Log: Genome Annotation Chrysomya putoria

tags: `Genome Annotation` `Cput`

Table of Contents

List of software (and where they are located in Rosalind):

- QUAST (/usr/local/bin/quast.py)
- RepeatModuler (/RepeatModeler-2.0.4)
- RepeatMasker(/dados/home/pedro/Programs/RepeatMasker/RepeatMasker)
- STAR (/dados/home/bruno/anaconda3/bin/STAR)
- samtools (/dados/home/bruno/anaconda3/bin/samtools /dados/home/bruno/anaconda3/bin/samtools.pl)
- Braker (docker image ID: 7772eca57cee)
    - https://hub.docker.com/r/teambraker/braker3
- TSEBRA (/)
    - https://github.com/Gaius-Augustus/TSEBRA

- AUGUSTUS (docker image ID: c0dfd27799fc)
  - https://github.com/Gaius-Augustus/Augustus
- busco (/usr/local/bin/busco)
- gFACS (/)
  - https://gitlab.com/PlantGenomicsLab/gFACs
- EnTAP (/)
  - https://entap.readthedocs.io/en/v0.8.0-beta/introduction.html

Installation for every program steps can be found in the *Chrysomya megacephala* log (https://hackmd.io/ziOztK1MQZecVm0_qDAr1g)

Genome annotation workflow:

- [x] 1- BUSCO
- [x] 2- QUAST
- [x] 3- Mithocondrial genome
- [x] 4- RepeatModeler
- [ ] 5- RepeatMasker
- [x] 6- RNA-seq
- [x] 7- Trinity
- [x] 8- STAR
- [ ] 9- Braker

https://github.com/CBC-UCONN/Genome_Assembly
https://github.com/CBC-UCONN/Structural-Annotation

Genome sequencing

We sequenced a pool of males at Dovetail (Pac-Bio).

ANOTAR AQUI DADOS DO SEQUENCIAMENTO

Number of reads 2639925
Coverage (x) 58
HPA Length (bp) 1287183393
HPA N50 (bp) 1685985
HPA N90 (bp) 71006

HPA L50 168
HPA L90 1953
FA Length (bp) 587066816
FA N50 (bp) 4322519
FA N90 (bp) 590301
FA L50 45
FA L90 179

bp = Base pair; HPA = Hifiasm (Cheng et al., 2021) primary assembly; FA = Final assembly. N50 = Sequence length of the smallest contig within those that sum up to 50% of the total genome's length; N90 =Sequence length of the smallest contig within those that sum up to 90% of the total genome's length; L50 = Smallest sequence number that together sum 50% of the total genome's length; L90 = Smallest sequence number that together sum 90% of the total genome's length.

0- Copy of the genome

We made an extra copy of the genome and the PacBio reads just to be safe.

```
# we are here
/home/blowflies/genome_annotation/cput

# new directory
mkdir 0-genome
cd 0-genome

cp /home/Reference_genomes/Cputoria/purged.fa .
cp /home/Reference_genomes/Cputoria/XDOVE_20221110_S64411e_PL100270437-1_D01
```

1-BUSCO first

Busco version 5.3.2

```
mkdir /home/blowflies/genome_annotation/cput/1-busco_first
cd /home/blowflies/genome_annotation/cput/1-busco_first

#sudo docker pull ezlabgva/busco:v5.4.4_cv1

cp ../0-genome/cput.fa .

sudo docker run -u $(id -u) -v $(pwd):/busco_wd ezlabgva/busco:v5.4.4_cv1 bus
```

```
---------------------------------------------------
|Results from dataset diptera_odb10               |
---------------------------------------------------
|C:98.6%[S:93.3%,D:5.3%],F:0.3%,M:1.1%,n:3285     |
|3241   Complete BUSCOs (C)                       |
|3066   Complete and single-copy BUSCOs (S)       |
|175    Complete and duplicated BUSCOs (D)        |
|10     Fragmented BUSCOs (F)                     |
|34     Missing BUSCOs (M)                        |
|3285   Total BUSCO groups searched               |
---------------------------------------------------
```

2- QUAST

```
# we are here
/home/blowflies/genome_annotation/cput

# new directory
mkdir 2-quast
cd 2-quast

# running
sudo quast.py ../0-genome/purged.fa -t 10 --eukaryote --large --rna-finding -
```

3- Mitochondrial genome

Cput mitochondrial genome on NCBI

https://www.ncbi.nlm.nih.gov/nuccore/AF352790.1

search terms: "chrysomya putoria[ORGN] AND mitochondrial genome"

AF352790.1

```
esearch -db nuccore -query AF352790.1 | efetch -format fasta > AF352790.1.fa
cd home/blowflies/genome_annotation/calb/3-mithocondrial_genome
mkdir 1-blastn
```

