Metagenomic analysis using NanoDJ Notebooks in CyVerse

Presented by: Héctor Rodríguez Pérez
Hospital Nuestra Señora de Candelaria, Spain

September 27th, 2019

Unión Europea
Fondo Europeo
de Desarrollo Regional
“Una manera de hacer Europa”

NanoDJ

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

AGENCIA ESTATAL DE INVESTIGACIÓN

Outline of Today’s Focus Forum

- Introduction to metagenomics using long reads.
- Classification with BLAST and custom databases.
- Rapid taxonomic classification using Kraken2.

NSF

Unión Europea
Fondo Europeo
de Desarrollo Regional
“Una manera de hacer Europa”

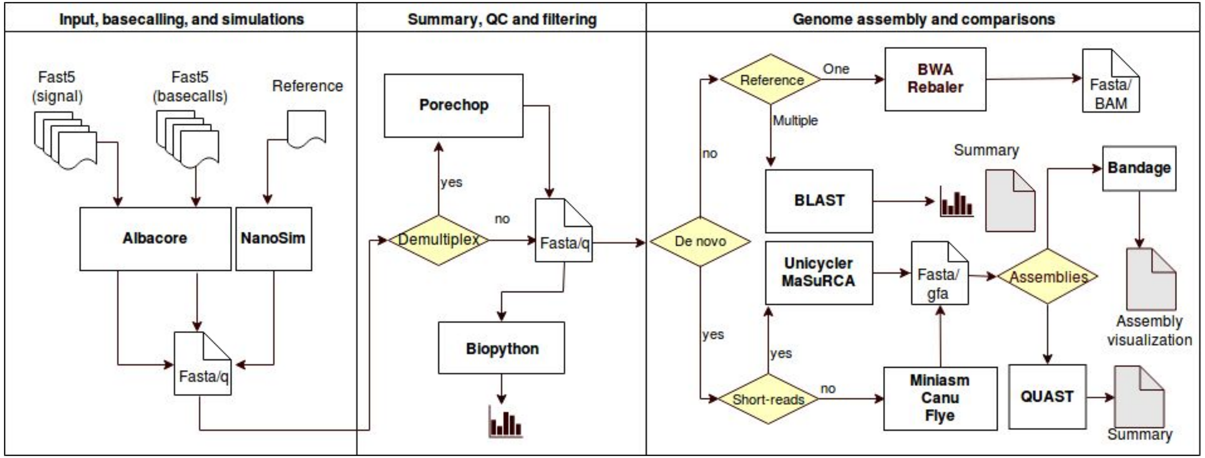
NanoDJ

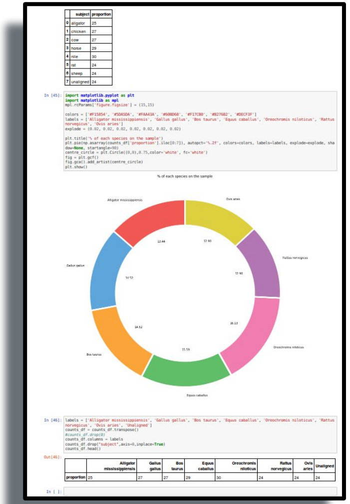
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

AGENCIA ESTATAL DE INVESTIGACIÓN

NanoDJ

- **Integration of bioinformatics tools for Nanopore data analysis.**
- **Jupyter notebooks with genome assembly and metagenomics pipelines step by step.**





