Calliphoridae Phylogeny

This notebook is for the description of the steps taken for the inference of a phylogeny for calliphorid species.

SPECIES (N = 15)

Species	Genome	Transcriptome	Pedro	Vanessa
Agra	X		x	
Bpan	X		x	
Chom	X		x	x
Cvom	X		x	
Lcup	X		x	x
Lser	X		x	x
Pazu	X		x	
Cput		X	x	x
Lexi		X	x	
Cmeg		X	x	x
Calb		X		x
Clop		X		x
Loch		X		x
Cbez		X		x
Cmac		X		x

*Cmac was added for Vanessa's study

0 - Promagrams

Program	Version
TransDecoder	5.7.0

cd-hit	4.8.1
BUSCO	5.4.7
MAFFT	7.505
Trimal	1.4.1
IQTree	2.2.2.6
FigTree	1.4.4

1 - Getting things that were ready before

We used the proteome from the species with genomes, and the transcriptomes from the othe ones. The BUSCO search for the proteomes was ready (Carol did it). We ran a new one for the proteins predicted from the transcriptomes.

```
# From the genomes
mkdir home/cunha/Phylogeny/0-Original_sequences

scp -r -P 2205 pedro@143.107.244.181:/home/blowflies/Gene_families/02-CDS_anc

mkdir home/cunha/Phylogeny/1-BUSCO
scp -r -P 2205 pedro@143.107.244.181:/home/blowflies/Gene_families/04-BUSCO/I

for i in *busco; do mv $i ${i%%_protein_longest_isof.fasta.busco}_busco; done

# From the transcriptomes
cd /home/cunha/Phylogeny/0-Original_sequences

cp /home/cunha/Genotype_Phenotype/3-CDS/*.pep .

for i in *pep; do mv $i ${i%%.fasta.transdecoder.pep}_protein.fasta; done

for i in *trinity*; do sed -i -r 's/ .*//' $i; done
```

OBS.: TransDecoder (Lexi)

We did not have a "proteome" for Lexi, so we ran TransDecoder (the same way that was done for the other transcriptomes)

```
mkdir /home/cunha/Phylogeny/Lexi_transdecoder

cp /home/cunha/03-RNA/02-Trinity/lexi_trinity.fasta .

~/Programs/TransDecoder-TransDecoder-v5.7.0/TransDecoder.LongOrfs -t lexi_tri

cd lexi_trinity.fasta.transdecoder_dir/

cp longest_orfs.pep /home/cunha/Phylogeny/0-Original_sequences/lexi_trinity_p
```

Removing unnecessary parts of the id

```
for i in *; do sed -i -r 's/ .*//' $i; done
```

2 - cdhit (Transcriptomes)

We use this to remove the redundancy from the peptide files that came from the transcriptomes.

```
for i in *trinity*; do cd-hit -i $i -o $i.cdhit -c 0.97 -n 5 -T 5; done

# Renaming
for i in *trinity*fasta; do mv $i $i.NO-CD-HIT; done

for i in *cdhit; do mv $i ${i%%.cdhit}; done
```

3 - BUSCO (Transcriptomes)

```
cd home/cunha/Phylogeny/0-Original_sequences

for i in *trinity*fasta; do docker run -u $(id -u) -v $(pwd):/busco_wd ezlabg

mv *busco ../1-BUSCO
```

4 - Selection of Orthologs

Before running the command line, it is necessary to rename every full_table file with the first 3 letters or numbers of the respective sample

> Step 1

This command will select every Complete gene in BUSCO within the full_table file and assemble a new table with them.

```
cd /home/cunha/Phylogeny/1-BUSCO

# create-complete-txt.sh
for d in *_busco/
do
    echo "$d"
    cd "$d"
    cd run*/
        for i in $(find . -name full*)
            do grep Complete $i >> ${i%%.tsv}_complete.txt
            done
    cd ~/Phylogeny/1-BUSCO
done
```

> Step 2

Add the first 4 letters or numbers of the samples in each line of the complete genes

```
for i in $(find . -name *_complete.txt); do awk '{print
substr(FILENAME,3,4),$0}' $i > ${i%%.txt}_named.txt; done
```

> Step 3.1 (TRANSCRIPTOMA)

Add the samples names in the trinity files

```
cd /home/cunha/Phylogeny/0-Original_sequences

# name-in-trinity.sh
BEGIN {FS="\>"}
{
        if ($2 ~ /TRINITY_/) {
                print ">" substr(FILENAME,1,4) "_" $2
        } else {
                print $0
        }
}
```