We renamed the directory

```
mv 3-mithocondrial_genome/ 3-mitochondrial_genome
```

Blastn

version 2.11

```
cd home/blowflies/genome_annotation/cput/3-mitochondrial_genome/1-blastn
#Making database using the genome
makeblastdb -in ../AF352790.1.fa -dbtype nucl -out cput_mit_database
#Running
blastn -task blastn -evalue 0.00001 -db ./cput_mit_database -query ../../0-ge
```

The best aligment was against scaffold 1221, which is a very big scaffold (0.5Mb). We isolated this scaffold and cut it in three pieces "before mitochondria" - "mitochondria" - "after mitochondria"

```
cd /home/blowflies/genome_annotation/cput/0-genome

# Remove scaffold from genome
seqkit grep -v -p "ptg001221l" cput.fa > cput_N_genome.fa

# Isolate scaffold
seqkit grep -p "ptg001221l" cput.fa > ptg001221l.fa

#Isolating the mitochondria
seqkit subseq ptg001221l.fa -r 563637:579472 > cput_mit.fa

#Removing the mitochondria and breaking the scaffold in two
seqkit  subseq ptg001221l.fa  -r 1:553681 > cput_ptg1221l_1.fa
seqkit  subseq ptg001221l.fa  -r 583500:585720 > cput_ptg1221l_2.fa

#Renaming scaffold ids
sed -i 's/>ptg001221l/>ptg001221l_1/' cput_ptg1221l_1.fa
sed -i 's/>ptg001221l/>ptg001221l_2/' cput_ptg1221l_2.fa
sed -i 's/>ptg001221l/>cput_mit/' cput_mit.fa

# Joining scaffolds to genome again
cat cput_N_genome.fa cput_ptg1221l_1.fa cput_ptg1221l_2.fa > cput_N_genome_fi

#Checking
grep ">ptg001221l" cput_N_genome_final.fa
grep -c ">" cput_N_genome_final.fa
```

MITOS Web Server

We ran the mithocondrial genome annotation using MITOS2 web server with all the default parameters but the genetic code, which was specified to be the invertebrate one.

We then downloaded the output files to a local computer and sent them to 2-MITOS_results (/home/blowflies/genome_annotation/cput/3-mitochondrial_genome/2-MITOS_results).

The job settings were:

Job ID: cput

Property	Value
Reference	RefSeq 63 Metazoa
Genetic Code	5
Proteins	True
tRNAs	True
rRNAs	True
OH	True
OL	True
Circular	True
Use Al Arab et al.	False
E-value Exponent	2.0
Final Maximum Overlap	50nt
Fragment Quality Factor	100.0
Standard Code	False
Cutoff	50.0%
Clipping Factor	10.0
Fragment Overlap	20.0%
Local only	True
Sensitive only	False
ncRNA overlap:	50 nt

TEST on pacbio genome before assembly

```
/home/blowflies/genome_annotation/cput/3-mitochondrial_genome
mkdir 2-blast_pacbio
# we copied the fastq with the raw pacbio sequences to 2-blast_pacbio
gzip -d cput_raw_genome.fastq.gz
blastn -task blastn
#converting fastq to fasta
seqkit fq2fa cmeg_raw_genome.fastq -o cmeg_raw_genome.fa
#blastn
blastn -task blastn -evalue 0.00001 -db ../1-blastn/cput_mit_database -query
```

4- RepeatModeler

Installing with docker

It was complicated installing all the programs, so we used a Docker container

```
#Always use this before repeat modeler
docker run -it --rm dfam/tetools:latest

container-ID: 918a311e45cb
container-name: goofy_buck
container-image: dfam/tetools:latest

#To attach the container and continue running press CTRL+P, then CTRL+Q
```

Creating database

```
# Moving the fasta file to docker container from the Rosalind server using:
# docker cp file.txt container-name:/path/to/copy/file.txt
docker cp ./cput_N_genome_final.fa goofy_buck:/home

# Getting inside the container
docker exec -it goofy_buck /bin/bash

#Database
BuildDatabase -name cput_database cput_N_genome_final.fa
```

Running

```
RepeatModeler -database cput_database -threads 20 -LTRStruct  >log 2>err

#Then to exit the container and continue running press CTRL+C
```

Copying the results

```
# Compressing files inside docker
tar -cjvf cput_other_files.tar.gz *

# now here
cd /home/cunha/01-RepeatModeler/cput

docker cp -a goofy_buck:/home/cput_database-families.stk ./
docker cp -a goofy_buck:/home/cput_database-families.fa ./
docker cp -a goofy_buck:/home/cput_database-rmod.log ./
docker cp -a goofy_buck:/home/log ./
docker cp -a goofy_buck:/home/err ./
docker cp -a goofy_buck:/home/cput_other_files.tar.gz ./
```

