runRBLAST.sh PATH/TO/QUERY/FILE PATH/TO/DB/FILE PATH/TO/OUTPUTS
#Usage ex: bash runRBLAST.sh PATH/TO/INPUT1/species1.fasta.transdecoder.pep PATH/TO/INPUT2/species2.fasta.transdecoder.pep PATH/TO/OUTPUTS


#Input query file
inputQuery=$1
#Input DB reciprocal file
inputDB=$2
#Path to output results
outputPath=$3
#Move to DB directory
queryPath=$(dirname $inputQuery)

module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/trinotate_env

#cd $queryPath
#Make blastable protein DB of the input query
makeblastdb -in $inputQuery -dbtype prot
```

```
#Move to query directory
#dbPath=$(dirname $inputDB)
#cd $dbPath
#Make blastable protein DB of the input DB
makeblastdb -in $inputDB -dbtype prot
#Output start status message
echo "Beginning reciprocal BLAST..."
#Move to outputs folder
#cd $outputPath
#Use blastp to search a database
blastp -query $inputQuery -db $inputDB -max_target_seqs 1 -outfmt 6 -evalue 1e-3 -num_threads 8 > blast.outfmt6
#Switch query and search paths for reciprocal search
blastp -query $inputDB -db $inputQuery -max_target_seqs 1 -outfmt 6 -evalue 1e-3 -num_threads 8 > blast_reciprocal.outfmt6
#Output end status message
echo "Finished reciprocal BLAST!"
```

```
qsub runRBLAST.sh /data/scratch/btx654/ann_VS_ann_oliveira2022/oliveira2022_SINGLE.fa /data/scratch/btx654/ann_VS_ann_oliveira2022/Rpac_SINGLE.fa /data/scratch/btx654/ann_VS_ann_oliveira2022/
```

findRBH.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/ann_VS_ann_oliveira2022/
#$ -o /data/scratch/btx654/ann_VS_ann_oliveira2022/
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=10G
#$ -l h_rt=36:0:0
#$ -l highmem
```

```
9
10 #Script to filter reciprocal blast results for best hits
11 #Usage: bash findRBH.sh PATH/TO/QUERY/BLAST/RESULTS PATH/TO/DB/B
   LAST/RESULTS
12 #Usage ex: bash findRBH.sh blast.outfmt6 blast_reciprocal.outfmt
   6
13 #Input query blast results file
14 queryPath=$1
15 #Input DB reciprocal blast results file
16 dbPath=$2
17 #Final output files
18 outFileRBH="blast_RBH.txt"
19 outFileSummary="blast_RBH_summary.txt"
20 #Add headers to output RBH files
21 echo "queryHit,dbHit" > $outFileRBH
22 echo "queryHits,dbHits,bestHits" > $outFileSummary
23 #Output start status message
24 echo "Recording RBH..."
25 #Loop over query blast results
26 while IFS=$'\t' read -r f1 f2 f3 f4 f5 f6 f7 f8 f9 f10 f11 f12
27 do
28 #Determine RBH to DB blast results
29 if grep -q "$f2"$'\t'"$f1"$'\t' $dbPath; then #RBH
30 echo "$f1,$f2" >> $outFileRBH
31 fi
32 done < $queryPath
33 #Output summary of RBH
34 queryHits=$(wc -l "$queryPath" | cut -d ' ' -f 1)
35 dbHits=$(wc -l "$dbPath" | cut -d ' ' -f 1)
36 bestHits=$(($(wc -l "$outFileRBH" | cut -d ' ' -f 1)-1))
37 echo "$queryHits","$dbHits","$bestHits" >> $outFileSummary
```

```
#Output end status message
echo "Finished recording RBH!"
```

```
qsub findRBH.sh blast.outfmt6 blast_reciprocal.outfmt6
```

reciprocal hits 17988

35282 proteins of our riftia had a match with the other riftia
23441 proteins of the other riftia had a match with our riftia

4 - Annotation vs previous other transcriptome 2019

```
conda create -n SRAtools_env
conda activate SRAtools_env
conda install -c bioconda sra-tools
```

```
fasterq-dump SRA_list.txt -O /data/scratch/btx654/riftia_hinzke_2019
```

sra.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/riftia_hinzke_2019
#$ -o /data/scratch/btx654/riftia_hinzke_2019
#$ -j y
#$ -pe smp 4
#$ -l h_vmem=5G
#$ -l h_rt=24:0:0
#$ -l highmem

module load anaconda3
conda activate SRAtools_env

while read line; do
```

```
   fastq-dump --defline-seq '@$sn[_$rn]/$ri' --split-files $line
-O /data/scratch/btx654/riftia_hinzke_2019
done < SRA_list.txt
```

SRA_TOOLKIT/fastq-dump --defline-seq '@$sn[_$rn]/$ri' --split-files file.sra

Installation:

```
module load anaconda3
conda create -n Trinity_env
source activate Trinity_env
conda install -c bioconda trinity
```

- Trinity-v2.8.5

```
ace     ace_R1  owe_ace_R1_r1__paired.fastq.gz  owe_ace_R1_r2_pa
ired.fastq.gz
1cell   1cell_R1        owe_1cell_R1_r1__paired.fastq.gz
 owe_1cell_R1_r2_paired.fastq.gz
2cell   2cell_R1        owe_2cell_R1_r1__paired.fastq.gz
 owe_2cell_R1_r2_paired.fastq.gz
4cell   4cell_R1        owe_4cell_R1_r1__paired.fastq.gz
 owe_4cell_R1_r2_paired.fastq.gz
8cell   8cell_R1        owe_8cell_R1_r1_paired.fastq.gz owe_8cel
l_R1_r2_paired.fastq.gz
3h      3h_R1   owe_3h_R1_r1__paired.fastq.gz   owe_3h_R1_r2_pai
red.fastq.gz
4h      4h_R1   owe_4h_R1_r1__paired.fastq.gz   owe_4h_R1_r2_pai
red.fastq.gz
5h      5h_R1   owe_5h_R1_r1__paired.fastq.gz   owe_5h_R1_r2_pai
red.fastq.gz
9h      9h_R1   owe_9h_R1_r1__paired.fastq.gz   owe_9h_R1_r2_pai
red.fastq.gz
13h     13h_R1  owe_13h_R1_r1__paired.fastq.gz  owe_13h_R1_r2_pa
```

```
ired.fastq.gz
18h     18h_R1  owe_18h_R1_r1__paired.fastq.gz  owe_18h_R1_r2_pa
ired.fastq.gz
27h     27h_R1  owe_27h_R1_r1__paired.fastq.gz  owe_27h_R1_r2_pa
ired.fastq.gz
```

```
#!/bin/bash
#$ -pe smp 20
#$ -l highmem
#$ -l h_vmem=10G
#$ -l h_rt=240:0:0
#$ -cwd
#$ -j y

module load trinity/2.4.0

Trinity --seqType fq --max_memory 200G --samples_file tissueLibr
aries_R1.txt --SS_lib_type RF --CPU 10 --output Oxford_Illumina_
trinity_R1
```

```
Trinity \
 --seqType fq \
 --left $sample1_r1,$sample2_r1 \
 --right $sample1_r2,$sample2_r2 \
 --SS_lib_type RF \
 --max_memory 400G \
 --CPU 20 \
 --output $output \
 --full_cleanup \
 --trimmomatic
```

```
SRR8949056	SRR8949056	SRR8949056_1.fastq	SRR8949056_2.fastq
SRR8949057	SRR8949057	SRR8949057_1.fastq	SRR8949057_2.fastq
SRR8949058	SRR8949058	SRR8949058_1.fastq	SRR8949058_2.fastq
SRR8949059	SRR8949059	SRR8949059_1.fastq	SRR8949059_2.fastq
SRR8949060	SRR8949060	SRR8949060_1.fastq	SRR8949060_2.fastq
SRR8949061	SRR8949061	SRR8949061_1.fastq	SRR8949061_2.fastq
SRR8949062	SRR8949062	SRR8949062_1.fastq	SRR8949062_2.fastq
SRR8949063	SRR8949063	SRR8949063_1.fastq	SRR8949063_2.fastq
SRR8949064	SRR8949064	SRR8949064_1.fastq	SRR8949064_2.fastq
SRR8949065	SRR8949065	SRR8949065_1.fastq	SRR8949065_2.fastq
SRR8949066	SRR8949066	SRR8949066_1.fastq	SRR8949066_2.fastq
SRR8949067	SRR8949067	SRR8949067_1.fastq	SRR8949067_2.fastq
SRR8949068	SRR8949068	SRR8949068_1.fastq	SRR8949068_2.fastq
SRR8949069	SRR8949069	SRR8949069_1.fastq	SRR8949069_2.fastq
SRR8949070	SRR8949070	SRR8949070_1.fastq	SRR8949070_2.fastq
SRR8949071	SRR8949071	SRR8949071_1.fastq	SRR8949071_2.fast
```

```
q
SRR8949072     SRR8949072  SRR8949072_1.fastq  SRR8949072_2.fast
q
SRR8949073     SRR8949073  SRR8949073_1.fastq  SRR8949073_2.fast
q
SRR8949074     SRR8949074  SRR8949074_1.fastq  SRR8949074_2.fast
q
SRR8949075     SRR8949075  SRR8949075_1.fastq  SRR8949075_2.fast
q
SRR8949076     SRR8949076  SRR8949076_1.fastq  SRR8949076_2.fast
q
SRR8949077     SRR8949077  SRR8949077_1.fastq  SRR8949077_2.fast
q
```

combine samples all together and then trinity on the combined sample

```
cat SRR8949056_1.fastq SRR8949057_1.fastq SRR8949058_1.fastq SRR
8949059_1.fastq SRR8949060_1.fastq SRR8949061_1.fastq SRR8949062
_1.fastq SRR8949063_1.fastq SRR8949064_1.fastq SRR8949065_1.fast
q SRR8949066_1.fastq SRR8949067_1.fastq SRR8949068_1.fastq SRR89
49069_1.fastq SRR8949070_1.fastq SRR8949071_1.fastq SRR8949072_
1.fastq SRR8949073_1.fastq SRR8949074_1.fastq SRR8949075_1.fastq
SRR8949076_1.fastq SRR8949077_1.fastq > combined_1.fastq
cat SRR8949056_2.fastq SRR8949057_2.fastq SRR8949058_2.fastq SRR
8949059_2.fastq SRR8949060_2.fastq SRR8949061_2.fastq SRR8949062
_2.fastq SRR8949063_2.fastq SRR8949064_2.fastq SRR8949065_2.fast
q SRR8949066_2.fastq SRR8949067_2.fastq SRR8949068_2.fastq SRR89
49069_2.fastq SRR8949070_2.fastq SRR8949071_2.fastq SRR8949072_
2.fastq SRR8949073_2.fastq SRR8949074_2.fastq SRR8949075_2.fastq
SRR8949076_2.fastq SRR8949077_2.fastq > combined_2.fastq
```

```
#!/bin/bash
#$ -wd /data/scratch/btx654/riftia_hinzke_2019
#$ -j y
```

```
#$ -o /data/scratch/btx654/riftia_hinzke_2019
#$ -pe smp 20
#$ -l h_vmem=20G
#$ -l h_rt=240:0:0
#$ -l highmem

output=riftia_hinzke_2019_trinity

module load anaconda3
source activate Trinity_env

Trinity \
 --seqType fq \
 --left combined_1.fastq \
 --right combined_2.fastq \
 --SS_lib_type RF \
 --max_memory 400G \
 --CPU 20 \
 --output $output \
 --full_cleanup \
 --trimmomatic


if [ -f "$output".Trinity.fasta ]
then
    if [ -s "$output".Trinity.fasta ]
    then
        echo $output".Trinity.fasta exists and not empty"
    else
        echo $output".Trinity.fasta exists but empty"
```

```
    fi
else
    echo $output".Trinity.fasta not exists"
fi
```

trinity_single.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/riftia_hinzke_2019
#$ -j y
#$ -o /data/scratch/btx654/riftia_hinzke_2019
#$ -pe smp 20
#$ -l h_vmem=20G
#$ -l h_rt=240:0:0
#$ -l highmem

input=$1
R1="$input"_1.fastq
R2="$input"_2.fastq
output="$input"_trinity

module load anaconda3
source activate Trinity_env

Trinity \
 --seqType fq \
 --left $R1 \
 --right $R2 \
 --SS_lib_type RF \
 --max_memory 400G \
 --CPU 20 \
 --output $output \
```

```
 --full_cleanup \
 --trimmomatic


if [ -f "$output".Trinity.fasta ]
then
    if [ -s "$output".Trinity.fasta ]
    then
        echo $output".Trinity.fasta exists and not empty"
    else
        echo $output".Trinity.fasta exists but empty"
    fi
else
    echo $output".Trinity.fasta not exists"
fi
```

now I need to merge all the transcriptomes into a single one with cd-hit

```
conda create -n cdhit_env
conda activate cdhit_env
conda install -c bioconda cd-hit
```

- cd hit version v4.8.1

first I need to merge all the transcriptome into one with cat

```
cat SRR8949056_trinity.Trinity.fasta SRR8949057_trinity.Trinity.
fasta SRR8949058_trinity.Trinity.fasta SRR8949059_trinity.Trinit
y.fasta SRR8949060_trinity.Trinity.fasta SRR8949061_trinity.Trin
ity.fasta SRR8949062_trinity.Trinity.fasta SRR8949063_trinity.Tr
inity.fasta SRR8949064_trinity.Trinity.fasta SRR8949065_trinity.
Trinity.fasta SRR8949066_trinity.Trinity.fasta SRR8949067_trinit
y.Trinity.fasta SRR8949068_trinity.Trinity.fasta SRR8949069_trin
ity.Trinity.fasta SRR8949070_trinity.Trinity.fasta SRR8949071_tr
inity.Trinity.fasta SRR8949072_trinity.Trinity.fasta SRR8949073_
trinity.Trinity.fasta SRR8949074_trinity.Trinity.fasta SRR894907
```

```
5_trinity.Trinity.fasta SRR8949076_trinity.Trinity.fasta SRR8949
077_trinity.Trinity.fasta > combined_trinity.Trinity.fasta
```

- there are 2284338 sequences in combined_trinity.Trinity.fasta

cdhit.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -pe smp 20
#$ -l h_vmem=20G
#$ -l h_rt=240:0:0
#$ -l highmem

/data/SBCS-MartinDuranLab/02-Chema/src/cdhit/cd-hit-est -i combi
ned_trinity.Trinity.fasta -o cdhit_90similarity -M 380000 -T 20
```

-i = input
-o = output
-c = cut-off
-n = word size:
n=5 for thresholds 0.7 ~ 1.0
n=4 for thresholds 0.6 ~ 0.7
n=3 for thresholds 0.5 ~ 0.6
n=2 for thresholds 0.4 ~ 0.5
-M = maximum available memory
-T threads

5 - PFAM barplot

generate a list of uniq pfam from riftia ann file:

```
head -5 riftia_annotation_Jan2021_TrinoPantherKO.xls | cut -f 8
| sed '/^\./d' | sed "s/\`/\n/g" | sed "s/\^.*//g"

cut -f 8 riftia_annotation_Jan2021_TrinoPantherKO.xls | sed 's/^
```

```
/\./g' | sed "s/\`/\n/g" | sed "s/\^.*//g" | sort | uniq | wc -l
cut -f 8 riftia_annotation_Jan2021_TrinoPantherKO.xls | sed "s/\
`/\n/g" | sed "s/\^.*//g" | sort | uniq | wc -l #then remove fir
st and last lines
```

16467 proteins annotated with pfam

5435 unique pfam

HMMER_Rpac.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/pfam
#$ -o /data/scratch/btx654/pfam
#$ -j y
#$ -l highmem
#$ -pe smp 12
#$ -l h_vmem=40G
#$ -l h_rt=36:0:0
#$ -l highmem

module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda
3/trinotate_env

hmmscan --cpu 12 --domtblout PFAM_Rpac.out /data/SBCS-MartinDura
nLab/03-Giacomo/db/trinotate/Pfam-A.hmm Rpac.fa > pfam_Rpac.log
```

HMMER_oliveira2022NR.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/pfam
#$ -o /data/scratch/btx654/pfam
#$ -j y
#$ -l highmem
```

```
#$ -pe smp 12
#$ -l h_vmem=40G
#$ -l h_rt=36:0:0
#$ -l highmem

module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/trinotate_env

hmmscan --cpu 12 --domtblout PFAM_oliveira2022NR.out /data/SBCS-MartinDuranLab/03-Giacomo/db/trinotate/Pfam-A.hmm RIFPA_final_gene_models_AUGUSTUS_v1_AGAT.prot.longestIsoform.fasta > pfam_oliveira2022NR.log
```

check how many proteins have been annotated with pfam:

```
grep -v "^#" PFAM_oliveira2022NR.out | grep -o "\sRIFPA.*t[0-9]"
  | sort | uniq | wc -l
```

20805

check how many different pfam in oliveira2022

```
grep -v "^#" PFAM_oliveira2022NR.out | grep -o "\sPF........."
 | sort | uniq | wc -l
```

14439 uniq pfam

I need to filter the output in a better way, there are too many overlapping pfam

```
conda create -n pfamScan_env
conda activate pfamScan_env
conda install -c bioconda pfam_scan
```

```
pfam_scan.pl -fasta <fasta_file> -dir <directory location of Pfam files>
```

```
Useful options are:
-outfile <file>    : output file, otherwise send to STDOUT
```

pfamScan_oliveira2022NR.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/pfam/pfamScan
#$ -o /data/scratch/btx654/pfam/pfamScan
#$ -j y
#$ -l highmem
#$ -pe smp 12
#$ -l h_vmem=40G
#$ -l h_rt=36:0:0
#$ -l highmem

module load anaconda3
conda activate pfamScan_env

pfam_scan.pl -cpu 12 -fasta ../RIFPA_final_gene_models_AUGUSTUS_v1_AGAT.prot.longestIsoform.fasta -dir /data/SBCS-MartinDuranLab/00-BlastDBs -outfile PFAMscan_oliveira2022NR.out
```

pfamScan_Rpac.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/pfam/pfamScan
#$ -o /data/scratch/btx654/pfam/pfamScan
#$ -j y
#$ -l highmem
#$ -pe smp 12
#$ -l h_vmem=40G
#$ -l h_rt=36:0:0
```

```
#$ -l highmem


module load anaconda3
conda activate pfamScan_env


pfam_scan.pl -cpu 12 -fasta ../Rpac.fa -dir /data/SBCS-MartinDur
anLab/00-BlastDBs -outfile PFAMscan_Rpac.out
```

check how many proteins have been annotated with pfam:

```
grep -v "^#" PFAMscan_oliveira2022NR.out | grep -o "RIFPA.*t[0-
9]"   | sort | uniq | wc -l
```

13179

check how many different pfam in oliveira2022

```
grep -v "^#" PFAMscan_oliveira2022NR.out | grep -o "\sP
F........."  | sort | uniq | wc -l
```

5079 uniq pfam

CHEMA

CD-hit merge transcriptomes

input file:

```
/data/SBCS-MartinDuranLab/03-Giacomo/data/07-Reviews/riftia_hinz
ke_2019/combined_trinity.Trinity.fasta
```

cdhit.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
```

```
#$ -pe smp 20
#$ -l h_vmem=20G
#$ -l h_rt=240:0:0
#$ -l highmem

/data/SBCS-MartinDuranLab/02-Chema/src/cdhit/cd-hit-est -i combined_trinity.Trinity.fasta -o cdhit_90similarity -M 380000 -T 20
```

Best Blast hit
GUIDE

use the merged cd-hit transcriptome as input

Our riftia vs oliveira 2022 directory with all the db files generated:

```
/data/SBCS-MartinDuranLab/03-Giacomo/data/07-Reviews/riftia_oliveira_2022/ann_VS_ann_oliveira2022/
```

runBLAST.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/ann_VS_ann_oliveira2022/
#$ -o /data/scratch/btx654/ann_VS_ann_oliveira2022/
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=36:0:0
#$ -l highmem

#Script to perform a reciprocal blast search
#Usage: bash runRBLAST.sh PATH/TO/QUERY/FILE PATH/TO/DB/FILE PATH/TO/OUTPUTS
#Usage ex: bash runRBLAST.sh PATH/TO/INPUT1/species1.fasta.trans
```

```
decoder.pep PATH/TO/INPUT2/species2.fasta.transdecoder.pep PATH/
TO/OUTPUTS
13
14
15 #Input query file
16 inputQuery=$1
17 #Input DB reciprocal file
18 inputDB=$2
19 #Path to output results
20 outputPath=$3
21 #Move to DB directory
22 queryPath=$(dirname $inputQuery)
23
24 module load anaconda3
25 conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda
3/trinotate_env
26
27 #cd $queryPath
28 #Make blastable protein DB of the input query
29 makeblastdb -in $inputQuery -dbtype prot
30 #Move to query directory
31 #dbPath=$(dirname $inputDB)
32 #cd $dbPath
33 #Make blastable protein DB of the input DB
34 makeblastdb -in $inputDB -dbtype prot
35 #Output start status message
36 echo "Beginning reciprocal BLAST..."
37 #Move to outputs folder
38 #cd $outputPath
39 #Use blastp to search a database
40 blastp -query $inputQuery -db $inputDB -max_target_seqs 1 -outfm
```

```
t 6 -evalue 1e-3 -num_threads 8 > blast.outfmt6
#Switch query and search paths for reciprocal search
blastp -query $inputDB -db $inputQuery -max_target_seqs 1 -outfm
t 6 -evalue 1e-3 -num_threads 8 > blast_reciprocal.outfmt6
#Output end status message
echo "Finished reciprocal BLAST!"
```

```
qsub runRBLAST.sh /data/scratch/btx654/ann_VS_ann_oliveira2022/o
liveira2022_SINGLE.fa /data/scratch/btx654/ann_VS_ann_oliveira20
22/Rpac_SINGLE.fa /data/scratch/btx654/ann_VS_ann_oliveira2022/
```

findRBH.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/ann_VS_ann_oliveira2022/
#$ -o /data/scratch/btx654/ann_VS_ann_oliveira2022/
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=10G
#$ -l h_rt=36:0:0
#$ -l highmem

#Script to filter reciprocal blast results for best hits
#Usage: bash findRBH.sh PATH/TO/QUERY/BLAST/RESULTS PATH/TO/DB/B
LAST/RESULTS
#Usage ex: bash findRBH.sh blast.outfmt6 blast_reciprocal.outfmt
6
#Input query blast results file
queryPath=$1
#Input DB reciprocal blast results file
dbPath=$2
#Final output files
outFileRBH="blast_RBH.txt"
```

```
outFileSummary="blast_RBH_summary.txt"
#Add headers to output RBH files
echo "queryHit,dbHit" > $outFileRBH
echo "queryHits,dbHits,bestHits" > $outFileSummary
#Output start status message
echo "Recording RBH..."
#Loop over query blast results
while IFS=$'\t' read -r f1 f2 f3 f4 f5 f6 f7 f8 f9 f10 f11 f12
do
#Determine RBH to DB blast results
if grep -q "$f2"$'\t'"$f1"$'\t' $dbPath; then #RBH
echo "$f1,$f2" >> $outFileRBH
fi
done < $queryPath
#Output summary of RBH
queryHits=$(wc -l "$queryPath" | cut -d ' ' -f 1)
dbHits=$(wc -l "$dbPath" | cut -d ' ' -f 1)
bestHits=$(($(wc -l "$outFileRBH" | cut -d ' ' -f 1)-1))
echo "$queryHits","$dbHits","$bestHits" >> $outFileSummary
#Output end status message
echo "Finished recording RBH!"
```

```
qsub findRBH.sh blast.outfmt6 blast_reciprocal.outfmt6
```

reciprocal hits 17988
reciprocal hits (transcriptome vs our Riftia) = 15469

I re-do this using the script: https://scriptomika.wordpress.com/2014/01/28/extract-best-reciprocal-blast-matches/

```
../get_RBH.py blast.outfmt6 blast_reciprocal.outfmt6 1 2 11 low
OliveiraVSGenome.hits.out
```

Genome vs mBIO == 15469
Genome vs Oliveira == 17981

35282 proteins of our riftia had a match with the other riftia
23441 proteins of the other riftia had a match with our riftia

PFAM

use the merged cd-hit transcriptome as input

pfamScan_Rpac.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/pfam/pfamScan
#$ -o /data/scratch/btx654/pfam/pfamScan
#$ -j y
#$ -l highmem
#$ -pe smp 12
#$ -l h_vmem=40G
#$ -l h_rt=36:0:0
#$ -l highmem

module load anaconda3
conda activate pfamScan_env

pfam_scan.pl -cpu 12 -fasta ../Rpac.fa -dir /data/SBCS-MartinDur
anLab/00-BlastDBs -outfile PFAMscan_Rpac.out
```

check how many proteins have been annotated with pfam:

```
grep -v "^#" PFAMscan_oliveira2022NR.out | grep -o "RIFPA.*t[0-
9]"   | sort | uniq | wc -l
```

13179
Trinity CD-HIT: 5187 unique PFAM (by Chema)

chema's code:

```
grep -v "^#" Rpac_cdhit_transD_pfamscan.out | tr "[:space:]" "\
n" | grep -E '^PF[0-9]' | sort | uniq | wc -l
```

Chapter 4

Gene family analyses

Broccoli

```
conda create --prefix /data/SBCS-MartinDuranLab/03-Giacomo/src/a
naconda3/broccoli_env python=3.6 ete3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda
3/broccoli_env
conda install -c bioconda fasttree
conda install -c bioconda diamond=0.9.35
cd /data/SBCS-MartinDuranLab/03-Giacomo/src/
git clone https://github.com/rderelle/Broccoli
```

broccoli_v1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 26
#$ -l h_vmem=20G
#$ -l h_rt=120:0:0
```

```
#$ -l highmem


module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/broccoli_env

mkdir broccoli
cd broccoli
cp -r /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/ ./

python /data/SBCS-MartinDuranLab/03-Giacomo/src/Broccoli/broccoli.py -dir ./NR_proteomes -threads 26
```

- the fasta file in the folder “NR_proteomes” should all have the expansion .fasta (e.g. “Ofus.fasta”)

Broccoli -

```
module load anaconda3
conda create --prefix /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/broccoli_verysensitive_env python=3.6 ete3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/broccoli_verysensitive_env
conda install -c bioconda fasttree
conda install -c bioconda diamond=2.0.6
cd /data/SBCS-MartinDuranLab/03-Giacomo/src/
cp -r Broccoli/ Broccoli_very_sensitive/
cp -r Broccoli/ Broccoli_ultra_sensitive/
#we need to modify the script Broccoli/scripts/broccoli_step2.py at line 238 changing --more-sensitive to --very-sensitive and --ultra-sensitive
```

broccoli_very_sensitive_v1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
```

```
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 26
#$ -l h_vmem=20G
#$ -l h_rt=120:0:0
#$ -l highmem

module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/broccoli_verysensitive_env

mkdir -p broccoli_very_sensitive
cd broccoli_very_sensitive
cp -r /data/scratch/btx654/broccoli/NR_proteomes/ ./

python /data/SBCS-MartinDuranLab/03-Giacomo/src/Broccoli_very_sensitive/broccoli.py -dir ./NR_proteomes -threads 26
```

broccoli_ultra_sensitive_v1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 26
#$ -l h_vmem=20G
#$ -l h_rt=120:0:0
#$ -l highmem

module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/broccoli_verysensitive_env
```

```

mkdir -p broccoli_ultra_sensitive
cd broccoli_ultra_sensitive
cp -r /data/scratch/btx654/broccoli/NR_proteomes/ ./

python /data/SBCS-MartinDuranLab/03-Giacomo/src/Broccoli_ultra_sensitive/broccoli.py -dir ./NR_proteomes -threads 26
```

OrthoFinder

/data/home/btx654/scripts/gene_family_evolution/orthofinder_Jan2021_v1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 12
#$ -l h_vmem=5G
#$ -l h_rt=240:0:0



cd /data/scratch/btx654/gene_family_evolution/NR_proteomes/OrthoFinder/Results_Dec15/
#Load anaconda and activate MMseqs2 environment
module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/orthofinder_env

echo "(Nvec,(Hmia,(((Skow,Spur),(Blan,(Locu,Hsap))),((Smar,Tcas),(Smed,((Lgig,(Cgig,(Myes,Bpla))),(((Ofus,(Dgyr,(((Lluy,(Oalv,Rpac)),Ofra),(Ctel,(Hrob,Eand))))))),(Ngen,(Paus,Lana)))))))));" > SpeciesTree_Jan2021.nwk
#Run orthofinder with mmseqs and inflation of 2
```

```
orthofinder -t 12 -S mmseqs -I 2 -s SpeciesTree_Jan2021.nwk -fg
/data/scratch/btx654/gene_family_evolution/NR_proteomes/OrthoFin
der/Results_Dec15/
```

orthofinder_env

```
module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda
3/orthofinder_env
conda install -c bioconda orthofinder
```

- v 2.5.2-hdfd78af_1

/data/home/btx654/scripts/gene_family_evolution/Jun2021
/orthofinder_ultra_sensitive.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 12
#$ -l h_vmem=20G
#$ -l h_rt=240:0:0
#$ -l highmem

mkdir -p gene_family_evolution
cd gene_family_evolution
mkdir -p orthofinder_ultra_sensitive_Jun2021
cd orthofinder_ultra_sensitive_Jun2021
module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda
3/orthofinder_env

cp -r /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/ ./
echo "(Nvec,(Hmia,(((Skow,Spur),(Blan,(Locu,Hsap))),((Smar,Tca
```

```
s),(Smed,(((Lgig,Gaeg),(Cgig,(Myes,Bpla))),(((Ofus,(Dgyr,(((Lluy,(Pech,(Oalv,Rpac))),Ofra),(Ctel,(Hrob,Eand)))))),(Ngen,(Paus,Lana)))))))));" > SpeciesTree.nwk
#Run orthofinder with mmseqs and inflation of 2
orthofinder -f NR_proteomes -t 12 -S diamond_ultra_sens -I 2 -s SpeciesTree.nwk
```

/data/home/btx654/scripts/gene_family_evolution/Jun2021/orthofinder_more_sensitive.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 12
#$ -l h_vmem=20G
#$ -l h_rt=240:0:0
#$ -l highmem

mkdir -p gene_family_evolution
cd gene_family_evolution
mkdir -p orthofinder_more_sensitive_Jun2021
cd orthofinder_ultra_more_Jun2021
module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/orthofinder_env

cp -r /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/ ./
echo "(Nvec,(Hmia,(((Skow,Spur),(Blan,(Locu,Hsap))),((Smar,Tcas),(Smed,(((Lgig,Gaeg),(Cgig,(Myes,Bpla))),(((Ofus,(Dgyr,(((Lluy,(Pech,(Oalv,Rpac))),Ofra),(Ctel,(Hrob,Eand)))))),(Ngen,(Paus,Lana)))))))));" > SpeciesTree.nwk
#Run orthofinder with mmseqs and inflation of 2
```

```
orthofinder -f NR_proteomes -t 12 -S diamond -I 2 -s SpeciesTre
e.nwk
```

Ferdi script
In Apple M1

```
conda install -c etetoolkit ete3
conda install numpy scipy statsmodels
conda install -c conda-forge tqdm
conda install -c anaconda ipykernel
python -m ipykernel install --user
conda install -c conda-forge notebook
conda install -c conda-forge ipywidgets
jupyter notebook # this will launch the browser app to actually
run the script
```

```
# first import
from collections import defaultdict
from collections import Counter
from ete3 import PhyloTree
import numpy as np
from tqdm import tqdm_notebook
import scipy.stats as stats
from statsmodels.stats.multitest import multipletests
import time
import csv
import copy

# broccoli
Fams,species={},[]
for line in open('table_OGs_protein_names_modified_OFUS.txt'):
```

```
    if line.startswith("#"): continue
    fid=line.strip().split("\t")[0]
    glt=[]
    for i in line.strip().split("\t")[1:]:
        glt.append(i)
    gbs=defaultdict(list)
    #print (fid)
    for g in glt:
        for j in g.split():
            #if j =='gene-ND4L': continue
            spc,gid=j.split('_',1)
            #print (spc)
            #print (gid)
            gbs[spc].append(gid)
    species.extend(gbs.keys())
    Fams[fid]=gbs
        #if i==5: break
species=list(set(species))
print(Fams['OG_1'])

# orthofinder
Fams={}
with open('Orthogroups.csv') as csvfile:
    species=[]
    for i,rc in enumerate(csv.reader(csvfile, delimiter='\t', quotechar='"')):
        #print (i)
        #print (rc)
        if i==0:
            species=[sp for sp in rc[1:]]
```

```
            continue
        #for j,spg in enumerate(rc[1:]):
            #for g in spg.split(','):
                #print (g)
                #print (g.split('|')[1])
        #print (i)
        #print (species)
        #print(rc[1:])
        #print(rc[19])
        #print("hello")
        genBySp=dict((species[j],[g.split('|')[1] for g in spg.s
plit(',')]) for j,spg in enumerate(rc[1:]) if not spg=='')
        #print (genBySp)
        #genBySp=
        Fams[rc[0]]=genBySp
        #if i==5: break

print(Fams['OG0000000'])

# PFAM domains
Fams={}
with open('Pfam_species_ferdi_OK.csv') as csvfile:
    species=[]
    for i,rc in enumerate(csv.reader(csvfile, delimiter='\t', qu
otechar='"')):
        #print (i)
        #print (rc)
        if i==0:
            species=[sp for sp in rc[1:]]
            continue
        #for j,spg in enumerate(rc[1:]):
```

```
            #for g in spg.split(','):
                #print (g)
                #print (g.split('|')[1])
        #print (i)
        #print (species)
        #print(rc[1:])
        #print(rc[19])
        #print("hello")
        genBySp=dict((species[j],[g.split('|')[1] for g in spg.s
plit(',')]) for j,spg in enumerate(rc[1:]) if not spg=='')
        #print (genBySp)
        #genBySp=
        Fams[rc[0]]=genBySp
        #if i==5: break

print(Fams['PF00010.27'])

#Creates non-redundant species set
spSet=set()
for fid,gbs in Fams.items():
    for sp in gbs.keys():
        spSet.add(sp)
print(spSet)

#Load species tree
spT=PhyloTree(open('tree.tree').read(), format=1)

# Testing family expansions
#only testing species expansion
def fishExp(gbs,tbs):
```

```
    cbs=dict((sp,len(gs)) for sp,gs in gbs.items())
    med=np.median(list(cbs.values()))
    #print(cbs,med)
    spval={}
    for sp,st in cbs.items():
        nis=sum(cbs.values())-st
        nit=sum(tbs.values())-tbs[sp]-sum(cbs.values())+st
        odds,pval=stats.fisher_exact([[st, tbs[sp]-st], [nis, nit]],alternative='greater')
        nmed=np.median([cbs[ssp] for ssp in cbs if not ssp==sp])
        tmed=np.median([tbs[ssp] for ssp in tbs if not ssp==sp])
        m_odds,m_pval=stats.fisher_exact([[st, tbs[sp]-st], [nmed, tmed]],alternative='greater')
        #print(sp,pval,[st, tbs[sp]-st], [nis, nit],m_pval,[st, tbs[sp]-st], [nmed, tmed])
        spval[sp]=(cbs[sp],m_pval)

    return spval

#calculate number of genes per species
tbs=defaultdict(int)
for fid,gbs in Fams.items():
    for sp in gbs:
        tbs[sp]+=len(gbs[sp])

toosmall=0
expRes=[]
for fid,gbs in tqdm_notebook(Fams.items()):
    cbs=dict((sp,len(gs)) for sp,gs in gbs.items())
    if len(cbs.keys())<3 or sum(cbs.values())<5:
        toosmall+=1
```

```
        continue
    try:
        spval=fishExp(gbs,tbs)
        expRes.append((fid,spval))
    except:
        print(fid,cbs)


print(len(Fams),len(expRes),toosmall)

spFam,spPval=defaultdict(list),defaultdict(list)
for fid,spval in expRes:
    for sp in spval:
        ct,pval=spval[sp]
        spFam[sp].append(fid)
        spPval[sp].append(pval)

famEnrich=defaultdict(dict)
for sp in spPval:
    #signif_pvals = sidak(spPval[sp], alpha=0.05)
    #before_adj=['True' if p < 0.05  else 'False' for p in spPva
l[sp] ]
    #1signif_pvals = lsu(np.array(spPval[sp]), q=0.05) #this is
benferroni-hochberg
    adj=multipletests(pvals=np.array(spPval[sp]), alpha=0.05, me
thod="fdr_bh")
    print (sp,Counter(adj[0]))
    for fam,pth,pval in zip(spFam[sp],adj[0],adj[1]):
        famEnrich[fam][sp]=(pth,pval)

with open('orthogroups_ultrasensitive_enrich_Jun2021.txt','w') a
```

```
s out:
159     for fid,gbs in Fams.items():
160         cbs=dict((sp,len(gs)) for sp,gs in gbs.items())
161         pvl=famEnrich[fid]
162         enriched=set([])
163         outList=[fid,]
164         out.write('\t'.join(outList))
165 #Gains and Losses
166 
167 def checkFam(fam,spTd):
168     fid,gbs=fam
169     spset=list(gbs.keys())
170     if len(spset)>1:
171         phtyp=spT.get_common_ancestor(list(gbs.keys()))
172     else:
173         phtyp=[l for l in spT.get_leaves() if l.name==spset
[0]][0]
174     #print phtyp.name
175     #phyloCt[phtyp.name]+=1
176     #gbs[Ai]
177     ndesc=len([l.name for l in phtyp.get_leaves()])
178     for leaf in spTd:
179         pv='1' if leaf.name in gbs else '0'
180         leaf.add_features(presence=pv)
181     lost=[node.name  for node in phtyp.get_monophyletic(values=
['0'], target_attr="presence")]
182     #print spTd.get_ascii(attributes=["name","presence"], show_i
nternal=True)
183     #for node in spTd.get_monophyletic(values=['0'], target_attr
="presence"):
184     #    print node.name
```

```
185     #    print node.get_ascii(attributes=["presence", "name"], s
    how_internal=False)
186     return fid,len(gbs.keys()),ndesc,phtyp.name,lost
187 
188 origins=defaultdict(int)
189 losses=defaultdict(int)
190 idGainLoss={}
191 for fam in Fams.items():
192     fid,nspec,ndesc,oritax,lost=checkFam(fam,spT)
193     idGainLoss[fid]=(nspec,ndesc,oritax,lost)
194     origins[oritax]+=1
195     for tax in lost:
196         losses[tax]+=1
197 
198 spTi=copy.deepcopy(spT)
199 for node in spTi.traverse("postorder"):
200     #print node.name,origins[node.name],losses[node.name]
201     #node.add_features(origins=origins[node.name],losses=losses
    [node.name])
202     node.name="{0}_{1}_{2}]".format(node.name,origins[node.nam
    e],losses[node.name])
203 
204 print(spTi.write(format=7))
205 #print(spTi.get_ascii(attributes=["name"], show_internal=True))
206 
207 #Output
208 
209 with open('orthofinder_ultrasensitive_stats_Jun2021.tsv','w') as
    out:
210     out.write("FID\tnb_genes\tnb_species\torigin\tlost_sp\tlost_
    taxa\texpanded_tax\n")
```

```
    for fid,gbs in Fams.items():
        nspec,ndesc,oritax,lost=idGainLoss[fid]
        ngn=sum([len(gl) for gl in gbs.values()])
        exp=[s for s,p in famEnrich[fid].items() if p[0]]
        #rpvals=zip(pvals.keys(),[round(-np.log10(l[1]),2) for l in pvals.values()])
        outL=[fid,str(ngn),str(nspec),oritax,str(ndesc-nspec),','.join(lost),','.join(exp)]
        out.write('\t'.join(outL)+'\n')
```

We need two input files for this:

- the .tsv output of OrthoFinder that we will need to transform into a csv (/orthofinder_Dec2020/Orthogroups/Orthogroups.tsv)
- and a newick format tree: tree.tree

```
(Nvec,(Hmia,(((Skow,Spur)Ambulacraria,(Blan,(Locu,Hsap)Vertebrata)Chordata)Deuterostomia,((Smar,Tcas)Arthropoda,(Smed,(((Lgig,Gaeg)Gastropoda,(Cgig,(Myes,Bpla)Bivalvia_cl1)Bivalvia)Mollusca,((Ofus,(Dgyr,(((Lluy,(Pech,(Oalv,Rpac)Vestimentifera_cl2)Vestimentifera_cl1)Vestimentifera,Ofra)Siboglinidae,(Ctel,(Hrob,Eand)Clitellata)Sedentaria_cl1)Sedentaria)Annelida_cl1)Annelida,(Ngen,(Paus,Lana)Lophophorata)Kryptotrochozoa)Lophotrochozoa_cl1)Lophotrochozoa)Spiralia)Protostomia)Nephrozoa)Bilateria)Eumetazoa;
```

First thing first let's edit the Orthogroups.tsv file:

```
cp Orthogroups.tsv Orthogroups_original.tsv
sed -i 's/OFUS/Ofus|OFUS/g' Orthogroups.tsv
sed -i 's/Blan_/Blan|/g' Orthogroups.tsv
sed -i 's/Bpla_/Bpla|/g' Orthogroups.tsv
```

```
sed -i 's/Cgig_/Cgig|/g' Orthogroups.tsv
sed -i 's/Ctel_/Ctel|/g' Orthogroups.tsv
sed -i 's/Dgyr_/Dgyr|/g' Orthogroups.tsv
sed -i 's/Eand_/Eand|/g' Orthogroups.tsv
sed -i 's/Hmia_/Hmia|/g' Orthogroups.tsv
sed -i 's/Hrob_/Hrob|/g' Orthogroups.tsv
sed -i 's/Hsap_/Hsap|/g' Orthogroups.tsv
sed -i 's/Lana_/Lana|/g' Orthogroups.tsv
sed -i 's/Lgig_/Lgig|/g' Orthogroups.tsv
sed -i 's/Lluy_/Lluy|/g' Orthogroups.tsv
sed -i 's/Locu_/Locu|/g' Orthogroups.tsv
sed -i 's/Myes_/Myes|/g' Orthogroups.tsv
sed -i 's/Ngen_/Ngen|/g' Orthogroups.tsv
sed -i 's/Nvec_/Nvec|/g' Orthogroups.tsv
sed -i 's/Oalv_/Oalv|/g' Orthogroups.tsv
sed -i 's/Ofra_/Ofra|/g' Orthogroups.tsv
sed -i 's/Ofus_/Ofus|/g' Orthogroups.tsv
sed -i 's/Paus_/Paus|/g' Orthogroups.tsv
sed -i 's/Rpac_/Rpac|/g' Orthogroups.tsv
sed -i 's/Skow_/Skow|/g' Orthogroups.tsv
sed -i 's/Smar_/Smar|/g' Orthogroups.tsv
sed -i 's/Smed_/Smed|/g' Orthogroups.tsv
sed -i 's/Spur_/Spur|/g' Orthogroups.tsv
sed -i 's/Tcas_/Tcas|/g' Orthogroups.tsv
sed -i 's/Spur_/Spur|/g' Orthogroups.tsv
sed -i 's/Gaeg_/Gaeg|/g' Orthogroups.tsv
sed -i 's/Pech_/Pech|/g' Orthogroups.tsv

mv Orthogroups.tsv Orthogroups.csv
remove "orthogroups" from the first line with nano
```

```

sed -i 's/gene-ND4L/Myes|gene-ND4L/g' Orthogroups.csv #this was creating a problem
```

Now we can launch the modified version of ferdi script and it will produce the results

```
cd /Users/giacomo/Jupyter_notebook/Ferdi_script/Jun2021/ultra_sensitive
cp /Users/giacomo/Jupyter_notebook/Ferdi_script/comp_genomics/Giacomo.ipynb /Users/giacomo/Jupyter_notebook/Ferdi_script/Jun2021/ultra_sensitive
conda activate Ferdi_env
jupyter notebook
cd /Users/giacomo/Jupyter_notebook/Ferdi_script/Jun2021/more_sensitive
cp /Users/giacomo/Jupyter_notebook/Ferdi_script/comp_genomics/Giacomo.ipynb /Users/giacomo/Jupyter_notebook/Ferdi_script/Jun2021/more_sensitive
```

Expansion, Gains and Losses

Expansions

orthogroups_annotations_expanded_oasisia.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/expansions/oasisia
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/expansions/oasisia
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
```

```
#$ -l h_rt=72:00:0
#$ -l highmem

cut -f 1,7 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep Oalv | cut -f 1 > gene_families_expanded_oasisia.txt #families expanded in oasisia
fgrep -f gene_families_expanded_oasisia.txt ../../Orthogroups.csv > gene_families_expanded_oasisia.csv

cut -f 1,19 gene_families_expanded_oasisia.csv > orthogroups_gene_IDs_expanded_oasisia.txt
sed 's/Oalv|//g' orthogroups_gene_IDs_expanded_oasisia.txt > orthogroups_gene_IDs_expanded_oasisia_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_expanded_oasisia.csv
while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OALV* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../oasisia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"oasisia"$'\t'$annotations >> orthogroups_annotations_expanded_oasisia.csv
```

```
        done
    else
    echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""" >> orthogroups_annotations_expanded_oasisia.csv
    fi
done < orthogroups_gene_IDs_expanded_oasisia_OK.txt

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_expanded_oasisia.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_expanded_oasisia.csv | cut -f 7 | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_expanded_oasisia.csv
done < gene_families_expanded_oasisia.csv
```

orthogroups_annotations_expanded_osedax.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/expansions/osedax
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/expansions/osedax
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=72:00:0
#$ -l highmem

cut -f 1,7 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep Ofra | cut -f 1 > gene_families_expanded_osedax.txt #famili
```

```
es expanded in oasisia
fgrep -f gene_families_expanded_osedax.txt ../../Orthogroups.csv
> gene_families_expanded_osedax.csv

cut -f 1,20 gene_families_expanded_osedax.csv > orthogroups_gene
_IDs_expanded_osedax.txt
sed 's/Ofra|//g' orthogroups_gene_IDs_expanded_osedax.txt > orth
ogroups_gene_IDs_expanded_osedax_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\
t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_nu
mber" > orthogroups_annotations_expanded_osedax.csv
while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFRA* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../osedax_annot
ation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"osedax"$'\t'$annotations >> ort
hogroups_annotations_expanded_osedax.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_expanded_osedax.csv
     fi
```

```
done < orthogroups_gene_IDs_expanded_osedax_OK.txt

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_expanded_osedax.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_expanded_osedax.csv | cut -f 7 | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_expanded_osedax.csv
done < gene_families_expanded_osedax.csv
```

orthogroups_annotations_expanded_riftia.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/expansions/riftia
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/expansions/riftia
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=72:00:0
#$ -l highmem

cut -f 1,7 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep Rpac | cut -f 1 > gene_families_expanded_riftia.txt #families expanded in riftia
fgrep -f gene_families_expanded_riftia.txt ../../Orthogroups.csv > gene_families_expanded_riftia.csv

```

```
cut -f 1,24 gene_families_expanded_riftia.csv > orthogroups_gene_IDs_expanded_riftia.txt
sed 's/Rpac|//g' orthogroups_gene_IDs_expanded_riftia.txt > orthogroups_gene_IDs_expanded_riftia_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_expanded_riftia.csv
while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == RPAC* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../riftia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"riftia"$'\t'$annotations >> orthogroups_annotations_expanded_riftia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_expanded_riftia.csv
     fi
done < orthogroups_gene_IDs_expanded_riftia_OK.txt

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_expanded_riftia.csv
```

```
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_expanded_riftia.csv | cut -f 7 | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_expanded_riftia.csv
done < gene_families_expanded_riftia.csv
```

orthogroups_annotations_expanded_lamellibrachia.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/expansions/lamellibrachia
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/expansions/lamellibrachia
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=72:00:0
#$ -l highmem

cut -f 1,7 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep Lluy | cut -f 1 > gene_families_expanded_lamellibrachia.txt #families expanded in riftia
fgrep -f gene_families_expanded_lamellibrachia.txt ../../Orthogroups.csv > gene_families_expanded_lamellibrachia.csv

cut -f 1,14 gene_families_expanded_lamellibrachia.csv > orthogroups_gene_IDs_expanded_lamellibrachia.txt
sed 's/Lluy|//g' orthogroups_gene_IDs_expanded_lamellibrachia.txt > orthogroups_gene_IDs_expanded_lamellibrachia_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\
```

```
t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_nu
mber" > orthogroups_annotations_expanded_lamellibrachia.csv
while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == FUN* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../lamellibrach
ia_annotation_Feb2021_TrinoPantherKO_OK.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"lamellibrachia"$'\t'$annotation
s >> orthogroups_annotations_expanded_lamellibrachia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_expanded_lamellibrachia.cs
v
     fi
done < orthogroups_gene_IDs_expanded_lamellibrachia_OK.txt

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbu
ndantAnnotation_expanded_lamellibrachia.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_exp
anded_lamellibrachia.csv | cut -f 7 | sort |  uniq -c | sort -r
| awk '{$1=""; print $0}' | head -1)
```

```
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_expanded_lamellibrachia.csv
done < gene_families_expanded_lamellibrachia.csv
```

orthogroups_annotations_expanded_paraescarpia.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/expansions/paraescarpia
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/expansions/paraescarpia
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=72:00:0
#$ -l highmem

cut -f 1,7 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep Pech | cut -f 1 > gene_families_expanded_paraescarpia.txt # families expanded in riftia
fgrep -f gene_families_expanded_paraescarpia.txt ../../Orthogroups.csv > gene_families_expanded_paraescarpia.csv

cut -f 1,23 gene_families_expanded_paraescarpia.csv > orthogroups_gene_IDs_expanded_paraescarpia.txt
sed 's/Pech|//g' orthogroups_gene_IDs_expanded_paraescarpia.txt > orthogroups_gene_IDs_expanded_paraescarpia_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_expanded_paraescarpia.csv
while read line; do
   genes=$(cut -f 2 <<< "$line")
```

```
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == nbis* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../paraescarpia_annotation_Jun2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"paraescarpia"$'\t'$annotations >> orthogroups_annotations_expanded_paraescarpia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_expanded_paraescarpia.csv
     fi
done < orthogroups_gene_IDs_expanded_paraescarpia_OK.txt

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_expanded_paraescarpia.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_expanded_paraescarpia.csv | cut -f 7 | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_expanded_paraescarpia.csv
done < gene_families_expanded_paraescarpia.csv
```

Gains

orthogroups_annotations_originated_oasisia.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/oasisia
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/oasisia
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=72:00:0
#$ -l highmem

cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Oalv | cut -f 1 > gene_families_originated_oasisia.txt # families originated in oasisia
fgrep -f gene_families_originated_oasisia.txt ../../Orthogroups.csv > gene_families_originated_oasisia.csv

cut -f 1,19 gene_families_originated_oasisia.csv > orthogroups_gene_IDs_originated_oasisia.txt
sed 's/Oalv|//g' orthogroups_gene_IDs_originated_oasisia.txt > orthogroups_gene_IDs_originated_oasisia_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_originated_oasisia.csv
while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OALV* ]]
```

```
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../oasisia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"oasisia"$'\t'$annotations >> orthogroups_annotations_originated_oasisia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_oasisia.csv
     fi
done < orthogroups_gene_IDs_originated_oasisia_OK.txt

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_originated_oasisia.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_originated_oasisia.csv | cut -f 7 | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_originated_oasisia.csv
done < gene_families_originated_oasisia.csv
```

orthogroups_annotations_originated_osedax.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/osedax
```

```
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/osedax
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=72:00:0
#$ -l highmem

cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Ofra | cut -f 1 > gene_families_originated_osedax.txt #families originated in oasisia
fgrep -f gene_families_originated_osedax.txt ../../Orthogroups.csv > gene_families_originated_osedax.csv

cut -f 1,20 gene_families_originated_osedax.csv > orthogroups_gene_IDs_originated_osedax.txt
sed 's/Ofra|//g' orthogroups_gene_IDs_originated_osedax.txt > orthogroups_gene_IDs_originated_osedax_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_originated_osedax.csv
while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFRA* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
```

```
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../osedax_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"osedax"$'\t'$annotations >> orthogroups_annotations_originated_osedax.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_osedax.csv
     fi
done < orthogroups_gene_IDs_originated_osedax_OK.txt

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_originated_osedax.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_originated_osedax.csv | cut -f 7 | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_originated_osedax.csv
done < gene_families_originated_osedax.csv
```

orthogroups_annotations_originated_riftia.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/riftia
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/riftia
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
```

```
#$ -l h_rt=72:00:0
#$ -l highmem

cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Rpac | cut -f 1 > gene_families_originated_riftia.txt #families originated in riftia
fgrep -f gene_families_originated_riftia.txt ../../Orthogroups.csv > gene_families_originated_riftia.csv

cut -f 1,24 gene_families_originated_riftia.csv > orthogroups_gene_IDs_originated_riftia.txt
sed 's/Rpac|//g' orthogroups_gene_IDs_originated_riftia.txt > orthogroups_gene_IDs_originated_riftia_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_originated_riftia.csv
while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == RPAC* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../riftia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"riftia"$'\t'$annotations >> orthogroups_annotations_originated_riftia.csv
```

```
        done
    else
    echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_riftia.csv
    fi
done < orthogroups_gene_IDs_originated_riftia_OK.txt

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_originated_riftia.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_originated_riftia.csv | cut -f 7 | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_originated_riftia.csv
done < gene_families_originated_riftia.csv
```

orthogroups_annotations_originated_lamellibrachia.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/lamellibrachia
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/lamellibrachia
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=72:00:0
#$ -l highmem

cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Lluy | cut -f 1 > gene_families_originated_lamellibrachi
```

```
a.txt #families originated in riftia
fgrep -f gene_families_originated_lamellibrachia.txt ../../Orthogroups.csv > gene_families_originated_lamellibrachia.csv

cut -f 1,14 gene_families_originated_lamellibrachia.csv > orthogroups_gene_IDs_originated_lamellibrachia.txt
sed 's/Lluy|//g' orthogroups_gene_IDs_originated_lamellibrachia.txt > orthogroups_gene_IDs_originated_lamellibrachia_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_originated_lamellibrachia.csv
while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == FUN* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../lamellibrachia_annotation_Feb2021_TrinoPantherKO_OK.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"lamellibrachia"$'\t'$annotations >> orthogroups_annotations_originated_lamellibrachia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_lamellibrachia.csv
```

```
     fi
done < orthogroups_gene_IDs_originated_lamellibrachia_OK.txt

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_originated_lamellibrachia.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_originated_lamellibrachia.csv | cut -f 7 | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_originated_lamellibrachia.csv
done < gene_families_originated_lamellibrachia.csv
```

orthogroups_annotations_originated_paraescarpia.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/paraescarpia
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/paraescarpia
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=72:00:0
#$ -l highmem

cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Pech | cut -f 1 > gene_families_originated_paraescarpia.txt #families originated in riftia
fgrep -f gene_families_originated_paraescarpia.txt ../../Orthogroups.csv > gene_families_originated_paraescarpia.csv

```

```
cut -f 1,23 gene_families_originated_paraescarpia.csv > orthogroups_gene_IDs_originated_paraescarpia.txt
sed 's/Pech|//g' orthogroups_gene_IDs_originated_paraescarpia.txt > orthogroups_gene_IDs_originated_paraescarpia_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_originated_paraescarpia.csv
while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == nbis* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../paraescarpia_annotation_Jun2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"paraescarpia"$'\t'$annotations >> orthogroups_annotations_originated_paraescarpia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_paraescarpia.csv
     fi
done < orthogroups_gene_IDs_originated_paraescarpia_OK.txt

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbu
```

```
ndantAnnotation_originated_paraescarpia.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_originated_paraescarpia.csv | cut -f 7 | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_originated_paraescarpia.csv
done < gene_families_originated_paraescarpia.csv
```

orthogroups_annotations_originated_siboglinidae.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/siboglinidae
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/siboglinidae
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=140:00:00
#$ -l highmem

cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Siboglinidae | cut -f 1 > gene_families_originated_siboglinidae.txt #families originated in oasisia
fgrep -f gene_families_originated_siboglinidae.txt ../../Orthogroups.csv > gene_families_originated_siboglinidae.csv

cut -f 1,19 gene_families_originated_siboglinidae.csv > orthogroups_gene_IDs_originated_siboglinidae_oasisia.txt
sed 's/Oalv|//g' orthogroups_gene_IDs_originated_siboglinidae_oasisia.txt > orthogroups_gene_IDs_originated_siboglinidae_oasisia_OK.txt
```

```
cut -f 1,20 gene_families_originated_siboglinidae.csv > orthogroups_gene_IDs_originated_siboglinidae_osedax.txt
sed 's/Ofra|//g' orthogroups_gene_IDs_originated_siboglinidae_osedax.txt > orthogroups_gene_IDs_originated_siboglinidae_osedax_OK.txt
cut -f 1,24 gene_families_originated_siboglinidae.csv > orthogroups_gene_IDs_originated_siboglinidae_riftia.txt
sed 's/Rpac|//g' orthogroups_gene_IDs_originated_siboglinidae_riftia.txt > orthogroups_gene_IDs_originated_siboglinidae_riftia_OK.txt
cut -f 1,14 gene_families_originated_siboglinidae.csv > orthogroups_gene_IDs_originated_siboglinidae_lamellibrachia.txt
sed 's/Lluy|//g' orthogroups_gene_IDs_originated_siboglinidae_lamellibrachia.txt > orthogroups_gene_IDs_originated_siboglinidae_lamellibrachia_OK.txt
cut -f 1,23 gene_families_originated_siboglinidae.csv > orthogroups_gene_IDs_originated_siboglinidae_paraescarpia.txt
sed 's/Pech|//g' orthogroups_gene_IDs_originated_siboglinidae_paraescarpia.txt > orthogroups_gene_IDs_originated_siboglinidae_paraescarpia_OK.txt


while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OALV* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
```

```
       annotations=$(cut -f 13,14,15,18,20,24 ../../oasisia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"oasisia"$'\t'$annotations >> orthogroups_annotations_originated_siboglinidae_oasisia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_siboglinidae_oasisia.csv
     fi
done < orthogroups_gene_IDs_originated_siboglinidae_oasisia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFRA* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../osedax_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"osedax"$'\t'$annotations >> orthogroups_annotations_originated_siboglinidae_osedax.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_siboglinidae_os
```

```
edax.csv
58      fi
59 done < orthogroups_gene_IDs_originated_siboglinidae_osedax_OK.tx
   t
60
61 while read line; do
62    genes=$(cut -f 2 <<< "$line")
63    echo $genes
64    orthogroup_ID=$(cut -f 1 <<< "$line")
65    echo $orthogroup_ID
66      if [[ "$genes" == RPAC* ]]
67      then
68       IFS=', '     # space is set as delimiter
69       read -ra ADDR <<< "$genes"   # str is read into an array a
   s tokens separated by IFS
70       for gene in "${ADDR[@]}"; do
71        annotations=$(cut -f 13,14,15,18,20,24 ../../riftia_annot
   ation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
72         echo $orthogroup_ID$'\t'"riftia"$'\t'$annotations >> ort
   hogroups_annotations_originated_siboglinidae_riftia.csv
73         done
74      else
75      echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
   ""$'\t'"" >> orthogroups_annotations_originated_siboglinidae_ri
   ftia.csv
76      fi
77 done < orthogroups_gene_IDs_originated_siboglinidae_riftia_OK.tx
   t
78
79 while read line; do
80    genes=$(cut -f 2 <<< "$line")
81    echo $genes
```

```
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == FUN* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../lamellibrachia_annotation_Feb2021_TrinoPantherKO_OK.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"lamellibrachia"$'\t'$annotations >> orthogroups_annotations_originated_siboglinidae_lamellibrachia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_siboglinidae_lamellibrachia.csv
     fi
done < orthogroups_gene_IDs_originated_siboglinidae_lamellibrachia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == nbis* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
```

```
106       for gene in "${ADDR[@]}"; do
107        annotations=$(cut -f 13,14,15,18,20,24 ../../paraescarpia_annotation_Jun2021_TrinoPantherKO.xls | fgrep $gene)
108         echo $orthogroup_ID$'\t'"paraescarpia"$'\t'$annotations >> orthogroups_annotations_originated_siboglinidae_paraescarpia.csv
109         done
110      else
111      echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_siboglinidae_paraescarpia.csv
112      fi
113 done < orthogroups_gene_IDs_originated_siboglinidae_paraescarpia_OK.txt
114
115 echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_originated_siboglinidae_Ofra_Oalv_Rpac_Lluy_Pech.csv
116 cat orthogroups_annotations_originated_siboglinidae_oasisia.csv orthogroups_annotations_originated_siboglinidae_osedax.csv orthogroups_annotations_originated_siboglinidae_riftia.csv orthogroups_annotations_originated_siboglinidae_lamellibrachia.csv orthogroups_annotations_originated_siboglinidae_paraescarpia.csv >> orthogroups_annotations_originated_siboglinidae_Ofra_Oalv_Rpac_Lluy_Pech.csv
117 sort orthogroups_annotations_originated_siboglinidae_Ofra_Oalv_Rpac_Lluy_Pech.csv > orthogroups_annotations_originated_siboglinidae_Ofra_Oalv_Rpac_Lluy_Pech_OK.csv
118
119 echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_originated_siboglinidae.csv
120 while read line; do
121    orthogroup_ID=$(cut -f 1 <<< "$line")
```

```
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_originated_siboglinidae_Ofra_Oalv_Rpac_Lluy_Pech_OK.csv | cut -f 7 | sed '/^$/d' | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_originated_siboglinidae.csv
done < gene_families_originated_siboglinidae.csv
```

orthogroups_annotations_originated_vestimentifera.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/vestimentifera
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/vestimentifera
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=140:00:0
#$ -l highmem

cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Vestimentifera | cut -f 1 > gene_families_originated_vestimentifera.txt #families originated in oasisia
fgrep -f gene_families_originated_vestimentifera.txt ../../Orthogroups.csv > gene_families_originated_vestimentifera.csv

cut -f 1,19 gene_families_originated_vestimentifera.csv > orthogroups_gene_IDs_originated_vestimentifera_oasisia.txt
sed 's/Oalv|//g' orthogroups_gene_IDs_originated_vestimentifera_oasisia.txt > orthogroups_gene_IDs_originated_vestimentifera_oasisia_OK.txt
cut -f 1,24 gene_families_originated_vestimentifera.csv > orthogroups_gene_IDs_originated_vestimentifera_riftia.txt
```

```
sed 's/Rpac|//g' orthogroups_gene_IDs_originated_vestimentifera_riftia.txt > orthogroups_gene_IDs_originated_vestimentifera_riftia_OK.txt
cut -f 1,14 gene_families_originated_vestimentifera.csv > orthogroups_gene_IDs_originated_vestimentifera_lamellibrachia.txt
sed 's/Lluy|//g' orthogroups_gene_IDs_originated_vestimentifera_lamellibrachia.txt > orthogroups_gene_IDs_originated_vestimentifera_lamellibrachia_OK.txt
cut -f 1,23 gene_families_originated_vestimentifera.csv > orthogroups_gene_IDs_originated_vestimentifera_paraescarpia.txt
sed 's/Pech|//g' orthogroups_gene_IDs_originated_vestimentifera_paraescarpia.txt > orthogroups_gene_IDs_originated_vestimentifera_paraescarpia_OK.txt


while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OALV* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../oasisia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"oasisia"$'\t'$annotations >> orthogroups_annotations_originated_vestimentifera_oasisia.csv
        done
     else
```

```
37      echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_originated_vestimentifera_
oasisia.csv
38      fi
39  done < orthogroups_gene_IDs_originated_vestimentifera_oasisia_O
K.txt
40
41  while read line; do
42     genes=$(cut -f 2 <<< "$line")
43     echo $genes
44     orthogroup_ID=$(cut -f 1 <<< "$line")
45     echo $orthogroup_ID
46       if [[ "$genes" == RPAC* ]]
47       then
48        IFS=', '      # space is set as delimiter
49        read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
50        for gene in "${ADDR[@]}"; do
51         annotations=$(cut -f 13,14,15,18,20,24 ../../riftia_annot
ation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
52          echo $orthogroup_ID$'\t'"riftia"$'\t'$annotations >> ort
hogroups_annotations_originated_vestimentifera_riftia.csv
53          done
54       else
55       echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_originated_vestimentifera_
riftia.csv
56       fi
57  done < orthogroups_gene_IDs_originated_vestimentifera_riftia_OK.
txt
58
59  while read line; do
```

```
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == FUN* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../lamellibrachia_annotation_Feb2021_TrinoPantherKO_OK.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"lamellibrachia"$'\t'$annotations >> orthogroups_annotations_originated_vestimentifera_lamellibrachia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_vestimentifera_lamellibrachia.csv
     fi
done < orthogroups_gene_IDs_originated_vestimentifera_lamellibrachia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == nbis* ]]
     then
      IFS=', '     # space is set as delimiter
```

```
85      read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
86      for gene in "${ADDR[@]}"; do
87       annotations=$(cut -f 13,14,15,18,20,24 ../../paraescarpia
_annotation_Jun2021_TrinoPantherKO.xls | fgrep $gene)
88        echo $orthogroup_ID$'\t'"paraescarpia"$'\t'$annotations
>> orthogroups_annotations_originated_vestimentifera_paraescarpi
a.csv
89        done
90     else
91     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_originated_vestimentifera_
paraescarpia.csv
92     fi
93 done < orthogroups_gene_IDs_originated_vestimentifera_paraescarp
ia_OK.txt
94
95 echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\
t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_nu
mber" > orthogroups_annotations_originated_vestimentifera_Oalv_R
pac_Lluy_Pech.csv
96 cat orthogroups_annotations_originated_vestimentifera_oasisia.cs
v orthogroups_annotations_originated_vestimentifera_riftia.csv o
rthogroups_annotations_originated_vestimentifera_lamellibrachia.
csv orthogroups_annotations_originated_vestimentifera_paraescarp
ia.csv >> orthogroups_annotations_originated_vestimentifera_Oalv
_Rpac_Lluy_Pech.csv
97 sort orthogroups_annotations_originated_vestimentifera_Oalv_Rpac
_Lluy_Pech.csv > orthogroups_annotations_originated_vestimentife
ra_Oalv_Rpac_Lluy_Pech_OK.csv
98
99 echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbu
ndantAnnotation_originated_vestimentifera.csv
100 while read line; do
```

```
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_originated_vestimentifera_Ofra_Oalv_Rpac_Lluy_Pech_OK.csv | cut -f 7 | sed '/^$/d' | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_originated_vestimentifera.csv
done < gene_families_originated_vestimentifera.csv
```

orthogroups_annotations_originated_vestimentifera_cl1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/vestimentifera_cl1
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/gains/vestimentifera_cl1
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=140:00:0
#$ -l highmem

cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Vestimentifera_cl1 | cut -f 1 > gene_families_originated_vestimentifera_cl1.txt #families originated in oasisia
fgrep -f gene_families_originated_vestimentifera_cl1.txt ../../Orthogroups.csv > gene_families_originated_vestimentifera_cl1.csv

cut -f 1,19 gene_families_originated_vestimentifera_cl1.csv > orthogroups_gene_IDs_originated_vestimentifera_cl1_oasisia.txt
sed 's/Oalv|//g' orthogroups_gene_IDs_originated_vestimentifera_cl1_oasisia.txt > orthogroups_gene_IDs_originated_vestimentifera_cl1_oasisia_OK.txt
cut -f 1,24 gene_families_originated_vestimentifera_cl1.csv > or
```

```
thogroups_gene_IDs_originated_vestimentifera_cl1_riftia.txt
sed 's/Rpac|//g' orthogroups_gene_IDs_originated_vestimentifera_cl1_riftia.txt > orthogroups_gene_IDs_originated_vestimentifera_cl1_riftia_OK.txt
cut -f 1,23 gene_families_originated_vestimentifera_cl1.csv > orthogroups_gene_IDs_originated_vestimentifera_cl1_paraescarpia.txt
sed 's/Pech|//g' orthogroups_gene_IDs_originated_vestimentifera_cl1_paraescarpia.txt > orthogroups_gene_IDs_originated_vestimentifera_cl1_paraescarpia_OK.txt


while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OALV* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../oasisia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"oasisia"$'\t'$annotations >> orthogroups_annotations_originated_vestimentifera_cl1_oasisia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_vestimentifera_cl1_oasisia.csv
```

```
     fi
done < orthogroups_gene_IDs_originated_vestimentifera_cl1_oasisia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == RPAC* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../riftia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"riftia"$'\t'$annotations >> orthogroups_annotations_originated_vestimentifera_cl1_riftia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_vestimentifera_cl1_riftia.csv
     fi
done < orthogroups_gene_IDs_originated_vestimentifera_cl1_riftia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
```

```
61    echo $orthogroup_ID
62      if [[ "$genes" == nbis* ]]
63      then
64       IFS=', '     # space is set as delimiter
65       read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
66       for gene in "${ADDR[@]}"; do
67        annotations=$(cut -f 13,14,15,18,20,24 ../../paraescarpia
_annotation_Jun2021_TrinoPantherKO.xls | fgrep $gene)
68         echo $orthogroup_ID$'\t'"paraescarpia"$'\t'$annotations
>> orthogroups_annotations_originated_vestimentifera_cl1_paraesc
arpia.csv
69         done
70      else
71      echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_originated_vestimentifera_
cl1_paraescarpia.csv
72      fi
73  done < orthogroups_gene_IDs_originated_vestimentifera_cl1_paraes
carpia_OK.txt
74
75  echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\
t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_nu
mber" > orthogroups_annotations_originated_vestimentifera_cl1_Oa
lv_Rpac_Pech.csv
76  cat orthogroups_annotations_originated_vestimentifera_cl1_oasisi
a.csv orthogroups_annotations_originated_vestimentifera_cl1_rift
ia.csv orthogroups_annotations_originated_vestimentifera_cl1_par
aescarpia.csv >> orthogroups_annotations_originated_vestimentife
ra_cl1_Oalv_Rpac_Pech.csv
77  sort orthogroups_annotations_originated_vestimentifera_cl1_Oalv_
Rpac_Pech.csv > orthogroups_annotations_originated_vestimentifer
a_cl1_Oalv_Rpac_Pech_OK.csv
```

```
78
79 echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbu
   ndantAnnotation_originated_vestimentifera_cl1.csv
80 while read line; do
81    orthogroup_ID=$(cut -f 1 <<< "$line")
82    annotation=$(fgrep $orthogroup_ID orthogroups_annotations_ori
   ginated_vestimentifera_cl1_Ofra_Oalv_Rpac_Lluy_Pech_OK.csv | cut
   -f 7 | sed '/^$/d' | sort |  uniq -c | sort -r | awk '{$1=""; pr
   int $0}' | head -1)
83         echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostA
   bundantAnnotation_originated_vestimentifera_cl1.csv
84 done < gene_families_originated_vestimentifera_cl1.csv
```

orthogroups_annotations_originated_vestimentifera_cl2.sh

```
1  #!/bin/bash
2  #$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/J
   ul2021/gains/vestimentifera_cl2
3  #$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Ju
   l2021/gains/vestimentifera_cl2
4  #$ -j y
5  #$ -pe smp 1
6  #$ -l h_vmem=100G
7  #$ -l h_rt=140:00:0
8  #$ -l highmem
9
10 cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv |
   grep -w Vestimentifera_cl2 | cut -f 1 > gene_families_originated
   _vestimentifera_cl2.txt #families originated in oasisia
11 fgrep -f gene_families_originated_vestimentifera_cl2.txt ../../O
   rthogroups.csv > gene_families_originated_vestimentifera_cl2.csv
12
13 cut -f 1,19 gene_families_originated_vestimentifera_cl2.csv > or
   thogroups_gene_IDs_originated_vestimentifera_cl2_oasisia.txt
```

```
sed 's/Oalv|//g' orthogroups_gene_IDs_originated_vestimentifera_cl2_oasisia.txt > orthogroups_gene_IDs_originated_vestimentifera_cl2_oasisia_OK.txt
cut -f 1,24 gene_families_originated_vestimentifera_cl2.csv > orthogroups_gene_IDs_originated_vestimentifera_cl2_riftia.txt
sed 's/Rpac|//g' orthogroups_gene_IDs_originated_vestimentifera_cl2_riftia.txt > orthogroups_gene_IDs_originated_vestimentifera_cl2_riftia_OK.txt



while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OALV* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../oasisia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"oasisia"$'\t'$annotations >> orthogroups_annotations_originated_vestimentifera_cl2_oasisia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_vestimentifera_cl2_oasisia.csv
     fi
```

```
done < orthogroups_gene_IDs_originated_vestimentifera_cl2_oasisia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == RPAC* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../riftia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"riftia"$'\t'$annotations >> orthogroups_annotations_originated_vestimentifera_cl2_riftia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_originated_vestimentifera_cl2_riftia.csv
     fi
done < orthogroups_gene_IDs_originated_vestimentifera_cl2_riftia_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_originated_vestimentifera_cl2_Oalv_Rpac.csv
cat orthogroups_annotations_originated_vestimentifera_cl2_oasisi
```

```
a.csv orthogroups_annotations_originated_vestimentifera_cl2_rift
ia.csv >> orthogroups_annotations_originated_vestimentifera_cl2_
Oalv_Rpac.csv
sort orthogroups_annotations_originated_vestimentifera_cl2_Oalv_
Rpac.csv > orthogroups_annotations_originated_vestimentifera_cl2
_Oalv_Rpac_OK.csv

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbu
ndantAnnotation_originated_vestimentifera_cl2.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_ori
ginated_vestimentifera_cl2_Oalv_Rpac_OK.csv | cut -f 7 | sed '/
^$/d' | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | he
ad -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostA
bundantAnnotation_originated_vestimentifera_cl2.csv
done < gene_families_originated_vestimentifera_cl2.csv
```

Losses

orthogroups_annotations_losses_osedax.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/J
ul2021/losses/osedax
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Ju
l2021/losses/osedax
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=140:00:0
#$ -l highmem

```

```
cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Ofra | cut -f 1 > gene_families_losses_osedax.txt #families losses in oasisia
fgrep -f gene_families_losses_osedax.txt ../../Orthogroups.csv > gene_families_losses_osedax.csv

cut -f 1,19 gene_families_losses_osedax.csv > orthogroups_gene_IDs_losses_osedax_oasisia.txt
sed 's/Oalv|//g' orthogroups_gene_IDs_losses_osedax_oasisia.txt > orthogroups_gene_IDs_losses_osedax_oasisia_OK.txt
cut -f 1,24 gene_families_losses_osedax.csv > orthogroups_gene_IDs_losses_osedax_riftia.txt
sed 's/Rpac|//g' orthogroups_gene_IDs_losses_osedax_riftia.txt > orthogroups_gene_IDs_losses_osedax_riftia_OK.txt
cut -f 1,14 gene_families_losses_osedax.csv > orthogroups_gene_IDs_losses_osedax_lamellibrachia.txt
sed 's/Lluy|//g' orthogroups_gene_IDs_losses_osedax_lamellibrachia.txt > orthogroups_gene_IDs_losses_osedax_lamellibrachia_OK.txt
cut -f 1,23 gene_families_losses_osedax.csv > orthogroups_gene_IDs_losses_osedax_paraescarpia.txt
sed 's/Pech|//g' orthogroups_gene_IDs_losses_osedax_paraescarpia.txt > orthogroups_gene_IDs_losses_osedax_paraescarpia_OK.txt
cut -f 1,21 gene_families_losses_osedax.csv > orthogroups_gene_IDs_losses_osedax_owenia.txt
sed 's/Ofus|//g' orthogroups_gene_IDs_losses_osedax_owenia.txt > orthogroups_gene_IDs_losses_osedax_owenia_OK.txt
cut -f 1,5 gene_families_losses_osedax.csv > orthogroups_gene_IDs_losses_osedax_capitella.txt
sed 's/Ctel|//g' orthogroups_gene_IDs_losses_osedax_capitella.txt > orthogroups_gene_IDs_losses_osedax_capitella_OK.txt


```

```
while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFUS* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
        cut -f 1,2,3,11,12,13 ../../Owenia_annotation_v250920.1_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 1 temp_file.txt)
        #gene_ID=$(cut -f 2 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 4 temp_file.txt)
        #GO_1=$(cut -f 5 temp_file.txt)
        #GO_1=$(cut -f 6 temp_file.txt)
        echo $orthogroup_ID$'\t'"owenia"$'\t'$(cut -f 4 temp_file.txt)$'\t'$(cut -f 5 temp_file.txt)$'\t'$(cut -f 6 temp_file.txt)$'\t'$(cut -f 2 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 1 temp_file.txt) >> orthogroups_annotations_losses_osedax_owenia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_osedax_owenia.csv
     fi
done < orthogroups_gene_IDs_losses_osedax_owenia_OK.txt

while read line; do
```

```
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == CapteT* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       cut -f 1,3,7,21,22 ../../Capitella_annotation_Feb2021_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 7 temp_file.txt)
        #gene_ID=$(cut -f 1 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 21 temp_file.txt)
        #GO_1=$(cut -f 22 temp_file.txt)
        #GO_1=$(cut -f 6 temp_file.txt) NONE
        echo $orthogroup_ID$'\t'"capitella"$'\t'$(cut -f 21 temp_file.txt)$'\t'$(cut -f 22 temp_file.txt)$'\t'""'\t'$(cut -f 1 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 7 temp_file.txt) >> orthogroups_annotations_losses_osedax_capitella.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""" >> orthogroups_annotations_losses_osedax_capitella.csv
     fi
done < orthogroups_gene_IDs_losses_osedax_capitella_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
```

```
77    echo $genes
78    orthogroup_ID=$(cut -f 1 <<< "$line")
79    echo $orthogroup_ID
80      if [[ "$genes" == OALV* ]]
81      then
82       IFS=', '     # space is set as delimiter
83       read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
84       for gene in "${ADDR[@]}"; do
85        annotations=$(cut -f 13,14,15,18,20,24 ../../oasisia_anno
tation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
86         echo $orthogroup_ID$'\t'"oasisia"$'\t'$annotations >> or
thogroups_annotations_losses_osedax_oasisia.csv
87         done
88      else
89      echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_losses_osedax_oasisia.csv
90      fi
91  done < orthogroups_gene_IDs_losses_osedax_oasisia_OK.txt
92
93  while read line; do
94    genes=$(cut -f 2 <<< "$line")
95    echo $genes
96    orthogroup_ID=$(cut -f 1 <<< "$line")
97    echo $orthogroup_ID
98      if [[ "$genes" == RPAC* ]]
99      then
100      IFS=', '     # space is set as delimiter
101      read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
102      for gene in "${ADDR[@]}"; do
```

```
       annotations=$(cut -f 13,14,15,18,20,24 ../../riftia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"riftia"$'\t'$annotations >> orthogroups_annotations_losses_osedax_riftia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_osedax_riftia.csv
     fi
done < orthogroups_gene_IDs_losses_osedax_riftia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == FUN* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../lamellibrachia_annotation_Feb2021_TrinoPantherKO_OK.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"lamellibrachia"$'\t'$annotations >> orthogroups_annotations_losses_osedax_lamellibrachia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_osedax_lamellibrachia.csv
     fi
```

```
done < orthogroups_gene_IDs_losses_osedax_lamellibrachia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == nbis* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../paraescarpia_annotation_Jun2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"paraescarpia"$'\t'$annotations >> orthogroups_annotations_losses_osedax_paraescarpia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_osedax_paraescarpia.csv
     fi
done < orthogroups_gene_IDs_losses_osedax_paraescarpia_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_losses_osedax_Oalv_Rpac_Lluy_Pech_Ofus_Ctel.csv
cat orthogroups_annotations_losses_osedax_oasisia.csv orthogroups_annotations_losses_osedax_riftia.csv orthogroups_annotations_losses_osedax_lamellibrachia.csv orthogroups_annotations_losses_o
```

```
sedax_paraescarpia.csv orthogroups_annotations_losses_osedax_owenia.csv orthogroups_annotations_losses_osedax_capitella.csv >> orthogroups_annotations_losses_osedax_Oalv_Rpac_Lluy_Pech_Ofus_Ctel.csv
sort orthogroups_annotations_losses_osedax_Oalv_Rpac_Lluy_Pech_Ofus_Ctel.csv > orthogroups_annotations_losses_osedax_Oalv_Rpac_Lluy_Pech_Ofus_Ctel_OK.csv

