Log: Genome Annotation Lucilia eximia

tags: `Genome Annotation` `Lexi`

Table of Contents

- Log: Genome Annotation Lucilia eximia
  - Table of Contents
  - List of software (and where they are located in Rosalind):
  - Genome annotation workflow:
- Genome sequencing
- 0- Copy of the genome
- 1- BUSCO First
- 2- QUAST
- 3- Mitochondrial genome
  - Blastn
  - MITOS Web Server
- 4- RepeatModeler
  - Installing with docker
  - Creating database
  - Running
  - Copying the results
- 5- RepeatMasker
  - 2.5.1 BUSCO
  - 2.5.2 QUAST
- 6- RNA-seq
  - Quality control of raw reads
  - Trimming
    - Quality control of trimmed reads
- Observation
- 7- Trinity
  - Running Trinity
  - Assembly statistics
  - BUSCO (transcriptome quality)
- 8- STAR - RNAseq alignment
  - index
  - STAR aligment
- 9- Braker3
  - BRAKER3 second run
  - Busco evaluation with protein sequences from restart Braker run
- 11- EnTAP
- 12-Final files
- OBSOLETE
- 10- gFACs
- 11- EnTAP
- 12. gFACs again (with EnTAP output)
- 13. Final annotation
- OBSOLETE STUFF
  - Trinity (first)

    - Moving trimmed reads to Darwin
    - Assembling the transcriptome
    - BUSCO (transcriptome quality)
- 14 - Submission to NCBI

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 2723669
Coverage (x) 65
HPA Length (bp) 1190603997
HPA N50 (bp) 2920813
HPA N90 (bp) 507262

HPA L50 118
HPA L90 467
FA Length (bp) 579241898
FA N50 (bp) 4214109
FA N90 (bp) 1077085
FA L50 42
FA L90 140

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/lexi

# new directory
mkdir 0-genome
cd 0-genome

cp /home/Reference_genomes/Leximia/purged.fa .
cp /home/Reference_genomes/Leximia/XDOVE_20221110_S64411e_PL100270436-1_C01.
```

1- BUSCO First

Busco version 5.3.2

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

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

cp ../0-genome/lexi.fa .

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

```
---------------------------------------------------
|Results from dataset diptera_odb10               |
---------------------------------------------------
|C:98.9%[S:65.2%,D:33.7%],F:0.3%,M:0.8%,n:3285    |
|3250   Complete BUSCOs (C)                       |
|2142   Complete and single-copy BUSCOs (S)       |
|1108   Complete and duplicated BUSCOs (D)        |
|9      Fragmented BUSCOs (F)                     |
|26     Missing BUSCOs (M)                        |
|3285   Total BUSCO groups searched               |
---------------------------------------------------
```

2- QUAST

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

# new directory
mkdir 2-qast
cd 2-quast

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

3- Mitochondrial genome

We had to use the NCBI's reference mitochondrial genome of Lucilia cuprina since there's not a reference mitochondrial genome for Lucilia eximia.
https://www.ncbi.nlm.nih.gov/nuccore/NC_002660.1
search terms: "lucilia cuprina [ORGN] AND mitochondrial genome"
NC_019573.1

```
esearch -db nuccore -query NC_019573.1 | efetch -format fasta > lcup_mitochor
cd home/blowflies/genome_annotation/lexi/3-mitochondrial_genome
mkdir 1-blastn
```

Blastn

version 2.11

```
cd home/blowflies/genome_annotation/lexi/3-mitochondrial_genome/1-blastn

#Making database using the genome
makeblastdb -in ../lcup_mitochondrial_genome.fa -dbtype nucl -out lexi_mit_da

#Running
blastn -task blastn -evalue 0.00001 -db ./lexi_mit_database -query ../../0-ge

#extract mitochondrial genome (scaffold with the bigger aligment with the mit
seqkit grep -p "ptg0010921" lexi.fa > ptg0010921.fa

#renaming
mv ptg0010921.fa lexi_mit_scaff.fa

#removing the mitochondria from the genome using grep invert matching
seqkit grep -v -p "ptg0010921" lexi.fa > lexi_N_genome.fa

#checking the sequences
grep -c ">" lexi*
#lexi.fa:369
#lexi_mit_scaff.fa:1
#lexi_N_genome.fa:368
```

Then we used the blast output table to

```
# removing duplicates - getting a new fasta from positions 13326:29418
seqkit subseq lexi_mit_scaff.fa -r 13326:29418 > lexi_mit_nodup.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/lexi/3-mitochondrial_genome/2-MITOS_results).