And run the following command line

```
for i in *trinity*fasta ; do awk -f name-in-trinity.sh $i >>
/home/cunha/Phylogeny/2-Named_oneline_sequences/${i%%.fasta}_named.fasta; done
```

> Step 3.2 (GENOMAS)

Add the samples names in the genome files:

```
ls *_gen_* | sed "s/_gen_protein.fasta//g" > list.txt
```

Create file name_in_genome.sh with this code:

```
while read file
    do
    sed "s/>/>$file/g" ${file}_gen_protein.fasta >> /home/cunha/Phylogeny/2-
Named_oneline_sequences/${file}_gen_protein_named.fasta
done < list.txt
```

> Step 4

Transforming sequences in one line files

```
cd /home/cunha/Phylogeny/2-Named_oneline_sequences

for file in *; do awk '/^>/ {printf("\n%s\n",$0);next; } { printf("%s",$0);}

for i in *oneline*; do sed -i -e 1d $i; done #remove first line

rm *named.fasta
```

> Step 5 (do in screen because it takes a while)

Select in the Hmmer output table the busco genes with evalue 0 or less than e^-100

```
cd /home/cunha/Phylogeny/1-BUSCO

for i in *_busco; do echo $i; done > dir.txt

# hmmr.sh (evalue < e-100)
while read file
do
    SPP=$(echo $file | sed -r 's/_.*//')

    cd /home/cunha/Phylogeny/1-BUSCO/$file/run_diptera_odb10/hmmer_output/in:

    for i in *.out*; do awk -v SPP2=$SPP '(NR == 4 && ($7 ~ "[0-9].[0-9]e-[0-

     cd /home/cunha/Phylogeny/1-BUSCO

done < dir.txt

mv *evalue* ../3-evalues/
```

> Step 7
> Assembling the final fasta files

```
cd /home/cunha/Phylogeny

# find - finds the full table with complete buscos with the species names
for i in $(find . -name *_complete_named.txt); do awk '{print $1 ".*" $4 ".*'

#find the genes with the evalue we selected
for i in $(find . -name *_grepfile.txt); do grep -hof $i $(find . -name *eval

# The files above appeared in the busco folders for each species

# This writes the command line to assemble the fasta files
cd /home/cunha/Phylogeny/2-Named_oneline_sequences

for i in $(find .. -name *_final.txt); do awk '{print "grep -h -A1 " $1 ".*"

screen
sh assemble_fasta_files.sh

mv *fas ../4-Orthologs_fasta
```

> Step 8

Filtering the orthogroups

```
cd /home/cunha/Phylogeny/4-Orthologs_fasta

mkdir All_ortho
mv *fas All_ortho/

mkdir min_7
mkdir min_8
mkdir min_9
mkdir min_10
mkdir min_11

#for Cmac
mkdir min_12

# filter.sh - filter by the number of sequences/species
for i in All_ortho/*
do
    n=$(grep -c ">" $i)

    if [ $n -gt 11 ]
    then
        cp $i min_12
    fi
done
```

- min_12: XXXX (cov 75%)