	subject	proportion
0	aligator	25
1	chicken	27
2	cow	27
3	horse	29
4	nile	30
5	rat	24
6	sheep	24
7	unaligned	24

```
In [45]: import matplotlib.pyplot as plt
         import matplotlib as mpl
         mpl.rcParams['figure.figsize'] = (15,15)

         colors = ['#F15854', '#5DA5DA', '#FAA43A', '#60BD68', '#F17CB0', '#B276B2', '#DECF3F']
         labels = ['Alligator mississippiensis', 'Gallus gallus', 'Bos taurus', 'Equus caballus', 'Oreochromis niloticus', 'Rattus
         norvegicus', 'Ovis aries']
         explode = (0.02, 0.02, 0.02, 0.02, 0.02, 0.02, 0.02)

         plt.title('% of each species on the sample')
         plt.pie(np.asarray(counts_df['proportion'].iloc[0:7]), autopct='%.2f', colors=colors, labels=labels, explode=explode, sha
         dow=None, startangle=90)
         centre_circle = plt.Circle((0,0),0.75,color='white', fc='white')
         fig = plt.gcf()
         fig.gca().add_artist(centre_circle)
         plt.show()
```

```
In [46]: labels = ['Alligator mississippiensis', 'Gallus gallus', 'Bos taurus', 'Equus caballus', 'Oreochromis niloticus', 'Rattus
         norvegicus', 'Ovis aries', 'Unaligned']
         counts_df = counts_df.transpose()
         #counts_df.drop(0)
         counts_df.columns = labels
         counts_df.drop("subject",axis=0,inplace=True)
         counts_df.head()
```

Out[46]:

	Alligator mississippiensis	Gallus gallus	Bos taurus	Equus caballus	Oreochromis niloticus	Rattus norvegicus	Ovis aries	Unaligned
proportion	25	27	27	29	30	24	24	24

In [ ]:

Unión Europea

Fondo Europeo
de Desarrollo Regional
"Una manera de hacer Europa"

GOBIERNO
DE ESPAÑA

MINISTERIO
DE CIENCIA, INNOVACIÓN
Y UNIVERSIDADES

Materials

- **NanoDJ (included in VICE)**: Integration of bioinformatics tools for Nanopore data analysis on top of a Jupyter Lab environment.
    - Python programming language (Biopython, Numpy, Pandas, Matplotlib).
    - Command line tools (BLAST tools, Kraken2, Bracken).
- **2 notebooks:**
    - Classification with BLAST and custom database.
    - Rapid taxonomic classification with Kraken2.
- **Example sequence datasets:**
    - MtDNA of different species.
    - 16S rRNA sequences of a saliva sample.

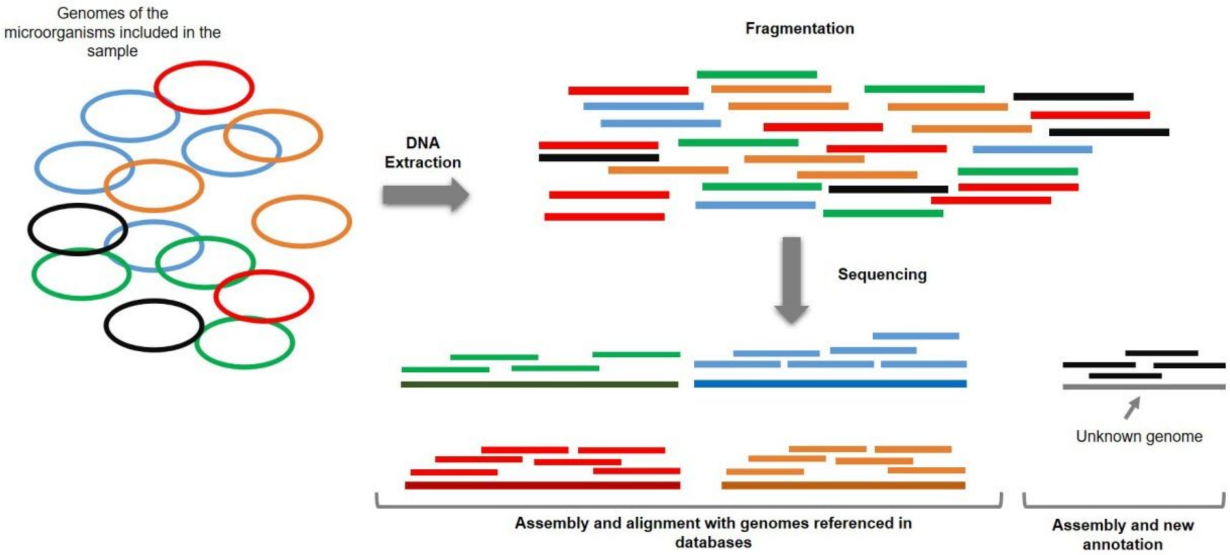
Unión Europea
Fondo Europeo
de Desarrollo Regional
"Una manera de hacer Europa"