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_losses_osedax.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_losses_osedax_Oalv_Rpac_Lluy_Pech_Ofus_Ctel_OK.csv | cut -f 7 | sed '/^$/d' | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_losses_osedax.csv
done < gene_families_losses_osedax.csv
```

orthogroups_annotations_losses_oasisia.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/losses/oasisia
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/losses/oasisia
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=140:00:0
#$ -l highmem

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv |
```

```
grep -w Oalv | cut -f 1 > gene_families_losses_oasisia.txt #families losses in oasisia
fgrep -f gene_families_losses_oasisia.txt ../../Orthogroups.csv > gene_families_losses_oasisia.csv

cut -f 1,20 gene_families_losses_oasisia.csv > orthogroups_gene_IDs_losses_oasisia_osedax.txt
sed 's/Ofra|//g' orthogroups_gene_IDs_losses_oasisia_osedax.txt > orthogroups_gene_IDs_losses_oasisia_osedax_OK.txt
cut -f 1,24 gene_families_losses_oasisia.csv > orthogroups_gene_IDs_losses_oasisia_riftia.txt
sed 's/Rpac|//g' orthogroups_gene_IDs_losses_oasisia_riftia.txt > orthogroups_gene_IDs_losses_oasisia_riftia_OK.txt
cut -f 1,14 gene_families_losses_oasisia.csv > orthogroups_gene_IDs_losses_oasisia_lamellibrachia.txt
sed 's/Lluy|//g' orthogroups_gene_IDs_losses_oasisia_lamellibrachia.txt > orthogroups_gene_IDs_losses_oasisia_lamellibrachia_OK.txt
cut -f 1,23 gene_families_losses_oasisia.csv > orthogroups_gene_IDs_losses_oasisia_paraescarpia.txt
sed 's/Pech|//g' orthogroups_gene_IDs_losses_oasisia_paraescarpia.txt > orthogroups_gene_IDs_losses_oasisia_paraescarpia_OK.txt
cut -f 1,21 gene_families_losses_oasisia.csv > orthogroups_gene_IDs_losses_oasisia_owenia.txt
sed 's/Ofus|//g' orthogroups_gene_IDs_losses_oasisia_owenia.txt > orthogroups_gene_IDs_losses_oasisia_owenia_OK.txt
cut -f 1,5 gene_families_losses_oasisia.csv > orthogroups_gene_IDs_losses_oasisia_capitella.txt
sed 's/Ctel|//g' orthogroups_gene_IDs_losses_oasisia_capitella.txt > orthogroups_gene_IDs_losses_oasisia_capitella_OK.txt


while read line; do
```

```
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFUS* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
        cut -f 1,2,3,11,12,13 ../../Owenia_annotation_v250920.1_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 1 temp_file.txt)
        #gene_ID=$(cut -f 2 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 4 temp_file.txt)
        #GO_1=$(cut -f 5 temp_file.txt)
        #GO_1=$(cut -f 6 temp_file.txt)
        echo $orthogroup_ID$'\t'"owenia"$'\t'$(cut -f 4 temp_file.txt)$'\t'$(cut -f 5 temp_file.txt)$'\t'$(cut -f 6 temp_file.txt)$'\t'$(cut -f 2 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 1 temp_file.txt) >> orthogroups_annotations_losses_oasisia_owenia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_oasisia_owenia.csv
     fi
done < orthogroups_gene_IDs_losses_oasisia_owenia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
```

```
53    echo $genes
54    orthogroup_ID=$(cut -f 1 <<< "$line")
55    echo $orthogroup_ID
56      if [[ "$genes" == CapteT* ]]
57      then
58       IFS=', '     # space is set as delimiter
59       read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
60       for gene in "${ADDR[@]}"; do
61        cut -f 1,3,7,21,22 ../../Capitella_annotation_Feb2021_Tri
noPantherKO.xls | fgrep $gene > temp_file.txt
62         #K0_number=$(cut -f 7 temp_file.txt)
63         #gene_ID=$(cut -f 1 temp_file.txt)
64         #Panther_annotation=$(cut -f 3 temp_file.txt)
65         #GO_1=$(cut -f 21 temp_file.txt)
66         #GO_1=$(cut -f 22 temp_file.txt)
67         #GO_1=$(cut -f 6 temp_file.txt) NONE
68         echo $orthogroup_ID$'\t'"capitella"$'\t'$(cut -f 21 temp
_file.txt)$'\t'$(cut -f 22 temp_file.txt)$'\t'""'\t'$(cut -f 1 t
emp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 7 temp_
file.txt) >> orthogroups_annotations_losses_oasisia_capitella.cs
v
69         done
70      else
71      echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_losses_oasisia_capitella.c
sv
72      fi
73  done < orthogroups_gene_IDs_losses_oasisia_capitella_OK.txt
74
75  while read line; do
76    genes=$(cut -f 2 <<< "$line")
```

```
77    echo $genes
78    orthogroup_ID=$(cut -f 1 <<< "$line")
79    echo $orthogroup_ID
80      if [[ "$genes" == OFRA* ]]
81      then
82       IFS=', '     # space is set as delimiter
83       read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
84       for gene in "${ADDR[@]}"; do
85        annotations=$(cut -f 13,14,15,18,20,24 ../../osedax_annot
ation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
86         echo $orthogroup_ID$'\t'"osedax"$'\t'$annotations >> ort
hogroups_annotations_losses_oasisia_osedax.csv
87         done
88      else
89      echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_losses_oasisia_osedax.csv
90      fi
91  done < orthogroups_gene_IDs_losses_oasisia_osedax_OK.txt
92
93  while read line; do
94    genes=$(cut -f 2 <<< "$line")
95    echo $genes
96    orthogroup_ID=$(cut -f 1 <<< "$line")
97    echo $orthogroup_ID
98      if [[ "$genes" == RPAC* ]]
99      then
100      IFS=', '     # space is set as delimiter
101      read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
102      for gene in "${ADDR[@]}"; do
```

```
       annotations=$(cut -f 13,14,15,18,20,24 ../../riftia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"riftia"$'\t'$annotations >> orthogroups_annotations_losses_oasisia_riftia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_oasisia_riftia.csv
     fi
done < orthogroups_gene_IDs_losses_oasisia_riftia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == FUN* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../lamellibrachia_annotation_Feb2021_TrinoPantherKO_OK.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"lamellibrachia"$'\t'$annotations >> orthogroups_annotations_losses_oasisia_lamellibrachia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_oasisia_lamellibrachia.csv
     fi
```

```
done < orthogroups_gene_IDs_losses_oasisia_lamellibrachia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == nbis* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../paraescarpia_annotation_Jun2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"paraescarpia"$'\t'$annotations >> orthogroups_annotations_losses_oasisia_paraescarpia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_oasisia_paraescarpia.csv
     fi
done < orthogroups_gene_IDs_losses_oasisia_paraescarpia_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_losses_oasisia_Ofra_Rpac_Lluy_Pech_Ofus_Ctel.csv
cat orthogroups_annotations_losses_oasisia_osedax.csv orthogroups_annotations_losses_oasisia_riftia.csv orthogroups_annotations_losses_oasisia_lamellibrachia.csv orthogroups_annotations_losses
```

```
_oasisia_paraescarpia.csv orthogroups_annotations_losses_oasisia
_owenia.csv orthogroups_annotations_losses_oasisia_capitella.csv
>> orthogroups_annotations_losses_oasisia_Ofra_Rpac_Lluy_Pech_Of
us_Ctel.csv
sort orthogroups_annotations_losses_oasisia_Ofra_Rpac_Lluy_Pech_
Ofus_Ctel.csv > orthogroups_annotations_losses_oasisia_Ofra_Rpac
_Lluy_Pech_Ofus_Ctel_OK.csv

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbu
ndantAnnotation_losses_oasisia.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_los
ses_oasisia_Ofra_Rpac_Lluy_Pech_Ofus_Ctel_OK.csv | cut -f 7 | se
d '/^$/d' | sort |  uniq -c | sort -r | awk '{$1=""; print $0}'
| head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostA
bundantAnnotation_losses_oasisia.csv
done < gene_families_losses_oasisia.csv
```

orthogroups_annotations_losses_riftia.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/J
ul2021/losses/riftia
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Ju
l2021/losses/riftia
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=140:00:0
#$ -l highmem

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv |
```

```
grep -w Rpac | cut -f 1 > gene_families_losses_riftia.txt #famil
ies losses in oasisia
fgrep -f gene_families_losses_riftia.txt ../../Orthogroups.csv >
gene_families_losses_riftia.csv

cut -f 1,20 gene_families_losses_riftia.csv > orthogroups_gene_I
Ds_losses_riftia_osedax.txt
sed 's/Ofra|//g' orthogroups_gene_IDs_losses_riftia_osedax.txt >
orthogroups_gene_IDs_losses_riftia_osedax_OK.txt
cut -f 1,14 gene_families_losses_riftia.csv > orthogroups_gene_I
Ds_losses_riftia_lamellibrachia.txt
sed 's/Lluy|//g' orthogroups_gene_IDs_losses_riftia_lamellibrach
ia.txt > orthogroups_gene_IDs_losses_riftia_lamellibrachia_OK.tx
t
cut -f 1,23 gene_families_losses_riftia.csv > orthogroups_gene_I
Ds_losses_riftia_paraescarpia.txt
sed 's/Pech|//g' orthogroups_gene_IDs_losses_riftia_paraescarpi
a.txt > orthogroups_gene_IDs_losses_riftia_paraescarpia_OK.txt
cut -f 1,21 gene_families_losses_riftia.csv > orthogroups_gene_I
Ds_losses_riftia_owenia.txt
sed 's/Ofus|//g' orthogroups_gene_IDs_losses_riftia_owenia.txt >
orthogroups_gene_IDs_losses_riftia_owenia_OK.txt
cut -f 1,5 gene_families_losses_riftia.csv > orthogroups_gene_ID
s_losses_riftia_capitella.txt
sed 's/Ctel|//g' orthogroups_gene_IDs_losses_riftia_capitella.tx
t > orthogroups_gene_IDs_losses_riftia_capitella_OK.txt
cut -f 1,19 gene_families_losses_riftia.csv > orthogroups_gene_I
Ds_losses_riftia_oasisia.txt
sed 's/Oalv|//g' orthogroups_gene_IDs_losses_riftia_oasisia.txt
> orthogroups_gene_IDs_losses_riftia_oasisia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
```

```
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OALV* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../oasisia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"oasisia"$'\t'$annotations >> orthogroups_annotations_losses_riftia_oasisia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_riftia_oasisia.csv
     fi
done < orthogroups_gene_IDs_losses_riftia_oasisia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFUS* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
```

```
        cut -f 1,2,3,11,12,13 ../../Owenia_annotation_v250920.1_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 1 temp_file.txt)
        #gene_ID=$(cut -f 2 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 4 temp_file.txt)
        #GO_1=$(cut -f 5 temp_file.txt)
        #GO_1=$(cut -f 6 temp_file.txt)
        echo $orthogroup_ID$'\t'"owenia"$'\t'$(cut -f 4 temp_file.txt)$'\t'$(cut -f 5 temp_file.txt)$'\t'$(cut -f 6 temp_file.txt)$'\t'$(cut -f 2 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 1 temp_file.txt) >> orthogroups_annotations_losses_riftia_owenia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_riftia_owenia.csv
     fi
done < orthogroups_gene_IDs_losses_riftia_owenia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == CapteT* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       cut -f 1,3,7,21,22 ../../Capitella_annotation_Feb2021_Tri
```

```
noPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 7 temp_file.txt)
        #gene_ID=$(cut -f 1 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 21 temp_file.txt)
        #GO_1=$(cut -f 22 temp_file.txt)
        #GO_1=$(cut -f 6 temp_file.txt) NONE
        echo $orthogroup_ID$'\t'"capitella"$'\t'$(cut -f 21 temp_file.txt)$'\t'$(cut -f 22 temp_file.txt)$'\t'""'\t'$(cut -f 1 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 7 temp_file.txt) >> orthogroups_annotations_losses_riftia_capitella.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_riftia_capitella.csv
     fi
done < orthogroups_gene_IDs_losses_riftia_capitella_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFRA* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../osedax_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
```

```
        echo $orthogroup_ID$'\t'"osedax"$'\t'$annotations >> orthogroups_annotations_losses_riftia_osedax.csv
        done
    else
    echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""" >> orthogroups_annotations_losses_riftia_osedax.csv
    fi
done < orthogroups_gene_IDs_losses_riftia_osedax_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == FUN* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../lamellibrachia_annotation_Feb2021_TrinoPantherKO_OK.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"lamellibrachia"$'\t'$annotations >> orthogroups_annotations_losses_riftia_lamellibrachia.csv
        done
    else
    echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""" >> orthogroups_annotations_losses_riftia_lamellibrachia.csv
    fi
done < orthogroups_gene_IDs_losses_riftia_lamellibrachia_OK.txt

```

```
while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == nbis* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../paraescarpia_annotation_Jun2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"paraescarpia"$'\t'$annotations >> orthogroups_annotations_losses_riftia_paraescarpia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_riftia_paraescarpia.csv
     fi
done < orthogroups_gene_IDs_losses_riftia_paraescarpia_OK.txt

echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_losses_riftia_Ofra_Oalv_Lluy_Pech_Ofus_Ctel.csv
cat orthogroups_annotations_losses_riftia_oasisia.csv orthogroups_annotations_losses_riftia_osedax.csv orthogroups_annotations_losses_riftia_lamellibrachia.csv orthogroups_annotations_losses_riftia_paraescarpia.csv orthogroups_annotations_losses_riftia_owenia.csv orthogroups_annotations_losses_riftia_capitella.csv >> orthogroups_annotations_losses_riftia_Ofra_Oalv_Lluy_Pech_Ofus_Ct
```

```
el.csv
148 sort orthogroups_annotations_losses_riftia_Ofra_Oalv_Lluy_Pech_Ofus_Ctel.csv > orthogroups_annotations_losses_riftia_Ofra_Oalv_Lluy_Pech_Ofus_Ctel_OK.csv
149
150 echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_losses_riftia.csv
151 while read line; do
152    orthogroup_ID=$(cut -f 1 <<< "$line")
153    annotation=$(fgrep $orthogroup_ID orthogroups_annotations_losses_riftia_Ofra_Oalv_Lluy_Pech_Ofus_Ctel_OK.csv | cut -f 7 | sed '/^$/d' | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
154         echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_losses_riftia.csv
155 done < gene_families_losses_riftia.csv
```

orthogroups_annotations_losses_lamellibrachia.sh

```
1  #!/bin/bash
2  #$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/losses/lamellibrachia
3  #$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/losses/lamellibrachia
4  #$ -j y
5  #$ -pe smp 1
6  #$ -l h_vmem=100G
7  #$ -l h_rt=140:00:0
8  #$ -l highmem
9
10 cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Lluy | cut -f 1 > gene_families_losses_lamellibrachia.txt #families losses in oasisia
11 fgrep -f gene_families_losses_lamellibrachia.txt ../../Orthogrou
```

```
ps.csv > gene_families_losses_lamellibrachia.csv

cut -f 1,20 gene_families_losses_lamellibrachia.csv > orthogroup
s_gene_IDs_losses_lamellibrachia_osedax.txt
sed 's/Ofra|//g' orthogroups_gene_IDs_losses_lamellibrachia_osed
ax.txt > orthogroups_gene_IDs_losses_lamellibrachia_osedax_OK.tx
t
cut -f 1,23 gene_families_losses_lamellibrachia.csv > orthogroup
s_gene_IDs_losses_lamellibrachia_paraescarpia.txt
sed 's/Pech|//g' orthogroups_gene_IDs_losses_lamellibrachia_para
escarpia.txt > orthogroups_gene_IDs_losses_lamellibrachia_paraes
carpia_OK.txt
cut -f 1,21 gene_families_losses_lamellibrachia.csv > orthogroup
s_gene_IDs_losses_lamellibrachia_owenia.txt
sed 's/Ofus|//g' orthogroups_gene_IDs_losses_lamellibrachia_owen
ia.txt > orthogroups_gene_IDs_losses_lamellibrachia_owenia_OK.tx
t
cut -f 1,5 gene_families_losses_lamellibrachia.csv > orthogroups
_gene_IDs_losses_lamellibrachia_capitella.txt
sed 's/Ctel|//g' orthogroups_gene_IDs_losses_lamellibrachia_capi
tella.txt > orthogroups_gene_IDs_losses_lamellibrachia_capitella
_OK.txt
cut -f 1,19 gene_families_losses_lamellibrachia.csv > orthogroup
s_gene_IDs_losses_lamellibrachia_oasisia.txt
sed 's/Oalv|//g' orthogroups_gene_IDs_losses_lamellibrachia_oasi
sia.txt > orthogroups_gene_IDs_losses_lamellibrachia_oasisia_OK.
txt
cut -f 1,24 gene_families_losses_lamellibrachia.csv > orthogroup
s_gene_IDs_losses_lamellibrachia_riftia.txt
sed 's/Rpac|//g' orthogroups_gene_IDs_losses_lamellibrachia_rift
ia.txt > orthogroups_gene_IDs_losses_lamellibrachia_riftia_OK.tx
t

while read line; do
```

```
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == RPAC* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../riftia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"riftia"$'\t'$annotations >> orthogroups_annotations_losses_lamellibrachia_riftia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""" >> orthogroups_annotations_losses_lamellibrachia_riftia.csv
     fi
done < orthogroups_gene_IDs_losses_lamellibrachia_riftia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OALV* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array a
```

```
s tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../oasisia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"oasisia"$'\t'$annotations >> orthogroups_annotations_losses_lamellibrachia_oasisia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_lamellibrachia_oasisia.csv
     fi
done < orthogroups_gene_IDs_losses_lamellibrachia_oasisia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFUS* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
        cut -f 1,2,3,11,12,13 ../../Owenia_annotation_v250920.1_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 1 temp_file.txt)
        #gene_ID=$(cut -f 2 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 4 temp_file.txt)
```

```
        #GO_1=$(cut -f 5 temp_file.txt)
        #GO_1=$(cut -f 6 temp_file.txt)
        echo $orthogroup_ID$'\t'"owenia"$'\t'$(cut -f 4 temp_file.txt)$'\t'$(cut -f 5 temp_file.txt)$'\t'$(cut -f 6 temp_file.txt)$'\t'$(cut -f 2 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 1 temp_file.txt) >> orthogroups_annotations_losses_lamellibrachia_owenia.csv
        done
    else
    echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_lamellibrachia_owenia.csv
    fi
done < orthogroups_gene_IDs_losses_lamellibrachia_owenia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == CapteT* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       cut -f 1,3,7,21,22 ../../Capitella_annotation_Feb2021_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 7 temp_file.txt)
        #gene_ID=$(cut -f 1 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 21 temp_file.txt)
```

```
        #GO_1=$(cut -f 22 temp_file.txt)
        #GO_1=$(cut -f 6 temp_file.txt) NONE
        echo $orthogroup_ID$'\t'"capitella"$'\t'$(cut -f 21 temp_file.txt)$'\t'$(cut -f 22 temp_file.txt)$'\t'""'\t'$(cut -f 1 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 7 temp_file.txt) >> orthogroups_annotations_losses_lamellibrachia_capitella.csv
        done
    else
    echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_lamellibrachia_capitella.csv
    fi
done < orthogroups_gene_IDs_losses_lamellibrachia_capitella_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFRA* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../osedax_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"osedax"$'\t'$annotations >> orthogroups_annotations_losses_lamellibrachia_osedax.csv
        done
```

```
123     else
124     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_losses_lamellibrachia_osed
ax.csv
125     fi
126 done < orthogroups_gene_IDs_losses_lamellibrachia_osedax_OK.txt
127 
128 while read line; do
129    genes=$(cut -f 2 <<< "$line")
130    echo $genes
131    orthogroup_ID=$(cut -f 1 <<< "$line")
132    echo $orthogroup_ID
133      if [[ "$genes" == nbis* ]]
134      then
135       IFS=', '     # space is set as delimiter
136       read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
137       for gene in "${ADDR[@]}"; do
138        annotations=$(cut -f 13,14,15,18,20,24 ../../paraescarpia
_annotation_Jun2021_TrinoPantherKO.xls | fgrep $gene)
139         echo $orthogroup_ID$'\t'"paraescarpia"$'\t'$annotations
>> orthogroups_annotations_losses_lamellibrachia_paraescarpia.cs
v
140         done
141     else
142     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_losses_lamellibrachia_para
escarpia.csv
143     fi
144 done < orthogroups_gene_IDs_losses_lamellibrachia_paraescarpia_O
K.txt
145 
```

```
echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_losses_lamellibrachia_Ofra_Oalv_Rpac_Pech_Ofus_Ctel.csv
cat orthogroups_annotations_losses_lamellibrachia_oasisia.csv orthogroups_annotations_losses_lamellibrachia_osedax.csv orthogroups_annotations_losses_lamellibrachia_riftia.csv orthogroups_annotations_losses_lamellibrachia_paraescarpia.csv orthogroups_annotations_losses_lamellibrachia_owenia.csv orthogroups_annotations_losses_lamellibrachia_capitella.csv >> orthogroups_annotations_losses_lamellibrachia_Ofra_Oalv_Rpac_Pech_Ofus_Ctel.csv
sort orthogroups_annotations_losses_lamellibrachia_Ofra_Oalv_Rpac_Pech_Ofus_Ctel.csv > orthogroups_annotations_losses_lamellibrachia_Ofra_Oalv_Rpac_Pech_Ofus_Ctel_OK.csv

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_losses_lamellibrachia.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_losses_lamellibrachia_Ofra_Oalv_Rpac_Pech_Ofus_Ctel_OK.csv | cut -f 7 | sed '/^$/d' | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_losses_lamellibrachia.csv
done < gene_families_losses_lamellibrachia.csv
```

orthogroups_annotations_losses_paraescarpia.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/losses/paraescarpia
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/losses/paraescarpia
#$ -j y
```

```
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=140:00:0
#$ -l highmem

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Pech | cut -f 1 > gene_families_losses_paraescarpia.txt
#families losses in oasisia
fgrep -f gene_families_losses_paraescarpia.txt ../../Orthogroups.csv > gene_families_losses_paraescarpia.csv

cut -f 1,20 gene_families_losses_paraescarpia.csv > orthogroups_gene_IDs_losses_paraescarpia_osedax.txt
sed 's/Ofra|//g' orthogroups_gene_IDs_losses_paraescarpia_osedax.txt > orthogroups_gene_IDs_losses_paraescarpia_osedax_OK.txt
cut -f 1,21 gene_families_losses_paraescarpia.csv > orthogroups_gene_IDs_losses_paraescarpia_owenia.txt
sed 's/Ofus|//g' orthogroups_gene_IDs_losses_paraescarpia_owenia.txt > orthogroups_gene_IDs_losses_paraescarpia_owenia_OK.txt
cut -f 1,5 gene_families_losses_paraescarpia.csv > orthogroups_gene_IDs_losses_paraescarpia_capitella.txt
sed 's/Ctel|//g' orthogroups_gene_IDs_losses_paraescarpia_capitella.txt > orthogroups_gene_IDs_losses_paraescarpia_capitella_OK.txt
cut -f 1,19 gene_families_losses_paraescarpia.csv > orthogroups_gene_IDs_losses_paraescarpia_oasisia.txt
sed 's/Oalv|//g' orthogroups_gene_IDs_losses_paraescarpia_oasisia.txt > orthogroups_gene_IDs_losses_paraescarpia_oasisia_OK.txt
cut -f 1,24 gene_families_losses_paraescarpia.csv > orthogroups_gene_IDs_losses_paraescarpia_riftia.txt
sed 's/Rpac|//g' orthogroups_gene_IDs_losses_paraescarpia_riftia.txt > orthogroups_gene_IDs_losses_paraescarpia_riftia_OK.txt
cut -f 1,14 gene_families_losses_paraescarpia.csv > orthogroups_
```

```
gene_IDs_losses_paraescarpia_lamellibrachia.txt
sed 's/Lluy|//g' orthogroups_gene_IDs_losses_paraescarpia_lamellibrachia.txt > orthogroups_gene_IDs_losses_paraescarpia_lamellibrachia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == FUN* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../lamellibrachia_annotation_Feb2021_TrinoPantherKO_OK.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"lamellibrachia"$'\t'$annotations >> orthogroups_annotations_losses_paraescarpia_lamellibrachia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_paraescarpia_lamellibrachia.csv
     fi
done < orthogroups_gene_IDs_losses_paraescarpia_lamellibrachia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
```

```
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == RPAC* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../riftia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"riftia"$'\t'$annotations >> orthogroups_annotations_losses_paraescarpia_riftia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_paraescarpia_riftia.csv
     fi
done < orthogroups_gene_IDs_losses_paraescarpia_riftia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OALV* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
```

```
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../oasisia_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"oasisia"$'\t'$annotations >> orthogroups_annotations_losses_paraescarpia_oasisia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_paraescarpia_oasisia.csv
     fi
done < orthogroups_gene_IDs_losses_paraescarpia_oasisia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFUS* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
        cut -f 1,2,3,11,12,13 ../../Owenia_annotation_v250920.1_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 1 temp_file.txt)
        #gene_ID=$(cut -f 2 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 4 temp_file.txt)
        #GO_1=$(cut -f 5 temp_file.txt)
```

```
        #GO_1=$(cut -f 6 temp_file.txt)
        echo $orthogroup_ID$'\t'"owenia"$'\t'$(cut -f 4 temp_file.txt)$'\t'$(cut -f 5 temp_file.txt)$'\t'$(cut -f 6 temp_file.txt)$'\t'$(cut -f 2 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 1 temp_file.txt) >> orthogroups_annotations_losses_paraescarpia_owenia.csv
        done
    else
    echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_paraescarpia_owenia.csv
    fi
done < orthogroups_gene_IDs_losses_paraescarpia_owenia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == CapteT* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       cut -f 1,3,7,21,22 ../../Capitella_annotation_Feb2021_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 7 temp_file.txt)
        #gene_ID=$(cut -f 1 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 21 temp_file.txt)
        #GO_1=$(cut -f 22 temp_file.txt)
```

```
        #GO_1=$(cut -f 6 temp_file.txt) NONE
        echo $orthogroup_ID$'\t'"capitella"$'\t'$(cut -f 21 temp_file.txt)$'\t'$(cut -f 22 temp_file.txt)$'\t'""'\t'$(cut -f 1 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 7 temp_file.txt) >> orthogroups_annotations_losses_paraescarpia_capitella.csv
        done
    else
    echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_paraescarpia_capitella.csv
    fi
done < orthogroups_gene_IDs_losses_paraescarpia_capitella_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFRA* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../osedax_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"osedax"$'\t'$annotations >> orthogroups_annotations_losses_paraescarpia_osedax.csv
        done
    else
    echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'
```

```
'""$'\t'"" >> orthogroups_annotations_losses_paraescarpia_osedax.csv
143       fi
144 done < orthogroups_gene_IDs_losses_paraescarpia_osedax_OK.txt
145
146
147 echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_losses_paraescarpia_Ofra_Oalv_Rpac_Lluy_Ofus_Ctel.csv
148 cat orthogroups_annotations_losses_paraescarpia_oasisia.csv orthogroups_annotations_losses_paraescarpia_osedax.csv orthogroups_annotations_losses_paraescarpia_riftia.csv orthogroups_annotations_losses_paraescarpia_lamellibrachia.csv orthogroups_annotations_losses_paraescarpia_owenia.csv orthogroups_annotations_losses_paraescarpia_capitella.csv >> orthogroups_annotations_losses_paraescarpia_Ofra_Oalv_Rpac_Lluy_Ofus_Ctel.csv
149 sort orthogroups_annotations_losses_paraescarpia_Ofra_Oalv_Rpac_Lluy_Ofus_Ctel.csv > orthogroups_annotations_losses_paraescarpia_Ofra_Oalv_Rpac_Lluy_Ofus_Ctel_OK.csv
150
151 echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_losses_paraescarpia.csv
152 while read line; do
153    orthogroup_ID=$(cut -f 1 <<< "$line")
154    annotation=$(fgrep $orthogroup_ID orthogroups_annotations_losses_paraescarpia_Ofra_Oalv_Rpac_Lluy_Ofus_Ctel_OK.csv | cut -f 7 | sed '/^$/d' | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
155         echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_losses_paraescarpia.csv
156 done < gene_families_losses_paraescarpia.csv
```

orthogroups_annotations_losses_vestimentifera_cl2.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/losses/vestimentifera_cl2
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/losses/vestimentifera_cl2
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=140:00:0
#$ -l highmem

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Vestimentifera_cl2 | cut -f 1 > gene_families_losses_vestimentifera_cl2.txt #families losses in oasisia
fgrep -f gene_families_losses_vestimentifera_cl2.txt ../../Orthogroups.csv > gene_families_losses_vestimentifera_cl2.csv

cut -f 1,20 gene_families_losses_vestimentifera_cl2.csv > orthogroups_gene_IDs_losses_vestimentifera_cl2_osedax.txt
sed 's/Ofra|//g' orthogroups_gene_IDs_losses_vestimentifera_cl2_osedax.txt > orthogroups_gene_IDs_losses_vestimentifera_cl2_osedax_OK.txt
cut -f 1,21 gene_families_losses_vestimentifera_cl2.csv > orthogroups_gene_IDs_losses_vestimentifera_cl2_owenia.txt
sed 's/Ofus|//g' orthogroups_gene_IDs_losses_vestimentifera_cl2_owenia.txt > orthogroups_gene_IDs_losses_vestimentifera_cl2_owenia_OK.txt
cut -f 1,5 gene_families_losses_vestimentifera_cl2.csv > orthogroups_gene_IDs_losses_vestimentifera_cl2_capitella.txt
sed 's/Ctel|//g' orthogroups_gene_IDs_losses_vestimentifera_cl2_capitella.txt > orthogroups_gene_IDs_losses_vestimentifera_cl2_capitella_OK.txt
cut -f 1,14 gene_families_losses_vestimentifera_cl2.csv > orthog
```

```
roups_gene_IDs_losses_vestimentifera_cl2_lamellibrachia.txt
sed 's/Lluy|//g' orthogroups_gene_IDs_losses_vestimentifera_cl2_lamellibrachia.txt > orthogroups_gene_IDs_losses_vestimentifera_cl2_lamellibrachia_OK.txt
cut -f 1,23 gene_families_losses_vestimentifera_cl2.csv > orthogroups_gene_IDs_losses_vestimentifera_cl2_paraescarpia.txt
sed 's/Pech|//g' orthogroups_gene_IDs_losses_vestimentifera_cl2_paraescarpia.txt > orthogroups_gene_IDs_losses_vestimentifera_cl2_paraescarpia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == nbis* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../paraescarpia_annotation_Jun2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"paraescarpia"$'\t'$annotations >> orthogroups_annotations_losses_vestimentifera_cl2_paraescarpia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_vestimentifera_cl2_paraescarpia.csv
     fi
done < orthogroups_gene_IDs_losses_vestimentifera_cl2_paraescarp
```

```
ia_OK.txt
41
42 while read line; do
43    genes=$(cut -f 2 <<< "$line")
44    echo $genes
45    orthogroup_ID=$(cut -f 1 <<< "$line")
46    echo $orthogroup_ID
47      if [[ "$genes" == FUN* ]]
48      then
49       IFS=', '     # space is set as delimiter
50       read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
51       for gene in "${ADDR[@]}"; do
52        annotations=$(cut -f 13,14,15,18,20,24 ../../lamellibrach
ia_annotation_Feb2021_TrinoPantherKO_OK.xls | fgrep $gene)
53         echo $orthogroup_ID$'\t'"lamellibrachia"$'\t'$annotation
s >> orthogroups_annotations_losses_vestimentifera_cl2_lamellibr
achia.csv
54         done
55      else
56      echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_losses_vestimentifera_cl2_
lamellibrachia.csv
57      fi
58 done < orthogroups_gene_IDs_losses_vestimentifera_cl2_lamellibra
chia_OK.txt
59
60 while read line; do
61    genes=$(cut -f 2 <<< "$line")
62    echo $genes
63    orthogroup_ID=$(cut -f 1 <<< "$line")
64    echo $orthogroup_ID
```

```
65      if [[ "$genes" == OFUS* ]]
66      then
67       IFS=', '     # space is set as delimiter
68       read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
69       for gene in "${ADDR[@]}"; do
70         cut -f 1,2,3,11,12,13 ../../Owenia_annotation_v250920.1_
TrinoPantherKO.xls | fgrep $gene > temp_file.txt
71         #K0_number=$(cut -f 1 temp_file.txt)
72         #gene_ID=$(cut -f 2 temp_file.txt)
73         #Panther_annotation=$(cut -f 3 temp_file.txt)
74         #GO_1=$(cut -f 4 temp_file.txt)
75         #GO_1=$(cut -f 5 temp_file.txt)
76         #GO_1=$(cut -f 6 temp_file.txt)
77         echo $orthogroup_ID$'\t'"owenia"$'\t'$(cut -f 4 temp_fil
e.txt)$'\t'$(cut -f 5 temp_file.txt)$'\t'$(cut -f 6 temp_file.tx
t)$'\t'$(cut -f 2 temp_file.txt)$'\t'$(cut -f 3 temp_file.tx
t)$'\t'$(cut -f 1 temp_file.txt) >> orthogroups_annotations_loss
es_vestimentifera_cl2_owenia.csv
78         done
79      else
80      echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_losses_vestimentifera_cl2_
owenia.csv
81      fi
82  done < orthogroups_gene_IDs_losses_vestimentifera_cl2_owenia_OK.
txt
83
84  while read line; do
85     genes=$(cut -f 2 <<< "$line")
86     echo $genes
87     orthogroup_ID=$(cut -f 1 <<< "$line")
```

```
   echo $orthogroup_ID
     if [[ "$genes" == CapteT* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       cut -f 1,3,7,21,22 ../../Capitella_annotation_Feb2021_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 7 temp_file.txt)
        #gene_ID=$(cut -f 1 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 21 temp_file.txt)
        #GO_1=$(cut -f 22 temp_file.txt)
        #GO_1=$(cut -f 6 temp_file.txt) NONE
        echo $orthogroup_ID$'\t'"capitella"$'\t'$(cut -f 21 temp_file.txt)$'\t'$(cut -f 22 temp_file.txt)$'\t'""'\t'$(cut -f 1 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 7 temp_file.txt) >> orthogroups_annotations_losses_vestimentifera_cl2_capitella.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_vestimentifera_cl2_capitella.csv
     fi
done < orthogroups_gene_IDs_losses_vestimentifera_cl2_capitella_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
```

```
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFRA* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../osedax_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"osedax"$'\t'$annotations >> orthogroups_annotations_losses_vestimentifera_cl2_osedax.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_vestimentifera_cl2_osedax.csv
     fi
done < orthogroups_gene_IDs_losses_vestimentifera_cl2_osedax_OK.txt


echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_losses_vestimentifera_cl2_Ofra_Pech_Lluy_Ofus_Ctel.csv
cat orthogroups_annotations_losses_vestimentifera_cl2_osedax.csv orthogroups_annotations_losses_vestimentifera_cl2_paraescarpia.csv orthogroups_annotations_losses_vestimentifera_cl2_lamellibrachia.csv orthogroups_annotations_losses_vestimentifera_cl2_owenia.csv orthogroups_annotations_losses_vestimentifera_cl2_capitella.csv >> orthogroups_annotations_losses_vestimentifera_cl2_Ofra_Pech_Lluy_Ofus_Ctel.csv
```

```
sort orthogroups_annotations_losses_vestimentifera_cl2_Ofra_Pech_Lluy_Ofus_Ctel.csv > orthogroups_annotations_losses_vestimentifera_cl2_Ofra_Pech_Lluy_Ofus_Ctel_OK.csv

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbundantAnnotation_losses_vestimentifera_cl2.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_losses_vestimentifera_cl2_Ofra_Pech_Lluy_Ofus_Ctel_OK.csv | cut -f 7 | sed '/^$/d' | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostAbundantAnnotation_losses_vestimentifera_cl2.csv
done < gene_families_losses_vestimentifera_cl2.csv
```

orthogroups_annotations_losses_vestimentifera_cl1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/losses/vestimentifera_cl1
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/losses/vestimentifera_cl1
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=140:00:0
#$ -l highmem

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Vestimentifera_cl1 | cut -f 1 > gene_families_losses_vestimentifera_cl1.txt #families losses in oasisia
fgrep -f gene_families_losses_vestimentifera_cl1.txt ../../Orthogroups.csv > gene_families_losses_vestimentifera_cl1.csv
```

```
12
13 cut -f 1,20 gene_families_losses_vestimentifera_cl1.csv > orthog
roups_gene_IDs_losses_vestimentifera_cl1_osedax.txt
14 sed 's/Ofra|//g' orthogroups_gene_IDs_losses_vestimentifera_cl1_
osedax.txt > orthogroups_gene_IDs_losses_vestimentifera_cl1_osed
ax_OK.txt
15 cut -f 1,21 gene_families_losses_vestimentifera_cl1.csv > orthog
roups_gene_IDs_losses_vestimentifera_cl1_owenia.txt
16 sed 's/Ofus|//g' orthogroups_gene_IDs_losses_vestimentifera_cl1_
owenia.txt > orthogroups_gene_IDs_losses_vestimentifera_cl1_owen
ia_OK.txt
17 cut -f 1,5 gene_families_losses_vestimentifera_cl1.csv > orthogr
oups_gene_IDs_losses_vestimentifera_cl1_capitella.txt
18 sed 's/Ctel|//g' orthogroups_gene_IDs_losses_vestimentifera_cl1_
capitella.txt > orthogroups_gene_IDs_losses_vestimentifera_cl1_c
apitella_OK.txt
19 cut -f 1,14 gene_families_losses_vestimentifera_cl1.csv > orthog
roups_gene_IDs_losses_vestimentifera_cl1_lamellibrachia.txt
20 sed 's/Lluy|//g' orthogroups_gene_IDs_losses_vestimentifera_cl1_
lamellibrachia.txt > orthogroups_gene_IDs_losses_vestimentifera_
cl1_lamellibrachia_OK.txt
21
22 while read line; do
23    genes=$(cut -f 2 <<< "$line")
24    echo $genes
25    orthogroup_ID=$(cut -f 1 <<< "$line")
26    echo $orthogroup_ID
27      if [[ "$genes" == FUN* ]]
28      then
29       IFS=', '     # space is set as delimiter
30       read -ra ADDR <<< "$genes"   # str is read into an array a
s tokens separated by IFS
31       for gene in "${ADDR[@]}"; do
```

```
32        annotations=$(cut -f 13,14,15,18,20,24 ../../lamellibrachia_annotation_Feb2021_TrinoPantherKO_OK.xls | fgrep $gene)
33        echo $orthogroup_ID$'\t'"lamellibrachia"$'\t'$annotations >> orthogroups_annotations_losses_vestimentifera_cl1_lamellibrachia.csv
34        done
35     else
36     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_vestimentifera_cl1_lamellibrachia.csv
37     fi
38 done < orthogroups_gene_IDs_losses_vestimentifera_cl1_lamellibrachia_OK.txt
39
40 while read line; do
41    genes=$(cut -f 2 <<< "$line")
42    echo $genes
43    orthogroup_ID=$(cut -f 1 <<< "$line")
44    echo $orthogroup_ID
45      if [[ "$genes" == OFUS* ]]
46      then
47       IFS=', '     # space is set as delimiter
48       read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
49       for gene in "${ADDR[@]}"; do
50         cut -f 1,2,3,11,12,13 ../../Owenia_annotation_v250920.1_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
51         #K0_number=$(cut -f 1 temp_file.txt)
52         #gene_ID=$(cut -f 2 temp_file.txt)
53         #Panther_annotation=$(cut -f 3 temp_file.txt)
54         #GO_1=$(cut -f 4 temp_file.txt)
55         #GO_1=$(cut -f 5 temp_file.txt)
```

```
        #GO_1=$(cut -f 6 temp_file.txt)
        echo $orthogroup_ID$'\t'"owenia"$'\t'$(cut -f 4 temp_file.txt)$'\t'$(cut -f 5 temp_file.txt)$'\t'$(cut -f 6 temp_file.txt)$'\t'$(cut -f 2 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 1 temp_file.txt) >> orthogroups_annotations_losses_vestimentifera_cl1_owenia.csv
        done
    else
    echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_vestimentifera_cl1_owenia.csv
    fi
done < orthogroups_gene_IDs_losses_vestimentifera_cl1_owenia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == CapteT* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       cut -f 1,3,7,21,22 ../../Capitella_annotation_Feb2021_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 7 temp_file.txt)
        #gene_ID=$(cut -f 1 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 21 temp_file.txt)
```

```
        #GO_1=$(cut -f 22 temp_file.txt)
        #GO_1=$(cut -f 6 temp_file.txt) NONE
        echo $orthogroup_ID$'\t'"capitella"$'\t'$(cut -f 21 temp_file.txt)$'\t'$(cut -f 22 temp_file.txt)$'\t'""'\t'$(cut -f 1 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 7 temp_file.txt) >> orthogroups_annotations_losses_vestimentifera_cl1_capitella.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_vestimentifera_cl1_capitella.csv
     fi
done < orthogroups_gene_IDs_losses_vestimentifera_cl1_capitella_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFRA* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../osedax_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"osedax"$'\t'$annotations >> orthogroups_annotations_losses_vestimentifera_cl1_osedax.csv
        done
```

```
101      else
102      echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_losses_vestimentifera_cl1_
osedax.csv
103      fi
104 done < orthogroups_gene_IDs_losses_vestimentifera_cl1_osedax_OK.
txt
105
106
107 echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\
t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_nu
mber" > orthogroups_annotations_losses_vestimentifera_cl1_Ofra_L
luy_Ofus_Ctel.csv
108 cat orthogroups_annotations_losses_vestimentifera_cl1_osedax.csv
orthogroups_annotations_losses_vestimentifera_cl1_lamellibrachi
a.csv orthogroups_annotations_losses_vestimentifera_cl1_owenia.c
sv orthogroups_annotations_losses_vestimentifera_cl1_capitella.c
sv >> orthogroups_annotations_losses_vestimentifera_cl1_Ofra_Llu
y_Ofus_Ctel.csv
109 sort orthogroups_annotations_losses_vestimentifera_cl1_Ofra_Lluy
_Ofus_Ctel.csv > orthogroups_annotations_losses_vestimentifera_c
l1_Ofra_Lluy_Ofus_Ctel_OK.csv
110
111 echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbu
ndantAnnotation_losses_vestimentifera_cl1.csv
112 while read line; do
113    orthogroup_ID=$(cut -f 1 <<< "$line")
114    annotation=$(fgrep $orthogroup_ID orthogroups_annotations_los
ses_vestimentifera_cl1_Ofra_Lluy_Ofus_Ctel_OK.csv | cut -f 7 | s
ed '/^$/d' | sort |  uniq -c | sort -r | awk '{$1=""; print $0}'
| head -1)
115         echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostA
bundantAnnotation_losses_vestimentifera_cl1.csv
116 done < gene_families_losses_vestimentifera_cl1.csv
```

orthogroups_annotations_losses_vestimentifera.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/losses/vestimentifera
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/losses/vestimentifera
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=140:00:0
#$ -l highmem