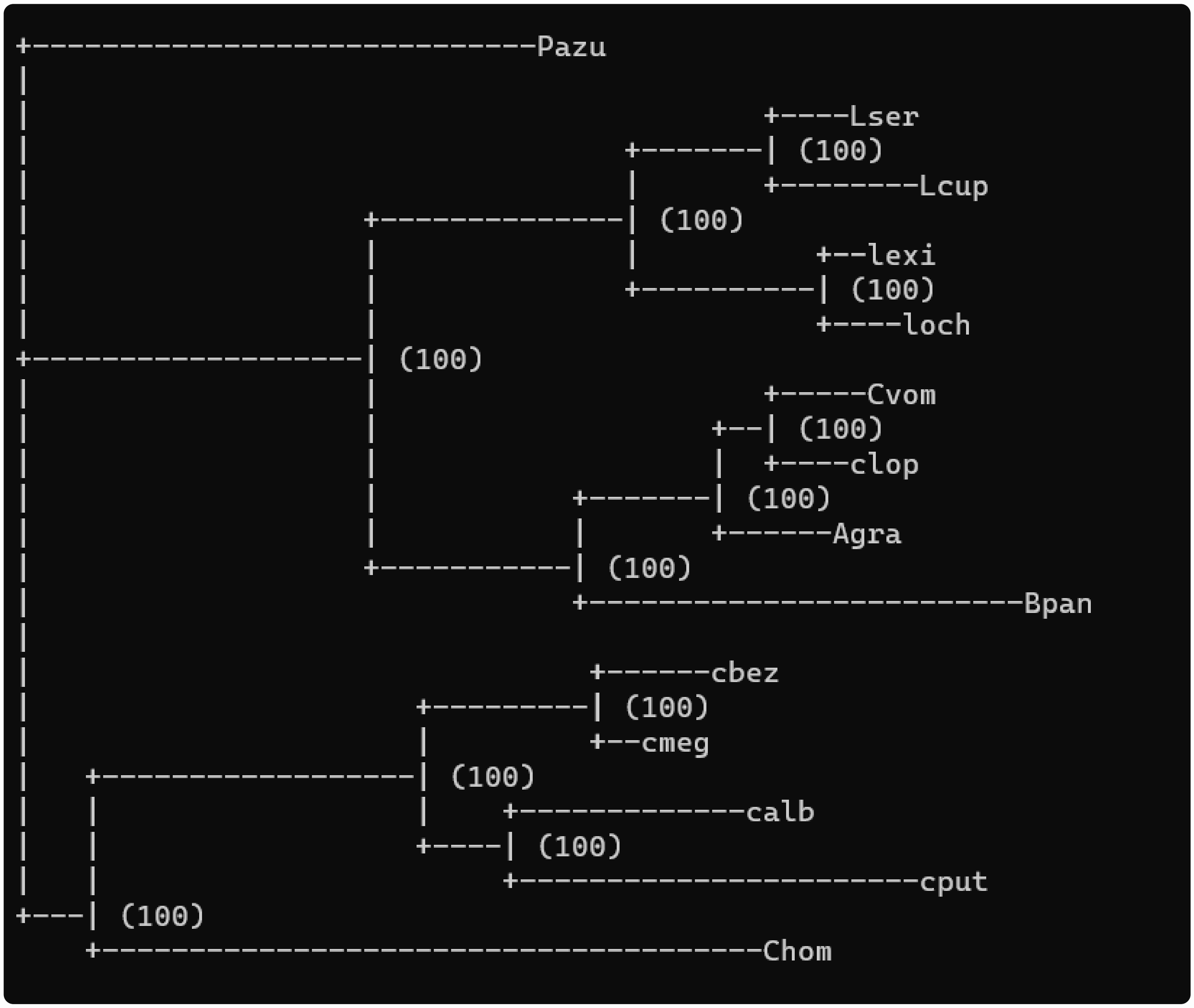
We will infer the tree with the 75% cov.

5 - Alignment

```
mkdir /home/cunha/Phylogeny/5-Alignments

cd /home/cunha/Phylogeny/4-Orthologs_fasta/min_11/

# alignment
for i in *fas; do mafft-ginsi --thread 25  --maxiterate 1000 --adjustdirecti

# cleaning the alignment
cd /home/cunha/Phylogeny/5-Alignments

for i in *fas; do trimal -in $i -out ${i%%.fas}_trim.fas -automated1; done

# fixing the sequence ids (leave just the species name)
for i in *trim*; do cat $i | sed 's/>Pazu.*/>Pazu/g' | sed 's/>Lser.*/>Lser/

mkdir 1-Mafft 2-Trimmed 3-Renamed
mv *renamed.fas 3-Renamed/
mv *trim.fas 2-Trimmed/
mv *aln.fas 1-Mafft/
```