NanoDJ

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Metagenomics

- Shotgun metagenomics
- Targeted metagenomics



Genomes of the microorganisms included in the sample

DNA Extraction

Fragmentation

Sequencing

Unknown genome

Assembly and alignment with genomes referenced in databases

Assembly and new annotation

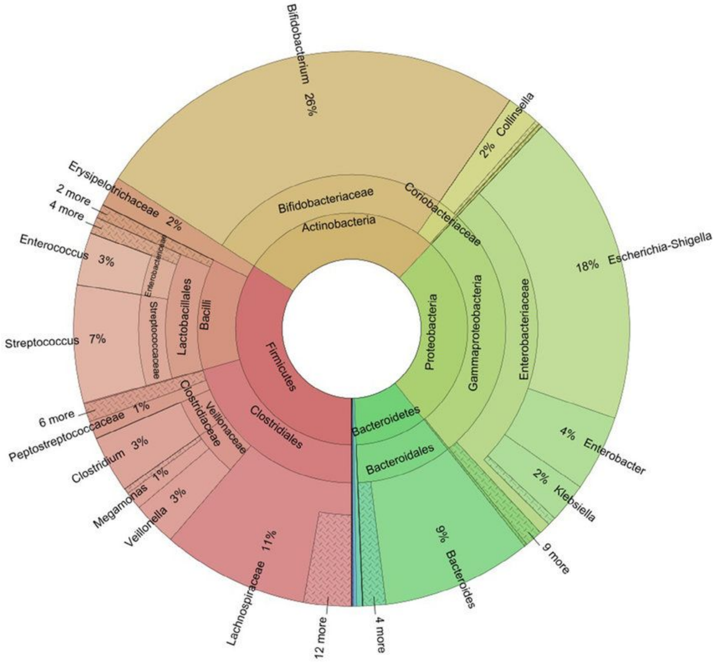
Unión Europea
Fondo Europeo
de Desarrollo Regional
“Una manera de hacer Europa”

NanoDJ

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Metagenomics applications

- Ecology.
- Agriculture.
- Medicine:
    - ○ Gut microbe characterization.
    - ○ Infectious disease diagnosis.



Unión Europea

Fondo Europeo
de Desarrollo Regional
"Una manera de hacer Europa"

NanoDJ

GOBIERNO DE ESPAÑA

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

AGENCIA ESTATAL DE INVESTIGACIÓN

Metagenomic analysis using long reads

- Cost-efficient approach to metagenomics.
- Rapid library preparation and real time data output.
- Long reads enable sequencing of informative genes (e.g. 16S rRNA) in their entirety (Full 16s on one read).

NSF

Unión Europea
Fondo Europeo
de Desarrollo Regional
“Una manera de hacer Europa”