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Vestimentifera | cut -f 1 > gene_families_losses_vestimentifera.txt #families losses in oasisia
fgrep -f gene_families_losses_vestimentifera.txt ../../Orthogroups.csv > gene_families_losses_vestimentifera.csv

cut -f 1,20 gene_families_losses_vestimentifera.csv > orthogroups_gene_IDs_losses_vestimentifera_osedax.txt
sed 's/Ofra|//g' orthogroups_gene_IDs_losses_vestimentifera_osedax.txt > orthogroups_gene_IDs_losses_vestimentifera_osedax_OK.txt
cut -f 1,21 gene_families_losses_vestimentifera.csv > orthogroups_gene_IDs_losses_vestimentifera_owenia.txt
sed 's/Ofus|//g' orthogroups_gene_IDs_losses_vestimentifera_owenia.txt > orthogroups_gene_IDs_losses_vestimentifera_owenia_OK.txt
cut -f 1,5 gene_families_losses_vestimentifera.csv > orthogroups_gene_IDs_losses_vestimentifera_capitella.txt
sed 's/Ctel|//g' orthogroups_gene_IDs_losses_vestimentifera_capitella.txt > orthogroups_gene_IDs_losses_vestimentifera_capitella
```

```
_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFUS* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
        cut -f 1,2,3,11,12,13 ../../Owenia_annotation_v250920.1_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 1 temp_file.txt)
        #gene_ID=$(cut -f 2 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 4 temp_file.txt)
        #GO_1=$(cut -f 5 temp_file.txt)
        #GO_1=$(cut -f 6 temp_file.txt)
        echo $orthogroup_ID$'\t'"owenia"$'\t'$(cut -f 4 temp_file.txt)$'\t'$(cut -f 5 temp_file.txt)$'\t'$(cut -f 6 temp_file.txt)$'\t'$(cut -f 2 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 1 temp_file.txt) >> orthogroups_annotations_losses_vestimentifera_owenia.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_vestimentifera_owenia.csv
     fi
```

```
done < orthogroups_gene_IDs_losses_vestimentifera_owenia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == CapteT* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       cut -f 1,3,7,21,22 ../../Capitella_annotation_Feb2021_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 7 temp_file.txt)
        #gene_ID=$(cut -f 1 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 21 temp_file.txt)
        #GO_1=$(cut -f 22 temp_file.txt)
        #GO_1=$(cut -f 6 temp_file.txt) NONE
        echo $orthogroup_ID$'\t'"capitella"$'\t'$(cut -f 21 temp_file.txt)$'\t'$(cut -f 22 temp_file.txt)$'\t'""'\t'$(cut -f 1 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 7 temp_file.txt) >> orthogroups_annotations_losses_vestimentifera_capitella.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_vestimentifera_capitella.csv
     fi
```

```
done < orthogroups_gene_IDs_losses_vestimentifera_capitella_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFRA* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       annotations=$(cut -f 13,14,15,18,20,24 ../../osedax_annotation_Jan2021_TrinoPantherKO.xls | fgrep $gene)
        echo $orthogroup_ID$'\t'"osedax"$'\t'$annotations >> orthogroups_annotations_losses_vestimentifera_osedax.csv
        done
     else
     echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_vestimentifera_osedax.csv
     fi
done < orthogroups_gene_IDs_losses_vestimentifera_osedax_OK.txt


echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_number" > orthogroups_annotations_losses_vestimentifera_Ofra_Ofus_Ctel.csv
cat orthogroups_annotations_losses_vestimentifera_osedax.csv ort
```

```
hogroups_annotations_losses_vestimentifera_owenia.csv orthogroup
s_annotations_losses_vestimentifera_capitella.csv >> orthogroups
_annotations_losses_vestimentifera_Ofra_Ofus_Ctel.csv
sort orthogroups_annotations_losses_vestimentifera_Ofra_Ofus_Cte
l.csv > orthogroups_annotations_losses_vestimentifera_Ofra_Ofus_
Ctel_OK.csv

echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbu
ndantAnnotation_losses_vestimentifera.csv
while read line; do
   orthogroup_ID=$(cut -f 1 <<< "$line")
   annotation=$(fgrep $orthogroup_ID orthogroups_annotations_los
ses_vestimentifera_Ofra_Ofus_Ctel_OK.csv | cut -f 7 | sed '/^$/d
' | sort |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -
1)
        echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostA
bundantAnnotation_losses_vestimentifera.csv
done < gene_families_losses_vestimentifera.csv
```

orthogroups_annotations_losses_siboglinidae.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/gene_family_evolution/ferdi_script/J
ul2021/losses/siboglinidae
#$ -o /data/scratch/btx654/gene_family_evolution/ferdi_script/Ju
l2021/losses/siboglinidae
#$ -j y
#$ -pe smp 1
#$ -l h_vmem=100G
#$ -l h_rt=140:00:0
#$ -l highmem

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv |
grep -w Siboglinidae | cut -f 1 > gene_families_losses_siboglini
```

```
dae.txt #families losses in oasisia
fgrep -f gene_families_losses_siboglinidae.txt ../../Orthogroups.csv > gene_families_losses_siboglinidae.csv

cut -f 1,21 gene_families_losses_siboglinidae.csv > orthogroups_gene_IDs_losses_siboglinidae_owenia.txt
sed 's/Ofus|//g' orthogroups_gene_IDs_losses_siboglinidae_owenia.txt > orthogroups_gene_IDs_losses_siboglinidae_owenia_OK.txt
cut -f 1,5 gene_families_losses_siboglinidae.csv > orthogroups_gene_IDs_losses_siboglinidae_capitella.txt
sed 's/Ctel|//g' orthogroups_gene_IDs_losses_siboglinidae_capitella.txt > orthogroups_gene_IDs_losses_siboglinidae_capitella_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == OFUS* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
        cut -f 1,2,3,11,12,13 ../../Owenia_annotation_v250920.1_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 1 temp_file.txt)
        #gene_ID=$(cut -f 2 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 4 temp_file.txt)
        #GO_1=$(cut -f 5 temp_file.txt)
```

```
        #GO_1=$(cut -f 6 temp_file.txt)
        echo $orthogroup_ID$'\t'"owenia"$'\t'$(cut -f 4 temp_file.txt)$'\t'$(cut -f 5 temp_file.txt)$'\t'$(cut -f 6 temp_file.txt)$'\t'$(cut -f 2 temp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 1 temp_file.txt) >> orthogroups_annotations_losses_siboglinidae_owenia.csv
        done
    else
    echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t'"" >> orthogroups_annotations_losses_siboglinidae_owenia.csv
    fi
done < orthogroups_gene_IDs_losses_siboglinidae_owenia_OK.txt

while read line; do
   genes=$(cut -f 2 <<< "$line")
   echo $genes
   orthogroup_ID=$(cut -f 1 <<< "$line")
   echo $orthogroup_ID
     if [[ "$genes" == CapteT* ]]
     then
      IFS=', '     # space is set as delimiter
      read -ra ADDR <<< "$genes"   # str is read into an array as tokens separated by IFS
      for gene in "${ADDR[@]}"; do
       cut -f 1,3,7,21,22 ../../Capitella_annotation_Feb2021_TrinoPantherKO.xls | fgrep $gene > temp_file.txt
        #K0_number=$(cut -f 7 temp_file.txt)
        #gene_ID=$(cut -f 1 temp_file.txt)
        #Panther_annotation=$(cut -f 3 temp_file.txt)
        #GO_1=$(cut -f 21 temp_file.txt)
        #GO_1=$(cut -f 22 temp_file.txt)
```

```
58          #GO_1=$(cut -f 6 temp_file.txt) NONE
59          echo $orthogroup_ID$'\t'"capitella"$'\t'$(cut -f 21 temp
_file.txt)$'\t'$(cut -f 22 temp_file.txt)$'\t'""'\t'$(cut -f 1 t
emp_file.txt)$'\t'$(cut -f 3 temp_file.txt)$'\t'$(cut -f 7 temp_
file.txt) >> orthogroups_annotations_losses_siboglinidae_capitel
la.csv
60          done
61      else
62      echo $orthogroup_ID$'\t'""$'\t'""$'\t'""$'\t'""$'\t'""$'\t
'""$'\t'"" >> orthogroups_annotations_losses_siboglinidae_capite
lla.csv
63      fi
64  done < orthogroups_gene_IDs_losses_siboglinidae_capitella_OK.txt
65
66
67  echo "Orthogroup"$'\t'"Species"$'\t'"GO_term1"$'\t'"GO_term2"$'\
t'"GO_term3"$'\t'"gene_ID"$'\t'"Panther_annotation"$'\t'"KEGG_nu
mber" > orthogroups_annotations_losses_siboglinidae_Ofus_Ctel.cs
v
68  cat orthogroups_annotations_losses_siboglinidae_owenia.csv ortho
groups_annotations_losses_siboglinidae_capitella.csv >> orthogro
ups_annotations_losses_siboglinidae_Ofus_Ctel.csv
69  sort orthogroups_annotations_losses_siboglinidae_Ofus_Ctel.csv >
orthogroups_annotations_losses_siboglinidae_Ofus_Ctel_OK.csv
70
71  echo "Orthogroup"$'\t'"Panther_annotation" > orthogroups_mostAbu
ndantAnnotation_losses_siboglinidae.csv
72  while read line; do
73     orthogroup_ID=$(cut -f 1 <<< "$line")
74     annotation=$(fgrep $orthogroup_ID orthogroups_annotations_los
ses_siboglinidae_Ofus_Ctel_OK.csv | cut -f 7 | sed '/^$/d' | sor
t |  uniq -c | sort -r | awk '{$1=""; print $0}' | head -1)
75          echo $orthogroup_ID$'\t'$annotation >> orthogroups_mostA
```

```
bundantAnnotation_losses_siboglinidae.csv
done < gene_families_losses_siboglinidae.csv
```

GO_terms

Base code:

```
install.packages("BiocManager")
BiocManager::install("topGO")
install.packages("ggpubr")

library(topGO)
library(ggplot2)
library(ggpubr)
library(cowplot)

# Import gene universe: whole (GO-annotated) genome
geneID2GO <- readMappings(file = "/Users/giacomo/Desktop/R/GO_enrichment/osedax/osedax_GO_universe.txt") ### 21108 transcripts have GO annotation
geneUniverse <- names(geneID2GO)

# Import and transform genes of interest: 8 clusters from step2a
cluster1 <- read.table("/Users/giacomo/Desktop/R/GO_enrichment/osedax/gene_IDs_originated_osedax_Siboglinidae.txt",header=FALSE)
cluster1 <- as.character(cluster1$V1)
cluster1genelist <- factor(as.integer(geneUniverse %in% cluster1))
names(cluster1genelist) <- geneUniverse

# fisher testing of GO term enrichment for Molecular Function (MF)
```

```
#cluster 1 - red genes
cluster1_GOdata_MF <- new("topGOdata", description="Cluster1",
                          ontology="MF", allGenes=cluster1geneli
st,
                          annot = annFUN.gene2GO, gene2GO = gene
ID2GO)
cluster1_resultFisher_MF <- runTest(cluster1_GOdata_MF,
                                    algorithm="classic", statist
ic="fisher")
cluster1_MF <- GenTable(cluster1_GOdata_MF, classicFisher = clus
ter1_resultFisher_MF,
                        orderBy = "resultFisher", ranksOf = "cla
ssicFisher", topNodes = 15)

cluster1_MF[cluster1_MF == "< 1e-30"] <- "1e-30"
cluster1_MF[cluster1_MF == "<1e-30"] <- "1e-30"

goEnrichment <- cluster1_MF
goEnrichment$classicFisher <- as.numeric(goEnrichment$classicFis
her)
goEnrichment <- goEnrichment[,c("GO.ID","Term","classicFisher")]
goEnrichment$Term <- gsub(" [a-z]*\\.\\.\\.$", "", goEnrichmen
t$Term)
goEnrichment$Term <- gsub("\\.\\.\\.$", "", goEnrichment$Term)
goEnrichment$Term <- paste(goEnrichment$GO.ID, goEnrichment$Ter
m, sep=", ")
goEnrichment$Term <- factor(goEnrichment$Term, levels=rev(goEnri
chment$Term))

#it could happen that the second line of the previous block will
give the error " Warning message:NAs introduced by coercion "
# a fix for that is " goEnrichment$classicFisher <- c(30, 30, 3
```

```
0, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30) "

cluster_1_plot <- ggplot(goEnrichment, aes(x=Term, y=-log10(clas
sicFisher))) +
  stat_summary(geom = "bar", fun = mean, position = "dodge") +
  xlab("Molecular Function") +
  ylab("-log10(p-value)") +
  ggtitle("GF gains in Siboglinidae (Osedax)") +
  scale_y_continuous(limits=c(0,30),breaks=round(seq(0,30, by =
2), 1)) +
  theme_classic() +
  theme(
    legend.position='none',
    legend.background=element_rect(),
    plot.title=element_text(angle=0, size=12, face="bold", vjust
=1),
    axis.text.x=element_text(angle=0, size=10, hjust=1.10),
    axis.text.y=element_text(angle=0, size=10, vjust=0.5),
    axis.title=element_text(size=12),
    legend.key=element_blank(),     #removes the border
    legend.key.size=unit(1, "cm"),      #Sets overall area/size
of the legend
    legend.text=element_text(size=12),  #Text size
    title=element_text(size=12)) +
  guides(colour=guide_legend(override.aes=list(size=2.5))) +
  coord_flip()

cluster_1_plot

cluster1_GOdata_BP <- new("topGOdata", description="Cluster1",
                          ontology="BP", allGenes=cluster1geneli
```

```
st,
                          annot = annFUN.gene2GO, gene2GO = gene
ID2GO)
cluster1_resultFisher_BP <- runTest(cluster1_GOdata_BP,
                                    algorithm="classic", statist
ic="fisher")
cluster1_BP <- GenTable(cluster1_GOdata_BP, classicFisher = clus
ter1_resultFisher_BP,
                        orderBy = "resultFisher", ranksOf = "cla
ssicFisher", topNodes = 15)

cluster1_BP[cluster1_BP == "< 1e-30"] <- "1e-30"
cluster1_BP[cluster1_BP == "<1e-30"] <- "1e-30"

goEnrichment <- cluster1_BP
goEnrichment$classicFisher <- as.numeric(goEnrichment$classicFis
her)
goEnrichment <- goEnrichment[,c("GO.ID","Term","classicFisher")]
goEnrichment$Term <- gsub(" [a-z]*\\.\\.\\.$", "", goEnrichmen
t$Term)
goEnrichment$Term <- gsub("\\.\\.\\.$", "", goEnrichment$Term)
goEnrichment$Term <- paste(goEnrichment$GO.ID, goEnrichment$Ter
m, sep=", ")
goEnrichment$Term <- factor(goEnrichment$Term, levels=rev(goEnri
chment$Term))

#it could happen that the second line of the previous block will
give the error " Warning message:NAs introduced by coercion "
# a fix for that is " goEnrichment$classicFisher <- c(30, 30, 3
0, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30) "

```

```
cluster_1BP_plot <- ggplot(goEnrichment, aes(x=Term, y=-log10(classicFisher))) +
  stat_summary(geom = "bar", fun = mean, position = "dodge") +
  xlab("Biological Process") +
  ylab("-log10(p-value)") +
  scale_y_continuous(limits=c(0,30),breaks=round(seq(0,30, by = 2), 1)) +
  theme_classic() +
  theme(
    legend.position='none',
    legend.background=element_rect(),
    plot.title=element_text(angle=0, size=12, face="bold", vjust=1),
    axis.text.x=element_text(angle=0, size=10, hjust=1.10),
    axis.text.y=element_text(angle=0, size=10, vjust=0.5),
    axis.title=element_text(size=12),
    legend.key=element_blank(),     #removes the border
    legend.key.size=unit(1, "cm"),      #Sets overall area/size of the legend
    legend.text=element_text(size=12),  #Text size
    title=element_text(size=12)) +
  guides(colour=guide_legend(override.aes=list(size=2.5))) +
  coord_flip()

cluster_1BP_plot

plot_grid(cluster_1_plot + rremove("x.title"),
          cluster_1BP_plot,
          ncol = 1, align="v")
```

- exported in pdf 7x8 inches

Riftia example

```
cd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/GO_terms/riftia
# I will use BlastX GO terms
cut -f2,13 ../../riftia_annotation_Jan2021_TrinoPantherKO.xls | tail -n +2 > riftia_GO_raw.txt # 38179 genes in riftia_GO_raw.txt
grep 'GO' riftia_GO_raw.txt > riftia_GO_only_raw.txt # 20737 riftia_GO_only_raw.txt

```

python.py

```
if __name__ == "__main__":

    import re

    i = open("riftia_GO_only_raw.txt", "r")
    o = open("riftia_GO_universe.txt", "w")

    regex = re.compile(r'GO:\d+')

    for line in i:
        GOmatches = regex.findall(line)
        Gene_ID = line.split("\t",1)[0]
        if not GOmatches == []:
            o.write(Gene_ID+"\t")
            for i, match in enumerate(GOmatches):
                if i+1 == len(GOmatches):
                    o.write(match.strip("'")+"\n")
                else:
                    o.write(match.strip("'")+", ")
```

```
module load python
python python.py
```

Now we have the file riftia_GO_universe.txt and we nned to select a subgroup of genes:

Expansions

```
cut -f 1,7 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Rpac | cut -f 1 > gene_families_expanded_riftia.txt #families expanded in riftia
fgrep -f gene_families_expanded_riftia.txt ../../Orthogroups.csv > gene_families_expanded_riftia.csv
cut -f 1,24 gene_families_expanded_riftia.csv | sed 's/Rpac|//g' | cut -f 2 | sed 's/, /\n/g' > gene_IDs_expanded_riftia.txt
```

R script:

```
install.packages("BiocManager")
BiocManager::install("topGO")
install.packages("ggpubr")

library(topGO)
library(ggplot2)
library(ggpubr)
library(cowplot)

# Import gene universe: whole (GO-annotated) genome
geneID2GO <- readMappings(file = "/Users/giacomo/Desktop/R/GO_enrichment/riftia_GO_universe.txt") ### 21108 transcripts have GO annotation
geneUniverse <- names(geneID2GO)

# Import and transform genes of interest: 8 clusters from step2a
cluster1 <- read.table("/Users/giacomo/Desktop/R/GO_enrichment/g
```

```
ene_IDs_expanded_riftia.txt",header=FALSE)
cluster1 <- as.character(cluster1$V1)
cluster1genelist <- factor(as.integer(geneUniverse %in% cluster
1))
names(cluster1genelist) <- geneUniverse

# fisher testing of GO term enrichment for Molecular Function (M
F)
#cluster 1 - red genes
cluster1_GOdata_MF <- new("topGOdata", description="Cluster1",
                          ontology="MF", allGenes=cluster1geneli
st,
                          annot = annFUN.gene2GO, gene2GO = gene
ID2GO)
cluster1_resultFisher_MF <- runTest(cluster1_GOdata_MF,
                                    algorithm="classic", statist
ic="fisher")
cluster1_MF <- GenTable(cluster1_GOdata_MF, classicFisher = clus
ter1_resultFisher_MF,
                        orderBy = "resultFisher", ranksOf = "cla
ssicFisher", topNodes = 15)

cluster1_MF[cluster1_MF == "< 1e-30"] <- "1e-30"

goEnrichment <- cluster1_MF
goEnrichment$classicFisher <- as.numeric(goEnrichment$classicFis
her)
goEnrichment <- goEnrichment[,c("GO.ID","Term","classicFisher")]
goEnrichment$Term <- gsub(" [a-z]*\\.\\.\\.$", "", goEnrichmen
t$Term)
goEnrichment$Term <- gsub("\\.\\.\\.$", "", goEnrichment$Term)
goEnrichment$Term <- paste(goEnrichment$GO.ID, goEnrichment$Ter
```

```
m, sep=", ")
goEnrichment$Term <- factor(goEnrichment$Term, levels=rev(goEnrichment$Term))


cluster_1_plot <- ggplot(goEnrichment, aes(x=Term, y=-log10(classicFisher))) +
  stat_summary(geom = "bar", fun = mean, position = "dodge") +
  xlab("Biological process") +
  ylab("-log10(p-value)") +
  ggtitle("GF expansions in Riftia") +
  scale_y_continuous(limits=c(0,30),breaks=round(seq(0,30, by = 2), 1)) +
  theme_classic() +
  theme(
    legend.position='none',
    legend.background=element_rect(),
    plot.title=element_text(angle=0, size=12, face="bold", vjust=1),
    axis.text.x=element_text(angle=0, size=10, hjust=1.10),
    axis.text.y=element_text(angle=0, size=10, vjust=0.5),
    axis.title=element_text(size=12),
    legend.key=element_blank(),     #removes the border
    legend.key.size=unit(1, "cm"),      #Sets overall area/size of the legend
    legend.text=element_text(size=12),  #Text size
    title=element_text(size=12)) +
  guides(colour=guide_legend(override.aes=list(size=2.5))) +
  coord_flip()

cluster_1_plot
```

- exported in pdf 4x8 inches

Gains

```
cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Rpac | cut -f 1 > gene_families_originated_riftia.txt #families expanded in riftia
fgrep -f gene_families_originated_riftia.txt ../../Orthogroups.csv > gene_families_originated_riftia.csv
cut -f 1,24 gene_families_originated_riftia.csv | sed 's/Rpac|//g' | cut -f 2 | sed 's/, /\n/g' > gene_IDs_originated_riftia.txt

cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Siboglinidae | cut -f 1 > gene_families_originated_riftia_siboglinidae.txt #families expanded in riftia
fgrep -f gene_families_originated_riftia_siboglinidae.txt ../../Orthogroups.csv > gene_families_originated_riftia_siboglinidae.csv
cut -f 1,24 gene_families_originated_riftia_siboglinidae.csv | sed 's/Rpac|//g' | cut -f 2 | sed 's/, /\n/g' | sed '/^$/d' > gene_IDs_originated_riftia_siboglinidae.txt

cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Vestimentifera | cut -f 1 > gene_families_originated_riftia_Vestimentifera.txt #families expanded in riftia
fgrep -f gene_families_originated_riftia_Vestimentifera.txt ../../Orthogroups.csv > gene_families_originated_riftia_Vestimentifera.csv
cut -f 1,24 gene_families_originated_riftia_Vestimentifera.csv | sed 's/Rpac|//g' | cut -f 2 | sed 's/, /\n/g' | sed '/^$/d' > gene_IDs_originated_riftia_Vestimentifera.txt

cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Vestimentifera_cl1 | cut -f 1 > gene_families_originated_riftia_Vestimentifera_cl1.txt #families expanded in riftia
```

```
fgrep -f gene_families_originated_riftia_Vestimentifera_cl1.txt ../../Orthogroups.csv > gene_families_originated_riftia_Vestimentifera_cl1.csv
cut -f 1,24 gene_families_originated_riftia_Vestimentifera_cl1.csv | sed 's/Rpac|//g' | cut -f 2 | sed 's/, /\n/g' | sed '/^$/d' > gene_IDs_originated_riftia_Vestimentifera_cl1.txt

cut -f 1,4 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Vestimentifera_cl2 | cut -f 1 > gene_families_originated_riftia_Vestimentifera_cl2.txt #families expanded in riftia
fgrep -f gene_families_originated_riftia_Vestimentifera_cl2.txt ../../Orthogroups.csv > gene_families_originated_riftia_Vestimentifera_cl2.csv
cut -f 1,24 gene_families_originated_riftia_Vestimentifera_cl2.csv | sed 's/Rpac|//g' | cut -f 2 | sed 's/, /\n/g' | sed '/^$/d' > gene_IDs_originated_riftia_Vestimentifera_cl2.txt
```

Losses

using Owenia to annotate the losses

```
cd /data/scratch/btx654/gene_family_evolution/ferdi_script/Jul2021/GO_terms/losses
cut -f2,11 ../../Owenia_annotation_v250920.1_TrinoPantherKO.xls | tail -n +2 | sed 's/"//g' > owenia_GO_raw.txt #
grep 'GO' owenia_GO_raw.txt > owenia_GO_only_raw.txt # 21108 osedax_GO_only_raw.txt
```

python.py

```
if __name__ == "__main__":

    import re

    i = open("owenia_GO_only_raw.txt", "r")
    o = open("owenia_GO_universe.txt", "w")
```

```
    regex = re.compile(r'GO:\d+')

    for line in i:
        GOmatches = regex.findall(line)
        Gene_ID = line.split("\t",1)[0]
        if not GOmatches == []:
            o.write(Gene_ID+"\t")
            for i, match in enumerate(GOmatches):
                if i+1 == len(GOmatches):
                    o.write(match.strip("'")+"\n")
                else:
                    o.write(match.strip("'")+", ")
```

```
module load python
python python.py
```

losses.sh

```
cut -f 1,21 ../../Orthogroups.csv | tail -n +2 | grep -w Ofus | cut -f 1 > owenia_all_GF

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Ofra | cut -f 1 > gene_families_lost_osedax.txt
grep -f gene_families_lost_osedax.txt owenia_all_GF > losses_osedax_owenia
fgrep -f losses_osedax_owenia ../../Orthogroups.csv > gene_families_lost_osedax_owenia.csv
cut -f 1,21 gene_families_lost_osedax_owenia.csv | sed 's/Ofus//g' | cut -f 2 | sed 's/, /\n/g' > gene_IDs_lost_osedax_owenia.txt

```

```
cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Oalv | cut -f 1 > gene_families_lost_oasisia.txt
grep -f gene_families_lost_oasisia.txt owenia_all_GF > losses_oasisia_owenia
fgrep -f losses_oasisia_owenia ../../Orthogroups.csv > gene_families_lost_oasisia_owenia.csv
cut -f 1,21 gene_families_lost_oasisia_owenia.csv | sed 's/Ofus|//g' | cut -f 2 | sed 's/, /\n/g' > gene_IDs_lost_oasisia_owenia.txt

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Rpac | cut -f 1 > gene_families_lost_riftia.txt
grep -f gene_families_lost_riftia.txt owenia_all_GF > losses_riftia_owenia
fgrep -f losses_riftia_owenia ../../Orthogroups.csv > gene_families_lost_riftia_owenia.csv
cut -f 1,21 gene_families_lost_riftia_owenia.csv | sed 's/Ofus|//g' | cut -f 2 | sed 's/, /\n/g' > gene_IDs_lost_riftia_owenia.txt

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Pech | cut -f 1 > gene_families_lost_paraescarpia.txt
grep -f gene_families_lost_paraescarpia.txt owenia_all_GF > losses_paraescarpia_owenia
fgrep -f losses_paraescarpia_owenia ../../Orthogroups.csv > gene_families_lost_paraescarpia_owenia.csv
cut -f 1,21 gene_families_lost_paraescarpia_owenia.csv | sed 's/Ofus|//g' | cut -f 2 | sed 's/, /\n/g' > gene_IDs_lost_paraescarpia_owenia.txt

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Lluy | cut -f 1 > gene_families_lost_lamellibrachia.txt
```

```
grep -f gene_families_lost_lamellibrachia.txt owenia_all_GF > losses_lamellibrachia_owenia
fgrep -f losses_lamellibrachia_owenia ../../Orthogroups.csv > gene_families_lost_lamellibrachia_owenia.csv
cut -f 1,21 gene_families_lost_lamellibrachia_owenia.csv | sed 's/Ofus|//g' | cut -f 2 | sed 's/, /\n/g' > gene_IDs_lost_lamellibrachia_owenia.txt

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Siboglinidae | cut -f 1 > gene_families_lost_Siboglinidae.txt
grep -f gene_families_lost_Siboglinidae.txt owenia_all_GF > losses_Siboglinidae_owenia
fgrep -f losses_Siboglinidae_owenia ../../Orthogroups.csv > gene_families_lost_Siboglinidae_owenia.csv
cut -f 1,21 gene_families_lost_Siboglinidae_owenia.csv | sed 's/Ofus|//g' | cut -f 2 | sed 's/, /\n/g' > gene_IDs_lost_Siboglinidae_owenia.txt

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Vestimentifera | cut -f 1 > gene_families_lost_Vestimentifera.txt
grep -f gene_families_lost_Vestimentifera.txt owenia_all_GF > losses_Vestimentifera_owenia
fgrep -f losses_Vestimentifera_owenia ../../Orthogroups.csv > gene_families_lost_Vestimentifera_owenia.csv
cut -f 1,21 gene_families_lost_Vestimentifera_owenia.csv | sed 's/Ofus|//g' | cut -f 2 | sed 's/, /\n/g' > gene_IDs_lost_Vestimentifera_owenia.txt

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -w Vestimentifera_cl1 | cut -f 1 > gene_families_lost_Vestimentifera_cl1.txt
grep -f gene_families_lost_Vestimentifera_cl1.txt owenia_all_GF
```

```
> losses_Vestimentifera_cl1_owenia
fgrep -f losses_Vestimentifera_cl1_owenia ../../Orthogroups.csv
> gene_families_lost_Vestimentifera_cl1_owenia.csv
cut -f 1,21 gene_families_lost_Vestimentifera_cl1_owenia.csv | s
ed 's/Ofus|//g' | cut -f 2 | sed 's/, /\n/g' > gene_IDs_lost_Ves
timentifera_cl1_owenia.txt

cut -f 1,6 ../../orthofinder_ultrasensitive_stats_Jun2021.tsv |
grep -w Vestimentifera_cl2 | cut -f 1 > gene_families_lost_Vesti
mentifera_cl2.txt
grep -f gene_families_lost_Vestimentifera_cl2.txt owenia_all_GF
> losses_Vestimentifera_cl2_owenia
fgrep -f losses_Vestimentifera_cl2_owenia ../../Orthogroups.csv
> gene_families_lost_Vestimentifera_cl2_owenia.csv
cut -f 1,21 gene_families_lost_Vestimentifera_cl2_owenia.csv | s
ed 's/Ofus|//g' | cut -f 2 | sed 's/, /\n/g' > gene_IDs_lost_Ves
timentifera_cl2_owenia.txt
```

Various

Piecharts of composition lost gene families

```
03s
mkdir piecharts
cd piecharts
cp /data/SBCS-MartinDuranLab/03-Giacomo/data/all_together/gene_f
amily_evolution/orthofinder_Jun2021/ultra_sensitive/Ferdi_result
/orthofinder_ultrasensitive_stats_Jun2021.tsv ./
cut -f 1,6 orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -
w Ofra | cut -f 1 > gene_families_lost_osedax.txt
grep -f gene_families_lost_osedax.txt orthofinder_ultrasensitive
_stats_Jun2021.tsv | cut -f 4 | sort | uniq -c > piechart_losses
_osedax

cut -f 1,6 orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -
```

```
w Dgyr | cut -f 1 > gene_families_lost_dimorphilus.txt
grep -f gene_families_lost_dimorphilus.txt orthofinder_ultrasens
itive_stats_Jun2021.tsv | cut -f 4 | sort | uniq -c > piechart_l
osses_dimorphilus

cut -f 1,6 orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -
w Oalv | cut -f 1 > gene_families_lost_oasisia.txt
grep -f gene_families_lost_oasisia.txt orthofinder_ultrasensitiv
e_stats_Jun2021.tsv | cut -f 4 | sort | uniq -c > piechart_losse
s_oasisia

cut -f 1,6 orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -
w Rpac | cut -f 1 > gene_families_lost_riftia.txt
grep -f gene_families_lost_riftia.txt orthofinder_ultrasensitive
_stats_Jun2021.tsv | cut -f 4 | sort | uniq -c > piechart_losses
_riftia

cut -f 1,6 orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -
w Pech | cut -f 1 > gene_families_lost_paraescarpia.txt
grep -f gene_families_lost_paraescarpia.txt orthofinder_ultrasen
sitive_stats_Jun2021.tsv | cut -f 4 | sort | uniq -c > piechart_
losses_paraescarpia

cut -f 1,6 orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -
w Lluy | cut -f 1 > gene_families_lost_lamellibrachia.txt
grep -f gene_families_lost_lamellibrachia.txt orthofinder_ultras
ensitive_stats_Jun2021.tsv | cut -f 4 | sort | uniq -c > piechar
t_losses_lamellibrachia

cut -f 1,6 orthofinder_ultrasensitive_stats_Jun2021.tsv | grep -
w Siboglinidae | cut -f 1 > gene_families_lost_siboglinidae.txt
grep -f gene_families_lost_siboglinidae.txt orthofinder_ultrasen
sitive_stats_Jun2021.tsv | cut -f 4 | sort | uniq -c > piechart_
losses_siboglinidae
```

```
ggplot(data, aes(x="", y=value, fill=group)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0)
```

```
cp /data/SBCS-MartinDuranLab/03-Giacomo/data/all_together/gene_family_evolution/orthofinder_Jun2021/ultra_sensitive/Results_Jun09/Orthogroups/Orthogroups.csv ./
grep "Ofra|" Orthogroups.csv | cut -f 1 > orthogroups_Ofra
grep -f orthogroups_Ofra orthofinder_ultrasensitive_stats_Jun2021.tsv | cut -f 4 | sort | uniq -c > piechart_Ofra

grep "Oalv|" Orthogroups.csv | cut -f 1 > orthogroups_Oalv
grep -f orthogroups_Oalv orthofinder_ultrasensitive_stats_Jun2021.tsv | cut -f 4 | sort | uniq -c > piechart_Oalv

grep "Rpac|" Orthogroups.csv | cut -f 1 > orthogroups_Rpac
grep -f orthogroups_Rpac orthofinder_ultrasensitive_stats_Jun2021.tsv | cut -f 4 | sort | uniq -c > piechart_Rpac

grep "Hrob|" Orthogroups.csv | cut -f 1 > orthogroups_Hrob
grep -f orthogroups_Hrob orthofinder_ultrasensitive_stats_Jun2021.tsv | cut -f 4 | sort | uniq -c > piechart_Hrob
```

PCA - orthogroups

input is a file generated by orthofinder: Orthogroups.GeneCount.tsv

```
library(stats)
library(factoextra)
library(R.utils)

```

```
# Whole dataset: 47,685 orthogroups
orthogroups <- read.csv("/Users/giacomo/Desktop/PCA/Orthogroups.GeneCount.tsv", header=T, sep='\t')
orthogroups_selection <- data.frame()

# We subset the 21,189 orthogroups that are not species-specific
# or the 14,373 orthogroups that are found in at least 3 species
for (i in 1:nrow(orthogroups)){
  if (sum(isZero(orthogroups[i,-c(1,30)])) < 26){
    orthogroups_selection <- rbind(orthogroups_selection,orthogroups[i,])
  }
}

# We transpose the orthogroups to analyse the individuals rather than the
# orthogroups in the downstream analysis```

individuals <- t(orthogroups_selection[-c(1,30)])
individuals_clean <- individuals[,which(apply(individuals, 2, var)!=0)]

pca_results <- prcomp(individuals_clean, scale = TRUE)
fviz_eig(pca_results)
fviz_pca_ind(pca_results, repel = TRUE) +
  theme_classic()
```

DNA repair

list

```
A-OGG1         PTHR10242
A-NTHL1         PTHR43286
A-NEIL         PTHR22993
A-UNG         PTHR11264
A-SMUG1         PTHR13235
A-MUTYH         PTHR42944
A-MPG         PTHR10429
A-MBD4         PTHR15074
A-TDG         PTHR12159
A-APEX         PTHR22748
A-ERCC1         PTHR12749
A-POL         PTHR11276
A-HMGB1         PTHR48112
A-POLD3         PTHR17598
A-POLE3         PTHR46172
A-PCNA         PTHR11352
A-LIG         PTHR45674
A-FEN1         PTHR11081
A-PARP2         PTHR10459
B-RBX1         PTHR11210
B-CUL4B         PTHR11932
B-DDB1         PTHR10644
B-DDB2         PTHR15169
B-XPC         PTHR12135
B-HR23B         PTHR10621
B-CETN2         PTHR23050
B-CSA         PTHR46202
B-CSB         PTHR45629
B-CDK7         PTHR24056
```

```
B-MNAT1        PTHR12683
B-CCNH        PTHR10026
B-XPB        PTHR11274
B-XPD        PTHR11472
B-TTDA        PTHR28580
B-TFIIH1        PTHR12856
B-TFIIH2        PTHR12695
B-TFIIH3        PTHR12831
B-TFIIH4        PTHR13152
B-XPG        PTHR16171
B-XPA        PTHR10142
B-RPA        PTHR13989
B-XPF        PTHR10150
B-ERCC1        PTHR12749
B-POLD3        PTHR17598
B-POLE3        PTHR46172
B-PCNA        PTHR11352
B-RFC1        PTHR23389
B-LIGI        PTHR45674
C-PMS2-MLH        PTHR10073
C-MSH        PTHR11361
C-RFC1        PTHR23389
C-RFC        PTHR11669
C-PCNA        PTHR11352
C-EXOI        PTHR11081
C-RPA        PTHR13989
C-POLD3        PTHR17598
C-LIGI        PTHR45674
D-ATM        PTHR11139
D-RAD50        PTHR18867
```

```
D-MRE11        PTHR10139
D-NBS1        PTHR12162
D-TOPB1        PTHR13561
D-CTIP        PTHR15107
D-BARD1        PTHR24171
D-BRCA1        PTHR13763
D-BRIP1        PTHR11472
D-PALB2        PTHR14662
D-BRCA2        PTHR11289
D-DSS1        PTHR16771
D-SYCP3        PTHR19368
D-ABRAXAS1        PTHR31728
D-RAP80        PTHR15932
D-NPA1        PTHR15660
D-BRE        PTHR15189
D-BRCC36        PTHR10410
D-RPA        PTHR13989
D-RAD51        PTHR22942
D-RAD52        PTHR12132
D-RAD54        PTHR45629
D-POLD3        PTHR17598
D-BLM        PTHR13710
D-MUS81        PTHR13451
D-TOP3        PTHR11390
D-EME1        PTHR21077
D-RAD51B        PTHR46456
D-RAD51C        PTHR46239
D-RAD51D        PTHR46457
D-XRCC2        PTHR46644
D-XRCC3        PTHR46487
```

```
E-KU         PTHR12604
E-RAD50        PTHR18867
E-MRE11        PTHR10139
E-ARTEMIS        PTHR23240
E-DNAPK        PTHR11139
E-RAD27        PTHR11081
E-POL-TDT        PTHR11276
E-LIG        PTHR45997
E-XRCC4        PTHR28559
E-XLF        PTHR32235
F-MHF        PTHR22980
F-FANCM        PTHR14025
F-FAAP24        PTHR31786
F-TELO2        PTHR15830
F-ATRIP        PTHR28594
F-ATR        PTHR11139
F-WDR48        PTHR19862
F-USP1        PTHR24006
F-FANCI        PTHR21818
F-FANCD2        PTHR32086
F-FANCD2OS        PTHR31036
F-FANCB        PTHR28450
F-FAAP100        PTHR14890
F-FANCA        PTHR12047
F-FANCL        PTHR13206
F-FANCC        PTHR16798
F-FANCE        PTHR32094
F-FANCG        PTHR15254
F-FANCF        PTHR14449
F-UBE2T        PTHR24068
```

```
F-HES1         PTHR10985
F-MUS81         PTHR13451
F-EME1         PTHR21077
F-ERCC1         PTHR12749
F-XRCC4         PTHR28559
F-SLX1A         PTHR20208
F-SLX4         PTHR21541
F-RMI1         PTHR14790
F-RMI2         PTHR33962
F-TOP3A         PTHR11390
F-BLM         PTHR13710
F-RPA         PTHR13989
F-REV1         PTHR45990
F-REV3L         PTHR45812
F-POLH         PTHR45873
F-POLI         PTHR46404
F-POLK         PTHR11076
F-POLN         PTHR10133
F-BRCA1         PTHR13763
F-BRIP1         PTHR11472
F-PALB2         PTHR14662
F-BRCA2         PTHR11289
F-RAD51         PTHR22942
F-RAD51C          PTHR46239
F-FAN1         PTHR15749
F-PMS2-MLH          PTHR10073
```

```
sed -i 's/  */\t/g' DNA_repair_pantherID.txt
```