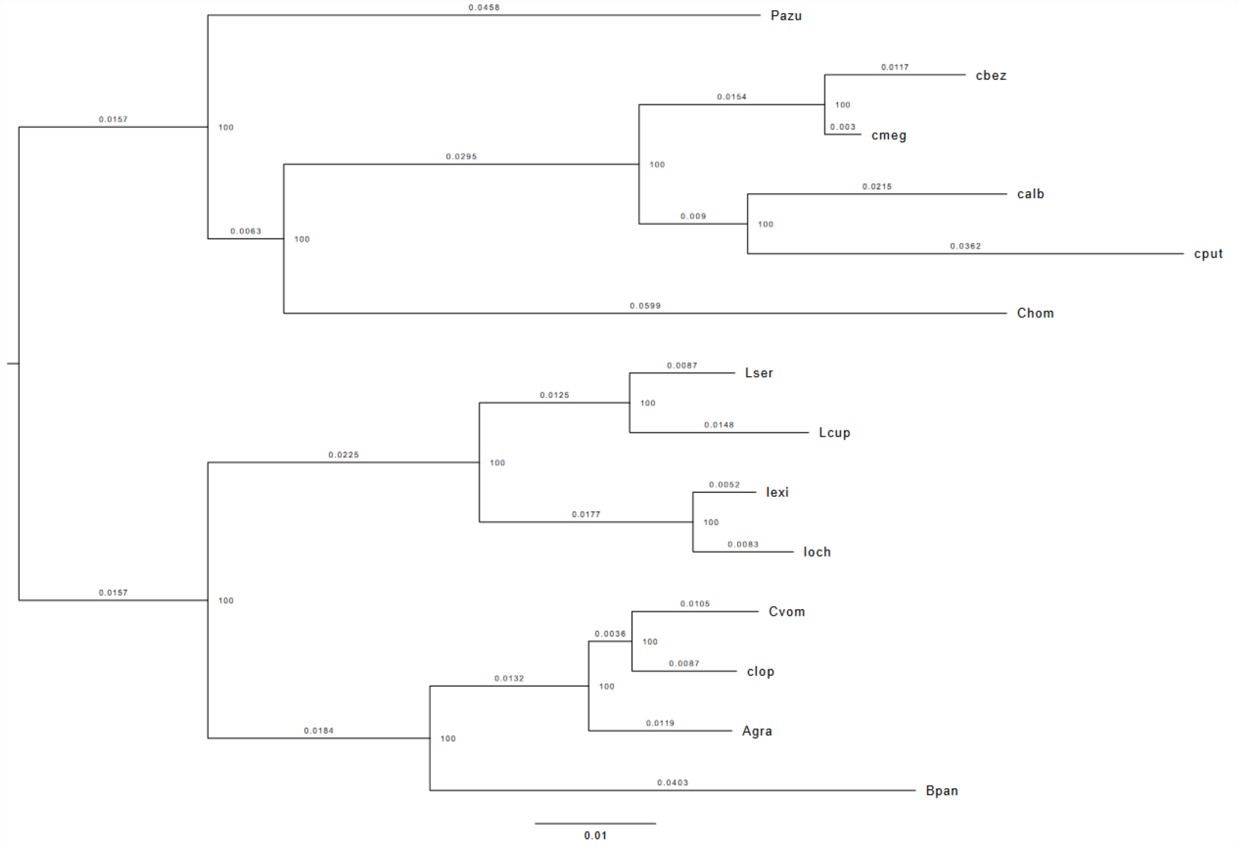
6 - Tree inference

```
cd /home/cunha/Phylogeny/5-Alignments

iqtree2 -T 30 -s /home/cunha/Phylogeny/5-Alignments/3-Renamed/ -m MFP -bb 10

# the output files were named like the directory, so it's better to rename t
for i in 3-Renamed.*; do mv $i Calliphoridae_tree.${i##3-Renamed.}; done

mv Calliphoridae* /home/cunha/Phylogeny/6-IQ-TREE
```



```
+-----------------------------Pazu
|
|                                          +----Lser
|                                  +-------| (100)
|                                  |       +--------Lcup
|                   +--------------| (100)
|                   |              |          +--lexi
|                   |              +----------| (100)
|                   |                         +----loch
+-------------------| (100)
|                   |                      +-----Cvom
|                   |                   +--| (100)
|                   |                   |  +----clop
|                   |           +-------| (100)
|                   |           |       +------Agra
|                   +-----------| (100)
|                               +-------------------------Bpan
|
|                                +------cbez
|                      +---------| (100)
|                      |         +--cmeg
|   +------------------| (100)
|   |                  |    +-------------calb
|   |                  +----| (100)
|   |                       +-----------------------cput
+---| (100)
    +--------------------------------------Chom
```

The tree matches the Dimensions phylogeny and a recently published one as well (Yan et al., 2021), with Chrysominae as a sister clade to Calliphorinae + Luciliinae. However, IQTREE infers an unrooted tree, so we rooted it in a way that the Pazu branch forms a clade with the other Chrysominae, using FigTree



Trees are here:

```
# Unrooted
/home/cunha/Phylogeny/6-IQ-TREE/Calliphoridae_tree.treefile
# Rooted
/home/cunha/Phylogeny/6-IQ-TREE/Calliphoridae_tree_root.tre
```