NanoDJ

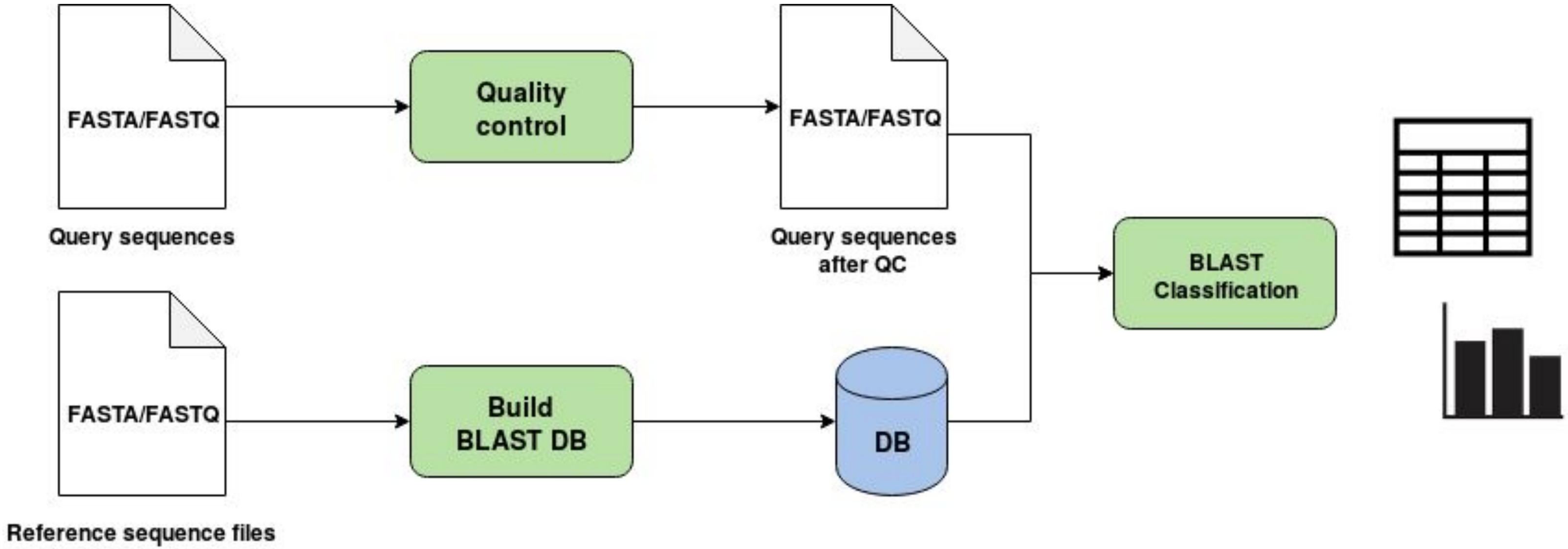
GOBIERNO
DE ESPAÑA

MINISTERIO
DE CIENCIA, INNOVACIÓN
Y UNIVERSIDADES

AGENCIA
ESTATAL DE
INVESTIGACIÓN

TM

Classification with custom database using BLAST



FASTA/FASTQ

Query sequences

Quality control

FASTA/FASTQ

Query sequences after QC

FASTA/FASTQ

Reference sequence files

Build BLAST DB

DB

BLAST Classification

NSF

Unión Europea
Fondo Europeo
de Desarrollo Regional
“Una manera de hacer Europa”