Test - LTR Harvest

First, we split the genome, to see if it works better

```
cd /home/blowflies/genome_annotation/cput/0-genome

seqkit split2 -p 5 cput_N_genome_final.fa

cd cput_N_genome_final.fa.split/

for i in *; do mv $i cput_${i##cput_N_genome_final.part_00}; done
```

Now, running LTRHarvest

```
# Doing it on Rosalind
# Container: cranky_haibt

cd /home/blowflies/genome_annotation/cput/0-genome/cput_N_genome_final.fa.spl

for i in *; do docker cp $i cranky_haibt:/home; done

docker exec -it cranky_haibt /bin/bash

cd home/
mkdir 1 2 3 4 5
mv *1.fa 1/
mv *2.fa 2/
mv *3.fa 3/
mv *4.fa 4/
mv *5.fa 5/

# LTR.sh
cd /home/1
/opt/genometools/bin/gt suffixerator -db cput_1.fa -indexname cput_1_db -tis
/opt/genometools/bin/gt ltrharvest -index cput_1_db -minlenltr 100 -maxlenltr
/opt/LTR_retriever/LTR_retriever -genome cput_1.fa -inharvest cput_1.harvest

cd /home/2
/opt/genometools/bin/gt suffixerator -db cput_2.fa -indexname cput_2_db -tis
/opt/genometools/bin/gt ltrharvest -index cput_2_db -minlenltr 100 -maxlenltr
/opt/LTR_retriever/LTR_retriever -genome cput_2.fa -inharvest cput_2.harvest

cd /home/3
/opt/genometools/bin/gt suffixerator -db cput_3.fa -indexname cput_3_db -tis
/opt/genometools/bin/gt ltrharvest -index cput_3_db -minlenltr 100 -maxlenltr
/opt/LTR_retriever/LTR_retriever -genome cput_3.fa -inharvest cput_3.harvest

cd /home/4
/opt/genometools/bin/gt suffixerator -db cput_4.fa -indexname cput_4_db -tis
/opt/genometools/bin/gt ltrharvest -index cput_4_db -minlenltr 100 -maxlenltr
/opt/LTR_retriever/LTR_retriever -genome cput_4.fa -inharvest cput_4.harvest

cd /home/5
/opt/genometools/bin/gt suffixerator -db cput_5.fa -indexname cput_5_db -tis
/opt/genometools/bin/gt ltrharvest -index cput_5_db -minlenltr 100 -maxlenltr
/opt/LTR_retriever/LTR_retriever -genome cput_5.fa -inharvest cput_5.harvest

cd /home
```

It worked! Now we need to download the final files back from the container, concatenate them with the output from the previous RepeatModeler run, and then use this final file to mask the genome

```
cd /home/blowflies/genome_annotation/cput/4-RepeatModeler
# cput_database-families.fa is here

docker cp -a cranky_haibt:/home/1/cput_1.fa.LTRlib.fa ./
docker cp -a cranky_haibt:/home/2/cput_2.fa.LTRlib.fa ./
docker cp -a cranky_haibt:/home/3/cput_3.fa.LTRlib.fa ./
docker cp -a cranky_haibt:/home/4/cput_4.fa.LTRlib.fa ./
docker cp -a cranky_haibt:/home/5/cput_5.fa.LTRlib.fa ./

docker cp -a cranky_haibt:/home/cput_LTR.tar.gz ./ # other outputs

docker stop cranky_haibt # kill container

cat *fa > cput_modeler_complete.fa

# remove redundancy
vsearch --cluster_fast cput_modeler_complete.fa -id 0.80 -threads 15 -centroi
```

5- RepeatMasker

Installing steps can be found in the *Chrysomya megacephala* log (https://hackmd.io/ziOztK1MQZecVm0_qDAr1g)

##Doing it on Rosalind

```
# from Darwin to Rosalind
cd /home/cunha/01-RepeatModeler/cput

scp -P 2205 * pedro@143.107.244.181:/home/blowflies/genome_annotation/cput/4-

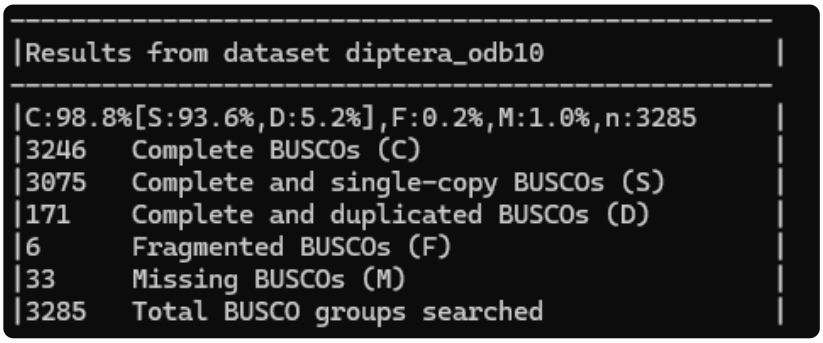
# running
cd /home/blowflies/genome_annotation/cput/5-RepeatMasker

RepeatMasker -lib /home/blowflies/genome_annotation/cput/4-RepeatModeler/cput
```