A - Base Excision Repair
B - Nucleotide Excision Repair

C - Mismatch repair
D - Homologous Recombination
E - Non Homologous End - Joining
F - Fanconi Anemia Pathway
G - Microhomology-end joining repair pathway

code

test.sh

```
#!/bin/bash


species=$1
xls=/data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/"$species"*xls
output2="$species"_DNArepair_count.txt

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   echo $Panther_ID
   count=$(fgrep "$Panther_ID" $xls | wc -l)
   echo $count >> $output2
done < DNA_repair_pantherID.txt
```

```
cut -f1 DNA_repair_pantherID.txt > DNA_repair_IDs
echo "gene"$'\t'"Riftia"$'\t'"Oasisia"$'\t'"Osedax"$'\t'"Paraescarpia"$'\t'"Lamellibrachia"$'\t'"Owenia"$'\t'"Capitella" > DNA_repair_siboglinidae.txt
paste DNA_repair_IDs riftia_DNArepair_count.txt oasisia_DNArepair_count.txt osedax_DNArepair_count.txt paraescarpia_DNArepair_count.txt lamellibrachia_DNArepair_count.txt owenia_DNArepair_count.txt capitella_DNArepair_count.txt >> DNA_repair_siboglinidae.txt
```

R plot

```r
library(tidyverse)
library(dplyr)
library(ggplot2)
library(data.table)
library(gplots)
library(pheatmap)
library(dendextend)
library(factoextra)
library(ComplexHeatmap)
library(RColorBrewer)
library(NbClust)
library(scales)

#Import data in matrix format
A <- read.delim("~/Desktop/DNA_repair/DNA_repair_NOisoforms/A", row.names=1)
B <- read.delim("~/Desktop/DNA_repair/DNA_repair_NOisoforms/B", row.names=1)
C <- read.delim("~/Desktop/DNA_repair/DNA_repair_NOisoforms/C", row.names=1)
D <- read.delim("~/Desktop/DNA_repair/DNA_repair_NOisoforms/D", row.names=1)
E <- read.delim("~/Desktop/DNA_repair/DNA_repair_NOisoforms/E", row.names=1)
F <- read.delim("~/Desktop/DNA_repair/DNA_repair_NOisoforms/F", row.names=1)

# Option 1. 0 to 1 relative abundance/expression (54 is the highest value in my dataset)
rescale_custom <- function(x) (x/14)
```

```
A_normalised <- t(apply(A, 1, rescale_custom))
rescale_custom <- function(x) (x/54)
B_normalised <- t(apply(B, 1, rescale_custom))
rescale_custom <- function(x) (x/11)
C_normalised <- t(apply(C, 1, rescale_custom))
rescale_custom <- function(x) (x/18)
D_normalised <- t(apply(D, 1, rescale_custom))
rescale_custom <- function(x) (x/18)
E_normalised <- t(apply(E, 1, rescale_custom))
rescale_custom <- function(x) (x/27)
F_normalised <- t(apply(F, 1, rescale_custom))


# To make 0 a different colour
# First create whatever gradient (e.g. RdBu)
heatmap_color <- colorRampPalette(brewer.pal(n = 7, name = "Reds"))(1000)
heatmap_color[1] <- rgb(1,1,1)
#column_labels = c("your","labels"),
#row_labels = c("your","labels"))

paletteLength <- 1000
# to go from 0 to max.value (e.g. 1):
myBreaks <- c(seq(1/paletteLength, 1, length.out=floor(paletteLength)))

pheatmap(A_normalised,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
```

```
          height = 25,
          width = 20,
          breaks = myBreaks)

pheatmap(B_normalised,
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          border_color = NA,
          color = heatmap_color,
          height = 25,
          width = 20,
          breaks = myBreaks)

pheatmap(C_normalised,
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          border_color = NA,
          color = heatmap_color,
          height = 25,
          width = 20,
          breaks = myBreaks)

pheatmap(D_normalised,
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          border_color = NA,
          color = heatmap_color,
          height = 25,
          width = 20,
          breaks = myBreaks)
```

```

pheatmap(E_normalised,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         height = 25,
         width = 20,
         breaks = myBreaks)

pheatmap(F_normalised,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         height = 25,
         width = 20,
         breaks = myBreaks)

```

- exported in PDF 8x6 inches

Developmental pathways

PFAM and panther tables

1st step

obtain the PFAM colums from the annotation files

```
cut -f 8 osedax_annotation_Jan2021_TrinoPantherKO.xls > PFAM_osedax
```

- for 5 sibo + owenia and capi

```
fgrep -f gene_families_originated_Vestimentifera.txt Orthogroups
```

```
_Jan2021.csv > gene_families_originated_Vestimentifera.csv
```

obtain transcript names from non redundant proteomes, this way we will get rid of the isoforms

```
grep ">" /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Ctel.fa | sed 's/>Ctel_//g' > capitella_isoform_names
fgrep -w -f capitella_isoform_names ../capitella_annotation_Feb2021_TrinoPantherKO.xls > capitella_isoform.xls
grep ">" /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Ofus.fa | sed 's/>//g' > owenia_isoform_names
fgrep -w -f owenia_isoform_names ../owenia_annotation_v250920.1_TrinoPantherKO.xls > owenia_isoform.xls
grep ">" /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Lluy.fa | sed 's/>Lluy_//g' > lamellibrachia_isoform_names
fgrep -w -f lamellibrachia_isoform_names ../lamellibrachia_annotation_Feb2021_TrinoPantherKO_OK.xls > lamellibrachia_isoform.xls
grep ">" /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Pech.fa | sed 's/>Pech_//g' | sed 's/nbis-mrna-/nbis_mrna_/g' > paraescarpia_isoform_names
fgrep -w -f paraescarpia_isoform_names ../paraescarpia_annotation_Jun2021_TrinoPantherKO.xls > paraescarpia_isoform.xls
grep ">" /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Ofra.fa | sed 's/>Ofra_//g' > osedax_isoform_names
fgrep -w -f osedax_isoform_names ../osedax_annotation_Jan2021_TrinoPantherKO.xls > osedax_isoform.xls
grep ">" /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Oalv.fa | sed 's/>Oalv_//g' > oasisia_isoform_names
fgrep -w -f oasisia_isoform_names ../oasisia_annotation_Jan2021_TrinoPantherKO.xls > oasisia_isoform.xls
grep ">" /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Rpac.fa | sed 's/>Rpac_//g' > riftia_isoform_names
fgrep -w -f riftia_isoform_names ../riftia_annotation_Jan2021_TrinoPantherKO.xls > riftia_isoform.xls
```

2nd step

First half:

PFAM_list.txt

```
T-box (PF00907) PF00907
Homeodomain (PF00046) PF0046
Fox (PF00250) PF00250
HMG-box (PF00505) PF00505
bHLH (PF00010) PF00010
bZIP (PF00170) PF00170
LAG (PF09271) PF09271
STAT (PF02864) PF02864
Mef2 (PF00319) PF00319
p53 (PF00870) PF00870
RHD (PF00554) PF00554
GATA (PF00320) PF00320
C2H2-Zn (PF00096) PF00096
GRH/LSF (PF04516) PF04516
Runt (PF00853) PF00853
zf-C4 (PF00105) PF00105
Myb (PF00249) PF00249
SMAD (PF03165) PF03165
```

Script:

test.sh

```
#!/bin/bash

species=$1
xls=/data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/"$species"*xls
output="$species"_FirstHalf_Development.txt

while read line; do
```

```
    PFAM_ID=$(cut -f 2 <<< "$line")
    echo $PFAM_ID
    count=$(fgrep "$PFAM_ID" $xls | wc -l)
    echo $count >> $output
done < PFAM_list.txt
```

Second half:
SecondHalf_list.txt

```
Notch (PF00066) PF00066
WNT (PF00110) PF00110
Fzl (PF01534) PF01534
HH (PF01085) PF01085
Patched PTHR46022(prot itself) PTHR10796(domain containing)
TGF-b (PF00019) PF00019
TGF-bR PTHR23255(TGF-BETA RECEPTOR TYPE-2 and 1) PTHR14002(TRANS
FORMING GROWTH FACTOR BETA RECEPTOR TYPE 3)
FGF (PF00167) PF00167
FGFR PTHR24416:SF131(type 1) (type 1) PTHR24416:SF130(type 2) PT
HR24416:SF505(type 3) PTHR24416:SF343(type 4)
VEGF (PF00341) PF00341
VEGFR PTHR24416:SF49(type 2) PTHR24416:SF45(type 3) PTHR24416:SF
390 (type 1)
EGF ligands TRANSFORMING GROWTH FACTOR ALPHA (PTHR10740) PTHR111
00:SF18 PTHR11100:SF20 PTHR11100:SF7 PTHR46513:SF5
EGFR PTHR24416:SF91 PTHR24416:SF566
```

test_SecondHalf.sh

```
#!/bin/bash

species=$1
xls=/data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_an
```

```
notations/"$species"*xls
output="$species"_SecondHalf_Development.txt

grep "PF00066" $xls | wc -l >> $output
grep "PF00110" $xls | wc -l >> $output
grep "PF01534" $xls | wc -l >> $output
grep "PF01085" $xls | wc -l >> $output
grep "PTHR46022\|PTHR10796" $xls | wc -l >> $output
grep "PF00019" $xls | wc -l >> $output
grep "PTHR23255\|PTHR14002" $xls | wc -l >> $output
grep "PF00167" $xls | wc -l >> $output
grep "PTHR24416:SF131\|PTHR24416:SF130\|PTHR24416:SF505\|PTHR24416:SF343" $xls | wc -l >> $output
grep "PF00341" $xls | wc -l >> $output
grep "PTHR24416:SF49\|PTHR24416:SF45\|PTHR24416:SF390" $xls | wc -l >> $output
grep "PTHR10740\|PTHR11100:SF18\|PTHR11100:SF20\|PTHR11100:SF7\|PTHR46513:SF5" $xls | wc -l >> $output
grep "PTHR24416:SF91\|PTHR24416:SF566" $xls | wc -l >> $output
```

3rd step

```
cut -f 1 PFAM_list.txt > list1
cut -f 1 SecondHalf_list.txt > list2

echo "gene"$'\t'"Riftia"$'\t'"Oasisia"$'\t'"Osedax"$'\t'"Lamellibrachia"$'\t'"Owenia"$'\t'"Capitella" > Developmental_pathways_table_FirstHalf
echo "gene"$'\t'"Riftia"$'\t'"Oasisia"$'\t'"Osedax"$'\t'"Lamellibrachia"$'\t'"Owenia"$'\t'"Capitella" > Developmental_pathways_table_SecondHalf

```

```
paste list1 riftia_FirstHalf_Development.txt oasisia_FirstHalf_Development.txt osedax_FirstHalf_Development.txt lamellibrachia_FirstHalf_Development.txt owenia_FirstHalf_Development.txt capitella_FirstHalf_Development.txt >> Developmental_pathways_table_FirstHalf
paste list2 riftia_SecondHalf_Development.txt oasisia_SecondHalf_Development.txt osedax_SecondHalf_Development.txt lamellibrachia_SecondHalf_Development.txt owenia_SecondHalf_Development.txt capitella_SecondHalf_Development.txt >> Developmental_pathways_table_SecondHalf
```

FirstHalf:

```
gene        Riftia        Oasisia        Osedax        Lamellibrachia        Owenia        Capitella
T-box (PF00907)        11        12        6        19        9        8
Homeodomain (PF00046)        99        110        90        117        132        151
Fox (PF00250)        28        30        25        31        34        42
HMG-box (PF00505)        27        29        34        28        30        23
bHLH (PF00010)        63        60        36        69        61        81
bZIP (PF00170)        17        19        9        17        13        16
LAG (PF09271)        1        1        1        1        1        1
STAT (PF02864)        2        1        1        2        2        6
Mef2 (PF00319)        2        2        2        2        2        2
p53 (PF00870)        1        1        1        1        4        1
```

```
RHD (PF00554)          3          3          3          3          2
 3
GATA (PF00320)         12         10          7         12          5
     9
C2H2-Zn (PF00096)         243         263         95         262
   563         142
GRH/LSF (PF04516)         2          2          2          2          2
     3
Runt (PF00853)         1          1          1          1          1
   1
zf-C4 (PF00105)         34         32         31         37         3
7         37
Myb (PF00249)         16         20         14         19         17
       17
SMAD (PF03165)         5          5          4          5          5
   4
```

SecondHalf:

```
gene         Riftia         Oasisia         Osedax         Lamellibr
achia         Owenia         Capitella
Notch (PF00066)         3         3         1         3         3
   5
WNT (PF00110)         13         14         5         21         11
     12
Fzl (PF01534)         6         4         4         7         4
 5
HH (PF01085)         2         2         1         3         1
 1
Patched         18         16         7         38         4         6
TGF-b         12         12         7         14         14         14
TGF-bR         15         15         5         15         24         9
FGF         4         4         3         4         4         2
FGFR         1         1         2         1         1         4
```

```
VEGF        1        1        4        1        2        2
VEGFR        2        1        1        1        1        2
EGF ligands        8        16        5        7        6
 5
EGFR        3        1        1        1        2        2
```

```
# convert to log(n)
awk 'NR>1{for(i=1;i<=NF;i++) $i=log($i)}1' riftia_PFAM_Developme
nt.txt > riftia_PFAM_Development_logN.txt
awk 'NR>1{for(i=1;i<=NF;i++) $i=log($i)}1' oasisia_PFAM_Developm
ent.txt > oasisia_PFAM_Development_logN.txt
awk 'NR>1{for(i=1;i<=NF;i++) $i=log($i)}1' osedax_PFAM_Developme
nt.txt > osedax_PFAM_Development_logN.txt
awk 'NR>1{for(i=1;i<=NF;i++) $i=log($i)}1' lamellibrachia_PFAM_D
evelopment.txt > lamellibrachia_PFAM_Development_logN.txt
awk 'NR>1{for(i=1;i<=NF;i++) $i=log($i)}1' owenia_PFAM_Developme
nt.txt > owenia_PFAM_Development_logN.txt
awk 'NR>1{for(i=1;i<=NF;i++) $i=log($i)}1' capitella_PFAM_Develo
pment.txt > capitella_PFAM_Development_logN.txt

awk 'NR>1{for(i=1;i<=NF;i++) $i=log($i)}1' riftia_Panther_Develo
pment.txt > riftia_Panther_Development_logN.txt
awk 'NR>1{for(i=1;i<=NF;i++) $i=log($i)}1' oasisia_Panther_Devel
opment.txt > oasisia_Panther_Development_logN.txt
awk 'NR>1{for(i=1;i<=NF;i++) $i=log($i)}1' osedax_Panther_Develo
pment.txt > osedax_Panther_Development_logN.txt
awk 'NR>1{for(i=1;i<=NF;i++) $i=log($i)}1' lamellibrachia_Panthe
r_Development.txt > lamellibrachia_Panther_Development_logN.txt
awk 'NR>1{for(i=1;i<=NF;i++) $i=log($i)}1' owenia_Panther_Develo
pment.txt > owenia_Panther_Development_logN.txt
awk 'NR>1{for(i=1;i<=NF;i++) $i=log($i)}1' capitella_Panther_Dev
```

```
elopment.txt > capitella_Panther_Development_logN.txt

echo "gene"$'\t'"Riftia"$'\t'"Oasisia"$'\t'"Osedax"$'\t'"Lamellibrachia"$'\t'"Owenia"$'\t'"Capitella" > Developmental_pathways_table_logN

paste list1 riftia_PFAM_Development_logN.txt oasisia_PFAM_Development_logN.txt osedax_PFAM_Development_logN.txt lamellibrachia_PFAM_Development_logN.txt owenia_PFAM_Development_logN.txt capitella_PFAM_Development_logN.txt >> Developmental_pathways_table_logN
paste list2 riftia_Panther_Development_logN.txt oasisia_Panther_Development_logN.txt osedax_Panther_Development_logN.txt lamellibrachia_Panther_Development_logN.txt owenia_Panther_Development_logN.txt capitella_Panther_Development_logN.txt >> Developmental_pathways_table_logN
```

result:

```
gene        Riftia        Oasisia        Osedax        Lamellibrachia        Owenia        Capitella
T-box (PF00907)        11        12        6        20        9        8
Homeodomain (PF0046)        3.09104        3.13549        3.04452        2.94444        3.04452        3.04452
Fox (PF00250)        3.3322        3.43399        3.21888        3.46574        3.68888        3.73767
HMG-box (PF00505)        3.58352        3.82864        3.55535        3.3673        3.68888        3.13549
bHLH (PF00010)        4.20469        4.29046        3.63759        4.23411        4.29046        4.39445
bZIP (PF00170)        2.99573        3.21888        2.30259        2.83321        2.77259        2.77259
LAG (PF09271)        0        0        0        0        1.09861        0
```

```
STAT (PF02864)        0.693147        0        0        0.693147
      1.09861        1.79176
Mef2 (PF00319)        0.693147        1.09861        1.09861
   0.693147        0.693147        0.693147
p53 (PF00870)        1.60944        0        0        0
 1.38629        0
RHD (PF00554)        1.09861        1.38629        1.38629
 1.09861        0.693147        1.09861
GATA (PF00320)        2.70805        2.63906        2.07944
   2.48491        2.3979        2.30259
C2H2-Zn (PF00096)        5.52943        5.63835        4.58497
     5.57215        6.44572        4.95583
GRH/LSF (PF04516)        1.09861        1.38629        0.693147
      0.693147        0.693147        1.09861
Runt (PF00853)        0        0        0        0        0
   0
zf-C4 (PF00105)        3.61092        3.52636        3.49651
   3.61092        3.95124        3.61092
Myb (PF00249)        2.94444        3.09104        2.77259
 2.94444        3.2581        2.89037
SMAD (PF03165)        1.79176        1.79176        1.60944
   1.60944        1.94591        1.38629
Notch (PF00066)        1.09861        1.09861        0        1.
09861        1.09861        1.60944
WNT (PF00110)        2.56495        2.63906        1.79176
 3.09104        2.77259        2.48491
Fzl (PF01534)        1.79176        1.38629        1.38629
 1.94591        1.38629        1.60944
HH (PF01085)        0.693147        0.693147        0        1.0
9861        0        0
TGF-b (PF00019)        2.48491        2.48491        1.94591
   2.63906        2.70805        2.63906
```

```
FGF (PF00167)        1.38629        1.60944        1.09861
 1.38629        1.79176        0.693147
VEGF (PF00341)        0        0        1.38629        0
 0.693147        0.693147
Patched        18        16        7        38        5        6
TGF-bR        2.70805        2.70805        1.60944        2.708
05        3.3673        2.19722
FGFR        0        0        0.693147        0        0
 1.38629
VEGFR        0.693147        0        0        0        0
 0.693147
EGF ligands        1.79176        2.48491        1.38629
 1.94591        1.60944        1.38629
EGFR        1.09861        0        0        0        1.09861
     0.693147
```

with updated result containing paraescarpia as well:

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(data.table)
library(gplots)
library(pheatmap)
library(dendextend)
library(factoextra)
library(ComplexHeatmap)
library(RColorBrewer)
library(NbClust)
library(scales)

#Import data in matrix format
```

```
Unique_table_log <- read.delim("~/Desktop/Developmental_pathways/no isoforms/Unique_table_log", row.names=1)


# Option 1. 0 to 1 relative abundance/expression (54 is the highest value in my dataset)



# To make 0 a different colour
# First create whatever gradient (e.g. RdBu)
heatmap_color <- colorRampPalette(rev(brewer.pal(n = 7, name = "Blues")))(100)
#heatmap_color[1] <- rgb(0,0,0) # here include the colour you're interested in
heatmap_color <- colorRampPalette(brewer.pal(n = 7, name = "RdBu"))(100)
heatmap_color <- heatmap_color[50:100]
#column_labels = c("your","labels"),
#row_labels = c("your","labels"))


# If you want common scale for different heatmaps:
# First define some "breaks"

pheatmap(Unique_table_log,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         height = 25,
```

```
          width = 20)


```

- exported in PDF 8x6 inches

Updated Plot
I will divide the plot in three parts:

- first half
- second half signals
- second half receptors

```
2 Notch (PF00066) Receptor
3 WNT (PF00110) Signal
4 Fzl (PF01534) Receptor
5 HH (PF01085) Signal
6 Patched Receptor
7 TGF-b Signal
8 TGF-bR Receptor
9 FGF SIgnal
10 FGFR Receptor
11 VEGF Signal
12 VEGFR Receptor
13 EGF ligands Signal
14 EGFR Receptor
```

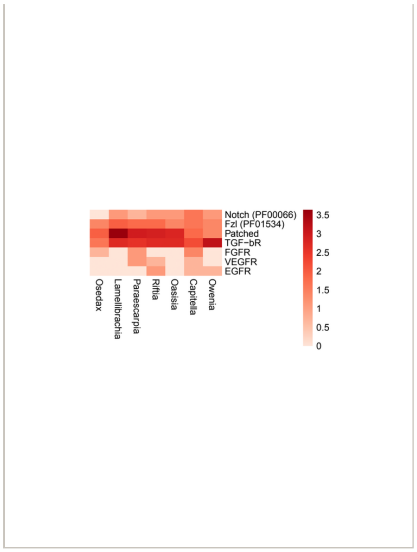
to divide the second half file

```
sed -e '2d;4d;6d;8d;10d;12d;14d' second_half_log > second_half_log_signal
sed -e '3d;5d;7d;9d;11d;13d' second_half_log > second_half_log_receptor
```

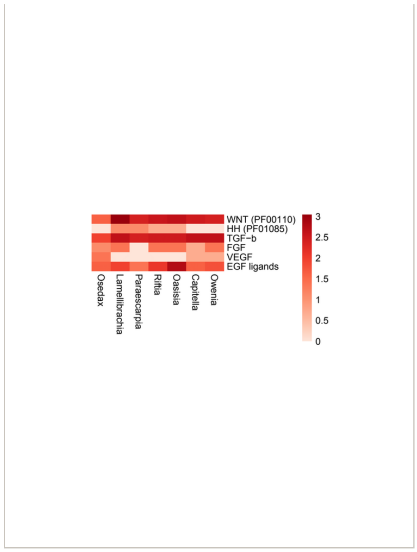
```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(data.table)
library(gplots)
library(pheatmap)
library(dendextend)
library(factoextra)
library(ComplexHeatmap)
library(RColorBrewer)
library(NbClust)
library(scales)

#Import data in matrix format
first_half_log <- read.delim("~/Desktop/Developmental_pathways/no isoforms/first_half_log_Paraescarpia_STAT=-inf", row.names=1)
second_half_log_signal <- read.delim("~/Desktop/Developmental_pathways/no isoforms/second_half_log_signal", row.names=1)
second_half_log_receptor <- read.delim("~/Desktop/Developmental_pathways/no isoforms/second_half_log_receptor", row.names=1)
# Option 1. 0 to 1 relative abundance/expression (54 is the highest value in my dataset)

heatmap_color <- colorRampPalette(brewer.pal(n = 7, name = "Reds"))(1000)

# If you want common scale for different heatmaps:
# First define some "breaks"

pheatmap(first_half_log,
```

```
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          border_color = NA,
          color = heatmap_color,
          cellheight = 10,
          cellwidth = 20)

pheatmap(second_half_log_signal,
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          border_color = NA,
          color = heatmap_color,
          cellheight = 10,
          cellwidth = 20)

pheatmap(second_half_log_receptor,
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          border_color = NA,
          color = heatmap_color,
          cellheight = 10,
          cellwidth = 20)

```

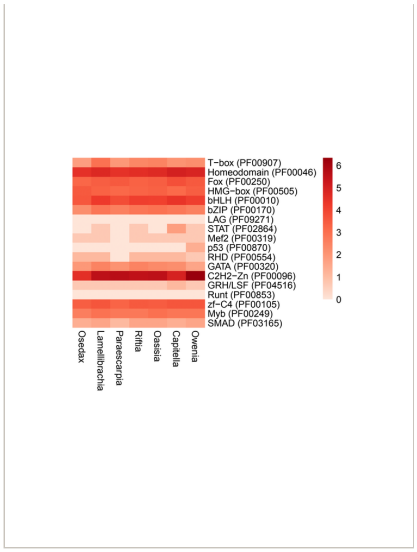
- exported in PDF 8x6 inches



PDF second_half_... • PDF document



PDF second_half_... • PDF document



PDF first_half_Par... • PDF document

Wnt, Fzl, BMP ligands & receptors

- **select candidates**

get transcripts id:

```
cd /data/scratch/btx654/developmental_pathways/WNT
grep "PF00110" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/capitella* | cut -f1 | sed 's/^/Ctel_/' > WNT_candidates_capitella
module load seqtk
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Ctel.fa WNT_candidates_capitella > WNT_candidates_capitella.fa
grep "PF00110" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/owenia*.xls | cut -f2  > WNT_candidates_owenia
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Ofus.fa WNT_candidates_owenia > WNT_candidates_owenia.fa
grep "PF00110" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/riftia* | cut -f2 | sed 's/^/Rpac_/' > WNT_candidates_riftia
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Rpac.fa WNT_candidates_riftia > WNT_candidates_riftia.fa
```

```
grep "PF00110" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/oasisia* | cut -f2 | sed 's/^/Oalv_/' > WNT_candidates_oasisia
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Oalv.fa WNT_candidates_oasisia > WNT_candidates_oasisia.fa
grep "PF00110" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/osedax* | cut -f2 | sed 's/^/Ofra_/' > WNT_candidates_osedax
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Ofra.fa WNT_candidates_osedax > WNT_candidates_osedax.fa
grep "PF00110" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/lamellibrachia* | cut -f2 | sed 's/^/Lluy_/' > WNT_candidates_lamellibrachia
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Lluy.fa WNT_candidates_lamellibrachia > WNT_candidates_lamellibrachia.fa
grep "PF00110" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/paraescarpia* | cut -f2 | sed 's/nbis_mrna_/nbis-mrna-/' | sed 's/^/Pech_/' > WNT_candidates_paraescarpia
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Pech.fa WNT_candidates_paraescarpia > WNT_candidates_paraescarpia.fa
```

- added these sequences to a fasta file containing wnt genes

```
"PF01534" Fzl
"PF01534" BMP ligands
"PTHR23255\|PTHR14002" BMP receptors
```

- **Phylogenetic reconstruction**

Step 1 - WNT phylogenetic tree
make an alignment using the txt file obtained before and the online tool MAFFT
don't select any additional options
export in fasta format and name the output:

```
wnt_sequences_MAFFT1.fasta
```

Step 2 - WNT phylogenetic tree
download and install Jalview
open the file:

```
wnt_sequences_MAFFT1.fasta
```

Cut away all the areas with no conservation. Basically we should maintain only the domain
and save the cutted file as:

```
wnt_sequences_Jalview.fasta
```

Step 3 - WNT phylogenetic tree
make an alignment again using the file “wnt_sequences_Jalview.fasta” obtained before and the online tool MAFFT
select the option “L-INS-i”
export in fasta format and name the output:

```
wnt_sequences_MAFFT2.fasta
```

Step 4 - WNT phylogenetic tree
use Jalview again. Open the file:

```
wnt_sequences_MAFFT2.fasta
```

No chopping this time, just open the file and export it to fasta as:

```
wnt_sequences_Jalview2.fasta
```

Step 5 - WNT phylogenetic tree

```
wnt_sequences_gblocks.fa
```

or use trimAl
on my personal pc

```
conda create -n fasttree_env
conda activate fasttree_env
conda install -c bioconda fasttree
conda install -c bioconda trimal
```

```
trimal -in 6a.fasta -out wnt_sequences_trimal.fa
```

Step 6 - WNT phylogenetic tree
make a tree

```
conda activate fasttree_env
FastTree wnt_sequences_trimal.fa > wnt_sequences_a.tree
```

and then open it with FigTree

PARAHOXES

```
GS HOMEOBOX 1 (PTHR24339:SF29) GSX1
GS HOMEOBOX 2 (PTHR47421:SF1) GSX2
PANCREAS/DUODENUM HOMEOBOX PROTEIN 1 (PTHR45664:SF12) PDX1
HOMEOBOX PROTEIN CDX-1 (PTHR24332:SF16)
CDX2 PTHR24332:SF27
CDX4 PTHR24332:SF15
```

parahoxes_Ofus_Ctel.fa

```
>CDX1_owenia
MVQEFGETLHYTAGTRQPTTMSLNPFLTTQQQGGYPQDFGPAFQISPMESQMQQQWGMYAGSRS
AAGLEEWQQAFAAQAQNAGAYAAHFNQTGHPAPMSASVNTTSPRQQNRAPFDWMKRQTYAAQPA
AGKTRTKDKYRVVYSDHQRLELEKEFHYSRYITIRRKAELAQALSLSERQVKIWFQNRRAKERK
CNKKKDEQNSILGKEDSELIVSEHLTQEHVHQALSAQQHVV
>GSX2_owenia
MSTSYFVDALLLKKPTQMSLQRELSTAMSRQSQITHLPPPAHNHTPMLPGQPLACYPRRPSELF
GGCCPLCIQTPGGHLICPSNAASNLSTMKHLLPTSSASALSSSSGFTPRLPLAINTVSRLPSRR
DSPSPPEYSAVDTRRIRYMNLGNIGMTRDSSDDLPSGKRIRTAFTSTQLLELEREFASNMYLSR
LRRIEIATYLNLSEKQVKIWFQNRRVKQKKEGTDEAPTHDKCRCLRTCASRNEKQDIECHGNDC
ESVNSPSSEIDSSDINSSDISQVSSQSITKDSQIPVDING
>PDX1_owenia
MDGSNPYCSQGMYGRDSYGGHTQQSMGPYNLPACVYDTNKQTSIGLDYTSQHGMAMVDHMVEQP
```

```
MVNSIPAHSLQPQPTPAHMQHGVNMNISNLNTNVPQQQSHPTSVPLQPPPAHQSQKPSNSNGGN
SSSSNNNNEKPLQFPWMKTTKSHARQWKAQWPGANFNIEDDNKRTRTAYTRLQLVELEKEFHYS
KYISRPRRIEIAAMLNLTERHIKIWFQNRRMKWKKDEAKRRPRPLSEEIDSKVAINTELLDKDG
AGSSPEIMTKEENPTFDDLSDSLSPSNTKPFIGTMKD
>CDX_capitella
MVQELSYLDSYSSNFPMFHHQPPSGGALRHQGPPYMHAPSATGLSGYGVQQPTNNMLMASQQYN
HSQDYSNYMDSTGRAQQTTCSPVQSSAGWSTGGMYAAGPVHTPRMDEWGNGFVSQPSLQIPPNQ
HSPGTAYPPYRSPDQLQQAAGGGGGGGGSGGFGVSPAPQACSQSPMPSQHMQPAQSPSPSSII
GAQGDGLSPGSPGAGGMRHQGQQNPRVPYDWMKKQNYQSMPLAMGKTRTKDKYRIVYSEYQKVE
LEKEYLYSKYITIQRKAELSRSIGLSERQVKIWFQNRRAKERKQKRKMEEALSTSDSSDKEHIK
VETMHEVHAPHMHSHAPH*
>GSX_capitella
MTSSTSFSVDTLLYKGKPPKSSTSQIHQDSSPRSAIQPALFARTPLLPPPRNLFSETDGDKLLS
RMLCCPICLTSGHYGQICPLTIPVSANRLPSPYPHHKPVFSLSPLHVTSPFPRTHHMTIPTRHF
TGNGGHVMTPQRPSERGSPPESSSPSGEGTKKGHSPPLPCPDDDADESSDAVKRMRTAFSSTQL
LELEREFASNMYLSRLRRIEIATYLSLSEKQVKIWFQNRRVKFKKEGAAHGSRDHPHCQCQLRS
CTSRDRKRDHVTEHTDIEVNVTDDSDEKCL*
>PDX1_capitella
MEELDPCFPPPNHAAMFSRDFNGFLPSSNPYSTSESPSCVYDTRMGTSYNSGMVEDTYNHQYEN
PLPHQHQHQHISRSRVYADPSVHLRSPNEHLVGGMAHRAMDADNHVMHHVGMAPSHHSAPDKVK
EQGKVHFPWMKTTKSHAHQWKANWSGANFQTFSENKRTRTAYTRAQLLELEKEFHFNRYITRPR
RVELAAHLNLTEQHIKIWFQNRRMKWKKDVDKKRPQQSEQDGADDDVSSDVTDVKKIEPKIESV
QDEITDEITDENVNGVQSDQSL*
```

parahoxes_universal.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

```

```
species=$1
parahox_genes=parahoxes_Ofus_Ctel.fa
parahox_genes_path=/data/scratch/btx654/HOX_genes/parahox/$parahox_genes

if [ "$species" == "riftia" ]; then
    non_redundant_prot=Rpac.fa
fi
if [ "$species" == "oasisia" ]; then
    non_redundant_prot=Oalv.fa
fi
if [ "$species" == "osedax" ]; then
    non_redundant_prot=Ofra.fa
fi
if [ "$species" == "lamellibrachia" ]; then
    non_redundant_prot=Lluy.fa
fi
if [ "$species" == "paraescarpia" ]; then
    non_redundant_prot=Pech.fa
fi
non_redundant_prot_path=/data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/$non_redundant_prot

echo "Working on "$species

cd /data/scratch/btx654/HOX_genes/parahox
mkdir -p $species
cd $species
cp $non_redundant_prot_path ./
cp $parahox_genes_path ./

```

```
#make a diamond BLAST database of this proteome and BLAST the consensi.fa.classified dataset against it, to find potential bona fide genes. To make sure we only get the real genes, the e-value is very stringent.
module load anaconda3
source activate diamond
diamond makedb --in $non_redundant_prot -d non_redundant_prot

diamond blastp -d non_redundant_prot -q $parahox_genes -o default.1e10.blastp -f 6 qseqid bitscore evalue stitle -k 25 -e 1e-10 -p 8
diamond blastp -d non_redundant_prot -q $parahox_genes -o ultra_sensitive.1e10.blastp --ultra-sensitive -f 6 qseqid bitscore evalue stitle -k 25 -e 1e-10 -p 8
```

tblastn_osedax_assembly.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/parahox/osedax
#$ -o /data/scratch/btx654/HOX_genes/parahox/osedax
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

module load blast+
makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/haploidization/purge_dups/osedax_purged.fa -dbtype nucl -out osedax_assembly
tblastn -db osedax_assembly -query parahoxes_Ofus_Ctel.fa -out osedax_assembly_tblastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db osedax_assembly -query parahoxes_Ofus_Ctel.fa -out o
```

```
sedax_assembly_tblastn_out.html -max_target_seqs 5 -evalue 1e-10
-num_threads 8 -html
```

tblastn_osedax_transcriptome.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/parahox/osedax
#$ -o /data/scratch/btx654/HOX_genes/parahox/osedax
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

cp /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/trinity/osed
ax_*/*.Trinity.fasta ./
cat *.Trinity.fasta > combined.trinity.fasta
rm *.Trinity.fasta
module load blast+
makeblastdb -in combined.trinity.fasta -dbtype nucl -out osedax_
transcriptome
tblastn -db osedax_transcriptome -query parahoxes_Ofus_Ctel.fa -
out osedax_transcriptome_tblastn_out -max_target_seqs 5 -evalue
1e-10 -num_threads 8 -outfmt 6
tblastn -db osedax_transcriptome -query parahoxes_Ofus_Ctel.fa -
out osedax_transcriptome_tblastn_out.html -max_target_seqs 5 -ev
alue 1e-10 -num_threads 8 -html
```

HOX

```
module load anaconda3
conda activate diamond
conda install -c bioconda diamond
```

/data/home/btx654/scripts/06-other_analyses/HOX_genes/diamond_blastp_universal.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/
#$ -o /data/scratch/btx654/
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

species=$1
hox_genes=hox_genes_owenia.fa
hox_genes_path=/data/SBCS-MartinDuranLab/03-Giacomo/data/06-other_analyses/HOX_genes/$hox_genes

if [ "$species" == "riftia" ]; then
    non_redundant_prot=Rpac.fa
fi
if [ "$species" == "oasisia" ]; then
    non_redundant_prot=Oalv.fa
fi
if [ "$species" == "osedax" ]; then
    non_redundant_prot=Ofra.fa
fi
if [ "$species" == "lamellibrachia" ]; then
    non_redundant_prot=Lluy.fa
fi
if [ "$species" == "paraescarpia" ]; then
    non_redundant_prot=Pech.fa
fi
```

```
non_redundant_prot_path=/data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/$non_redundant_prot

echo "Working on "$species


mkdir -p HOX_genes
cd HOX_genes
mkdir -p $species
cd $species
cp $non_redundant_prot_path ./
cp $hox_genes_path ./

#make a diamond BLAST database of this proteome and BLAST the consensi.fa.classified dataset against it, to find potential bona fide genes. To make sure we only get the real genes, the e-value is very stringent.
module load anaconda3
source activate diamond
diamond makedb --in $non_redundant_prot -d non_redundant_prot

diamond blastp -d non_redundant_prot -q $hox_genes -o default.1e10.blastp -f 6 qseqid bitscore evalue stitle -k 25 -e 1e-10 -p 8
diamond blastp -d non_redundant_prot -q $hox_genes -o ultra_sensitive.1e10.blastp --ultra-sensitive -f 6 qseqid bitscore evalue stitle -k 25 -e 1e-10 -p 8
```

organising the output of blast

```
sort -u -k 1,1 ultra_sensitive.1e10.blastp > best_hit_riftia
sort -u -k 1,1 ultra_sensitive.1e10.blastp > best_hit_oasisia
sort -u -k 1,1 ultra_sensitive.1e10.blastp > best_hit_osedax
sort -u -k 1,1 ultra_sensitive.1e10.blastp > best_hit_lamellibra
```

```
chia
sort -u -k 1,1 ultra_sensitive.1e10.blastp > best_hit_paraescarpia
```

Step 0 - HOX phylogenetic tree
create a txt file with hox genes sequences from many species and the candidate genes identified by blast

```
cut -f 4 ultra_sensitive.1e10.blastp | sort | uniq > riftia_candidates
cut -f 4 ultra_sensitive.1e10.blastp | sort | uniq > oasisia_candidates
cut -f 4 ultra_sensitive.1e10.blastp | sort | uniq > osedax_candidates
cut -f 4 ultra_sensitive.1e10.blastp | sort | uniq > lamellibrachia_candidates
cut -f 4 ultra_sensitive.1e10.blastp | sort | uniq > paraescarpia_candidates
```

```
module load seqtk
seqtk subseq Rpac.fa riftia_candidates > riftia_candidates.fa
seqtk subseq Oalv.fa oasisia_candidates > oasisia_candidates.fa
seqtk subseq Ofra.fa osedax_candidates > osedax_candidates.fa
seqtk subseq Lluy.fa lamellibrachia_candidates > lamellibrachia_candidates.fa
seqtk subseq Pech.fa paraescarpia_candidates > paraescarpia_candidates.fa
cd ..
cat */*candidates.fa > candidates.fa
```

- then I have appended this sequences to a txt file cointaining the hox genes sequences of many different organisms. I used nano to edit on my personal pc a file sent by Oceane

Step 1 - HOX phylogenetic tree
make an alignment using the txt file obtained before and the online tool MAFFT

don't select any additional options
export in fasta format and name the output:

```
hoxes_sequences_MAFFT1.fasta
```

Step 2 - HOX phylogenetic tree
download and install Jalview
open the file:

```
hoxes_sequences_MAFFT1.fasta
```

Cut away all the areas with no conservation. Basically we should maintain only the domain
and save the cutted file as:

```
hoxes_sequences_Jalview.fasta
```

Step 3 - HOX phylogenetic tree
make an alignment again using the file "hoxes_sequences_Jalview.fasta" obtained before and the online tool MAFFT
select the option "L-INS-i"
export in fasta format and name the output:

```
hoxes_sequences_MAFFT2.fasta
```

Step 4 - HOX phylogenetic tree
use Jalview again. Open the file:

```
hoxes_sequences_MAFFT2.fasta
```

No chopping this time, just open the file and export it to fasta as:

```
hoxes_sequences_Jalview2.fasta
```

Step 5 - HOX phylogenetic tree
use the online tool Gblocks.
Load the file "hoxes_sequences_Jalview2.fasta"
select all the 3 options for less stringent
copy paste the result in fasta format as:

