EBAii Assemblage & Annotation

Part 2: construction and analysis of a prokaryotic genomic dataset

H. Chiapello & V. Loux

Helene.chiapello@inrae.fr
https://orcid.org/0000-0001-5102-0632

Valentin.loux@inrae.fr
https://orcid.org/0000-0002-8268-915X

2. Construction and analysis of prokaryotic genomic dataset

Outline

> 2.1 Constructing a genome dataset

> 2.2 Analyzing the genome dataset

> 2.3 Comparing and dereplicating the dataset

Many slides from the "***Bioinformatique par la pratique***" migale training cycle "Comparison of microbial genomes" module

https://migale.inrae.fr/trainings

Hélène Chiapello

Training

Valentin Loux

Technical coordinator

2.3 Comparing and dereplicating a genome dataset

Why?

> To deal with

- The huge number or public genomes for some taxonomical groups including very similar or identical ones
  - Ex : *E. coli, S. enterica*
- The heterogenous quality of sequencing and assembly of these data

> To design a relevant comparative strategy adapted to the dataset

Back to genome diversity evaluation

Two main methods

- Alignment based approaches (ANI)
  - slow (need pairwise comparisons)
  - Robust to genome incompleteness
- k-mer based approaches (MASH)
  - Rapid (hash technics)
  - Not robust to genome incompleteness
  - Only provides an estimate of ANI

    > Become very approximative for very divergent genomes

Comparing and dereplicating a genome dataset

The dRep tool

> dRep is a python program which performs **rapid pairwise genome comparisons** using genomic distances

> it can be used for genome **dereplication**: identification of the 'same' genomes from a large set + determination of the highest quality genome in each replicate set

Very good documentation:
https://drep.readthedocs.io/en/latest/

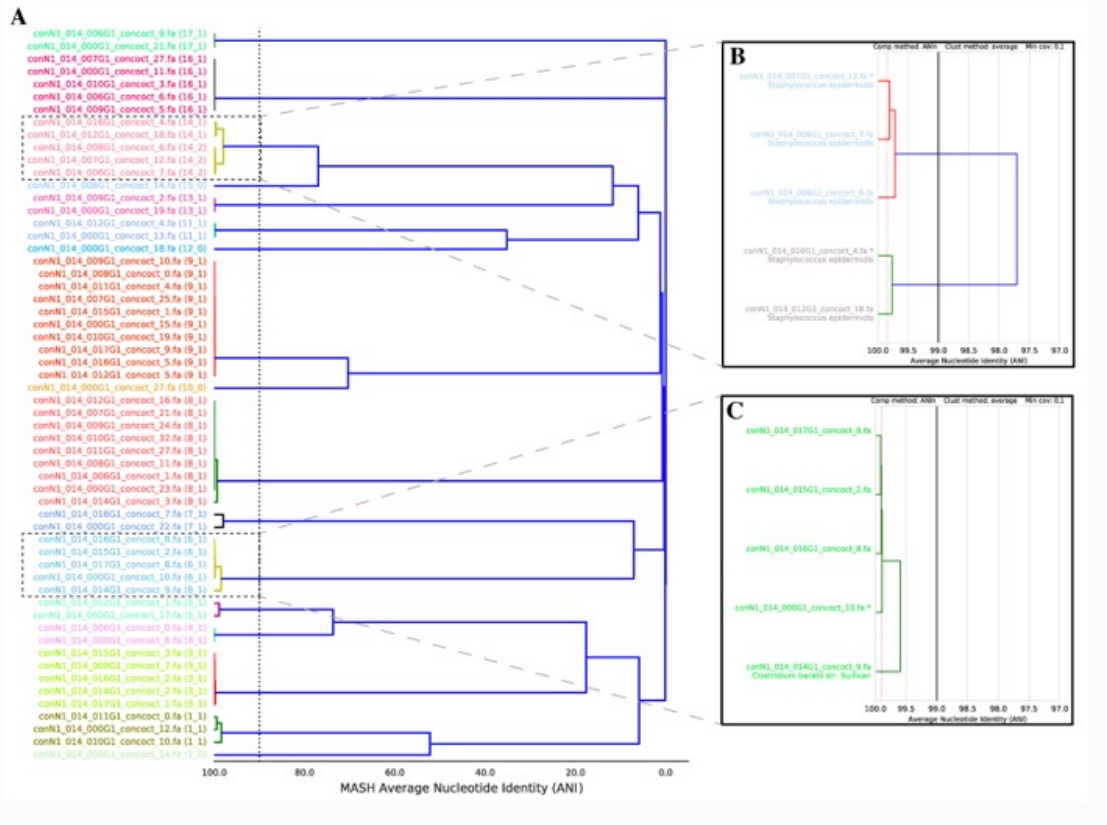
Comparing and dereplicating a genome dataset

The dRep tool

dREP uses 2 main steps:

1. a first (rapid) clustering of genomes using MASH similarity (90% by default)

2. a second more sensitive step based on ANI on pairs of genomes that have at least a minimum level of "MASH" similarity (99% by default)



Very good documentation:
https://drep.readthedocs.io/en/latest/

dRep important concepts

1. dRep primary clustering use a greedy algorithm, i.e. an algorithm that take shortcuts to run faster and generally produces "quasi-optimal" solutions. Genomes that are not on the same MASH primary clustering will never be compared with ANI

2. Importance of genome completeness: MASH is very sensitive to genome completness. the more incomplete of genomes you allow into your genome list, the more you must decrease the primary cluster threshold.