2.5.1 BUSCO

```
cd /home/blowflies/genome_annotation/cput/5-RepeatMasker
mkdir 1-BUSCO
cp cput_N_genome_final.fa.masked 1-BUSCO/
cd 1-BUSCO
sudo docker run -u $(id -u) -v $(pwd):/busco_wd ezlabgva/busco:v5.4.7_cv1 bus
```

#Results:



2.5.2 QUAST

```
#comparing with the file /home/blowflies/genome_annotation/cput/0-genome/cput
cd /home/blowflies/genome_annotation/cmeg/5-RepeatMasker
mkdir 2-QUAST
cp cput_N_genome_final.fa.masked 2-QUAST/
cd 2-QUAST
mkdir cput.fa
cd cput.fa
sudo quast.py ../cput_N_genome_final.fa.masked -t 20 --eukaryote --large --rr
```

```
#Results:
cd /home/blowflies/genome_annotation/cput/5-RepeatMasker/2-QUAST/cput.fa/quas
cat report.txt
```

Assembly	cput_N_genome_final.fa.maske
# contigs (>= 0 bp)	653
# contigs (>= 1000 bp)	653
# contigs (>= 5000 bp)	652
# contigs (>= 10000 bp)	650
# contigs (>= 25000 bp)	551
# contigs (>= 50000 bp)	446
Total length (>= 0 bp)	587036998
Total length (>= 1000 bp)	587036998
Total length (>= 5000 bp)	587034777
Total length (>= 10000 bp)	587016365
Total length (>= 25000 bp)	585212355
Total length (>= 50000 bp)	581400726
# contigs	652
Largest contig	15138138
Total length	587034777
Estimated reference length	500000000
GC (%)	28.98
N50	4322519
NG50	4997734
N75	1939160
NG75	2854789
L50	45
LG50	35
L75	95
LG75	68
# total reads	768
# left	0
# right	0
Mapped (%)	100.0
Properly paired (%)	0.0
Avg. coverage depth	1
Coverage >= 1x (%)	100.0
# N's per 100 kbp	0.07
# predicted rRNA genes	238 + 31 part

6- RNA-seq

We extracted RNA from:

- 50 eggs
- 10 L1
- 5 L2
- 2 L3
- 1 pupae
- 1 virgin female
- 1 gravid female
- 1 male

Then, we pooled all the samples (2ug of RNA from each sample) and sequenced it.

- RNAseq Illumina 20M reads paired end PE150 Q30>85%

```
#Coping files
cp -r /home/Raw_seqs/cput_pool_RNA /home/blowflies/genome_annotation/cput/
#Renaming
cd /home/blowflies/genome_annotation/cput/
mv cput_pool_RNA 8-cput_pool_RNA
#checking md5
cd 8-cput_pool_RNA/
cat MD5.txt
#8ecb8a84268de8259d50a885c316d6a9  Cput_1.fq.gz
#d6c794f0578c674b16f3f65c5da3933b  Cput_2.fq.gz
md5sum Cput*
#8ecb8a84268de8259d50a885c316d6a9  Cput_1.fq.gz
#d6c794f0578c674b16f3f65c5da3933b  Cput_2.fq.gz
mkdir 0-raw_reads
mv Cput* 0-raw_reads
mv MD5.txt 0-raw_reads
```

Quality control of raw reads

We ran FastQC and, then MultiQC.

> Don't need to unzip raw read files because fastqc can cope with zipped files (.gz).

FastQC will will process one sample at a time and give you an output report for each sample separately. MultiQC will combine all the outputs from FastQC analysis and give you one QC report for all processed samples, making them more easily comparable.
–> nice webpage on fastqc and multiqc: https://stab.st-andrews.ac.uk/wiki/index.php/FASTQC_and_MultiQC
–> https://multiqc.info/

```
cd /home/blowflies/genome_annotation/cput/8-cput_pool_RNA
mkdir 1-QC
cd 0-raw_reads
fastqc *fq.gz #v0.11.9
multiqc . #Version 1.11
mv *.html ../1-QC
mv *.zip ../1-QC
mv multiqc_data ../1-QC
```

Coping multiqc report to a local computer

```
scp -P 2205 vanessa@143.107.244.181:/home/blowflies/genome_annotation/cput/8-cput_pool_RNA/1-QC/multiqc_report.html /mnt/c/Users/vansc/Downloads
```

Results:
https://drive.google.com/file/d/1s44TfwIPpu2_-2_vLqoiHMEY9jyUq3-I/view?usp=share_link

Trimming

Processing raw reads to trimming (remove only bad quality bases).