The job settings were:

Job ID: lexi_mit_genome

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

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: fd973780a3fd
container-name: friendly_archimedes
container-image: dfam/tetools:latest

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

```
(base) cunha@darwin:~/00-Genomes$ docker ps
CONTAINER ID   IMAGE                COMMAND   CREATED          STATUS          PORTS     NAMES
fd973780a3fd   dfam/tetools:latest   "bash"    22 minutes ago   Up 14 minutes             friendly_archimedes
```

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 ./lexi_N_genome.fa friendly_archimedes:/home

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

#Database
BuildDatabase -name lexi_database lexi_N_genome.fa
```

Running

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

Copying the results

```
cd /home/cunha/01-RepeatModeler/lexi

docker cp -a friendly_archimedes:/home/lexi/lexi_database-families.stk ./
docker cp -a friendly_archimedes:/home/lexi/lexi_database-families.fa ./
docker cp -a friendly_archimedes:/home/lexi/lexi_database-rmod.log ./
```

5- RepeatMasker

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

```
# we start here
/home/cunha/02-RepeatMasker/lexi

# running
RepeatMasker -lib /home/cunha/01-RepeatModeler/lexi/lexi_database-families.fa
```

2.5.1 BUSCO

```
cd /home/blowflies/genome_annotation/lexi/5-RepeatMasker
mkdir 1-BUSCO
cp /home/blowflies/genome_annotation/lexi/9-STAR/Lexi_masked.fasta 1-BUSCO/
cd 1-BUSCO
md5sum Lexi_masked.fasta
sudo docker run -u $(id -u) -v $(pwd):/busco_wd ezlabgva/busco:v5.4.7_cv1 bus
```

#Results

```
---------------------------------------------------
|Results from dataset diptera_odb10               |
---------------------------------------------------
|C:99.0%[S:65.5%,D:33.5%],F:0.3%,M:0.7%,n:3285    |
|3251   Complete BUSCOs (C)                       |
|2151   Complete and single-copy BUSCOs (S)       |
|1100   Complete and duplicated BUSCOs (D)        |
|10     Fragmented BUSCOs (F)                     |
|24     Missing BUSCOs (M)                        |
|3285   Total BUSCO groups searched               |
---------------------------------------------------
```

2.5.2 QUAST

```
#comparing with the file /home/blowflies/genome_annotation/cmeg/0-genome/cmeg
cd /home/blowflies/genome_annotation/lexi/5-RepeatMasker
mkdir 2-QUAST
cp lexi_N_genome_final.fa.masked 2-QUAST/
cd 2-QUAST
mkdir lexi.fa
cd lexi.fa
sudo quast.py ../Lexi_masked.fasta -t 20 --eukaryote --large --rna-finding --
```

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