3. The secondary ANI threshold (default value: 99%, limit: 99.99%) indicates how similar genomes need to be to be considered the “same”. Depending on the application,you may modify this parameter, i.e.: 95% ANI for species-level de- replication or 98% ANI to generate a set of genomes that are distinct when mapping short reads.

4. A score is used to pick representative genomes takes into account several parameters such as Completeness, Contamination, strain heterogeneity and centrality (a measure of how similar a genome is to all other genomes in it’s cluster).

dRep commands and parameters

1. dREp compare: compare and cluster a set of genomes using one or two clustering steps.

2. dREp dereplicate: compare, cluster and dereplicate a set of genomes. During dereplication the first step is identifying groups of similar genomes, and the second step is picking a Representative Genome (RG) for each cluster

Parameters of primary and secondary clustering may have to be adjusted depending on the diversity of the dataset and on the objective of the comparison/dereplication

Default values of dRep clustering parameters:

```
-pa P_ANI, --P_ani P_ANI
                      ANI threshold to form primary (MASH) clusters
                      (default: 0.9)
-sa S_ANI, --S_ani S_ANI
                      ANI threshold to form secondary clusters (default:
                      0.99)
```

dRep tools and result files

dRep rely on several other programs:

1. **Mash**: to build the primary clusters
2. **Mummer**: to perform the ANI computation on pairwise genome alignements (used by default but **fastANI** or **gANI** may also be used)
3. **checkM** (Parks et al. 2015) to determine contamination and completeness of genomes
4. **Prodigal** (Hyatte et al. 2010): to predict genes (used by checkM and gANI)
5. **cipy** (Jones et al. 2001) to produce a final hierarchical clustering.

```
workDirectory
./data
...../checkM/
...../Clustering_files/
...../gANI_files/
...../MASH_files/
...../ANIn_files/
...../prodigal/
./data_tables
...../Bdb.csv  # Sequence locations and filenames
...../Cdb.csv  # Genomes and cluster designations
...../Chdb.csv # CheckM results for Bdb
...../Mdb.csv  # Raw results of MASH comparisons
...../Ndb.csv  # Raw results of ANIn comparisons
...../Sdb.csv  # Scoring information
...../Wdb.csv  # Winning genomes
...../Widb.csv # Winning genomes' checkM information
./dereplicated_genomes
./figures
./log
...../cluster_arguments.json
...../logger.log
...../warnings.txt
```

Output files of dRep

dRep practice

Use dREP-dreplicate to explore the Salmonella genome dataset diversity and completenes and dereplicate the dataset