I used Trimmometic to trimming version 0.39
–> nice webpage on how to use Trimmomatics: http://www.usadellab.org/cms/index.php?page=trimmomatic
https://datacarpentry.org/wrangling-genomics/03-trimming/

```
cd /home/blowflies/genome_annotation/cput/8-cput_pool_RNA/0-raw-reads
mkdir ../2-trimming

TrimmomaticPE Cput_1.fq.gz Cput_2.fq.gz -threads 8 -baseout
/home/blowflies/genome_annotation/cput/8-cput_pool_RNA/2-
trimming/cput.trimmed.fq.gz ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10
SLIDINGWINDOW:4:15 MINLEN:36
```

Quality control of trimmed reads

```
# in this directory -> /home/blowflies/genome_annotation/cput/4-cput_pool_RNA/2-tr
# QC
fastqc *.gz

# here -> /home/blowflies/genome_annotation/calb/4-calb_pool_RNA/1-QC
mkdir trimmed_reads_qc

# in this directory -> /home/blowflies/genome_annotation/cput/4-cput_pool_RNA/2-tr
mv *.html ../1-QC/trimmed_reads_qc/
mv *.zip ../1-QC/trimmed_reads_qc/
multiqc .
```

Observation

We renamed the genome file from this step onwards.

```
cd /home/blowflies/genome_annotation/cput/0-genome

mv purged.fa cput.fa
```

6.1 - New RNA sequencing

As we did not have a great representation of BUSCO genes and low quality in general in our previous sequencing, we did it again. The new data can be found at:

```
/home/blowflies/genome_annotation/cput/8-cput_pool_RNA/new_seq
```

Quality control of raw reads

We ran FastQC and, then MultiQC.

> Don't need to unzip raw read files because fastqc can cope with zipped files (.gz).

FastQC will will process one sample at a time and give you an output report for each sample separately. MultiQC will combine all the outputs from FastQC analysis and give you one QC report for all processed samples, making them more easily comparable.
–> nice webpage on fastqc and multiqc: https://stab.st-andrews.ac.uk/wiki/index.php/FASTQC_and_MultiQC
–> https://multiqc.info/

```
cd /home/blowflies/genome_annotation/cput/8-cput_pool_RNA/new_seq
mkdir 1-QC
cd 0-raw_reads
fastqc *fastq.gz #v0.11.9
multiqc . #Version 1.11
mv *.html ../1-QC
mv *.zip ../1-QC
mv multiqc_data ../1-QC
```

Coping multiqc report to a local computer

```
scp -P 2205 diniz@143.107.244.181:/home/blowflies/genome_annotation/cput/8-cput_pool_RNA/new_seq/1-QC/multiqc_report.html /Users/diniz/Desktop/
```

Results:
https://drive.google.com/drive/folders/1g_VW5o6HiXAKChcWKe_GDxqSjqAMdGAn

Trimming

Processing raw reads to trimming (remove only bad quality bases).

I used Trimmometic to trimming version 0.39
–> nice webpage on how to use Trimmomatics: http://www.usadellab.org/cms/index.php?page=trimmomatic
https://datacarpentry.org/wrangling-genomics/03-trimming/

```
cd /home/blowflies/genome_annotation/cput/8-cput_pool_RNA/0-raw-reads
mkdir ../2-trimming

TrimmomaticPE Cupt1_NGS629_S31_L001_R1_001.fastq.gz
Cupt1_NGS629_S31_L001_R2_001.fastq.gz -threads 8 -baseout
/home/blowflies/genome_annotation/cput/8-cput_pool_RNA/new_seq/2-
```

```
trimming/cput.trimmed.fq.gz ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10
SLIDINGWINDOW:4:15 MINLEN:36
```

Quality control of trimmed reads

```
# in this directory -> /home/blowflies/genome_annotation/cput/4-cput_pool_RNA/2-tr
# QC
fastqc *.gz

# here -> /home/blowflies/genome_annotation/calb/4-calb_pool_RNA/1-QC
mkdir trimmed_reads_qc

# in this directory -> /home/blowflies/genome_annotation/cput/4-cput_pool_RNA/2-tr
mv *.html ../1-QC/trimmed_reads_qc/
mv *.zip ../1-QC/trimmed_reads_qc/
mv multiqc_data ../1-QC/trimmed_reads_qc/
multiqc .
```

Assembling the transcriptome ("old + new" RNA-seq)