```
Assembly                    Lexi_masked
# contigs (>= 0 bp)         368
# contigs (>= 1000 bp)      368
# contigs (>= 5000 bp)      368
# contigs (>= 10000 bp)     368
# contigs (>= 25000 bp)     325
# contigs (>= 50000 bp)     306
Total length (>= 0 bp)      579203135
Total length (>= 1000 bp)   579203135
Total length (>= 5000 bp)   579203135
Total length (>= 10000 bp)  579203135
Total length (>= 25000 bp)  578414946
Total length (>= 50000 bp)  577733876
# contigs                   368
Largest contig              15451354
Total length                579203135
Estimated reference length  500000000
GC (%)                      29.30
N50                         4214109
NG50                        4760762
N75                         2285326
NG75                        3010171
L50                         42
LG50                        33
L75                         88
LG75                        66
# total reads               729
# left                      0
# right                     0
Mapped (%)                  100.0
Properly paired (%)         0.0
Avg. coverage depth         1
Coverage >= 1x (%)          99.99
# N's per 100 kbp           0.06
# predicted rRNA genes      182 + 17 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/lexi_pool_RNA /home/blowflies/genome_annotation/lexi/
#Renaming
cd /home/blowflies/genome_annotation/lexi/
mv lexi_pool_RNA 8-lexi_pool_RNA
#checking md5
cd 8-lexi_pool_RNA/
cat MD5.txt
#856d439819ae45beefd5e4c5785bd4a0  Lexi_1.fq.gz
#cdbb764449f3096589f36afbc0bad1a3  Lexi_2.fq.gz
md5sum Lexi*
#856d439819ae45beefd5e4c5785bd4a0  Lexi_1.fq.gz
#cdbb764449f3096589f36afbc0bad1a3  Lexi_2.fq.gz
mkdir 0-raw_reads
mv *.fq.gz 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/lexi/4-lexi_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/lexi/4-lexi_pool_RNA/1-QC/multiqc_report.html /mnt/c/Users/vansc/Downloads
```

Results:
https://drive.google.com/file/d/11K0LqiLvPrOrGrpvFY7DZRB5OosV_vzs/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/lexi/4-lexi_pool_RNA/0-raw-reads
mkdir ../2-trimming
screen
TrimmomaticPE Lexi_1.fq.gz Lexi_2.fq.gz -threads 8 -baseout
/home/blowflies/genome_annotation/lexi/8-lexi_pool_RNA/2-
trimming/lexi.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/lexi/4-lexi_pool_RNA/2-tr

# QC
fastqc *.gz

# in this directory -> /home/blowflies/genome_annotation/lexi/4-lexi_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/lexi/0-genome

mv purged.fa lexi.fa
```

7- Trinity

We had to send the files from Rosalind to Darwin then rename all of them to make it easier to run trinity

```
#renaming them
# rename files
for f in *fastq.gz; do mv -- "$f" "${f%.fastq.gz}.fq.gz"; done
```

Running Trinity

```
#!/bin/bash

#SBATCH --job-name trinity_lexi ## nome que aparecerá na fila
#SBATCH --output trinity_lexi.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 v
#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 --seq
```

The final assembly is here: /home/cunha/03-RNA/02-Trinity (Darwin). And it was renamed to lexi_trinity.fasta (md5checked)

Assembly statistics

```
file                      format  type  num_seqs      sum_len  min_len  avg_
trinity_all.Trinity.fasta  FASTA   DNA    136,730  139,151,237      176  1,0
```

BUSCO (transcriptome quality)

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

docker pull ezlabgva/busco:v5.4.7_cv1 # just because we didn't have busco on

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

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/lexi
mkdir 8-STAR
chmod 777 8-STAR
cd 8-STAR
mkdir star_index
chmod 777 star_index

STAR --runThreadN 8 --runMode genomeGenerate --genomeDir /home/blowflies/gen
```

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

```
cat *_1P.fq.gz > lexi_all_1P.fq.gz
cat *_2P.fq.gz > lexi_all_2P.fq.gz
```

STAR aligment

```
cd /home/blowflies/genome_annotation/lexi/8-lexi_pool_RNA/2-trimming/all

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

#started at: Feb 10 18:48:37
#finished at: Feb 10 19:17:44
```

9- Braker3

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

Final outputs are here: /home/blowflies/genome_annotation/lexi/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 Lexi_2
cp /home/blowflies/genome_annotation/lexi/10-BRAKER3 /home/diniz/programs/bra
cd /home/diniz/programs/braker/Lexi_2/10-BRAKER3
mkdir restart
# Run
singularity exec /home/diniz/programs/braker/braker3.sif braker.pl --genome=
```

Total transcripts: 28884

Busco evaluation with protein sequences from restart Braker run

```
cd /home/diniz/programs/braker/Lexi_2/10-BRAKER3/restart
mv braker.gtf lexi_braker.gtf
mv braker.aa lexi_braker.aa

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

```
# 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/lexi_braker.aa
# BUSCO was run in mode: proteins

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

        C:98.0%[S:56.5%,D:41.5%],F:0.9%,M:1.1%,n:3285
        3217    Complete BUSCOs (C)
        1855    Complete and single-copy BUSCOs (S)
        1362    Complete and duplicated BUSCOs (D)
        29      Fragmented BUSCOs (F)
        39      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/Lexi
cd /home/martins/EnTAP_restart