```
hoxes_sequences_gblocks.fa
```

or use trimAl
on my personal pc

```
conda activate fasttree_env
```

```
conda install -c bioconda trimal
```

```
trimal -in hoxes_sequences_Jalview2.fasta -out hoxes_sequences_trimal.fa
```

Step 6 - HOX phylogenetic tree
This part on Apocrita

```
scp -i ~/.ssh/id_rsa_apocrita hoxes_sequences_gblocks.fa btx654@login.hpc.qmul.ac.uk:/data/scratch/btx654/HOX_genes
```

raxml.sh

```
#!/bin/bash
#$ -pe smp 5
#$ -l highmem
#$ -l h_vmem=10G
#$ -l h_rt=240:0:0
#$ -cwd
#$ -j y
module load raxml
raxmlHPC -f a -b 476 -p 903 -x 12345 -# autoMRE -m PROTGAMMAAUTO -s hoxes_sequences_gblocks.fa -n hoxes_trial1.tre
```

the final output tree in newick format is:

```
hoxes_trial1.tre
```

```
cd /data/SBCS-MartinDuranLab/03-Giacomo/data/03-other_annotations/paraescarpia/step9
cp /data/SBCS-MartinDuranLab/03-Giacomo/data/03-other_annotations/paraescarpia/step8/annotation_report.xls ./
sort paraescarpia_Panther > panther_sorted
cut -f 1 panther_sorted > IDs_panther
```

```
cut -f 2 annotation_report.xls | tail -n +2 > IDs_all
fgrep -w -v -f IDs_panther IDs_all > IDs_absentPanther ### There are oasisia:8632 osedax:4052 riftia:8182 genes without Panther annotation
awk '{print $0"\t""NO PTHR""\t""NO HIT"}' IDs_absentPanther > PANTHER_nohits
cat panther_sorted PANTHER_nohits | sort -k 1,1 > Panther_sorted_allgenes
# now we need to remove duplicated lines from panther all genes
awk '!a[$1]++' Panther_sorted_allgenes > Panther_sorted_allgenes_noduplicates
## use vim to add a header in Owenia_Panther_sorted_allgenes so that it matches Trinotate file
# #gene_id        transcript_id        sprot_Top_BLASTX_hit        RNAMMER        prot_id        prot_coords        sprot_Top_BLASTP_hit        Pfam        SignalP        TmHMM        eggnog        Kegg        gene_ontology_BLASTX        gene_ontology_BLASTP        gene_ontology_Pfam        transcript        peptide
awk 'FNR == NR { lineno[$1] = NR; next} {print lineno[$1], $0;}' annotation_report.xls Panther_sorted_allgenes_noduplicates | sort -k 1,1n | cut -d' ' -f2- > Panther_sorted_allgenes_rightorder
echo "ID"$'\t'"Panther_1"$'\t'"Panther_2"$'\t'"Panther_3"$'\t'"Panther_4"$'\t'"Panther_5" > header
cat header Panther_sorted_allgenes_rightorder > Panther_sorted_allgenes_rightorder_ok
paste annotation_report.xls Panther_sorted_allgenes_rightorder_ok > paraescarpia_annotation_Jun2021_TrinoPanther.xls
awk '{print $2}' paraescarpia_KAAS_custom_SBH.txt > only_KOnumbers.txt
#add a line at the top of this file saying "KO_number"
nano only_KOnumbers.txt
paste paraescarpia_annotation_Jun2021_TrinoPanther.xls only_KOnumbers.txt > paraescarpia_annotation_Jun2021_TrinoPantherKO.xls
```

fasttree_env
on my personal computer

```
conda create -n fasttree_env
conda activate fasttree_env
conda install -c bioconda fasttree
```

```
FastTree alignment.file > tree_file

```

Step 7 - HOX phylogenetic tree
I have removed many siboglinids sequences from the tree that were clustering in a weird way. Now I can blast the siboglinid Hox genes against the mRNA of the 5 species in order to fill the gaps of the missing hoxes

first I need to obtain the the mRNA of the 5 species

```
scp -i ~/.ssh/id_rsa_apocrita /Volumes/5T\ hard-disk/Data/riftia/Annotation/steps/step1/softmasking/riftia_softmasked.fa btx654@login.hpc.qmul.ac.uk:/data/SBCS-MartinDuranLab/03-Giacomo/data/riftia/annotation/
```

```
cd /data/scratch/btx654/HOX_genes/further
module load anaconda3
source activate augustus

gffread -w riftia_mRNA.fa -g /data/SBCS-MartinDuranLab/03-Giacomo/data/riftia/annotation/riftia_softmasked.fa /data/SBCS-MartinDuranLab/03-Giacomo/data/riftia/annotation/New_annotation_Dec2020/step6/riftia_annotation_v101220.gff3
gffread -w osedax_mRNA.fa -g /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/annotation/softmasking/osedax_softmasked.fa /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/annotation/New_annotation_Dec2020/step6/osedax_annotation_v101220.gff3
```

```
gffread -w oasisia_mRNA.fa -g /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/annotation/new_annotation_Nov2020/step1/softmasking/oasisia_Nov2020_softmasked.fa /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/annotation/New_annotation_Dec2020/step6/oasisia_annotation_v101220.gff3
gffread -w lamellibrachia_mRNA.fa -g /data/SBCS-MartinDuranLab/03-Giacomo/data/03-other_annotations/lamellibrachia/new_multifasta_lamellibrachia.fa /data/SBCS-MartinDuranLab/03-Giacomo/data/03-other_annotations/lamellibrachia/lamellibrachia_lociMerged_longestIsoform.gff
cp /data/SBCS-MartinDuranLab/03-Giacomo/data/03-other_annotations/paraescarpia/step8/paraescarpia_mRNA.fa ./
gffread -y lamellibrachia_proteins.fa -g /data/SBCS-MartinDuranLab/03-Giacomo/data/03-other_annotations/lamellibrachia/new_multifasta_lamellibrachia.fa /data/SBCS-MartinDuranLab/03-Giacomo/data/03-other_annotations/lamellibrachia/lamellibrachia_lociMerged_longestIsoform.gff
cp /data/SBCS-MartinDuranLab/03-Giacomo/data/03-other_annotations/paraescarpia/step8/paraescarpia_proteins.fa ./
```

Then I need to obtain the hox sequences to use for my search:
list

```
FUN_032673-T1
FUN_032670-T1
FUN_032669-T1
FUN_005888-T1
FUN_038049-T1
FUN_038047-T1
FUN_030985-T1
FUN_038044-T1
nbis-mrna-10157
nbis-mrna-10158
```

extract the sequences with seqtk

```
module load seqtk
```

```
seqtk subseq lamellibrachia_proteins.fa list > Hoxes_single_prot.fa
seqtk subseq paraescarpia_proteins.fa list >> Hoxes_single_prot.fa
fold -w 60 Hoxes_single_prot.fa > Hoxes_prot.fa #header should not be longer than 60 charachters
cat /your/path/to/folder/*.fa > newname.fa
```

Make blast databases

tblastn.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further
#$ -o /data/scratch/btx654/HOX_genes/further
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

species=$1
proteins="$species"_proteins.fa
database="$species"_db
output="$species"_tblastn_out

module load blast+
#makedb $mRNA -o $database
makeblastdb -in $mRNA -dbtype nucl -out $database
tblastn -db $database -query Hoxes.fa -out $output -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -outfmt 6
```

to add
FUN_033852-T1 lamellibrachia

Step 8 - HOX phylogenetic tree
what I can do now is re-run single trees for all my species using the Hox genes Oceane sent me and the candidate genes obtained from tblastn one species per time

```
cut -f 2 paraescarpia_tblastn_out | sort | uniq > paraescarpia_candidates_list
sed -i 's/^/Pech_/' paraescarpia_candidates_list
seqtk subseq ../paraescarpia/Pech.fa paraescarpia_candidates_list > paraescarpia_candidates.fa
cut -f 2 riftia_tblastn_out | sort | uniq > riftia_candidates_list
sed -i 's/^/Rpac_/' riftia_candidates_list
seqtk subseq ../riftia/Rpac.fa riftia_candidates_list > riftia_candidates.fa
cut -f 2 osedax_tblastn_out | sort | uniq > osedax_candidates_list
sed -i 's/^/Ofra_/' osedax_candidates_list
seqtk subseq ../osedax/Ofra.fa osedax_candidates_list > osedax_candidates.fa
cut -f 2 oasisia_tblastn_out | sort | uniq > oasisia_candidates_list
sed -i 's/^/Oalv_/' oasisia_candidates_list
seqtk subseq ../oasisia/Oalv.fa oasisia_candidates_list > oasisia_candidates.fa
cut -f 2 lamellibrachia_tblastn_out | sort | uniq > lamellibrachia_candidates_list
sed -i 's/^/Lluy_/' lamellibrachia_candidates_list
seqtk subseq ../lamellibrachia/Lluy.fa lamellibrachia_candidates_list > lamellibrachia_candidates.fa
```

- then copy/paste *_candidates.fa at the end of the file Oceane sent with all the

Hoxes sequences

Now do step 1 to 6 for each single species

```
scp -i ~/.ssh/id_rsa_apocrita -r btx654@login.hpc.qmul.ac.uk:/data/scratch/btx654/HOX_genes/further/Oasisia_lox/*html /Users/giacomo/Desktop/H
```

```
cut -f 2 paraescarpia_tblastn_out | sort | uniq > paraescarpia_candidates_list
sed -i 's/^/Pech_/' paraescarpia_candidates_list
seqtk subseq ../paraescarpia/Pech.fa paraescarpia_candidates_list > paraescarpia_candidates.fa
```

Using Lox2 and 4 from Oasisia to search in lamellibrachia
list

```
Oalv_OALVG00000019412.1
Oalv_OALVG00000019413.1
```

```
seqtk subseq ../../oasisia/Oalv.fa list > Hoxes_prot.fa
```

using owenia hoxes
tblastn_riftia_transcriptomes.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further/owenia_hoxes/transcriptomes_riftia
#$ -o /data/scratch/btx654/HOX_genes/further/owenia_hoxes/transcriptomes_riftia
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem
```

```

module load blast+
makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/riftia/trinity/riftia_crown_trinity/riftia_crown_trinity.Trinity.fasta -dbtype nucl -out riftia_crown
tblastn -db riftia_crown -query ../hox_genes_owenia.fa -out riftia_crown_tblastn_out -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db riftia_crown -query ../hox_genes_owenia.fa -out riftia_crown_tblastn_out.html -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -html

makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/riftia/trinity/riftia_trunk_wall_trinity/riftia_trunk_wall_trinity.Trinity.fasta -dbtype nucl -out riftia_trunkwall
tblastn -db riftia_trunkwall -query ../hox_genes_owenia.fa -out riftia_trunkwall_tblastn_out -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db riftia_trunkwall -query ../hox_genes_owenia.fa -out riftia_trunkwall_tblastn_out.html -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -html
```

tblastn_oasisia_transcriptomes.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further/owenia_hoxes/transcriptomes_oasisia
#$ -o /data/scratch/btx654/HOX_genes/further/owenia_hoxes/transcriptomes_oasisia
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem
```

```

module load blast+
makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/trinity/oasisia_crown_trinity/oasisia_crown_trinity.Trinity.fasta -dbtype nucl -out oasisia_crown
tblastn -db oasisia_crown -query hox_genes_owenia.fa -out oasisia_crown_tblastn_out -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db oasisia_crown -query hox_genes_owenia.fa -out oasisia_crown_tblastn_out.html -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -html

makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/trinity/oasisia_opistosoma_trinity/oasisia_opistosoma_trinity.Trinity.fasta -dbtype nucl -out oasisia_opistosoma
tblastn -db oasisia_opistosoma -query hox_genes_owenia.fa -out oasisia_opistosoma_tblastn_out -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db oasisia_opistosoma -query hox_genes_owenia.fa -out oasisia_opistosoma_tblastn_out.html -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -html

makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/trinity/oasisia_trophosome_trinity/oasisia_trophosome_trinity.Trinity.fasta -dbtype nucl -out oasisia_trophosome
tblastn -db oasisia_trophosome -query hox_genes_owenia.fa -out oasisia_trophosome_tblastn_out -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db oasisia_trophosome -query hox_genes_owenia.fa -out oasisia_trophosome_tblastn_out.html -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -html
```

```
cat *out | cut -f 2 | sort | uniq > oasisia_candidates_list
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/t
```

```
rinity/oasisia_crown_trinity/oasisia_crown_trinity.Trinity.fasta
oasisia_candidates_list > oasisia_crown_candidates.fa
sed -i 's/\s.*$//' oasisia_crown_candidates.fa
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/t
rinity/oasisia_opistosoma_trinity/oasisia_opistosoma_trinity.Tri
nity.fasta oasisia_candidates_list > oasisia_opistosoma_candidat
es.fa
sed -i 's/\s.*$//' oasisia_opistosoma_candidates.fa
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/t
rinity/oasisia_trophosome_trinity/oasisia_trophosome_trinity.Tri
nity.fasta oasisia_candidates_list > oasisia_trophosome_candidat
es.fa
sed -i 's/\s.*$//' oasisia_trophosome_candidates.fa
```

```
module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda
3/seqkit_env
seqkit concat oasisia*candidates.fa > oasisia_concat_candidates.
fa
```

tblastx_oasisia_concat.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further/owenia_hoxes/trans
criptomes_oasisia
#$ -o /data/scratch/btx654/HOX_genes/further/owenia_hoxes/transc
riptomes_oasisia
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

```

```
module load blast+
#makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/trinity/oasisia_crown_trinity/oasisia_crown_trinity.Trinity.fasta -dbtype nucl -out oasisia_crown
tblastx -db ../../oasisia_db -query oasisia_concat_candidates.fa -out oasisia_concat_tblastx_out -max_target_seqs 1 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastx -db ../../oasisia_db -query oasisia_concat_candidates.fa -out oasisia_concat_tblastx_out.html -max_target_seqs 1 -evalue 1e-10 -num_threads 8 -html
```

```
cut -f 2 oasisia_concat_tblastx_out | sort | uniq > oasisia_concat_candidates_IDs
sed -i 's/^/Oalv_/' oasisia_concat_candidates_IDs
seqtk subseq ../../../oasisia/Oalv.fa oasisia_concat_candidates_IDs > oasisia_concat_candidates_OK.fa
```

- from this last fasta I remove manually the sequences of already assigned hox genes

using paraescarpia hoxes
tblastn_oasisia_transcriptomes.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further/paraescarpia_hoxes/transcriptomes_oasisia
#$ -o /data/scratch/btx654/HOX_genes/further/paraescarpia_hoxes/transcriptomes_oasisia
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

module load blast+
```

```
makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/trinity/oasisia_crown_trinity/oasisia_crown_trinity.Trinity.fasta -dbtype nucl -out oasisia_crown
tblastn -db oasisia_crown -query hox_genes_paraescarpia.fa -out oasisia_crown_tblastn_out -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db oasisia_crown -query hox_genes_paraescarpia.fa -out oasisia_crown_tblastn_out.html -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -html

makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/trinity/oasisia_opistosoma_trinity/oasisia_opistosoma_trinity.Trinity.fasta -dbtype nucl -out oasisia_opistosoma
tblastn -db oasisia_opistosoma -query hox_genes_paraescarpia.fa -out oasisia_opistosoma_tblastn_out -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db oasisia_opistosoma -query hox_genes_paraescarpia.fa -out oasisia_opistosoma_tblastn_out.html -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -html

makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/trinity/oasisia_trophosome_trinity/oasisia_trophosome_trinity.Trinity.fasta -dbtype nucl -out oasisia_trophosome
tblastn -db oasisia_trophosome -query hox_genes_paraescarpia.fa -out oasisia_trophosome_tblastn_out -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db oasisia_trophosome -query hox_genes_paraescarpia.fa -out oasisia_trophosome_tblastn_out.html -max_target_seqs 25 -evalue 1e-10 -num_threads 8 -html
```

```
cat *out | cut -f 2 | sort | uniq > oasisia_candidates_list
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/trinity/oasisia_crown_trinity/oasisia_crown_trinity.Trinity.fasta oasisia_candidates_list > oasisia_crown_candidates.fa
```

```
sed -i 's/\s.*$//' oasisia_crown_candidates.fa
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/trinity/oasisia_opistosoma_trinity/oasisia_opistosoma_trinity.Trinity.fasta oasisia_candidates_list > oasisia_opistosoma_candidates.fa
sed -i 's/\s.*$//' oasisia_opistosoma_candidates.fa
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/trinity/oasisia_trophosome_trinity/oasisia_trophosome_trinity.Trinity.fasta oasisia_candidates_list > oasisia_trophosome_candidates.fa
sed -i 's/\s.*$//' oasisia_trophosome_candidates.fa
```

```
module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/seqkit_env
seqkit concat oasisia*candidates.fa > oasisia_concat_candidates.fa
```

tblastx_oasisia_concat.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further/paraescarpia_hoxes/transcriptomes_oasisia
#$ -o /data/scratch/btx654/HOX_genes/further/paraescarpia_hoxes/transcriptomes_oasisia
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

module load blast+
#makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/oasis
```

```
ia/trinity/oasisia_crown_trinity/oasisia_crown_trinity.Trinity.fasta -dbtype nucl -out oasisia_crown
tblastx -db ../../oasisia_db -query oasisia_concat_candidates.fa -out oasisia_concat_tblastx_out -max_target_seqs 1 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastx -db ../../oasisia_db -query oasisia_concat_candidates.fa -out oasisia_concat_tblastx_out.html -max_target_seqs 1 -evalue 1e-10 -num_threads 8 -html
```

```
cut -f 2 oasisia_concat_tblastx_out | sort | uniq > oasisia_concat_candidates_IDs
sed -i 's/^/Oalv_/' oasisia_concat_candidates_IDs
seqtk subseq ../../../oasisia/Oalv.fa oasisia_concat_candidates_IDs > oasisia_concat_candidates_OK.fa
```

- from this last fasta I remove manually the sequences of already assigned hox genes

```
fgrep -w -f list_test lamellibrachia_annotation_Feb2021_TrinoPantherKO_OK.xls > IDs_absentPanther
cut -f 2,20 lamellibrachia_annotation_Feb2021_TrinoPantherKO_OK.xls | fgrep
```

single Hoxes
blast against final assemblies using single hoxes from different species
hox1.fa

```
>Owenia_fusiformis_Hox1
MNSASDYTICNLDNNTYSNNFTTDTAPYSCYANINNAGIESDSYRGGYSENNLSHHHHHHHQHQHPQQLSVEAHLSGTPHNTSLSLNVYQHGTHSYPIQPSPPQGNFYYEDAMIPGGELAECNAYPYPDNSPTSHQIAYQASEHHHIQQQPPQQCDTGQQQQQSPVAQYKWMQVKRNVPKPVTDYKLTDFTYVNPGNNLGRTNFTNKQLTELEKEFHFNKYLTRARRIEIAAALGLNEVQVKIWFQNRRMKQKKRLKENKLAQTVNGEHENGDDLSPGPTTPTSTDEQIS
>Capitella_tellata_Lab
MAGEYTLCNLDNHTYTSPYNGTEGANYNGYTGAEYGVHHGAGPGPPGASLELHSPAGLGYGEAG
```

```
GMCDGEPAHSQAVFLHSDGQGYGAIACGGGSSAPQQHQGYPAPHAGYYGHGMTFNGGIADMPPH
GLHSNGGYLGYPDPTNCNPLLGNPNASNTGYCLSPHEHVGLSSSPGSDQGPVTTYKWMTVKRGT
PKTSKAPGAGDFSVFAGQPNMGRTNFTNKQLTELEKEFHFNKYLTRARRIEIAASLGLNETQVK
IWFQNRRMKQKKRLKENTSTTPVSDSSQDGISGDLNEEAS
>Rpac_RPACG00000019294.1
MTALFGRLRQEQSIHCDLLAAGFRRLPTTAKHSEYTSYSHGAVFGSGGGAGNGTSPGGAVSAAG
QPNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLGLNETQVKIWFQNRRMKQKKRLKEG
HVCGSTNRMNDDAKSDIASLQQTADAIS
>Oalv_OALVG00000019490.1
MNFYAARSRGQWSSPVSNYVSIVSAKHTEYTSYNHGTVYGSAPGAGNGATPGGAVSAAGHPNLG
RTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLGLNETQVKIWFQNRRMKQKKRLKEGHVCGS
TNRMNDDAKSDIASLQQTADAIS
>Pech_nbis-mrna-10151
KHSEYTSYSHGAPYGSGGGVGNGATPGGAVSAVGQPNLGRTNFTNKQLTELEKEFHFNRYLTRA
RRIEIAASLGLNETQVKIWFQNRRMKQKKRLKEGHVCGLTNRMNDDAKSDMASLQQTADAIS
>Lluy_FUN_032673-T1
MNKCPLLSHHSSSPFDGTAGTYDSVTCPVTAGTKRPVRPRRSGREMSNRRFSFASKHSEYTSYN
HAGVYGNGGGGNGATPGGAVSAAGQPNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLG
LNETQVKIWFQNRRMKQKKRLKEGHVCGSTSRMNDDAKSEISSLQHTADAIS
```

tblastn_osedax_assembly.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further/assembly/Hox1
#$ -o /data/scratch/btx654/HOX_genes/further/assembly/Hox1
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

module load blast+
makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax
/haploidization/purge_dups/osedax_purged.fa -dbtype nucl -out os
```

```
edax_assembly
tblastn -db osedax_assembly -query hox1.fa -out osedax_assembly_tblastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db osedax_assembly -query hox1.fa -out osedax_assembly_tblastn_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -html

```

```
>osedax_1_owe
PQPTDFHPGEFGFEQKRTRQTYTRYQTLELEKEFHYNRYLTRRRRIEIAHSLGLSERQIKIWFQNRRMKWKKENNLPKLTGPNGNDQPADSTPV
>osedax_2_owe
RTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHMLTLTERQIKIWFQNRRMKWKK
>osedax_3_owe
RTAYTSAQLVELEKEFHFNRYLCRPRRIEMASLLSLSERQIKIWFQNRRMKFKK
>osedax_4_owe
RTAYTRHQVLELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRRMKWKK
>osedax_5_owe
GRQTYSRYQTLELEKEFQFNHYLTRKRRIEIAHVLCLTERQIKIWFQNRRMKLKKEKQQIKDLNDITRREHDLSPLP
>osedax_1_cap
RNPQPTDFHPGEFGFEQKRTRQTYTRYQTLELEKEFHYNRYLTRRRRIEIAHSLGLSERQIKIWFQNRRMKWKK
>osedax_2_cap
RTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHMLTLTERQIKIWFQNRRMKWKK
>osedax_3_cap
ARTAYTSAQLVELEKEFHFNRYLCRPRRIEMASLLSLSERQIKIWFQNRRMKFKK
>osedax_4_cap
RTAYTRHQVLELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRRMKWKK
>osedax_5_cap
```

```
SSQRRRGRQTYSRYQTLELEKEFQFNHYLTRKRRIEIAHVLCLTERQIKIWFQNRRMKLKK
>osedax_1_rif
RQTYTRYQTLELEKEFHYNRYLTRRRRIEIAHSLGLSERQIKIWFQNRRMKWKKENNLPKLTGP
NGNDQPAD
>osedax_2_rif
RTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHMLTLTERQIKIWFQNRRMKWKK
>osedax_3_rif
RTAYTSAQLVELEKEFHFNRYLCRPRRIEMASLLSLSERQIKIWFQNRRMKFKKEQRGGIGVVG
S
>osedax_4_rif
SNNPRRLRTAYTNTQLLELEKEFHFNKYLCRPRRIEIASTLDLTERQV
>osedax_5_rif
SRTAYTRHQVLELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRRMKWKK
>osedax_1_oas
RTAYTSAQLVELEKEFHFNRYLCRPRRIEMASLLSLSERQIKIWFQNRRMKFKKEQRGGIGVVG
S
>osedax_2_oas
SNNPRRLRTAYTNTQLLELEKEFHFNKYLCRPRRIEIASTLDLTERQV
>osedax_3_oas
RQTYTRYQTLELEKEFHYNRYLTRRRRIEIAHSLGLSERQIKIWFQNRRMKWKKENNLPKLTGP
NGNDQPAD
>osedax_4_oas
RTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHMLTLTERQIKIWFQNRRMKWKK
>osedax_5_oas
RTAYTRHQVLELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRRMKWKK
>osedax_1_par
RQTYTRYQTLELEKEFHYNRYLTRRRRIEIAHSLGLSERQIKIWFQNRRMKWKKENNLPKLTGP
NGNDQPAD
>osedax_2_par
RTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHMLTLTERQIKIWFQNRRMKWKK
>osedax_3_par
```

```
RTAYTSAQLVELEKEFHFNRYLCRPRRIEMASLLSLSERQIKIWFQNRRMKFKKEQRGG
>osedax_4_par
PRRLRTAYTNTQLLELEKEFHFNKYLCRPRRIEIASTLDLTERQV
>osedax_5_par
RTAYTRHQVLELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRRMKWKK
>osedax_1_lam
RQTYTRYQTLELEKEFHYNRYLTRRRRIEIAHSLGLSERQIKIWFQNRRMKWKK
>osedax_2_lam
RTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHMLTLTERQIKIWFQNRRMKWKK
>osedax_3_lam
RTAYTSAQLVELEKEFHFNRYLCRPRRIEMASLLSLSERQIKIWFQNRRMKFKKEQRGGIGVVG
S
>osedax_4_lam
SNNPRRLRTAYTNTQLLELEKEFHFNKYLCRPRRIEIASTLDLTERQV
>osedax_5_lam
RTAYTRHQVLELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRRMKWKK
```

tblastn_osedax_transcriptome.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further/assembly/Hox1
#$ -o /data/scratch/btx654/HOX_genes/further/assembly/Hox1
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

cp /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/trinity/osed
ax_*/*.Trinity.fasta ./
cat *.Trinity.fasta > combined.trinity.fasta
rm *.Trinity.fasta
```

```
module load blast+
makeblastdb -in combined.trinity.fasta -dbtype nucl -out osedax_transcriptome
tblastn -db osedax_transcriptome -query hox1.fa -out osedax_transcriptome_tblastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db osedax_transcriptome -query hox1.fa -out osedax_transcriptome_tblastn_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -html
```

```
>osedax_1_owe
VATYKWMTVKRNAPKTVKQTPQSSDYNGNSSTTSGACCASGSHFRSSPLSPSHSPSSIGSGCGGGNLSGNGLPPNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGHAAQWTVDTE
>osedax_2_owe
NLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGHAAQ
>osedax_3_owe
NLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGHAAQWTVDTE
>osedax_4_owe
GSANFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKE
>osedax_5_owe
RARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKE
>osedax_1_cap
VATYKWMTVKRNAPKTVKQTPQSSDYNGNSSTTSGACCASGSHFRSSPLSPSHSPSSIGSGCGGGNLSGNGLPPNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKE
>osedax_2_cap
GNLSGNGLPPNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKE
>osedax_3_cap
PNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKE
```

```
>osedax_4_cap
GSANFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKE
>osedax_5_cap
RARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKE
>osedax_1_rif
PNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGH
AAQWT
>osedax_2_rif
PNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGH
A
>osedax_3_rif
PNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRL
>osedax_4_rif
GSANFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGH
>osedax_5_rif
RARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGH
>osedax_1_oas
PNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGH
A
>osedax_2_oas
PNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGH
A
>osedax_3_oas
PNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGH
A
>osedax_4_oas
GSANFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGH
>osedax_5_oas
RARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGH
>osedax_1_par
PNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGH
```

```
AAQWT
>osedax_2_par
PNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGH
AAQWT
>osedax_3_par
PNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGH
AAQWT
>osedax_4_par
GSANFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGHDVL
K*RRPCALTRR
>osedax_5_par
RTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGHAAQW
T
>osedax_1_lam
PNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGH
AAQWT
>osedax_2_lam
RTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGHAAQW
T
>osedax_3_lam
GSANFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMKQKKRLKEGHAAQ
WT
>osedax_4_lam
RTAYTSAQLVELEKEFHFNRYLCRPRRIEMASLLSLSERQIKIWFQNRRMKFKKEQRGGIGVVG
S
>osedax_5_lam
RTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHMLTLTERQIKIWFQNRRMKWKK
```

lox2.fa

```
>Owenia_fusiformis_Lox2
MSSYFPQGQAGDMGPHDAGSSAVSEGSFNRESCTSTDFKSPGYVAPGSNYNDFTCRMPAAFQSR
TGEFRNGLPNNNFLASQYGQSGLGGQGFADIDCGLGTATSLSHCSPVSPPPRVTPFYPWMSIVG
PNSNQRRRGRQTYTRFQTLELEKEFKFNRYLTRRRRIELSHMLCLTERQIKIWFQNRR
```

```
>Capitella_tellata_Lox2
MSYFNSESSTRLNGPSSVEDAGGCSLTPTVDAGLSTPPSSRLSEPGQTPTPADSVRIVSSQGYV
PSQTHFQDYHCGTGSGISRMYESYNQHVHSNNNYLYNASQGGHLAAAAAAAAAGGSGQPYMDLS
VPLNCMPGSGGIPGCGGRMPGAGMNGPMYPWMSIVGPNSNQRRRGRQTYTRYQTLELEKEFKFN
RYLTRRRRIELSHMLCLTERQIKIWFQNRRMKEKKEIQAIKELNEKEKTKPNSVPNPTTVVD
>Oalv_OALVG00000019413.1
LPLALPRSYPSGITATAVAVAVAATATRREMSSFFSDHDRVEHQLRLARHAVGATAGDGPGLLG
GADHMTSVRPPDDCGSNCAMNCRPTSGLCAPVSTFQDFACSIPVFQSRQASADYNGGYLYSQPP
PSSSASLSSSMTGAITAPTAHESQCHPGGQHSQTLLESHSGLAGINCGALGAANINNCGARMPP
PQSLTAPVMYPWMSIVGPNSNQRRRGRQTYTRYQTLELEKEFKFNRYLTRRRRIELSHMLCLTE
RQIKIWFQNRRMKEKKEIQAIKELNDKEKAKTTSATVMPSAK
>Pech_nbis-mrna-10158
PNSNQRRRGRQTYTRYQTLELEKEFKFNRYLTRRRRIELSHMLCLTERQIKIWFQNRRMKEKKE
IQAIKELNEKEKMKGTSTTVLPTAK
```

tblastn_osedax_assembly.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further/assembly/Lox2
#$ -o /data/scratch/btx654/HOX_genes/further/assembly/Lox2
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

module load blast+
makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax
/haploidization/purge_dups/osedax_purged.fa -dbtype nucl -out os
edax_assembly
tblastn -db osedax_assembly -query lox2.fa -out osedax_assembly_
tblastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -out
fmt 6
tblastn -db osedax_assembly -query lox2.fa -out osedax_assembly_
```

```
tblastn_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8
-html
```

```
>osedax_1_owe
YSLLLNVGPNSSQRRRGRQTYSRYQTLELEKEFQFNHYLTRKRRIEIAHVLCLTERQIKIWFQN
RR
>osedax_2_owe
PQPTDFHPGEFGFEQKRTRQTYTRYQTLELEKEFHYNRYLTRRRRIEIAHSLGLSERQIKIWFQ
NRR
>osedax_3_owe
KRTRTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHMLTLTERQIKIWFQNRR
>osedax_4_owe
GFNGVDSKRSRTAYTRHQVLELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRR
>osedax_5_owe
KRARTAYTSAQLVELEKEFHFNRYLCRPRRIEMASLLSLSERQIKIWFQNRR
>osedax_6_owe
LSAESARQRKKRKPYTRYQTIMLEEEFKRNSYITRQKRWEISCKLQLSERQVKVWFQNRR
>osedax_1_cap
YSLLLNVGPNSSQRRRGRQTYSRYQTLELEKEFQFNHYLTRKRRIEIAHVLCLTERQIKIWFQN
RRMKLKKEKQQIKDLNDITRREHDLSPLP
>osedax_1_cap
LELEKEFHFNRYLTRRRRIEIAHMLTLTERQIKIWFQNRRMKWKKEHKA
>osedax_2_cap
LELEKEFHYNRYLTRRRRIEIAHSLGLSERQIKIWFQNRRMKWKKE
>osedax_3_cap
LELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRRMKWKKD
>osedax_4_cap
VELEKEFHFNRYLCRPRRIEMASLLSLSERQIKIWFQNRRMKFKKE
>osedax_1_oas
YSLLLNVGPNSSQRRRGRQTYSRYQTLELEKEFQFNHYLTRKRRIEIAHVLCLTERQIKIWFQN
RRMKLKKEKQQIKDLND
```

```
>osedax_2_oas
LELEKEFHFNRYLTRRRRIEIAHMLTLTERQIKIWFQNRRMKWKKEHKA
>osedax_3_oas
LELEKEFHYNRYLTRRRRIEIAHSLGLSERQIKIWFQNRRMKWKKE
>osedax_4_oas
LELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRRMKWKKD
>osedax_5_oas
VELEKEFHFNRYLCRPRRIEMASLLSLSERQIKIWFQNRRMKFKKE
>osedax_1_par
PNSSQRRRGRQTYSRYQTLELEKEFQFNHYLTRKRRIEIAHVLCLTERQIKIWFQNRRMKLKKE
KQQIKDLND
>osedax_2_par
LELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRRMKWKKD
>osedax_3_par
LELEKEFHFNRYLTRRRRIEIAHMLTLTERQIKIWFQNRRMKWKKEHKA
>osedax_4_par
LELEKEFHYNRYLTRRRRIEIAHSLGLSERQIKIWFQNRRMKWKKE
>osedax_5_par
VELEKEFHFNRYLCRPRRIEMASLLSLSERQIKIWFQNRRMKFKKE
```

tblastn_osedax_transcriptome.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further/assembly/Lox2
#$ -o /data/scratch/btx654/HOX_genes/further/assembly/Lox2
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

cp /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/trinity/osed
```

```
ax_*/*.Trinity.fasta ./
cat *.Trinity.fasta > combined.trinity.fasta
rm *.Trinity.fasta
module load blast+
makeblastdb -in combined.trinity.fasta -dbtype nucl -out osedax_transcriptome
tblastn -db osedax_transcriptome -query lox2.fa -out osedax_transcriptome_tblastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db osedax_transcriptome -query lox2.fa -out osedax_transcriptome_tblastn_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -html
```

```
>osedax_1_owe
ASEPPSPNVMYPWMSIVGPNSNQRRRGRQTYTRYQTLELEKEFKYNRYLTRRRRIELSHTLCLTERQIKIWFQNRR
>osedax_2_owe
GRQTYTRYQTLELEKEFKYNRYLTRRRRIELSHTLCLTERQIKIWFQNRR
>osedax_1_cap
YPWMSIVGPNSNQRRRGRQTYTRYQTLELEKEFKYNRYLTRRRRIELSHTLCLTERQIKIWFQNRRMKEKKEIQAIKELNAKEQ
>osedax_2_cap
LELEKEFKYNRYLTRRRRIELSHTLCLTERQIKIWFQNRRMKEKKEIQAIKELNAKEQ
>osedax_1_oas
NSDNGMASEPPSPNVMYPWMSIVGPNSNQRRRGRQTYTRYQTLELEKEFKYNRYLTRRRRIELSHTLCLTERQIKIWFQNRRMKEKKEIQAIKELNAKEQQST
>osedax_2_oas
WYLV*VGPNSNQRRRGRQTYTRYQTLELEKEFKYNRYLTRRRRIELSHTLCLTERQIKIWFQNRRMKEKKEIQAIKELNAKEQQST
>osedax_1_par
PNSNQRRRGRQTYTRYQTLELEKEFKYNRYLTRRRRIELSHTLCLTERQIKIWFQNRRMKEKKEIQAIKELNAKEQ
```

```
>osedax_2_par
LELEKEFKYNRYLTRRRRIELSHTLCLTERQIKIWFQNRRMKEKKEIQAIKELNAKEQ
```

antp.fa

```
>Owenia_fusiformis_Antp
MSYYHNGSYMTTEHFAGHNSPMTTNYQNSPRTATIYDDTSQPAYPRFPPYDRLDIRPIQSNQQP
QGGYYNQNTIARDNRDDYSHTNGQQLSPIQRYSSCKISDDTVESYLAGAVADHTTPLYENNNSP
PLQTSISPPQPQAPQTETPNQNQSQNQNSTQQQIPIYPWMRSQFGPDRKRGRQTYTRFQTLELE
KEFHFNKYLTRRRRIEIAHSLCLTERQIKIWFQNRRMKWKKENKQIEALKSPESDEKSEPPSPS
SPTEDDDELKKEKEDDDGDKPTTPDTL
>Capitella_tellata_Antp
MNNWTKTKMLRIQTKLSKKNQLKAGPERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHA
LCLTERQIKIWFQNRRMKWKKENRQIEVLRQHTDDDLDFR
```

tblastn_osedax_both.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further/assembly/Antp/osedax
#$ -o /data/scratch/btx654/HOX_genes/further/assembly/Antp/osedax
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

module load blast+
makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/haploidization/purge_dups/osedax_purged.fa -dbtype nucl -out osedax_assembly
tblastn -db osedax_assembly -query antp.fa -out osedax_assembly_tblastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
```

```
tblastn -db osedax_assembly -query antp.fa -out osedax_assembly_tblastn_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -html

cp /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/trinity/osedax_*/*.Trinity.fasta ./
cat *.Trinity.fasta > combined.trinity.fasta
rm *.Trinity.fasta
makeblastdb -in combined.trinity.fasta -dbtype nucl -out osedax_transcriptome
tblastn -db osedax_transcriptome -query antp.fa -out osedax_transcriptome_tblastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db osedax_transcriptome -query antp.fa -out osedax_transcriptome_tblastn_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -html
```

```
>osedax_1_assembly_owe
GEFGFEQKRTRQTYTRYQTLELEKEFHYNRYLTRRRRIEIAHSLGLSERQIKIWFQNRRMKWKKEN
>osedax_2_assembly_owe
GIDSKRTRTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHMLTLTERQIKIWFQNRRMKWKKEHK
>osedax_3_assembly_owe
GVDSKRSRTAYTRHQVLELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRRMKWKKDHK
>osedax_1_assembly_cap
GFEQKRTRQTYTRYQTLELEKEFHYNRYLTRRRRIEIAHSLGLSERQIKIWFQNRRMKWKKENNLPKLTGPNGND
>osedax_2_assembly_cap
GIDSKRTRTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHMLTLTERQIKIWFQNRRMKWKKEHKAKNQISLLGSHK*NEKSF
>osedax_3_assembly_cap
LNVGPNSSQRRRGRQTYSRYQTLELEKEFQFNHYLTRKRRIEIAHVLCLTERQIKIWFQNRRMKLKKEKQQIKDL
```

```
>osedax_1_transcriptome_owe
PFYPWM-GVVGPNSSQRRRGRQTYSRYQTLELEKEFQFNHYLTRKRRIEIAHVLCLTERQIKIW
FQNRRMKLKKEKQQIKDL
>osedax_2_transcriptome_owe
PIFPWMRRMHLDGIDGIDSKRTRTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHMLTLTERQIK
IWFQNRRMKWKKEHK
>osedax_3_transcriptome_owe
IFPWMKKVHNGTSNGGFNGVDSKRSRTAYTRHQVLELEKEFHFNRYLTRRRRIEIAHTLCLTER
QIKIWFQNRRMKWKKDHK
>osedax_1_transcriptome_cap
GPNSNQRRRGRQTYTRYQTLELEKEFKYNRYLTRRRRIELSHTLCLTERQIKIWFQNRRMKEKK
EIQAIKEL
>osedax_2_transcriptome_cap
GPNSNQRRRGRQTYTRYQTLELEKEFKYNRYLTRRRRIELSHTLCLTERQIKIWFQNRRMKEKK
EIQAIKEL
>osedax_3_transcriptome_cap
GPNSNQRRRGRQTYTRYQTLELEKEFKYNRYLTRRRRIELSHTLCLTERQIKIWFQNRRMKEKK
EIQAIKEL
```

check hoxes found yesterday in osedax in its genome:
osedax.fa

```
>osedax_Lox2
ASEPPSPNVMYPWMSIVGPNSNQRRRGRQTYTRYQTLELEKEFKYNRYLTRRRRIELSHTLCLT
ERQIKIWFQNRR
>osedax_Hox1
VATYKWMTVKRNAPKTVKQTPQSSDYNGNSSTTSGACCASGSHFRSSPLSPSHSPSSIGSGCGG
GNLSGNGLPPNLGRTNFTNKQLTELEKEFHFNRYLTRARRIEIAASLCLNETQVKIWFQNRRMK
QKKRLKEGHAAQWTVDTE
>osedax_Hox4
RTAYTRHQVLELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRRMKWKK
```

osedax_nucl.fa
used the previous sequences and translated them into nucleotide with this online tool

```
>osedax_Lox2
GCCAGCGAGCCCCCCAGCCCCAACGTGATGTACCCCTGGATGAGCATCGTGGGCCCCAAC
AGCAACCAGAGGAGGAGGGGCAGGCAGACCTACACCAGGTACCAGACCCTGGAGCTGGAG
AAGGAGTTCAAGTACAACAGGTACCTGACCAGGAGGAGGAGGATCGAGCTGAGCCACACC
CTGTGCCTGACCGAGAGGCAGATCAAGATCTGGTTCCAGAACAGGAGG
>osedax_Hox1
GTGGCCACCTACAAGTGGATGACCGTGAAGAGGAACGCCCCCAAGACCGTGAAGCAGACC
CCCCAGAGCAGCGACTACAACGGCAACAGCAGCACCACCAGCGGCGCCTGCTGCGCCAGC
GGCAGCCACTTCAGGAGCAGCCCCCTGAGCCCCAGCCACAGCCCCAGCAGCATCGGCAGC
GGCTGCGGCGGCGGCAACCTGAGCGGCAACGGCCTGCCCCCCAACCTGGGCAGGACCAAC
TTCACCAACAAGCAGCTGACCGAGCTGGAGAAGGAGTTCCACTTCAACAGGTACCTGACC
AGGGCCAGGAGGATCGAGATCGCCGCCAGCCTGTGCCTGAACGAGACCCAGGTGAAGATC
TGGTTCCAGAACAGGAGGATGAAGCAGAAGAAGAGGCTGAAGGAGGGCCACGCCGCCCAG
TGGACCGTGGACACCGAG
>osedax_Hox4
AGGACCGCCTACACCAGGCACCAGGTGCTGGAGCTGGAGAAGGAGTTCCACTTCAACAGG
TACCTGACCAGGAGGAGGAGGATCGAGATCGCCCACACCCTGTGCCTGACCGAGAGGCAG
ATCAAGATCTGGTTCCAGAACAGGAGGATGAAGTGGAAGAAG
```

check_osedax_assembly.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further/assembly/check/osedax
#$ -o /data/scratch/btx654/HOX_genes/further/assembly/check/osedax
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

```