In /home/cunha/03-RNA/01-Reads/cput/new_seq

```
#!/bin/bash

#SBATCH --job-name trinity ## nome que aparecerá na fila
#SBATCH --output trinity_calb.out ## nome do arquivo de saída; o %j é igual a
#SBATCH --ntasks=1 ## número de tarefas (análises) a serem executadas
#SBATCH --cpus-per-task=20 ## o número de threads alocados para cada tarefa
#SBATCH --mem-per-cpu=1000M # memória por núcleo da CPU
#SBATCH --partition=long ## as partições a serem executadas (separadas por vi
#SBATCH --time=10-00:00:00 ## hora para análise (dia-hora:min:seg)
#SBATCH --error=err

srun docker run --rm -v`pwd`:`pwd` trinityrnaseq/trinityrnaseq Trinity --seqT
```

BUSCO of the new transcriptome

In rosalind (Darwin was not working) /home/diniz/cput_all_busco

```
docker run -u $(id -u) -v $(pwd):/busco_wd ezlabgva/busco:v5.4.7_cv1 busco -i
```

7- Trinity

Transcriptome assembly

Moving trimmed reads to Darwin

```
mkdir /home/cunha/03-RNA/01-Reads/cmeg

scp -P 4988 /home/blowflies/genome_annotation/cput/8-cput_pool_RNA/2-trimming
```

Assembling the transcriptome

```
cd /home/cunha/03-RNA/01-Reads/cput

docker run --rm -v `pwd`:`pwd` trinityrnaseq/trinityrnaseq Trinity --seqType

mv cput_trinity.* /home/cunha/03-RNA/02-Trinity/cput_trinity

# Transcriptome size
grep -c ">" cput_trinity.Trinity.fasta # 28967
```

BUSCO (transcriptome quality)

```
mkdir /home/cunha/03-RNA/03-Busco

docker pull ezlabgva/busco:v5.4.4_cv1 # just because we didn't have busco on

# we copied all transcriptomes in this directory and ran everything at once
for i in *; do docker run -u $(id -u) -v $(pwd):/busco_wd ezlabgva/busco:v5.4
```

Transcriptome	Complete (all)	Complete Single	Complete Dup.	Fragmented	Missing
only old	24.9	17.7	7.2	9.1	66
only new	67.7	50	17.7	8.4	23.9
old + new	69.4	45.1	24.3	8.5	22.1

8- STAR - RNAseq alignment

index

> All genome FASTA files **cannot** be zipped

```
# unziping files
cd
gzip -dk *P*

# We need to create a directory where the genome indexes will be stored befo
cd /home/blowflies/genome_annotation/cput
mkdir 9-STAR
chmod 777 9-STAR
cd 9-STAR
mkdir star_index
chmod 777 star_index

STAR --runThreadN 20 --runMode genomeGenerate --genomeDir /home/blowflies/ge
```

Before the alignment itself we had to concatenate all the fastq files available. The files are in Rosalind (/home/blowflies/genome_annotation/cput/8-cput_pool_RNA/2-trimming/all)

```
cat *_1P.fq.gz > cput_all_1P.fq.gz
cat *_2P.fq.gz > cput_all_2P.fq.gz
```

STAR aligment

```
cd /home/blowflies/genome_annotation/cput/8-cput_pool_RNA/all

for i in *_1P.fq.gz; do STAR --runMode alignReads --readFilesCommand zcat --
```

9- Braker3

All information is in the Calb file (https://hackmd.io/KMzfBC2aQ9qliTy11f_QhQ)

Final outputs are here: /home/blowflies/genome_annotation/cput/10-BRAKER3

BRAKER3 second run

```
# need to run first
export BRAKER_SIF=/home/diniz/programs/braker/braker3.sif

#We copied the file 10-BRAKER3 to the braker file
cd /home/diniz/programs/braker/
mkdir Cput_2
cp /home/blowflies/genome_annotation/cput/10-BRAKER3 /home/diniz/programs/bra
cd /home/diniz/programs/braker/Cput_2/10-BRAKER3
mkdir restart
# Run
singularity exec /home/diniz/programs/braker/braker3.sif braker.pl --genome=/
```

Total transcripts: 21313

Busco evaluation with protein sequences from restart Braker run

```
cd /home/diniz/programs/braker/Cput_2/10-BRAKER3/restart
mv braker.gtf cput_braker.gtf
mv brker.aa cput_braker.aa

docker run -u $(id -u) -v $(pwd):/busco_wd ezlabgva/busco:v5.4.4_cv1 busco -i
```

```
# BUSCO version is: 5.4.4
# The lineage dataset is: diptera_odb10 (Creation date: 2020-08-05, number of genc
# Summarized benchmarking in BUSCO notation for file /busco_wd/cput_braker.aa
# BUSCO was run in mode: proteins

        ***** Results: *****

        C:97.7%[S:83.7%,D:14.0%],F:0.9%,M:1.4%,n:3285
        3211    Complete BUSCOs (C)
        2750    Complete and single-copy BUSCOs (S)
        461     Complete and duplicated BUSCOs (D)
        29      Fragmented BUSCOs (F)
        45      Missing BUSCOs (M)
        3285    Total BUSCO groups searched
```