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

mv entap_outfiles/ lexi/

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

Checking md5

```
cd /home/blowflies/genome_annotation/lexi/11-EnTAP/entap_outfiles/final_resul
md5sum entap_results.tsv
#b520b5c11281d98e4c3919f68e1316e4  entap_results.tsv
cd /home/diniz/programs/braker/Lexi_2/10-BRAKER3/restart
md5sum lexi_braker.gtf
#08af39a581366b0c871fcaa3bbff7e7c  lexi_braker.gtf
```

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

- /home/blowflies/genome_annotation/lexi/11-EnTAP/entap_outfiles/final_results/entap_results.tsv #b520b5c11281d98e4c3919f68e1316e4 entap_results.tsv
- /home/diniz/programs/braker/Lexi_2/10-BRAKER3/restart/lexi_braker.gtf #08af39a581366b0c871fcaa3bbff7e7c

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 = "lexi_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 = "lexi_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: 21835 (out of 28884)

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

```
cd /Users/diniz/Desktop
md5 Lexi_annot.gtf
#aef125d4a204649fc62f08a36c66716a
scp -P 2205 Lexi_annot.gtf diniz@143.107.244.181:/home/blowflies/genome_annot

#On Rosalind:
cd /home/blowflies/genome_annotation/lexi/12-final_files/
md5sum Lexi_annot.gtf
#aef125d4a204649fc62f08a36c66716a
```

12-Final files

```
cd /home/blowflies/genome_annotation/lexi/12-final_files
#aef125d4a204649fc62f08a36c66716a  Lexi_annot.gtf
#6e24cf78ab05fd2947eecb89f6eb85b6  Lexi_cds.fa
#4117a959ee8e7f397e3648cf53e8153e  Lexi_genome.fa
#a4df9e2040357393e1f62715a0ccf4d1  Lexi_protein.aa

```

type	original/copy	file	Path	md5
Raw genome	original	XDOVE_20221110_S64411e_PL100270436-1_C01.ccs.fastq.gz	/home/Reference_genomes/Leximia	5a7e8
Raw genome	original	purged.fa	/home/Reference_genomes/Leximia	5f5dc
Raw genome	copy	lexi.fa	/home/blowflies/genome_annotation/lexi/0-genome	5f5dc

type	original/copy	file	Path	md5
Mitocondrial genome	original	lexi_mit_nodup.fa	/home/blowflies/genome_annotation/lexi/0-genome	faff2f427a88295b14b4bdf
Nuclear genome	original	lexi_N_genome.fa	/home/blowflies/genome_annotation/lexi/0-genome	b7241aa3c751fc7a1c66df1

type	original/copy	file	Path	md5
RNA-seq	original	Lexi_1.fq.gz	/home/Raw_seqs/lexi_pool_RNA	856d439819ae45beefd5e
RNA-seq	original	Lexi_2.fq.gz	/home/Raw_seqs/lexi_pool_RNA	cdbb764449f3096589f36
RNA-seq unzipped unzipped	original	Lexi_1.fq	/home/Raw_seqs/lexi_pool_RNA	36cb7c42b1e8784c4678b
RNA-seq unzipped	original	Lexi_2.fq	/home/Raw_seqs/lexi_pool_RNA	b68d49d023625ba791b01
Trimmed reads zipped	original	lexi_all_1P.fq.gz	/home/blowflies/genome_annotation/lexi/8-lexi_pool_RNA/2-trimming/all	a3013b0d773c2f49dbc43
Trimmed reads zipped	original	lexi_all_2P.fq.gz	/home/blowflies/genome_annotation/lexi/8-lexi_pool_RNA/2-trimming/all	958d073075ad63426e3da
Transcriptome	original	lexi_trinity.fasta	/home/cunha/03-RNA/02-Trinity	404eecbe03bd747996b48

type	original/copy	file	Path	md5
Masked genome	original	lexi_N_genome.fa.masked	/home/cunha/02-RepeatMasker/lexi	4117a959ee8e7f39
Masked genome	copy	Lexi_masked.fasta	/home/blowflies/genome_annotation/lexi/9-STAR	4117a959ee8e7f39