```
module load blast+
makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/haploidization/purge_dups/osedax_purged.fa -dbtype nucl -out osedax_assembly
tblastn -db osedax_assembly -query osedax.fa -out osedax_assembly_tblastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db osedax_assembly -query osedax.fa -out osedax_assembly_tblastn_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -html

blastn -db osedax_assembly -query osedax_nucl.fa -out osedax_assembly_blastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
blastn -db osedax_assembly -query osedax_nucl.fa -out osedax_assembly_blastn_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -html
```

```
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/trinity/osedax_*/*.Trinity.fasta list_transcript > list_transcript.fa
```

check_osedax_blastn.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further/assembly/check/osedax
#$ -o /data/scratch/btx654/HOX_genes/further/assembly/check/osedax
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem
```

```

module load blast+

blastn -db osedax_assembly -query list_transcript.fa -out osedax_assembly_blastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
blastn -db osedax_assembly -query list_transcript.fa -out osedax_assembly_blastn_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -html
```

hox5.fa

```
>Owenia_fusiformis_Hox5
MSLYSLKSPAAYNSFMSDSGGGGHREFTHSENPYRAYTSGYPYTSHTAAPSGSTHQNGTPTDYSSFSNPATQRLIHPSYNREDSPTPVNNNKPMPAATTSVITSTSPQDYSIKSRTTTEFATTKSSSQITDRDSAVDSPSPTPGSVGSPVSPGQPNKDNDKYVKEEEDGSDREDRDGEGGADNPNIQIYPWMRRVHLGHDQNGAETKRTRTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHSLNLTERQIKIWFQNRRMKWKKEHKLAHLAKSQAKMLDLALAQRAAEAKMHHHAAHGHTLHL

```

tblastn_osedax_assembly.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/HOX_genes/further/assembly/Hox1
#$ -o /data/scratch/btx654/HOX_genes/further/assembly/Hox1
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

module load blast+
makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/haploidization/purge_dups/osedax_purged.fa -dbtype nucl -out os
```

```
edax_assembly
tblastn -db osedax_assembly -query hox1.fa -out osedax_assembly_tblastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db osedax_assembly -query hox1.fa -out osedax_assembly_tblastn_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -html
```

Hox result table

GENE	Osedax	Riftia	Oasisia	Paraescarpia	Lamellibrachia
HOX1	signal found in the trinity transcriptome VATYKWMTVKRNAPKTVKQTPQSSDYNGNSSTTSGACCASGSHFRSSPLSPSHSPSSIGSGCGGGNLSGNGLPPNLGRTNFTNKQLTELEKEFHFNRYLTRARIEIAASLC	Rpac_RPACG00000019294.1/66-123	Oalv_OALVG00000019490.1/61-118	Pech_nbis-mrna-10151/36-93	Lluy_FUN_032673-T1/90-147

	LNETQVKI WFQNRRM KQKKRLKE GHAAQWT VDTE TRINITY_D N22632_c0 _g1_i6 TRINITY_D N16248_c0 _g1_i5 no traces in the assembly				
HOX2	Ofra_OFRA G0000001 1248.1/15 5-212	signal found in the trinity transcripto me YSSGAVIN MCGPVAP PPASGGG GLTTPGH PRRLRTAY TNTQLLEL EKEFHFNK YLCRPRRI EIAASLDL TERQVKV WFQNRR	Oalv_OALV G0000001 9775.1/14 0-197	Pech_nbis-mrna-1015 3/29-86	Lluy_FUN_0 32670-T1/134-19 1

		MKFKRQT QPKSSDG VAMPGDD DFGSPAID STTVSDD SHSPLGVS SVGDKDA PSGDTCG DDCDKTP GVKSGND RSDAADS VSLGQRS MTDVDDS AIKCEDMS RGKAPSV DASPSPTD AVPFSNPL ARCDSRV DAQREFN MSQHMG APLQPPVL THLTGSD GMQTVRN VCHPYVTP GNVDTSL ANPSRPM LHGPPSLP HSEQNAR TSFPPTGH TSSLHPAD VPARQAM RGPSLPQT IYPPTAGIH RLAAPHTK			

		HNSYLGS ATSRRHAP YIDISNMS ASDDRRM DNYFPSN GSDTCVG STSQYGV TPVYVRH DMQYARD HRPPQQQ SCYGDVT QQQNFTR NQDMLNV PFQNNVN CMPFSQP ATGDSHA YTNLAGC YSQPGMT NNNYYQG AAQYGAD TGSIGGCE TDYTSGY DGRYMTS HINSAETT PPDGDGV SSSFPSLS EFCQITNY NYL >TRINITY_ DN39953_c 0_g1_i1			

		>TRINITY_DN28744_c0_g1_i1 no traces in the assembly			
HOX3	Ofra_OFRAG00000011246.1/109-166	Rpac_RPACG00000019878.1/239-296	Oalv_OALVG00000019773.1/208-265	nbis_mrna_10152	Lluy_FUN_032669-T1/164-221
HOX4	signal found in the haploid assembly RTAYTRHQVLELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRRMKWKK tig00004046|arrow|arrow|pilon|pilon - 108:269	signal found in the haploid assembly AFVNTV*WRGCVSVSGSTGSFNGDNKRTRTAYTRHQVLELEKEFHFNRYLTRRRRIEIAHTLCLTERQIKIWFQNRRMKWKK >tig00021207_pech	Oalv_OALVG00000019768.1/181-238	Pech_nbis-mrna-10154/167-224	Lluy_FUN_005888-T1/165-222
HOX5	Ofra_OFRA	signal found in	signal found in	Pech_nbis-mrna-1015	Lluy_FUN_0

	G00000000993.1/119-176	the trinity transcripto me RRMHLGHDGVNGVETKRTRTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHMLNLTERQIKIWFQNRRMKWKKEHKMAHLAKAQAQKLETQMHVGSADMTRKS >TRINITY_DN18618_c0_g2_i2 >TRINITY_DN18618_c0_g1_i1 >TRINITY_DN18618_c0_g2_i1 no traces in the assembly	the transcripto me RRMHLGHDGVNGVETKRTRTSYTRHQTLELEKEFHFNRYLTRRRRIEIAHMLNLTERQIKIWFQNRRMKWKKEHKMAHLAKAQAQKLETQLHVGTADMTRKS TRINITY_DN28412_c0_g1_i1 TRINITY_DN15117_c0_g3_i1 TRINITY_DN9861_c0_g1_i2 TRINITY_DN9861_c0_g1_i1	6/22-79	38049-T1/12-69

			no traces in the assembly		
Lox2	signal found in the trinity transcriptome ASEPPSPNVMYPWMSIVGPNSNQRRRGRQTYTRYQTLELEKEFKYNRYLTRRRRIELSHTLCLTERQIKIWFQNRR TRINITY_DN2605_c0_g1_i5 tig00000951|arrow|arrow|pilon|pilon	signal found in the trinity transcriptome PNSNQRRRGRQTYTRYQTLELEKEFKFNRYLTRRRRIELSHMLCLTERQIKIWFQNRRMKEKKEIQAIKELNEKEKTKGTPTTVMPTAK >TRINITY_DN72223_c0_g1_i1 no traces in the assembly	Oalv_OALVG00000019413.1/215-272	Pech_nbis-mrna-10158/6-63	FUN_038045-T1 ? very high blast hit with Oasisia but not clustering well
Lox4	Ofra_OFRAG00000011363.1/62-	signal found in the trinity transcripto	Oalv_OALVG00000019412.1/17	Pech_nbis-mrna-10157/6-63	FUN_038046-T1 ? very high

	119	me PFYPWMG VVGPNSS QRRRGRQ TYSRYQTL ELEKEFQF NHYLTRKR RIEIAHALC LTERQIKI WFQNRR >TRINITY_ DN726_c1_ g2_i3 no traces in the assembly	5-232		blast hit with Oasisia but not clustering well
Lox5	Ofra_OFRA G0000000 0992.1/10- 67	signal found in the trinity transcripto me GYEQKRT RQTYTRY QTLELEKE FHYNRYLT RRRRIEIAH ALGLSER QIKIWFQN RRMKWKK ENNLSKLT	Oalv_OALV G0000001 9872.1/18 8-245	Pech_nbis- mrna-1015 5/184-241	Lluy_FUN_0 38047- T1/185-24 2

		GPNGNDQPVESTNGSVE >TRINITY_DN38983_c0_g1_i1 >TRINITY_DN2094_c1_g1_i2 no traces in the assembly			
Post1	Ofra_OFRAG00000009996.1/48-112	signal found in the haploid assembly ESDPFRVRKPRHVLARPTPLRPRKKRKPYTKEQISDLEQEYLDTTYITRPKRTEIAKRLHLTERQVKIWFQNRRMKEKKTNNKNVNLFEI	signal found in the haploid assembly EPDPFRARKPRHILARPTPLRPRKKRKPYTKEQISELEQEYLDTTYITRPKRTEIAKRLHLTERQVKIWFQNRRMKEKKTNNKNINLYEV	Pech_nbis-mrna-9378/22-79	Lluy_FUN_030985-T1/70-127

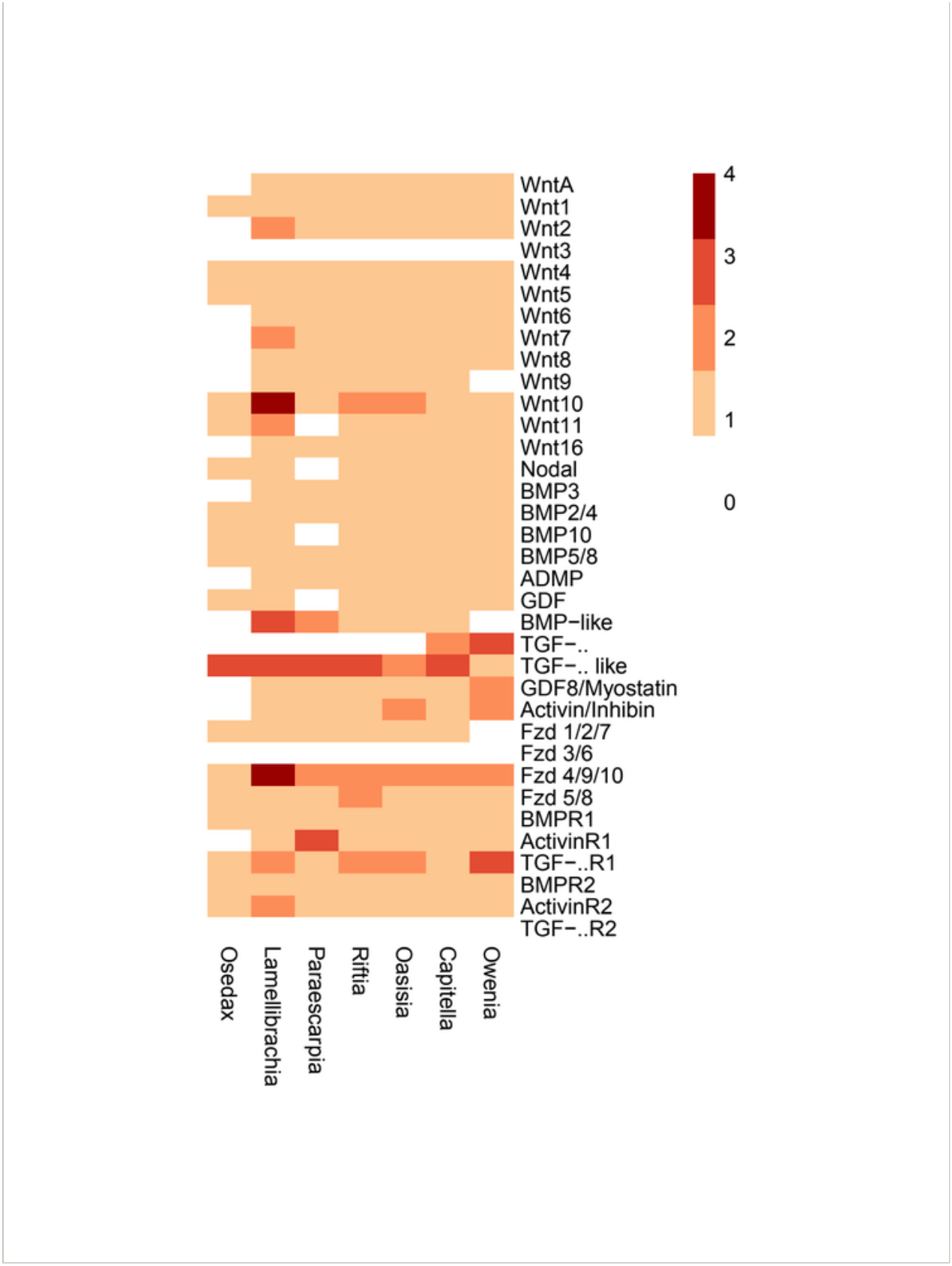
		>tig00035475_pech	>tig00001829 - 619630:619899		
Post2	Ofra_OFRAG00000014621.1/332-389	MISSING checked the assembly and the transcripto me	Oalv_OALVG00000019415.1/172-229	Pech_nbis-mrna-10159/1-45	Lluy_FUN_038044-T1/172-229
Antp	No Antp detected both in the haploid assembly and in the transcripto me	No Antp detected both in the haploid assembly and in the transcripto me	No Antp detected both in the haploid assembly and in the transcripto me	No Antp detected in non redundant proteome and mRNA	No Antp detected in non redundant proteome and mRNA

Plots

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(data.table)
library(gplots)
library(pheatmap)
```

```
library(dendextend)
library(factoextra)
library(ComplexHeatmap)
library(RColorBrewer)
library(NbClust)
library(scales)

#Import data in matrix format
pathways <- read.delim("~/Desktop/pathways_updated.txt", row.names=1)

# Option 1. 0 to 1 relative abundance/expression (54 is the highest value in my dataset)

heatmap_color <- colorRampPalette(brewer.pal(n = 7, name = "OrRd"))(6)
heatmap_color[1] <- rgb(1,1,1)

# If you want common scale for different heatmaps:
# First define some "breaks"

pheatmap(pathways,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         cellheight = 10,
         cellwidth = 20)

```

- exported in PDF 8x6 inches



PDF Dev_pathways • PDF document

Summary plot

sort.sh

```
cut -f1 Developmental_pathways_table_unique_ok > header
cut -f2 Developmental_pathways_table_unique_ok > riftia
cut -f3 Developmental_pathways_table_unique_ok > oasisia
cut -f4 Developmental_pathways_table_unique_ok > osedax
cut -f5 Developmental_pathways_table_unique_ok > lamellibrachia
cut -f6 Developmental_pathways_table_unique_ok > owenia
cut -f7 Developmental_pathways_table_unique_ok > capitella
cut -f8 Developmental_pathways_table_unique_ok > paraescarpia

paste header osedax lamellibrachia paraescarpia riftia oasisia c
apitella owenia > Developmental_pathways_table_unique_perfect
rm Developmental_pathways_table_unique_ok
rm header
rm osedax
rm lamellibrachia
rm paraescarpia
rm riftia
rm oasisia
rm capitella
rm owenia
```

Now in my folder I have only 7 files each containing a set of genes without header (eg Wnt genes)

sum.sh

```
input=$1
cut -f 2 $input | paste -sd+ - | bc > temp1
cut -f 3 $input | paste -sd+ - | bc > temp2
cut -f 4 $input | paste -sd+ - | bc > temp3
cut -f 5 $input | paste -sd+ - | bc > temp4
```

```
cut -f 6 $input | paste -sd+ - | bc > temp5
cut -f 7 $input | paste -sd+ - | bc > temp6
cut -f 8 $input | paste -sd+ - | bc > temp7
paste temp1 temp2 temp3 temp4 temp5 temp6 temp7 > "$input"_sum
rm temp*
```

tot number of genes per each group:

```
      12 BMPlig.txt
       6 BMPrec.txt
       4 Fzd.txt
       6 Ligands.txt
       7 Receptors.txt
      18 TranscriptionFactors.txt
      13 Wnt.txt
```

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(data.table)
library(gplots)
library(pheatmap)
library(dendextend)
library(factoextra)
library(ComplexHeatmap)
library(RColorBrewer)
library(NbClust)
library(scales)

#Import data in matrix format
first_half_log <- read.delim("~/Desktop/Developmental_pathways/n
```

```
o isoforms/first_half_log_Paraescarpia_STAT=-inf", row.names=1)
second_half_log_signal <- read.delim("~/Desktop/Developmental_pathways/no isoforms/second_half_log_signal", row.names=1)
second_half_log_receptor <- read.delim("~/Desktop/Developmental_pathways/no isoforms/second_half_log_receptor", row.names=1)
# Option 1. 0 to 1 relative abundance/expression (54 is the highest value in my dataset)

heatmap_color <- colorRampPalette(brewer.pal(n = 7, name = "Reds"))(1000)

# If you want common scale for different heatmaps:
# First define some "breaks"

pheatmap(first_half_log,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         cellheight = 10,
         cellwidth = 20)

pheatmap(second_half_log_signal,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         cellheight = 10,
         cellwidth = 20)

pheatmap(second_half_log_receptor,
```

```
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          border_color = NA,
          color = heatmap_color,
          cellheight = 10,
          cellwidth = 20)
```

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(data.table)
library(gplots)
library(pheatmap)
library(dendextend)
library(factoextra)
library(ComplexHeatmap)
library(RColorBrewer)
library(NbClust)
library(scales)

#Import data in matrix format
A <- read.delim("/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/04-Losses/03-DevelopmentalPathways/Summary/Wnt.txt_sum", row.names=1)
B <- read.delim("/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/04-Losses/03-DevelopmentalPathways/Summary/Fzd.txt_sum", row.names=1)
C <- read.delim("/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/04-Losses/03-DevelopmentalPathways/Summary/BMPlig.txt_sum", row.names=1)
D <- read.delim("/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/04-Losses/03-DevelopmentalPathways/Summary/BMPrec.txt_su
```

```
m", row.names=1)
E <- read.delim("/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/04-Losses/03-DevelopmentalPathways/Summary/TranscriptionFactors.txt_sum", row.names=1)
F <- read.delim("/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/04-Losses/03-DevelopmentalPathways/Summary/Ligands.txt_sum", row.names=1)
G <- read.delim("/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/04-Losses/03-DevelopmentalPathways/Summary/Receptors.txt_sum", row.names=1)

# Option 1. 0 to 1 relative abundance/expression (54 is the highest value in my dataset)
rescale_custom <- function(x) (x/18)
A_normalised <- t(apply(A, 1, rescale_custom))
rescale_custom <- function(x) (x/6)
B_normalised <- t(apply(B, 1, rescale_custom))
rescale_custom <- function(x) (x/14)
C_normalised <- t(apply(C, 1, rescale_custom))
rescale_custom <- function(x) (x/7)
D_normalised <- t(apply(D, 1, rescale_custom))
rescale_custom <- function(x) (x/920)
E_normalised <- t(apply(E, 1, rescale_custom))
rescale_custom <- function(x) (x/50)
F_normalised <- t(apply(F, 1, rescale_custom))
rescale_custom <- function(x) (x/66)
G_normalised <- t(apply(G, 1, rescale_custom))

# To make 0 a different colour
# First create whatever gradient (e.g. RdBu)
heatmap_color <- colorRampPalette(brewer.pal(n = 7, name = "Reds"))(1000)
```

```
heatmap_color[1] <- rgb(1,1,1)
#column_labels = c("your","labels"),
#row_labels = c("your","labels"))

paletteLength <- 1000
# to go from 0 to max.value (e.g. 1):
myBreaks <- c(seq(1/paletteLength, 1, length.out=floor(paletteLength)))

pheatmap(A_normalised,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         height = 25,
         width = 20,
         breaks = myBreaks)

pheatmap(B_normalised,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         height = 25,
         width = 20,
         breaks = myBreaks)

pheatmap(C_normalised,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
```

```
          border_color = NA,
          color = heatmap_color,
          height = 25,
          width = 20,
          breaks = myBreaks)

pheatmap(D_normalised,
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          border_color = NA,
          color = heatmap_color,
          height = 25,
          width = 20,
          breaks = myBreaks)

pheatmap(E_normalised,
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          border_color = NA,
          color = heatmap_color,
          height = 25,
          width = 20,
          breaks = myBreaks)

pheatmap(F_normalised,
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          border_color = NA,
          color = heatmap_color,
          height = 25,
```

```
          width = 20,
          breaks = myBreaks)

pheatmap(G_normalised,
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          border_color = NA,
          color = heatmap_color,
          height = 25,
          width = 20,
          breaks = myBreaks)
```

1.txt

```
grouptot_genes
Wnt13
BMPlig12
BMPrec6
Fzd4
```

2.txt

```
grouptot_genes
TranscriptionFactors18
Ligands6
Receptors7
```

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(data.table)
library(gplots)
library(pheatmap)
```

```
library(dendextend)
library(factoextra)
library(ComplexHeatmap)
library(RColorBrewer)
library(NbClust)
library(scales)

#Import data in matrix format
total_1 <- read.delim("/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/04-Losses/03-DevelopmentalPathways/Summary/1.txt", row.names=1)
total_2 <- read.delim("/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/04-Losses/03-DevelopmentalPathways/Summary/2.txt", row.names=1)

# Option 1. 0 to 1 relative abundance/expression (54 is the highest value in my dataset)
rescale_custom <- function(x) (x/13)
total_1_normalised <- t(apply(total_1, 1, rescale_custom))
rescale_custom <- function(x) (x/18)
total_2_normalised <- t(apply(total_2, 1, rescale_custom))

# To make 0 a different colour
# First create whatever gradient (e.g. RdBu)
heatmap_color <- colorRampPalette(brewer.pal(n = 7, name = "Reds"))(1000)
heatmap_color[1] <- rgb(1,1,1)
#column_labels = c("your","labels"),
#row_labels = c("your","labels"))

paletteLength <- 1000
```

```
# to go from 0 to max.value (e.g. 1):
myBreaks <- c(seq(1/paletteLength, 1, length.out=floor(paletteLength)))

pheatmap(total_1_normalised,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         height = 25,
         width = 20,
         breaks = myBreaks)

pheatmap(total_2_normalised,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         height = 25,
         width = 20,
         breaks = myBreaks)
```

Mega sum
summing everything together and then plotting bubbleplots
count_not_missing_genes.sh

```
input=$1
cut -f 2 $input | grep -cwv "0"  > temp1
cut -f 3 $input | grep -cwv "0" > temp2
cut -f 4 $input | grep -cwv "0" > temp3
cut -f 5 $input | grep -cwv "0" > temp4
```

```
cut -f 6 $input | grep -cwv "0" > temp5
cut -f 7 $input | grep -cwv "0" > temp6
cut -f 8 $input | grep -cwv "0" > temp7
paste temp1 temp2 temp3 temp4 temp5 temp6 temp7 > "$input"_not_missing_genes
rm temp*
```

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(data.table)
library(gplots)
library(pheatmap)
library(dendextend)
library(factoextra)
library(ComplexHeatmap)
library(RColorBrewer)
library(NbClust)
library(scales)

#Import data in matrix format
Averages <- read.delim("/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/04-Losses/03-DevelopmentalPathways/Summary/mega_sum/all_together.txt_sum", row.names=1)

# Option 1. 0 to 1 relative abundance/expression (54 is the highest value in my dataset)
rescale_custom <- function(x) (x/1032)
Averages_normalised <- t(apply(Averages, 1, rescale_custom))

```

```
# To make 0 a different colour
# First create whatever gradient (e.g. RdBu)
heatmap_color <- colorRampPalette(brewer.pal(n = 7, name = "Reds"))(1000)
heatmap_color[1] <- rgb(1,1,1)
#column_labels = c("your","labels"),
#row_labels = c("your","labels"))

paletteLength <- 1000
# to go from 0 to max.value (e.g. 1):
myBreaks <- c(seq(1/paletteLength, 1, length.out=floor(paletteLength)))

pheatmap(Averages_normalised,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         height = 25,
         width = 20,
         breaks = myBreaks)
```

Chapter 5

Matrix MetalloProteinases

Intro

Gene	Name	Aliases	Location	Description

MMP1	Interstitial collagenase	CLG, CLGN	secreted	Substrates include Col I, II, III, VII, VIII, X, gelatin
MMP2	Gelatinase-A, 72 kDa gelatinase		secreted	Substrates include Gelatin, Col I, II, III, IV, Vii, X
MMP3	Stromelysin 1	CHDS6, MMP-3, SL-1, STMY, STMY1, STR1	secreted	Substrates include Col II, IV, IX, X, XI, gelatin
MMP7	Matrilysin, PUMP 1	MMP-7, MPSL1, PUMP-1	secreted	membrane associated through binding to cholesterol sulfate in cell membranes, substrates include: fibronectin, laminin, Col IV, gelatin
MMP8	Neutrophil collagenase	CLG1, HNC, MMP-8, PMNL-CL	secreted	Substrates include Col I, II, III, VII, VIII, X, aggrecan, gelatin
MMP9	Gelatinase-B, 92 kDa gelatinase	CLG4B, GELB, MANDP2, MMP-9	secreted	Substrates include Gelatin, Col IV, V

MMP10	Stromelysin 2	SL-2, STMY2	secreted	Substrates include Col IV, laminin, fibronectin, elastin
MMP11	Stromelysin 3	SL-3, ST3, STMY3	secreted	MMP-11 shows more similarity to the MT-MMPs, is convertase-activatable and is secreted therefore usually associated to convertase-activatable MMPs. Substrates include Col IV, fibronectin, laminin, aggrecan
MMP12	Macrophage metalloelastase	HME, ME, MME, MMP-12	secreted	Substrates include elastin, fibronectin, Col IV
MMP13	Collagenase 3	CLG3, MANDP1, MMP-13	secreted	Substrates include Col I, II, III, IV, IX, X, XIV, gelatin

MMP14	MT1-MMP	MMP-14, MMP-X1, MT-MMP, MT-MMP 1, MT1-MMP, MT1MMP, MTMMP1, WNCHRS	membrane-associated	type-I transmembrane MMP; substrates include gelatin, fibronectin, laminin
MMP15	MT2-MMP	MT2-MMP, MTMMP2, SMCP-2, MMP-15, MT2MMP	membrane-associated	type-I transmembrane MMP; substrates include gelatin, fibronectin, laminin
MMP16	MT3-MMP	C8orf57, MMP-X2, MT-MMP2, MT-MMP3, MT3-MMP	membrane-associated	type-I transmembrane MMP; substrates include gelatin, fibronectin, laminin
MMP17	MT4-MMP	MT4-MMP, MMP-17, MT4MMP, MTMMP4	membrane-associated	glycosyl phosphatidylinositol-attached; substrates include fibrinogen, fibrin
MMP18	Collagenase 4, xcol4, xenopus		–	No known human

	collagenase			orthologue
MMP19	RASI-1, occasionally referred to as stromelysin-4	MMP18, RASI-1, CODA	–	
MMP20	Enamelysin	AI2A2, MMP-20	secreted	
MMP21	X-MMP	MMP-21, HTX7	secreted	Our findings suggest that MMP-21 functions in embryogenesis and tumor progression.
MMP23A	CA-MMP		membrane-associated	type-II transmembrane cysteine array
MMP23B	–	MIFR, MIFR-1, MMP22, MMP23A	membrane-associated	type-II transmembrane cysteine array
MMP24	MT5-MMP	MMP-24, MMP25, MT-MMP 5, MT-MMP5, MT5-MMP, MT5MMP, MTMMP5	membrane-associated	type-I transmembrane MMP
MMP25	MT6-MMP	MMP-25, MMP20, MMP20A, MMPL1, MT-MMP 6, MT-	membrane-associated	glycosyl phosphatidylinositol-attached

		MMP6, MT6-MMP, MT6MMP, MTMMP6		
MMP26	Matrilysin-2, endometase		–	
MMP27	MMP-22, C-MMP	MMP-27	–	
MMP28	Epilysin	EPILYSIN, MM28, MMP-25, MMP-28, MMP25	secreted	Discovered in 2001 and given its name due to have been discovered in human keratinocytes. Unlike other MMPs this enzyme is constitutivley expressed in many tissues (Highly expressed in testis and at lower levels in lung, heart, brain, colon, intestine, placenta, salivary glands, uterus, skin). A threonine

				replaces proline in its cysteine switch (PRCGVTD).[14]

Focus
Col I is

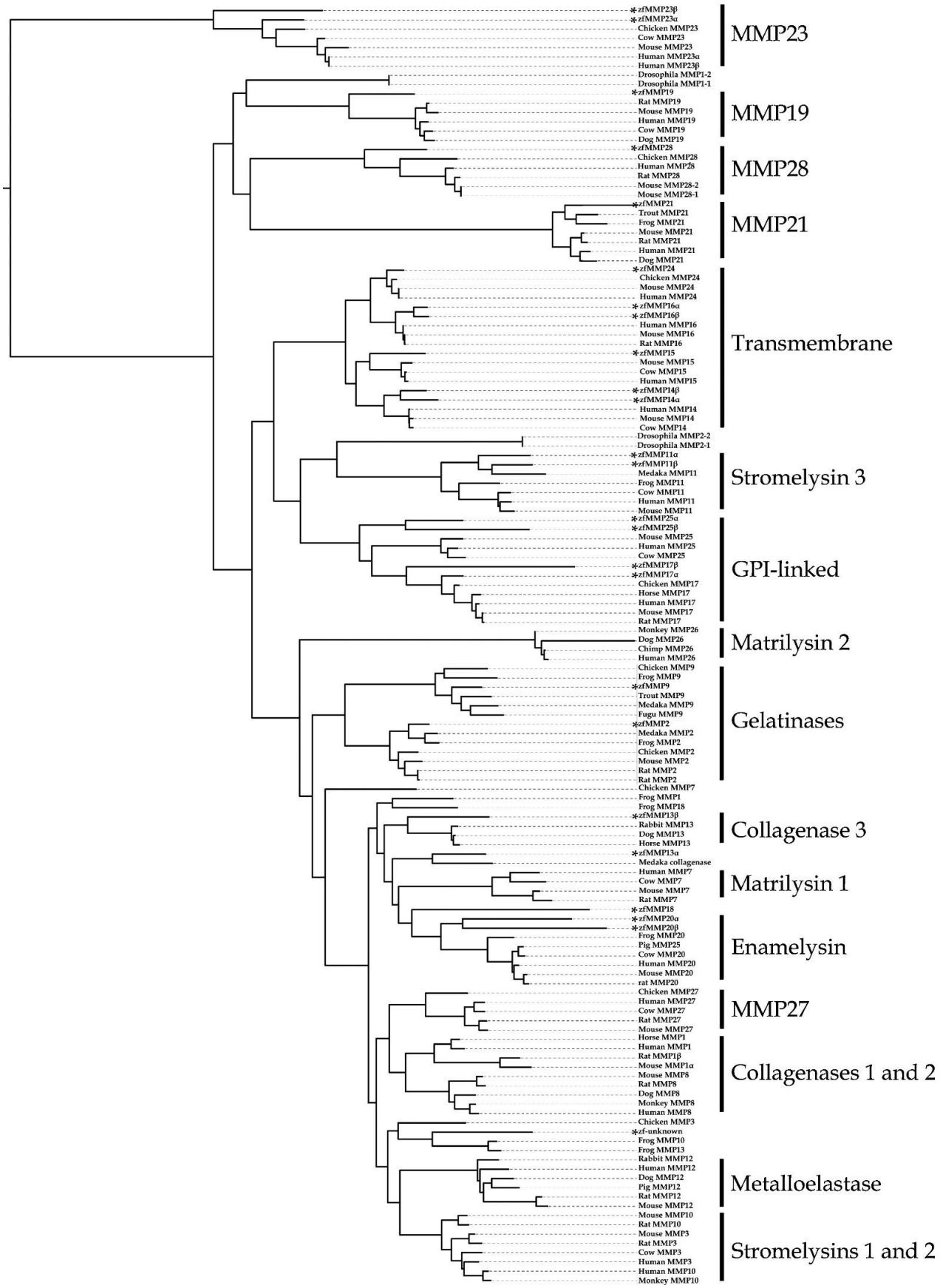
The gained collagenases related GO terms in Osedax are:

```
GO:0030574 collagen catabolic process (way more present in osedax)
GO:0032963 collagen metabolic process (not too much more present in osedax)
```

Collagen related Panther are:

```
PTHR10201 MATRIX METALLOPROTEINASE (way more present in osedax. even more than GOterm)
PTHR12411 CYSTEINE PROTEASE FAMILY C1-RELATED (not too much more present. cathepsin)
```

From this article I will be able to download the Matrix Metallo Proteases (MMP) sequences of Human, Mouse and Zebrafish: Link to additional material
I think I will have to manually collect the sequences from here
I can use this as a template



The sequences I have to collect are:

- MMP23
- MMP19
- MMP28
- MMP21
- MMP24
- MMP16
- MMP15
- MMP14
- MMP11
- MMP25
- MMP17
- MMP26
- MMP9
- MMP2
- MMP13
- MMP7
- MMP20
- MMP27
- MMP1
- MMP8
- MMP12
- MMP10
- MMP3
- MMP18 (one from osedax japonicus) NOT PRESENT IN HUMAN

CODE

test.sh

```
#!/bin/bash


species=osedax
xls="$species"_isoform.xls
output="$species"_MMPs.txt
output2="$species"_MMPs_count.txt

```

```
cut -f 18,19,20 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | fgrep "PTHR10201" | cut -f 1,2,3 > $output
cut -f 18,19,20 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | fgrep "PTHR10201" | wc -l > $output2


species=oasisia
xls="$species"_isoform.xls
output="$species"_MMPs.txt
output2="$species"_MMPs_count.txt

cut -f 18,19,20 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | fgrep "PTHR10201" | cut -f 1,2,3 > $output
cut -f 18,19,20 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | fgrep "PTHR10201" | wc -l > $output2


species=riftia
xls="$species"_isoform.xls
output="$species"_MMPs.txt
output2="$species"_MMPs_count.txt

cut -f 18,19,20 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | fgrep "PTHR10201" | cut -f 1,2,3 > $output
cut -f 18,19,20 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | fgrep "PTHR10201" | wc -l > $output2
```

```
28
29
30 species=lamellibrachia
31 xls="$species"_isoform.xls
32 output="$species"_MMPs.txt
33 output2="$species"_MMPs_count.txt
34
35 cut -f 18,19,20 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL
   _isoforms_annotations/$xls | fgrep "PTHR10201" | cut -f 1,2,3 >
   $output
36 cut -f 18,19,20 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL
   _isoforms_annotations/$xls | fgrep "PTHR10201" | wc -l > $output
   2
37
38
39 species=paraescarpia
40 xls="$species"_isoform.xls
41 output="$species"_MMPs.txt
42 output2="$species"_MMPs_count.txt
43
44 cut -f 18,19,20 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL
   _isoforms_annotations/$xls | fgrep "PTHR10201" | cut -f 1,2,3 >
   $output
45 cut -f 18,19,20 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL
   _isoforms_annotations/$xls | fgrep "PTHR10201" | wc -l > $output
   2
46
47
48 species=owenia
49 xls="$species"_isoform.xls
50 output="$species"_MMPs.txt
51 output2="$species"_MMPs_count.txt
```

```
52
53 cut -f 2,3,10 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | fgrep "PTHR10201" | cut -f 1,2 > $output
54 cut -f 2,3,10 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | fgrep "PTHR10201" | wc -l > $output2
55
56 species=capitella
57 xls="$species"_isoform.xls
58 output="$species"_MMPs.txt
59 output2="$species"_MMPs_count.txt
60
61 cut -f 1,2,3 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | fgrep "PTHR10201" | cut -f 1,2 > $output
62 cut -f 1,2,3 /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | fgrep "PTHR10201" | wc -l > $output2
```

seqtk.sh

```
1 #!/bin/bash
2 grep "PTHR10201" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/capitella* | cut -f1 | sed 's/^/Ctel_/' > MMPs_candidates_capitella
3 module load seqtk
4 seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Ctel.fa MMPs_candidates_capitella > MMPs_candidates_capitella.fa
5 grep "PTHR10201" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/owenia*.xls | cut -f2  > MMPs_candidates_owenia
6 seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Ofus.fa MMPs_candidates_owenia > MMPs_candidates_owenia.fa
7 grep "PTHR10201" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/riftia* | cut -f2 | sed 's/^/Rpac_/' > MM
```

```
Ps_candidates_riftia
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Rpac.fa MMPs_candidates_riftia > MMPs_candidates_riftia.fa
grep "PTHR10201" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/oasisia* | cut -f2 | sed 's/^/Oalv_/' > MMPs_candidates_oasisia
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Oalv.fa MMPs_candidates_oasisia > MMPs_candidates_oasisia.fa
grep "PTHR10201" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/osedax* | cut -f2 | sed 's/^/Ofra_/' > MMPs_candidates_osedax
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Ofra.fa MMPs_candidates_osedax > MMPs_candidates_osedax.fa
grep "PTHR10201" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/lamellibrachia* | cut -f2 | sed 's/^/Lluy_/' > MMPs_candidates_lamellibrachia
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Lluy.fa MMPs_candidates_lamellibrachia > MMPs_candidates_lamellibrachia.fa
grep "PTHR10201" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/paraescarpia* | cut -f2 | sed 's/nbis_mrna_/nbis-mrna-/' | sed 's/^/Pech_/' > MMPs_candidates_paraescarpia
seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Pech.fa MMPs_candidates_paraescarpia > MMPs_candidates_paraescarpia.fa
```

```
bash seqtk.sh
cat MMPs_candidates*.fa > MMPs_candidates_all.fa
```

Phylogenetic Tree

Step 1 - MMPs phylogenetic tree

make an alignment using the txt file obtained before and the online tool MAFFT

don't select any additional options

export in fasta format and name the output:

```
MMPs_sequences_MAFFT1.fasta
```

Step 2 - MMPs phylogenetic tree
download and install Jalview
open the file:

```
MMPs_sequences_MAFFT1.fasta
```

Cut away all the areas with no conservation. Basically we should maintain only the domain
and save the cutted file as:

```
MMPs_sequences_Jalview.fasta
```

Step 3 - MMPs phylogenetic tree
make an alignment again using the file “MMPs_sequences_Jalview.fasta” obtained before and the online tool MAFFT
select the option “L-INS-i”
export in fasta format and name the output:

```
MMPs_sequences_MAFFT2.fasta
```

Step 4 - MMPs phylogenetic tree
use Jalview again. Open the file:

```
MMPs_sequences_MAFFT2.fasta
```

No chopping this time, just open the file and export it to fasta as:

```
MMPs_sequences_Jalview2.fasta
```

Step 5 - MMPs phylogenetic tree

```
MMPs_sequences_gblocks.fa
```

or use trimAl
on my personal pc

```
conda create -n fasttree_env
conda activate fasttree_env
conda install -c bioconda fasttree
conda install -c bioconda trimal
```

```
trimal -in 4a.fasta -out MMPs_sequences_trimal.fa
```