11- EnTAP

We did it on Darwin using the output from the restart Bracker run

```
#We copied the aminoacid file from Rosalind (/home/diniz/programs/braker/Cput
cd /home/martins/EnTAP_restart

EnTAP --runP -i /home/martins/EnTAP_restart/Proteomes/cput_braker.aa -d /home

mv entap_outfiles/ cput/

scp -r -P 2205 entap_outfiles/ pedro@143.107.244.181:/home/blowflies/genome_a
```

Checking md5

```
cd /home/blowflies/genome_annotation/cput/11-EnTAP/entap_outfiles/final_resul
md5sum entap_results.tsv
#ded21bc806abf83077034a891b2e5cdd  entap_results.tsv
cd /home/diniz/programs/braker/Cput_2/10-BRAKER3/restart
md5sum cput_braker.gtf
#ec60e5bfed4626446c95fba0d56a4bd3  cput_braker.gtf
```

Making an unique gtf file with augustus and ENTAP outputs files:

- /home/blowflies/genome_annotation/cput/11-EnTAP/entap_outfiles/final_results/entap_results.tsv #ded21bc806abf83077034a891b2e5cdd
- /home/diniz/programs/braker/Cput_2/10-BRAKER3/restart/cput_braker.gtf #ec60e5bfed4626446c95fba0d56a4bd3

In R:

```
# matching AUGUSTUS and ENTAP output into a unique gtf

# libraries
library(data.table)
library(dplyr)

# reading the files
tsv <- fread(file = "entap_results.tsv", header = FALSE)
tsv <- tsv[-1,]
tsv <- tsv[,c(1,13)]
gtf <- fread(file = "cput_final.gtf")

# updated gtf
new_gtf <- left_join(gtf, tsv, by = c("V9" = "V1"))

# write gtf
fwrite(x = new_gtf, quote = FALSE, sep = '\t', row.names = FALSE,
       col.names = FALSE, file = "cput_entap_final.gtf")

# to know how many annotated transcripts are (annoted proteins)
ann_tra <- na.omit(tsv$V13) # look the number of elements in this and compare
```

Total annotated transcripts: 16843 (out of 21313)

We copied the final gtf file to Rosalind server and checked md5:

```
cd /Users/diniz/Desktop
md5 Cput_annot.gtf
#0386ed7e643c421d139327e33a81f699
scp -P 2205 Cput_annot.gtf diniz@143.107.244.181:/home/blowflies/genome_annot

#On Rosalind:
cd /home/blowflies/genome_annotation/cput/12-final_files/
md5sum Cput_annot.gtf
#74d1c8756206130b81c9db811ad7713a
```

12-Final files

```
cd /home/blowflies/genome_annotation/cput/12-final_files
md5sum *
#0386ed7e643c421d139327e33a81f699  Cput_annot.gtf
#87fd55c764a84b7229abf3e493a8c0e5  Cput_cds.fa
#2213621d218d62c90311483ddedc2ceb  Cput_genome.fa
#67d462183e79c464cb5305e593e5f669  Cput_protein.aa
```

type	original/copy	file	Path	md5
Raw genome	original	XDOVE_20221110_S64411e_PL100270437-1_D01.ccs.fastq.gz
Raw genome	original	purged.fa	/home/Reference_genomes/Cputoria	314250c4
Raw genome	copy	cput.fa	/home/blowflies/genome_annotation/cput/0-genome

type	original/copy	file	Path	md5
Mitocondrial genome	original	cput_mit.fa	/home/blowflies/genome_annotatic
Nuclear genome	original	cput_N_genome_final.fa	/home/blowflies/genome_a

type	original/copy	file	Path	md5
RNA-seq	original	Cput_1.fq.gz	/home/Raw_seqs/cput_pool_RNA	8ecb8a84268c
RNA-seq	original	Cput_2.fq.gz	/home/Raw_seqs/cput_pool_RNA	d6c794f0578c
RNA-seq unzipped unzipped	original	Cput_1.fq	/home/Raw_seqs/cput_pool_RN
RNA-seq unzipped	original	Cput_2.fq	/home/Raw_seqs/cput_pool_RNA	8c13e3
Trimmed reads zipped	original	cput_all_1P.fq.gz	/home/blowflies/genome_a
Trimmed reads zipped	original	cput_all_2P.fq.gz	/home/blowflies/genome_a

