Metagenomic analysis using NanoDJ Notebooks in CyVerse

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

September 27th, 2019
Outline of Today’s Focus Forum

• Introduction to metagenomics using long reads.

• Classification with BLAST and custom databases.

• Rapid taxonomic classification using Kraken2.
NanoDJ

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

- Ecology.
- Agriculture.
- Medicine:
  - Gut microbe characterization.
  - Infectious disease diagnosis.
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).
Classification with custom database using BLAST

1. FASTA/FASTQ (Query sequences)
2. FASTA/FASTQ (Reference sequence files)
3. Quality control
4. Build BLAST DB
5. BLAST Classification

Flowchart:
- Query sequences after QC
- FASTA/FASTQ
- Build BLAST DB
- DB

Diagram:
- FastA/FastQ
- Reference sequence files
- Quality control
- Build BLAST DB
- BLAST Classification

Diagram elements:
- Arrows indicating flow
- Rectangles for each step
- Tables and graphs on the right side
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.
Rapid taxonomic classification of complete 16S reads
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.

★ Check NanoDJ repository for updates. More notebooks coming!

https://github.com/genomicsITER/NanoDJ
Thank you for attending

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

FI17/00177
PI17/00610
PI14/00844

RTC-2017-6471-1