NanoDJ

GOBIERNO DE ESPAÑA

MINISTERIO
DE CIENCIA, INNOVACIÓN
Y UNIVERSIDADES

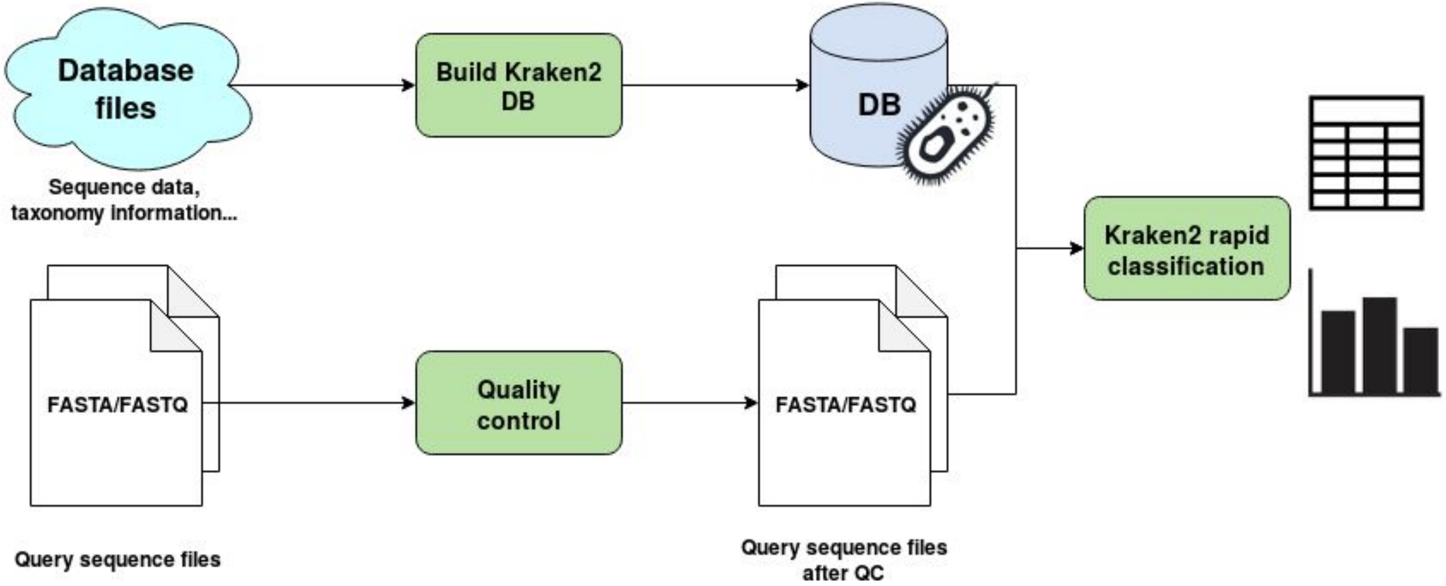
Rapid taxonomic classification of long read data

Why is rapid taxonomic classification of long reads challenging?

- Much longer reads (300bp vs <1kbp) and higher error-ratio.
- BLAST scales badly in time and space. Not suitable for high-throughput data.
- Databases are not perfect (not even close) on accuracy and species diversity representation.

NSF
Unión Europea
Fondo Europeo
de Desarrollo Regional
"Una manera de hacer Europa"
NanoDJ
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
AGENCIA ESTATAL DE INVESTIGACIÓN

Rapid taxonomic classification of complete 16S reads



Database files

Sequence data, taxonomy information...

Build Kraken2 DB

DB

FASTA/FASTQ

Query sequence files

Quality control

FASTA/FASTQ

Query sequence files after QC

Kraken2 rapid classification

Unión Europea
Fondo Europeo
de Desarrollo Regional
"Una manera de hacer Europa"

NanoDJ

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Things to keep in mind

- Classification of long reads and high-throughput sequencing data can be done using rapid approaches. Modern tools focus on increasing classification speed and reducing memory consumption.

- Database choice is important. Kraken2 includes commands for building NCBI RefSeq databases, 16S databases and custom databases.

- This tools are included in NanoDJ and work fine with Nanopore long read data but they also work also with short read data.

★ <u>Check NanoDJ repository for updates. More notebooks coming!</u>

`https://github.com/genomicsITER/NanoDJ`

NSF

Unión Europea
Fondo Europeo
de Desarrollo Regional
“Una manera de hacer Europa”

NanoDJ

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

AGENCIA ESTATAL DE INVESTIGACIÓN

TM

Thank you for attending

TM

CyVerse is supported by the National Science Foundation under Grants No. DBI-0735191, DBI-1265383 and DBI-1743442.

MINISTERIO
DE CIENCIA, INNOVACIÓN
Y UNIVERSIDADES

Instituto
de Salud
Carlos III

FI17/00177
PI17/00610
PI14/00844

Servicio
Canario de la Salud

Universidad
de La Laguna

ITER
Instituto Tecnológico y de
Energías Renovables

CABILDO DE TENERIFE

RTC-2017-6471-1

NSF

GOBIERNO
DE ESPAÑA

MINISTERIO
DE CIENCIA, INNOVACIÓN
Y UNIVERSIDADES

AGENCIA
ESTATAL DE
INVESTIGACIÓN

THE UNIVERSITY
OF ARIZONA®

TACC
TEXAS ADVANCED
COMPUTING CENTER

CSH
Cold
Spring
Harbor
Laboratory

Unión Europea
Fondo Europeo
de Desarrollo Regional
"Una manera de hacer Europa"

TM