Step 6 - MMPs phylogenetic tree
make a tree

```
conda activate fasttree_env
FastTree MMPs_sequences_trimal.fa > MMPs_sequences_a.tree
```

and then open it with FigTree

Improved
I need to use human MMPs to find the gene families corresponding to those genes in all the species I have used for Orthofinder

Step1 - Blast

First I need to blast the MMPs sequences against the Human non redundant proteome to find the correct gene name
test.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/MMPs/blast
#$ -o /data/scratch/btx654/MMPs/blast
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=10G
#$ -l h_rt=120:0:0
#$ -l highmem

$non_redundant_prot=/data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Hsap.fa
cp /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/Hsap.fa ./

#make a diamond BLAST database of this proteome and BLAST the consensi.fa.classified dataset against it, to find potential bona fide genes. To make sure we only get the real genes, the e-value
```

```
is very stringent.
module load anaconda3
source activate diamond
diamond makedb --in $Hsap.fa -d Hsap_non_redundant_prot

diamond blastp -d Hsap_non_redundant_prot -q MMPs_Hsap_genes.txt -o default.1e10.blastp -f 6 qseqid bitscore evalue stitle -k 25 -e 1e-10 -p 8
diamond blastp -d Hsap_non_redundant_prot -q MMPs_Hsap_genes.txt -o ultra_sensitive.1e10.blastp --ultra-sensitive -f 6 qseqid bit score evalue stitle -k 25 -e 1e-10 -p 8
```

		OG0000071	OG0032603	OG0007560	OG0001105
MMP23-	Hsap_ENST00000356026		0		
MMP19-	Hsap_ENST00000322569	0			
MMP28-	Hsap_ENST00000605424	0			
MMP21-	Hsap_ENST00000368808			0	

MMP24-	Hsap_ENST00000246186	0			
MMP16-	Hsap_ENST00000286614	0			
MMP15-	Hsap_ENST00000219271	0			
MMP14-	Hsap_ENST00000311852	0			
MMP11-	Hsap_ENST00000215743	0			
MMP25-	Hsap_ENST00000336577	0			
MMP17-	Hsap_ENST00000360564	0			
MMP26-	Hsap_ENST00000380390	0			

MMP9-	Hsap_ENST00000372330				0
MMP2-	Hsap_ENST00000219070				0
MMP13-	Hsap_ENST00000340273	0			
MMP7-	Hsap_ENST00000260227	0			
MMP20-	Hsap_ENST00000260228	0			
MMP27-	Hsap_ENST00000260229	0			
MMP1-	Hsap_ENST00000315274	0			
MMP8-	Hsap_ENST00000236826	0			

MMP12-	Hsap_ENST00000571244	0			
MMP10-	Hsap_ENST00000279441	0			
MMP3-	Hsap_ENST00000299855	0			

OG0000071

Hsap|ENST00000215743, Hsap|ENST00000219271, Hsap|ENST00000236826, Hsap|ENST00000246186, Hsap|ENST00000260227, Hsap|ENST00000260228, Hsap|ENST00000260229, Hsap|ENST00000279441, Hsap|ENST00000286614, Hsap|ENST00000299855, Hsap|ENST00000311852, Hsap|ENST00000315274, Hsap|ENST00000322569, Hsap|ENST00000336577, Hsap|ENST00000340273, Hsap|ENST00000360564, Hsap|ENST00000380390, Hsap|ENST00000571244, Hsap|ENST00000605424

```
OG0000071        275        28        Eumetazoa        0
       Hrob,Ofra
```

OG0001105

Hsap|ENST00000219070, Hsap|ENST00000339841, Hsap|ENST00000344839, Hsap|ENST00000372330

Step2 - get the sequences

test.sh

```
1 module load seqtk
2
```

```
head -1 /data/SBCS-MartinDuranLab/03-Giacomo/data/all_together/gene_family_evolution/orthofinder_Jun2021/ultra_sensitive/Results_Jun09/Orthogroups/Orthogroups.csv | cut -f 2-29 | sed 's/\t\t*/\n/g' > species_list.txt

grep "OG0000071\|OG0032603\|OG0007560\|OG0001105" /data/SBCS-MartinDuranLab/03-Giacomo/data/all_together/gene_family_evolution/orthofinder_Jun2021/ultra_sensitive/Results_Jun09/Orthogroups/Orthogroups.csv | cut -f 2-29 > MMPs_Orthogroups.csv

for run in {1..28}; do
species=$(cat species_list.txt | head -"$run" | tail -1)
echo $species

cut -f "$run" MMPs_Orthogroups.csv > "$species"_MMPs_Orthogroups.csv
sed 's/|/_/g' "$species"_MMPs_Orthogroups.csv | sed 's/, /\n/g' | sed -r '/^\s*$/d' > "$species"_gene_names.txt
rm "$species"_MMPs_Orthogroups.csv

seqtk subseq /data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/"$species".fa "$species"_gene_names.txt > "$species"_genes.fa

done
```

```
bash test.sh
```

Step3 - Tree

Collect all the sequences extracted in step2 and add them to those downloaded from uniprot (exept those from 5 sibo + owenia and capi and those from human)

```
cat Blan_genes.fa Bpla_genes.fa Cgig_genes.fa Dgyr_genes.fa Eand_genes.fa Gaeg_genes.fa Hmia_genes.fa Hrob_genes.fa Lana_genes.f
```

```
a Lgig_genes.fa Locu_genes.fa Myes_genes.fa Ngen_genes.fa Nvec_g
enes.fa Paus_genes.fa Skow_genes.fa Smar_genes.fa Smed_genes.fa
Spur_genes.fa Tcas_genes.fa > ALL_genes.fa
```

Then I am adding these sequences to “2.fasta” obtained after aligning and trimming only the MMPs sequences downloaded from uniprot. I DID NOT ADD THE SEQUENCES FROM SIBO AND CAPI-OWENIA OBTAINED IN THE SECTION CODE

The tree is still a bit messy:
I am going to name some of the clades that came up in the tree:
MMP21_like
MMP11_17_25_like
MMP
MMPA (is the branch containing osedax japonicus)
MMPB
MMPC
MMPD
MMPE
MMPF
MMPG

removed:

```
>OFUSG03511.1
>Pech_nbis-mrna-5872
>Lluy_FUN_007880-T1
```

raxml.sh

```
#!/bin/bash
#$ -pe smp 5
#$ -l highmem
#$ -l h_vmem=10G
#$ -l h_rt=240:0:0
#$ -cwd
```

```
#$ -j y
module load raxml
raxmlHPC -f a -b 476 -p 903 -x 12345 -# autoMRE -m PROTGAMMAAUTO
-s MMPs_sequences_trimal_ok.fa -n MMPs_raxml_tree.tre
```

removed: (after cropping with Jalview there were no ammino acids left)

```
Nvec_EDO29979_MMP2_9_like
Nvec_EDO33047_MMP2_9_like
Nvec_EDO32111_MMP2_9_like
```

I am now aligning just the ZnMc domain which contains the active site:
after trimal some proteins result to not have that domain so I will remove them:

```
>Ofra_OFRAG00000002650.1
>Lluy_FUN_026149-T1_MMP2_9_like
>Blan_BL22912_cuf1_MMP21_like
>Ofra_OFRAG00000014433.1
>Nvec_EDO45372_MMP2_9_like
Nvec_EDO45377_MMP2_9_like
>Paus_g3879.t1_MMP2_9_like
>Ofra_OFRAG00000000712.1
>Ofra_OFRAG00000000713.1
>Ofra_OFRAG00000000714.1
>Ofra_OFRAG00000000715.1
>Ofra_OFRAG00000000716.1
>Ofra_OFRAG00000000717.1
>Ofra_OFRAG00000000718.1
>Ofra_OFRAG00000000719.1
>Ofra_OFRAG00000000720.1
>Ofra_OFRAG00000000721.1
>Ofra_OFRAG00000000722.1
```

```
>Ofra_OFRAG00000014432.1
>Nvec_EDO29980_MMPB
>Ngen_g12182.t1_MMP?
>Lana_g4594.t1_MMP2_9_like
>Ngen_g7295.t1_MMP2_9_like
>Ngen_g37516.t1_MMP2_9_like
>Ofra_OFRAG00000014430.1
>Ofra_OFRAG00000014431.1
>Locu_ENSLOCT00000011484_MMP2_9_like
>Ngen_g22040.t1_MMP2_9_like
>Nvec_EDO47418_MMP2_9_like
>Paus_g15030.t1_MMP2_9_like
>Ngen_g21330.t1_MMP2_9_like
>Blan_BL96786_cuf2_MMP2_9_like
>Blan_BL10256_cuf0_MMP2_9_like
>Paus_g10720.t1_MMP2_9_like
>Ofra_OFRAG00000002369.1
>Ngen_g23379.t1_MMP2_9_like
>Nvec_EDO35188_MMP?
>Nvec_EDO45821_MMP2_9_like
>Ofra_OFRAG00000002286.1
>Ngen_g34577.t1_MMPB
>Ofra_OFRAG00000007211.1
>Ofra_OFRAG00000007212.1
>Ofra_OFRAG00000007213.1
>Ofra_OFRAG00000008179.1
>Blan_BL10324_cuf0_MMP2_9_like
>Ofra_OFRAG00000013465.1

Nvec_EDO29979_MMP2_9_like
```

```
Nvec_EDO33047_MMP2_9_like
Nvec_EDO32111_MMP2_9_like

removed from MMPs_sequences_trimal_ok.fa:
>Lana_g26508.t1_MMP2_9_like
>Ngen_g7296.t1_MMP2_9_like
>Locu_ENSLOCT00000017075_MMP23
>Blan_BL06531_evm0_MMP?
>OFUSG11076.1_MMP?
```

Raxml-ng

on the cluster

```
module load anaconda3
conda create --prefix /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/raxml_env
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/raxml_env
conda install -c bioconda raxml
```

- raxml-8.2.12

raxml.sh

```
#!/bin/bash
#$ -pe smp 7
#$ -l highmem
#$ -l h_vmem=20G
#$ -l h_rt=240:0:0
#$ -cwd
#$ -j y

module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/raxml_env
```

```
raxmlHPC -f a -b 476 -p 903 -# autoMRE -m PROTGAMMAAUTO -s MMPs_sequences_trimal_ok.fa -n MMPs_raxml_tree.tre
```

raxml_oceane.sh

```
#!/bin/bash
#$ -pe smp 7
#$ -l highmem
#$ -l h_vmem=20G
#$ -l h_rt=240:0:0
#$ -cwd
#$ -j y

module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/raxml_env

raxmlHPC -f a -b 476 -p 903 -x 12345 -# autoMRE -m PROTGAMMAAUTO -s MMPs_sequences_trimal_ok.fa -n MMPs_raxml_tree_oceane.tre
```

```
module load anaconda3
conda create --prefix /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/raxml_ng_env
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/raxml_ng_env
conda install -c bioconda raxml-ng
```

on command line:

```
raxml-ng --check --msa MMPs_sequences_trimal_ok.fa --model LG+G8+F --prefix T1
raxml-ng --parse --msa T1.raxml.reduced.phy --model LG+G8+F --prefix T2
```

raxml_ng.sh

```
#!/bin/bash
#$ -pe smp 20
#$ -l highmem
#$ -l h_vmem=20G
#$ -l h_rt=240:0:0
#$ -cwd
#$ -j y

module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/raxml_ng_env

raxml-ng --all --msa T2.raxml.rba --model LG+G8+F --bs-trees 1000 --threads 20 --prefix MMPs_raxml_ng
```

- LG+G8+F Perform an all-in-one analysis (ML tree search + non-parametric bootstrap) (10 randomized parsimony starting trees, fixed empirical substitution matrix (LG), empirical aminoacid frequencies from alignment, 8 discrete GAMMA categories, 1000 bootstrap replicates):

```
module load anaconda3
conda create --prefix /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/modeltest_env
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/modeltest_env
conda install -c bioconda modeltest-ng
```

modeltest_ng.sh

```
#!/bin/bash
#$ -pe smp 20
#$ -l highmem
#$ -l h_vmem=20G
```

```
#$ -l h_rt=240:0:0
#$ -cwd
#$ -j y

module load anaconda3
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/modeltest_env

/data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/modeltest_env/bin/modeltest-ng -i MMPs_sequences_trimal_ok.fa --datatype aa --processes 20 --template raxml -t ml --output MMPs_modeltest_ng_output
```

IQtree

```
module load anaconda3
conda create --prefix /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/IQtree_env
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/IQtree_env
conda install -c bioconda iqtree
```

iqtree.sh

```
#!/bin/bash
#$ -pe smp 20
#$ -l highmem
#$ -l h_vmem=20G
#$ -l h_rt=240:0:0
#$ -cwd
#$ -j y

module load anaconda3
```

```
conda activate /data/SBCS-MartinDuranLab/03-Giacomo/src/anaconda3/IQtree_env

iqtree -s MMPs_sequences_trimal_ok.fas -m MFP -B 1000
```

Blast_pacbio

in order to confirm the duplication events I am going to blast the ZnMc domain of Osedax against the raw pacbio reads

```
module load seqtk
seqtk seq -a osedax_pb_raw.fastq > osedax_pb_raw.fa
```

ZnMc_domain_osedax.fa

```
>ZnMc_domain_osedax
GFVWKHLNITYKITEYTRKVSHTHIDEAAAKALNFWGEVTQLNFRQVSPYSKADIDIKFVVGDH
GDGLPFDGKGGIIGHAFPPEYGLAHFDDAESWVIGECDDASINILQVMTHEFGHSLGLAHSFNR
SNVMFPSYKGYEPNFALSGDDIKGIQSLYG
```

blast.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=20G
#$ -l h_rt=120:0:0
#$ -l highmem

module load blast+
makeblastdb -in osedax_pb_raw.fa -dbtype nucl -out osedax_pacbio
tblastn -db osedax_pacbio -query ZnMc_domain_osedax.fa -out osedax_pacbio_tblastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db osedax_pacbio -query ZnMc_domain_osedax.fa -out osedax_pacbio_tblastn_out.html -max_target_seqs 5 -evalue 1e-10 -num
```

```
_threads 8 -html
```

used the previous sequences and translated them into nucleotide with this online tool

ZnMc_domain_osedax_nucl.fa

```
>ZnMc_domain_osedax
GGCTTCGTGTGGAAGCACCTGAACATCACCTACAAGATCACCGAGTACACCCGCAAGGTG
AGCCACACCCACATCGATGAGGCCGCCGCCAAGGCCCTGAACTTCTGGGGCGAGGTGACC
CAGCTGAACTTCCGCCAGGTGAGCCCCTACAGCAAGGCCGATATCGATATCAAGTTCGTG
GTGGGCGATCACGGCGATGGCCTGCCCTTCGATGGCAAGGGCGGCATCATCGGCCACGCC
TTCCCCCCCGAGTACGGCCTGGCCCACTTCGATGATGCCGAGAGCTGGGTGATCGGCGAG
TGCGATGATGCCAGCATCAACATCCTGCAGGTGATGACCCACGAGTTCGGCCACAGCCTG
GGCCTGGCCCACAGCTTCAACCGCAGCAACGTGATGTTCCCCAGCTACAAGGGCTACGAG
CCCAACTTCGCCCTGAGCGGCGATGATATCAAGGGCATCCAGAGCCTGTACGGC
```

blast_nucl.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=20G
#$ -l h_rt=120:0:0
#$ -l highmem

module load blast+
blastn -db osedax_pacbio -query ZnMc_domain_osedax_nucl.fa -out osedax_pacbio_blastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
blastn -db osedax_pacbio -query ZnMc_domain_osedax_nucl.fa -out osedax_pacbio_blastn_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -html
```

no hits found

blast_assembly.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -pe smp 8
#$ -l h_vmem=20G
#$ -l h_rt=120:0:0
#$ -l highmem

module load blast+
makeblastdb -in /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/haploidization/purge_dups/osedax_purged.fa -dbtype nucl -out osedax_assembly
tblastn -db osedax_assembly -query ZnMc_domain_osedax.fa -out osedax_pacbio_tblastn_out -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -outfmt 6
tblastn -db osedax_assembly -query ZnMc_domain_osedax.fa -out osedax_pacbio_tblastn_out.html -max_target_seqs 5 -evalue 1e-10 -num_threads 8 -html
```

```
module load anaconda3
conda create -n genomePolishing python=2.7 anaconda
conda activate genomePolishing
conda install -c bioconda genomicconsensus
conda install -c bioconda pbmm2
conda install -c bioconda pbgcpp
conda install pbbam
```

First Round:

Align raw PB reads to the assembly in a bam file
Code to run pbmm2:

```
#!/bin/bash
#$ -wd /data/scratch/btx604/Oasisia/genomePolishing/pbalign/step1
#$ -j y
#$ -o /data/SBCS-MartinDuranLab/03-Giacomo/logs/genomePolishing
#$ -pe smp 5
#$ -l h_vmem=8G
#$ -l h_rt=72:0:0

cd /data/scratch/btx604/Oasisia/genomePolishing/pbalign/step1

module load anaconda2
source activate genomePolishing

pbmm2 align \
 /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/canu/Oasisia.contigs.fasta  \
 /data/SBCS-MartinDuranLab/03-Giacomo/data/oasisia/00-pacbio/data2/pb/r64044_20190812_215729/1_A01/m64044_190812_220643.subreads.bam \
 /data/scratch/btx604/Oasisia/genomePolishing/pbalign/step1/oasisia_pbalign_step1.bam

module load samtools/1.9

samtools sort /data/scratch/btx604/Oasisia/genomePolishing/pbalign/step1/oasisia_pbalign_step1.bam -o /data/scratch/btx604/Oasisia/genomePolishing/pbalign/step1/oasisia_sorted_step1.bam
samtools index /data/scratch/btx604/Oasisia/genomePolishing/pbalign/step1/oasisia_sorted_step1.bam
```

- Sorted output can be generated using `--sort`

code to run minimap2
minimap2.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -pe smp 10
#$ -l h_vmem=8G
#$ -l h_rt=120:0:0

module load minimap2
minimap2 -ax map-pb /data/SBCS-MartinDuranLab/03-Giacomo/data/osedax/haploidization/purge_dups/osedax_purged.fa /data/scratch/btx654/MMPs/blast_pacbio/osedax_pb_raw.fastq > raw_pacbio_vs_purged_assembly.sam
```

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -pe smp 5
#$ -l h_vmem=4G
#$ -l h_rt=120:0:0
module load samtools
samtools view -S -b raw_pacbio_vs_purged_assembly.sam > raw_pacbio_vs_purged_assembly.bam
samtools sort raw_pacbio_vs_purged_assembly.bam -o raw_pacbio_vs_purged_assembly_sorted.bam
samtools index raw_pacbio_vs_purged_assembly_sorted.bam
samtools view raw_pacbio_vs_purged_assembly.bam "tig00006601|arrow|arrow|pilon|pilon" > tig00006601.bam
```

R studio

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(data.table)
library(gplots)
library(pheatmap)
library(dendextend)
library(factoextra)
library(ComplexHeatmap)
library(RColorBrewer)
library(NbClust)
library(scales)

#Import data in matrix format
df1 <- read.delim("~/Dropbox/02-OweniaGenome/02-Figures/00-Clutter/02-Fran&YanFigures/Urechis_Hox_genes_16_stages.txt")
df2
#Name columns and stuff
names_row <- raw_data[,1]
df_curated <- data.matrix(raw_data[,2:ncol(raw_data)])
rownames(df_curated) <- names_row

# Option 1. Deviation from the mean (z-score).
# Normalise your data around the mean (0) and then look deviation around mean.
df_normalised <- t(scale(t(df_curated)))
df_normalised <- na.omit(df_normalised)

# Option 1. 0 to 1 relative abundance/expression
rescale_custom <- function(x) (x/(max(x)))
```

```
df_normalised <- t(apply(df_curated, 1, rescale_custom))
df_normalised <- na.omit(df_normalised)

# Colours
# Option 1. RdBu gradient.
heatmap_color <- colorRampPalette(rev(brewer.pal(n = 7, name = "
RdBu")))(100)

# Option 2. Reds gradient.
heatmap_color <- colorRampPalette(rev(brewer.pal(n = 7, name = "
Reds")))(100)

# Option 3. 1/2 RdBu gradient.
heatmap_color <- colorRampPalette(rev(brewer.pal(n = 7, name = "
RdBu")))(100)
heatmap_color <- heatmap_color[50:100]

# Option 4. RdGy gradient.
heatmap_color <- colorRampPalette(rev(brewer.pal(n = 7, name = "
RdGy")))(100)

# To make 0 a different colour
# First create whatever gradient (e.g. RdBu)
heatmap_color <- colorRampPalette(rev(brewer.pal(n = 7, name = "
RdBu")))(100)
heatmap_color[1] <- rgb(0,0,0) # here include the colour you're
interested in

# Conversion from dataframe to datamatrix (if necessary)
dm <- data.matrix(df)

```

```
# Heatmaps
# ComplexHeatmap package
ComplexHeatmap::Heatmap(dm,
                        cluster_columns = FALSE,
                        cluster_rows = FALSE,
                        col = heatmap_color,
                        heatmap_legend_param = list(color_bar =
"continuous"))

#column_labels = c("your","labels"),
#row_labels = c("your","labels"))

# pheatmap
pheatmap(df,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color)

# If you want common scale for different heatmaps:
# First define some "breaks"

paletteLength <- 100
# to go from 0 to max.value (e.g. 1):
myBreaks <- c(seq(max.value/paletteLength, max.value, length.out
=floor(paletteLength)))
# e.g. if it was actually 1, then:
myBreaks <- c(seq(1/paletteLength, 1, length.out=floor(paletteLe
ngth)))

# Then use the following:
```

```
pheatmap(df,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         breaks = myBreaks)

# To set sizes of heatmaps:
# Option 1: Cell height, each cell is the same size
pheatmap(df,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         cellheight = 20,
         cellwidth = 20)

# Option 2: Heatmap height, the whole heatmap will have this size
pheatmap(df,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         height = 20,
         width = 20)
```

Expression

```
sed 's/^.*>Ofra_//g' osedax_IDs_only > osedax_IDs_only_ok
grep -f osedax_IDs_only_ok osedax_kallisto_tpm.tsv > osedax_MMPs_expression
```

```
for S in $(cat osedax_IDs_only_ok | awk '{print $1}'); do grep $S osedax_MMPs_expression; done > right_order_osedax_MMPs_expression
```

```
#scale all the gene values then subset
#in R
A <- read.delim("/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/05-Expanssions/01-Collagenases/RNAseq/osedax_kallisto_tpm.tsv", row.names=1)
B <- data.matrix(scale(A))
write.csv(B, file="/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/05-Expanssions/01-Collagenases/RNAseq/osedax_kallisto_tpm_scaled.tsv")
#in shell
sed 's/\,/\t/g' osedax_kallisto_tpm_scaled.tsv | sed 's/\"//g' > osedax_kallisto_tpm_scaled_ok.tsv
cut -f 1,2,3 osedax_kallisto_tpm_scaled_ok.tsv > osedax_kallisto_tpm_scaled_ok_ok.tsv
grep -f osedax_IDs_only_ok osedax_kallisto_tpm_scaled_ok_ok.tsv > osedax_MMPs_expression_ALLscaled
for S in $(cat osedax_IDs_only_ok | awk '{print $1}'); do grep $S osedax_MMPs_expression_ALLscaled; done > right_order_osedax_MMPs_expression_ALLscaled
```

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(data.table)
library(gplots)
library(pheatmap)
library(dendextend)
library(factoextra)
```

```
library(ComplexHeatmap)
library(RColorBrewer)
library(NbClust)
library(scales)

#Import data in matrix format
A <- read.delim("/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/05-Expanssions/01-Collagenases/RNAseq/right_order_osedax_MMPs_expression", row.names=1)
A_scaled <- data.matrix(scale(A))
# Option 1. 0 to 1 relative abundance/expression (54 is the highest value in my dataset)
B <- read.delim("/Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/05-Expanssions/01-Collagenases/RNAseq/right_order_osedax_MMPs_expression_ALLscaled", row.names=1)

# To make 0 a different colour
# First create whatever gradient (e.g. RdBu)
heatmap_color <- colorRampPalette(brewer.pal(n = 7, name = "Reds"))(1000)
heatmap_color[1] <- rgb(1,1,1)
#column_labels = c("your","labels"),
#row_labels = c("your","labels"))

paletteLength <- 1000
# to go from 0 to max.value (e.g. 1):
myBreaks <- c(seq(1/paletteLength, 1, length.out=floor(paletteLength)))

pheatmap(A_scaled,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
```

```
          border_color = NA,
          color = heatmap_color,
          height = 5,
          width = 25)

pheatmap(B,
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          border_color = NA,
          color = heatmap_color,
          height = 5,
          width = 25)
```

B is definitely the best one!

Panther Metabolism

Nomenclature:
A_II_1 (compound_pathwayNumber_stepNumber)
A_II_1A (uppercase final letters in case more than one enzyme can catalise the reaction “OR”)
A_II_1a (lowercase final letters in case more than one enzyme are needed to catalise the reaction “AND”)
A_II_1_3 (in case the same enzyme can catalise different steps)

CODE

```
#!/bin/bash

species=$1
xls="$species"*xls
output="$species"_metabolism.txt

cut -f 18,19 $xls > "$species"_annotations
```

```
annotations="$species"_annotations

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   echo $Panther_ID
   genes=$(fgrep "$Panther_ID" $annotations | cut -f 1)
   echo $genes
   echo $genes >> $output
done < metabolism_pantherID.txt
```

test_owenia_capitella.sh

```
#!/bin/bash

species=owenia
xls=Owenia_annotation_v250920.1_TrinoPantherKO.xls
output="$species"_metabolism.txt
#cut -f 18,19 for the 5 siboglinidae, -f 1,2 for capitella
cut -f 2,10 $xls > "$species"_annotations
annotations="$species"_annotations

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   echo $Panther_ID
   genes=$(fgrep "$Panther_ID" $annotations | cut -f 1)
   echo $genes
   echo $genes >> $output
done < metabolism_pantherID.txt

species=capitella
xls=Capitella_annotation_Feb2021_TrinoPantherKO.xls
output="$species"_metabolism.txt
```

```
#cut -f 18,19 for the 5 siboglinidae, -f 1,2 for capitella
cut -f 1,2 $xls > "$species"_annotations
annotations="$species"_annotations

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   echo $Panther_ID
   genes=$(fgrep "$Panther_ID" $annotations | cut -f 1)
   echo $genes
   echo $genes >> $output
done < metabolism_pantherID.txt
```

Glycine degradation

GlycineDegradation_pantherID.txt

```
GD_1        PTHR46120
GD_2        PTHR13847:SF187
GD_3        PTHR13847:SF200
GD_4        PTHR11680
```

script.sh

```
#!/bin/bash

species=osedax
xls="$species"_isoform.xls
output="$species"_GlycineDegradation.txt
output2="$species"_GlycineDegradation_count.txt

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
```

```
   genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | cut -f 18)
   echo $genes >> $output
   count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | wc -l)
   echo $count >> $output2
done < GlycineDegradation_pantherID.txt

species=oasisia
xls="$species"_isoform.xls
output="$species"_GlycineDegradation.txt
output2="$species"_GlycineDegradation_count.txt

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | cut -f 18)
   echo $genes >> $output
   count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | wc -l)
   echo $count >> $output2
done < GlycineDegradation_pantherID.txt

species=riftia
xls="$species"_isoform.xls
output="$species"_GlycineDegradation.txt
output2="$species"_GlycineDegradation_count.txt

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
```

```
omo/data/00-ALL_isoforms_annotations/$xls | cut -f 18)
   echo $genes >> $output
   count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
omo/data/00-ALL_isoforms_annotations/$xls | wc -l)
   echo $count >> $output2
done < GlycineDegradation_pantherID.txt

species=lamellibrachia
xls="$species"_isoform.xls
output="$species"_GlycineDegradation.txt
output2="$species"_GlycineDegradation_count.txt

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
omo/data/00-ALL_isoforms_annotations/$xls | cut -f 18)
   echo $genes >> $output
   count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
omo/data/00-ALL_isoforms_annotations/$xls | wc -l)
   echo $count >> $output2
done < GlycineDegradation_pantherID.txt

species=paraescarpia
xls="$species"_isoform.xls
output="$species"_GlycineDegradation.txt
output2="$species"_GlycineDegradation_count.txt

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
omo/data/00-ALL_isoforms_annotations/$xls | cut -f 18)
```

```
63    echo $genes >> $output
64    count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
omo/data/00-ALL_isoforms_annotations/$xls | wc -l)
65    echo $count >> $output2
66 done < GlycineDegradation_pantherID.txt
67
68 species=owenia
69 xls="$species"_isoform.xls
70 output="$species"_GlycineDegradation.txt
71 output2="$species"_GlycineDegradation_count.txt
72
73 while read line; do
74    Panther_ID=$(cut -f 2 <<< "$line")
75    genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
omo/data/00-ALL_isoforms_annotations/$xls | cut -f 2)
76    echo $genes >> $output
77    count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
omo/data/00-ALL_isoforms_annotations/$xls | wc -l)
78    echo $count >> $output2
79 done < GlycineDegradation_pantherID.txt
80
81 species=capitella
82 xls="$species"_isoform.xls
83 output="$species"_GlycineDegradation.txt
84 output2="$species"_GlycineDegradation_count.txt
85
86 while read line; do
87    Panther_ID=$(cut -f 2 <<< "$line")
88    genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
omo/data/00-ALL_isoforms_annotations/$xls | cut -f 1)
89    echo $genes >> $output
```

```
   count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
omo/data/00-ALL_isoforms_annotations/$xls | wc -l)
   echo $count >> $output2
done < GlycineDegradation_pantherID.txt
```

```
cut -f 1 GlycineDegradation_pantherID.txt > GlycineDegradation_p
antherID_FirstColumn
paste GlycineDegradation_pantherID_FirstColumn owenia_GlycineDeg
radation_count.txt capitella_GlycineDegradation_count.txt lamell
ibrachia_GlycineDegradation_count.txt paraescarpia_GlycineDegrad
ation_count.txt oasisia_GlycineDegradation_count.txt riftia_Glyc
ineDegradation_count.txt osedax_GlycineDegradation_count.txt > G
lycineDegradation_count_table.txt
```

Glycine Cleavage

GlycineCleavageSystem_pantherID.txt

```
GCS_1        PTHR22912:SF151
GCS_2        PTHR11773:SF1
GCS_3        PTHR43757:SF2
```

script.sh

```
#!/bin/bash

species=osedax
xls="$species"_isoform.xls
output="$species"_GlycineCleavageSystem.txt
output2="$species"_GlycineCleavageSystem_count.txt

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
```

```
omo/data/00-ALL_isoforms_annotations/$xls | cut -f 18)
   echo $genes >> $output
   count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
omo/data/00-ALL_isoforms_annotations/$xls | wc -l)
   echo $count >> $output2
done < GlycineCleavageSystem_pantherID.txt

species=oasisia
xls="$species"_isoform.xls
output="$species"_GlycineCleavageSystem.txt
output2="$species"_GlycineCleavageSystem_count.txt

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
omo/data/00-ALL_isoforms_annotations/$xls | cut -f 18)
   echo $genes >> $output
   count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
omo/data/00-ALL_isoforms_annotations/$xls | wc -l)
   echo $count >> $output2
done < GlycineCleavageSystem_pantherID.txt

species=riftia
xls="$species"_isoform.xls
output="$species"_GlycineCleavageSystem.txt
output2="$species"_GlycineCleavageSystem_count.txt

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
omo/data/00-ALL_isoforms_annotations/$xls | cut -f 18)
```

```
   echo $genes >> $output
   count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | wc -l)
   echo $count >> $output2
done < GlycineCleavageSystem_pantherID.txt

species=lamellibrachia
xls="$species"_isoform.xls
output="$species"_GlycineCleavageSystem.txt
output2="$species"_GlycineCleavageSystem_count.txt

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | cut -f 18)
   echo $genes >> $output
   count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | wc -l)
   echo $count >> $output2
done < GlycineCleavageSystem_pantherID.txt

species=paraescarpia
xls="$species"_isoform.xls
output="$species"_GlycineCleavageSystem.txt
output2="$species"_GlycineCleavageSystem_count.txt

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | cut -f 18)
   echo $genes >> $output
```

```
   count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | wc -l)
   echo $count >> $output2
done < GlycineCleavageSystem_pantherID.txt

species=owenia
xls="$species"_isoform.xls
output="$species"_GlycineCleavageSystem.txt
output2="$species"_GlycineCleavageSystem_count.txt

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | cut -f 2)
   echo $genes >> $output
   count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | wc -l)
   echo $count >> $output2
done < GlycineCleavageSystem_pantherID.txt

species=capitella
xls="$species"_isoform.xls
output="$species"_GlycineCleavageSystem.txt
output2="$species"_GlycineCleavageSystem_count.txt

while read line; do
   Panther_ID=$(cut -f 2 <<< "$line")
   genes=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giacomo/data/00-ALL_isoforms_annotations/$xls | cut -f 1)
   echo $genes >> $output
   count=$(fgrep "$Panther_ID" /data/SBCS-MartinDuranLab/03-Giac
```

```
omo/data/00-ALL_isoforms_annotations/$xls | wc -l)
   echo $count >> $output2
done < GlycineCleavageSystem_pantherID.txt
```

```
cut -f 1 GlycineCleavageSystem_pantherID.txt > GlycineCleavageSystem_pantherID_FirstColumn
paste GlycineCleavageSystem_pantherID_FirstColumn owenia_GlycineCleavageSystem_count.txt capitella_GlycineCleavageSystem_count.txt lamellibrachia_GlycineCleavageSystem_count.txt paraescarpia_GlycineCleavageSystem_count.txt oasisia_GlycineCleavageSystem_count.txt riftia_GlycineCleavageSystem_count.txt osedax_GlycineCleavageSystem_count.txt > GlycineCleavageSystem_count_table.txt
echo "step"$'\t'"owenia"$'\t'"capitella"$'\t'"lamellibrachia"$'\t'"paraescarpia"$'\t'"oasisia"$'\t'"riftia"$'\t'"osedax" > header.txt
cat header.txt GlycineCleavageSystem_count_table.txt > GlycineCleavageSystem_count_table_OK.txt
```

Visualisation

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(data.table)
library(gplots)
library(pheatmap)
library(dendextend)
library(factoextra)
library(ComplexHeatmap)
library(RColorBrewer)
library(NbClust)
library(scales)

```

```
A <- read.delim("/Users/giacomo/Desktop/GlycineDegradation/GlycineDegradation_count_table.txt ", row.names=1)

heatmap_color <- colorRampPalette(brewer.pal(n = 7, name = "Reds"))(3)
#column_labels = c("your","labels"),
#row_labels = c("your","labels"))

paletteLength <- 3
pheatmap(A,
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          border_color = NA,
          color = heatmap_color,
          height = 25,
          width = 20)
```

Expression

```
cd /Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/06-Metabolism/01-AA/expression
```

```
while read line; do
echo $line
annotations=$(cut -f 1,2,3 /Users/giacomo/Dropbox/11-Siboglinids/05-PAPER/00-DATA/06-Metabolism/01-AA/expression/osedax_kallisto_tpm_scaled_ok_ok.tsv | fgrep $line)
cat $annotations
kallisto_body=$(cut -f 2 <<< $annotations)
cat $kallisto_body
kallisto_roots=$(cut -f 3 <<< $annotations)
cat $kallisto_roots
```

```
echo $kallisto_body$'\t'$kallisto_roots >> GlycineDegradation_kallisto_tpm_ALLscaled.tsv
done < GlycineDegradation_JUSTgeneIDs_osedax.txt
```

```
cut -f 1 GlycineDegradation_kallisto_tpm_ALLscaled.tsv | sed -e 's/ /\t/g' | cut -f 2,3 > GlycineDegradation_kallisto_tpm_ALLscaled_ONLY.tsv
cut -f 1 GlycineDegradation_geneIDs_osedax.txt > firstColumn
paste firstColumn GlycineDegradation_kallisto_tpm_ALLscaled_ONLY.tsv > FINAL_GlycineDegradation_osedax_kallisto_tpm_ALLscaled.tsv
```

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(data.table)
library(gplots)
library(pheatmap)
library(dendextend)
library(factoextra)
library(ComplexHeatmap)
library(RColorBrewer)
library(NbClust)
library(scales)

#Import data in matrix format
B <- read.delim("/Users/giacomo/Desktop/GlycineDegradation/FINAL_GlycineDegradation_osedax_kallisto_tpm_ALLscaled.tsv", row.names=1)

# To make 0 a different colour
# First create whatever gradient (e.g. RdBu)
```

```
heatmap_color <- colorRampPalette(brewer.pal(n = 7, name = "Reds"))(1000)
heatmap_color[1] <- rgb(1,1,1)
#column_labels = c("your","labels"),
#row_labels = c("your","labels"))

paletteLength <- 1000

pheatmap(B,
         cluster_rows = FALSE,
         cluster_cols = FALSE,
         border_color = NA,
         color = heatmap_color,
         height = 5,
         width = 25)
```

KEGG Metabolism

KofamKoala_universal_v1.sh

```
#!/bin/bash
#$ -wd /data/scratch/btx654/metabolism
#$ -o /data/scratch/btx654/metabolism
#$ -j y
#$ -pe smp 12
#$ -l h_vmem=40G
#$ -l h_rt=72:0:0
#$ -l highmem
```

```

species=$1
NR_proteome="$species".fa
NR_proteome_path=/data/SBCS-MartinDuranLab/03-Giacomo/NR_proteomes/$NR_proteome
output_kofam="$species"_kofam_result.txt

echo "Working on "$species

module load anaconda3
conda activate kofam_env

mkdir $species
cd $species

exec_annotation \
 --profile=/data/SBCS-MartinDuranLab/03-Giacomo/db/kofam/profiles/ \
 --ko-list=/data/SBCS-MartinDuranLab/03-Giacomo/db/kofam/ko_list \
 --cpu=12 \
 --format=mapper \
 --report-unannotated \
 -o $output_kofam \
 $NR_proteome_path
```

```
qsub KofamKoala_universal_v1.sh Ofra
```

select max 10000 sequences to submit to the online kofamkoala

```
awk -v RS='>' 'NR>1 { gsub("\n", ";", $0); sub(";$", "", $0); print ">"$0 }' Ofra.fa | head -n 10000 | tr ',' '\n' | sed "s/;/\n/g" > Ofra_first10000.fa
```

```
awk -v RS='>' 'NR>1 { gsub("\n", ";", $0); sub(";$", "", $0); print ">"$0 }' Ofra.fa | tail -n 8024 | tr ',' '\n' | sed "s/;/\n/g" > Ofra_last8024.fa
```

I got the results of kofamKOALA but they represent less protein compared to KAAS annotated ones: although I just noticed that we managed to annotate just 5522 proteins with kofamkoala whereas we annotated 8514 using KAAS
so I will try to combine both the methods, kofamKOALA first then KAAS

```
awk '{print $2}' riftia_KAAS_custom_SBH.txt > only_KOnumbers.txt
#add a line at the top of this file saying "KO_number"
nano only_KOnumbers.txt
paste riftia_annotation_Dec2020_TrinoPanther.xls only_KOnumbers.txt > riftia_annotation_Jan2021_TrinoPanther.xls

#awk '{print $1"\t"$2}' Ofra_kofam_result.txt > Ofra_kofam_result_ok.txt
cut -f 1 Ofra > Ofra_IDs_KAAS
grep "\sK" Ofra_kofam_result.txt | awk '{print $1"\t"$2}' | sed 's/Ofra_//g' > Ofra_kofam_result_ok.txt
cut -f 1 Ofra_kofam_result_ok.txt > Ofra_IDs_kofam
#select the genes annotated with KAAS which don't have an annotation with kofam
fgrep -vf Ofra_IDs_kofam  Ofra | tail -n +2 > Ofra_KAAS_only
#combine the annotations using kofam as the primary one

cat Ofra_kofam_result_ok.txt Ofra_KAAS_only > Ofra_combined #9324 entries
```

combine.sh

```
#!/bin/bash
```

```
species=$1

grep "\sK" "$species"_kofam_result.txt | awk '{print $1"\t"$2}' | sed "s/$species//g" | sed "s/_//g" > "$species"_kofam_result_ok.txt
cut -f 1 "$species"_kofam_result_ok.txt > "$species"_IDs_kofam
#select the genes annotated with KAAS which don't have an annotation with kofam
fgrep -vf "$species"_IDs_kofam  ../$species | tail -n +2 > "$species"_KAAS_only
#combine the annotations using kofam as the primary one
cat "$species"_kofam_result_ok.txt "$species"_KAAS_only > "$species"_combined
```

```
cat Ofra_combined Rpac_combined Oalv_combined Lluy_combined Pech_combined Ofus_combined Ctel_combined > KOlist_OfraRpacOalvLluyPechOfusCtel_combined.txt
```

reconstruct pathways

RESULTS
