MaCuMBA

Grupo de Genómica Evolutiva
Evolutionary Genomics Group

Practical Session 4– December 12th
Assembling metagenomes

Rohit Ghai, Carolina Mizuno
Mario López-Pérez, Aitor Gonzaga,
Ana-Belen Martin-Cuadrado, Inmaculada Garcia-Heredia, Nikole Kimes

µgen

Gobierno de España
Ministerio de Ciencia e Innovación

Consolider

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 311957.

MaCuMBA

Program

9-12th December

- **Session 1:** Introduction to UNIX ✓
  (Carolina Mizuno and Rohit Ghai)
- **Session 2:** Genome assembly and annotation ✓
  (Mario Lopez-Perez and Aitor Gonzaga)
- **Session 3:** Metagenomic reads ✓
  (Ana-Belen Martin-Cuadrado and Inmaculada Garcia-Heredia)
- **Session 4:** Assembling metagenomes
  (Rohit Ghai and Carolina Mizuno)
- **Session 5:** Sampling Cruise

μgen

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN

Consolider

Assembling metagenomes

Programs

- illumina2mira
- trimfastq
- velvet-shuffleSequences_fastq.pl
- velveth
- velvetg
- contig_arrange
- faslice
- run-prodigal.pl
- blast
- attach_blast_results.pl
- gbk2table

Assembling metagenomes

General instructions

- **Ok, rule to follow for every command in today's exercise. Run the "ls" and "pwd" command after EVERY COMMAND. You should always be working in YOUR home directory /home/guest/group1/username/DCM. If you are somewhere else, please go back to your own folder.**

STEP1: Getting the raw data from the data folder

Make a fresh directory for today's work, the name of the directory will be **DCM** (deep chlorophyll maximum) and it should be in YOUR folder.

Copy raw metagenomic data (illumina hiseq) from this folder

/home/guest/data/day4

There are 2 files (remember paired-ends?)

cp /home/guest/data/day4/DCM1.fastq

cp /home/guest/data/day4/DCM2.fastq

STEP2: preparing the files for assembly

Change read names from this type

@HWI-ST1207:173:C14H3ACXX:6:1101:1488:1970 1:N:0:CGATGT

to

@DCM-1/1

Using illumina2mira (use –h option to see help)

/home/guest/bin/illumina2mira –h

/home/guest/bin/illumina2mira -i DCM1.fastq -o dcm_1.fastq -pair 1 -rename "DCM-" -pattern="\@HWI-ST" &

/home/guest/bin/illumina2mira -i DCM2.fastq -o dcm_2.fastq -pair 2 -rename "DCM-" -pattern="\@HWI-ST" &

Quality trimming

/home/guest/bin/trimfastq dcm_1.fastq

Output file= dcm_1.fastq.trimmed

/home/guest/bin/trimfastq dcm_2.fastq

Output file= dcm_2.fastq.trimmed

Combining both these files:

- /home/guest/bin/velvet-shuffleSequences_fastq.pl dcm_1.fastq.trimmed dcm_2.fastq.trimmed dcm-combined.fastq

STEP3: Running velvet

velveth DCM-K51 51 -shortPaired -fastq dcm-combined.fastq

velvetg DCM-K51 -ins_length 300 -cov_cutoff auto

RESULTS DIRECTORY

cd DCM-K51

Sequences
Roadmaps
PreGraph
Graph
contigs.fa <<- this file contains the final assembled contigs in fasta format
stats.txt
LastGraph
Log

STEP4: organizing your contigs

/home/guest/bin/contig_arrange -i contigs.fa -o CONTIGS.FA -label "DCM-C"

cat CONTIGS.FA | lenseq fasta | head

STEP5: annotation

Taking the top 5 longest contigs

cat CONTIGS.FA | faslice 1 5 > top5.fna

Gene prediction

/home/guest/bin/run-prodigal.pl -i top5.fna -meta -prefix TOP5

This will produce 3 output files
TOP5.gbk <-- genbank file
TOP5.faa <-- fasta file containing all the protein sequences
TOP5.ffn <-- fasta file containing all the gene sequences

BLAST AGAINST NR DATABASE

blastall -i TOP5.faa -m 8 -d nr -e 1e-5 -b 1 -v 1 -a 5 -o TOP5.faa.blastp -p blastp

Output file: TOP5.faa.blastp

ATTACH BLAST RESULTS TO GENBANK FILE

/home/guest/bin/attach_blast_results.pl -i TOP5.gbk -b TOP5.faa.blastp -o TOP5_final.gbk

MAKE A TABLE FROM THE GENBANK FILE

/home/guest/bin/gbk2table -i TOP5_final.gbk -o TOP5_table.txt

OPEN TABLE IN EXCEL

Thank you!