Masked genome	copy	Lexi_masked.fasta	/home/blowflies/genome_annotation/lexi/10-BRAKER3	4117a959ee8e7f39
Masked genome	copy	Lexi_masked.fasta	/home/pedro/Non_Coding_Element_Evolution/3-Masking/2-RepeatMasker	4117a959ee8e7f39

type	original/copy	file	Path	md5
Proteome	original	lexi_braker.aa	/home/diniz/programs/braker/Lexi_2/10-BRAKER3/restart	a4df9e2040357393e1f62715a0ccf4
Proteome	copy	lexi_braker.aa	/home/martins/EnTAP_restart/Proteomes	a4df9e2040357393e1f62715a0ccf4
Proteome	copy	lexi_braker.aa	/home/00-Sequences/Leximia/01-Genomic_data/2023	bd5a03ebc9b56a26e9fb15921dc514

type	original/copy	file	Path	md5
gtf output Braker3 second run	original	lexi_braker.gtf	/home/diniz/programs/braker/Lexi_2/10-BRAKER3/restart	08af39a581366b0c871fcaa3bbff7e7c
gtf output Braker3 second run	copy		**Computador do Diniz**	
Final gtf	original		**Computador do Diniz**	
Final gtf	copy		**Rosalind**	

type	original/copy	file	Path	md5
Condingseq	original	braker.codingseq	/home/diniz/programs/braker/Lexi_2/10-BRAKER3/restart	6e24cf78ab05fd2947eecb
Condingseq	copy	lexi_braker.codingseq	/home/00-Sequences/Leximia/01-Genomic_data/2023	6cda82d7deb70bf4b6bf17

FALTA: conferir md5 dos arquivos usados no script do R do entap e cmeg_entap_final.gtf que está no computador do Diniz

PROBLEMAS:

- **arquivo gtf que está na pasta 00-sequences não é o final**
- **arquivo codingseq da pasta 00-sequences está com md5 diferente do original**
- **arquivo proteoma da pasta 00-sequences está com md5 diferente do original**

OBSOLETE

10- gFACs

https://gfacs.readthedocs.io/en/latest/Flags/index.html

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

Results:
Number of genes (Augustus/BRAKER): 29174

Number of genes (gFACs): 28339

11- EnTAP

We did it on Darwin

```
cd /home/martins/EnTAP

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

mkdir lexi
cd lexi
EnTAP --runP -i /home/martins/EnTAP/Proteomes/lexi_genes.fasta.faa -d /home/
```

I copied the output folder into Rosalind

```
cd /home/blowflies/genome_annotation/lexi

mkdir 12-EnTAP
cd 12-EnTAP
scp -r -P 4988 martins@lem.ib.usp.br:/home/martins/EnTAP/lexi/entap_outfiles/
```

12. gFACs again (with EnTAP output)

```
cd /home/blowflies/genome_annotation/lexi/
mkdir 13-gfacs_entap
cd /home/blowflies/genome_annotation/lexi/13-gfacs_entap
mkdir results
sudo perl /gFACs-master/gFACs.pl -f gFACs_gene_table -p lexi --rem-all-incomp
```

13. Final annotation

Final files are here:

```
/dados/home/blowflies/genome_annotation/lexi/14-final_annot
# 24543 gene models
```

Final busco

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

```
***** Results: *****

C:97.6%[S:57.6%,D:40.0%],F:0.8%,M:1.6%,n:3285
3206    Complete BUSCOs (C)
1891    Complete and single-copy BUSCOs (S)
1315    Complete and duplicated BUSCOs (D)
25      Fragmented BUSCOs (F)
54      Missing BUSCOs (M)
3285    Total BUSCO groups searched
```

OBSOLETE STUFF

Trinity (first)

Transcriptome assembly

Moving trimmed reads to Darwin

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

scp -P 4988 /home/blowflies/genome_annotation/lexi/8-lexi_pool_RNA/2-trimmin
```

Assembling the transcriptome

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

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

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

# Transcriptome size
grep -c ">" lexi_trinity.Trinity.fasta # 36184
```

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.
```

Complete (all)	Complete Single	Complete Dup.	Fragmented	Missing
80	37.4	42.6	3.4	16.6

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=Lucilia eximia]/' Lexi_genome.fa
singularity run agat_1.0.0--pl5321hdfd78af_0.sif
agat_convert_sp_gxf2gxf.pl --gtf Lexi_annot.gtf --output Lexi_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 Lexi_annot.gff \
-i Lexi_genome.fa \
-locus-tag-prefix Lexi \
-o lexi.sqn \
-Z \
-V b
```