==**Não sei qual é o arquivo correto**==
==|Transcriptome| original| ???| /home/cunha/03-RNA/02-Trinity| ???|==

type	original/copy	file	Path	md5
Masked genome	original	cput_N_genome_final.fa.masked	/home/blowflies/g
Masked genome	copy	Cput_masked.fasta	/home/blowflies/genome_annotatic
Masked genome	copy	Cput_masked.fasta	/home/blowflies/genome_annotation/c
Masked genome	copy	Cput_masked.fasta	/home/diniz/programs/braker/Cput_2/
Masked genome	copy	Cput_masked.fasta	/home/pedro/Non_Coding_Element

type	original/copy	file	Path	md5
Proteome	original	cput_braker.aa	/home/diniz/programs/braker/Cput_2/10-
Proteome	copy	cput_braker.aa	/home/martins/EnTAP_restart/Proteomes	67
Proteome	copy	cput_braker.aa	/home/00-Sequences/Cputoria/01-Genomic_dat

type	original/copy	file	Path	md5
Condingseq	original	braker.codingseq	/home/diniz/programs/braker/Cput_2
Condingseq	copy	cput_braker.codingseq	/home/00-Sequences/Cputoria/01-Gen

type	original/copy	file	Path	md5
gtf output Braker3 second run	original	cput_braker.gtf	/home/diniz/progr
gtf output Braker3 second run	copy		**Computador do Diniz**	
Final gtf	original		**Computador do Diniz**	
Final gtf	copy		**Rosalind**	

**FALTA: conferir md5 dos arquivos usados no script do R do entap e cmeg_ent

```
**PROBLEMAS:**
- **arquivo proteoma da pasta 00-sequences está com md5 diferente do origina
- **arquivo codingseq da pasta 00-sequences está com md5 diferente do origin
- **arquivo gtf que está na pasta 00-sequences não é o final**

_______
# OBSOLETE
# 10- gFACs
https://gfacs.readthedocs.io/en/latest/Flags/index.html

```bash=
cd /home/blowflies/genome_annotation/cput/
mkdir 11-gfacs
cd /home/blowflies/genome_annotation/cput/11-gfacs
mkdir results
sudo perl /gFACs-master/gFACs.pl -f braker_2.1.2_gtf -p cput --rem-all-incom
```

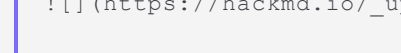
Results:
Number of genes (Augustus/BRAKER): 21450
Number of genes (gFACs): 20894

11- EnTAP
We did it on Darwin
```bash=
cd /home/martins/EnTAP

# The gFACs outputs for all species are here (*_genes.fasta.faa)
mkdir Proteomes

mkdir cput
cd cput
EnTAP --runP -i /home/martins/EnTAP/Proteomes/cput_genes.fasta.faa -d /home/
```
12. gFACs again (with EnTAP output)
```bash=
cd /home/blowflies/genome_annotation/cput/
mkdir 13-gfacs_entap
cd /home/blowflies/genome_annotation/cput/13-gfacs_entap
mkdir results
sudo perl /gFACs-master/gFACs.pl -f gFACs_gene_table -p cput --rem-all-incom
```

13. Final annotation
Final files are here:
```bash=
/dados/home/blowflies/genome_annotation/cput/14-final_annot
# 24543 gene models
```

Final busco
```bash=
docker run -u $(id -u) -v $(pwd):/busco_wd ezlabgva/busco:v5.4.4_cv1 busco -
```


OBSOLETE STUFF
RepeatMasker
```

```bash=
# we start here
/home/blowflies/genome_annotation/cput

# new directory
mkdir 7-RepeatMasker
cd 7-RepeatMasker/
mkdir With_drosophila
cd With_drosophila/

# running
RepeatMasker -species "Drosophila melanogaster"  -dir . -pa 8 -a -xsmall -s
```

14 - Submission to NCBI

NCBI require a sqn file for assembly submission. I followed the step described in https://www.ncbi.nlm.nih.gov/genbank/genomes_gff/

Then I had to rename the headers in the genome fasta and convert from gtf to gff

```
sed -i '/^>/ s/$/ [organism=Chrysomya putoria]/' Cput_genome.fa
singularity run agat_1.0.0--pl5321hdfd78af_0.sif
agat_convert_sp_gxf2gxf.pl --gtf Cput_annot.gtf --output Cput_annot.gff
```

FInally I ran table2asn to get the sqn file

```
/home/diniz/programs/linux64.table2asn -M n \
-J \
-c w \
-euk \
-gaps-min 10 \
-f Cput_annot.gff \
-i Cput_genome.fa \
-locus-tag-prefix Cput \
-o cput.sqn \
-Z \
-V b
```