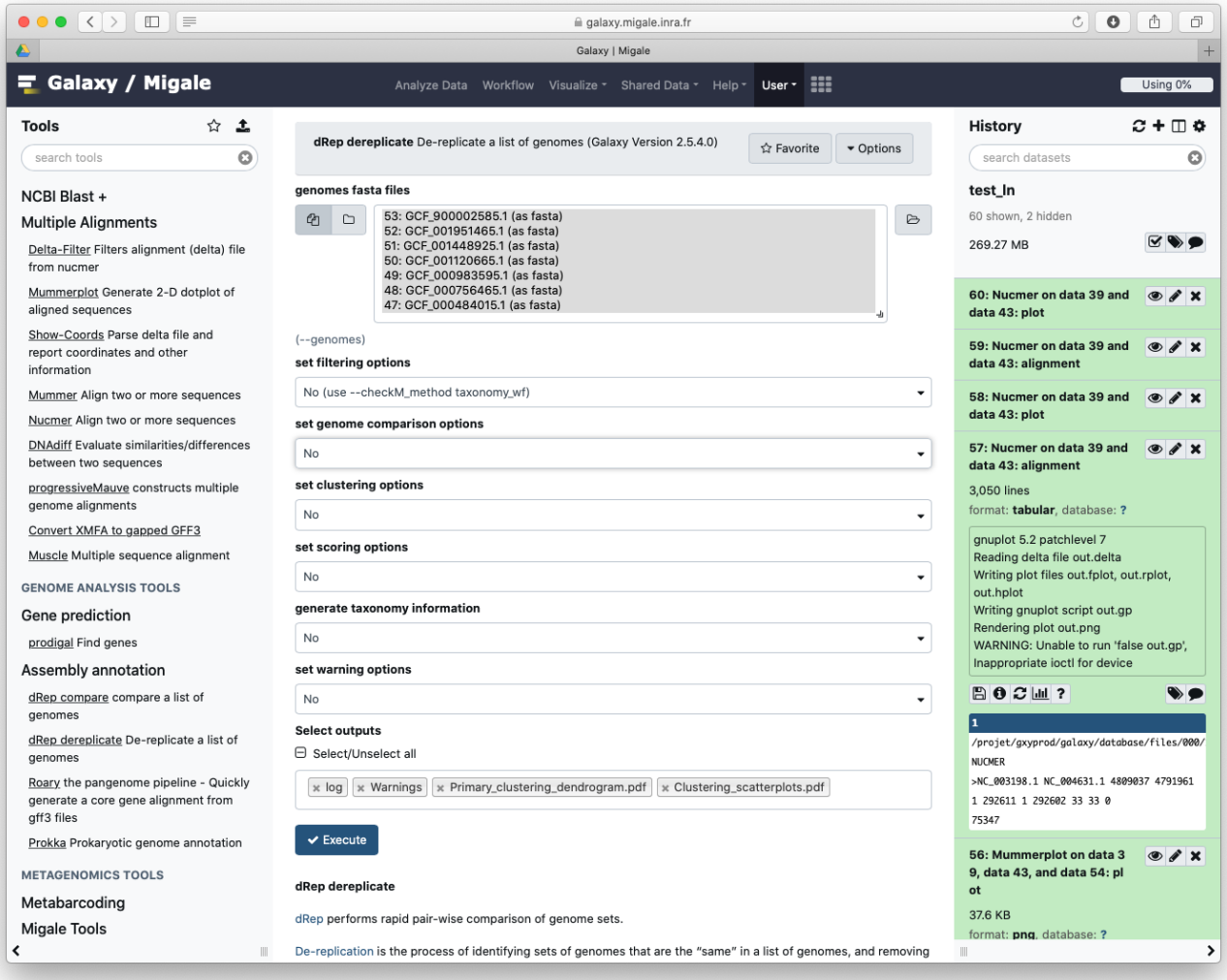
> input : 16 Salmonella genome fasta files

> Default parameters

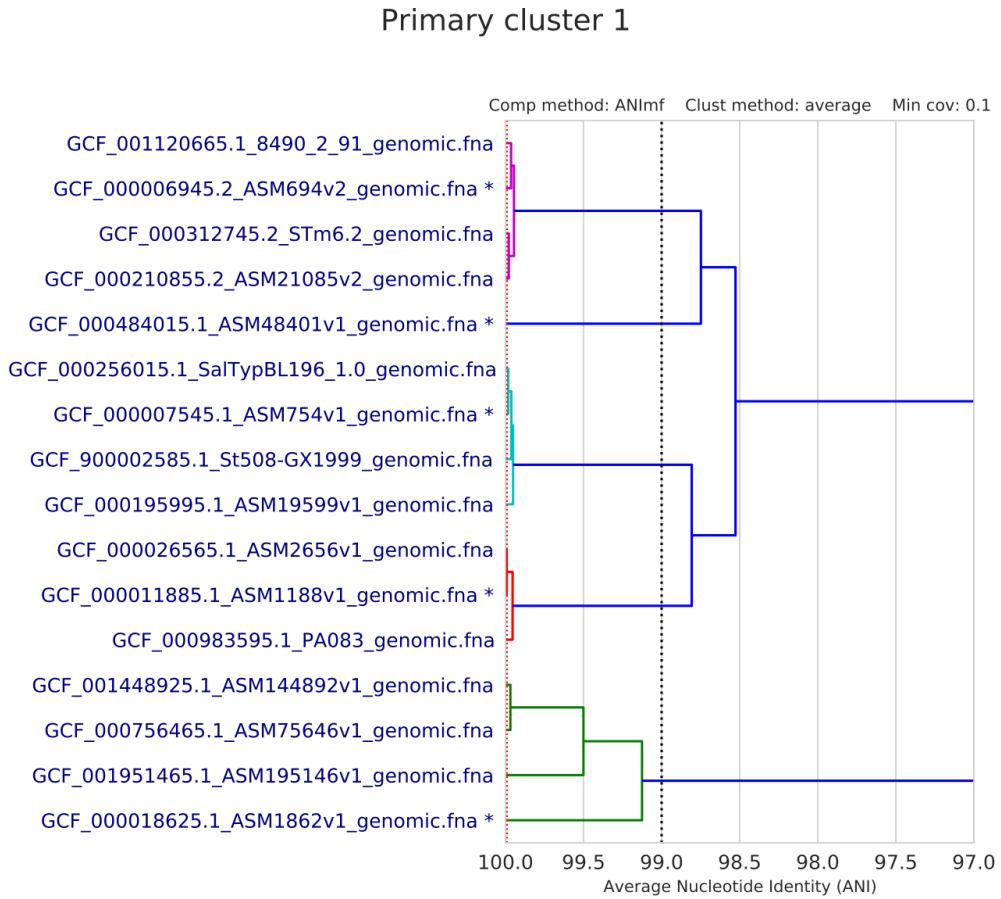
Explore and interpret results

Important:

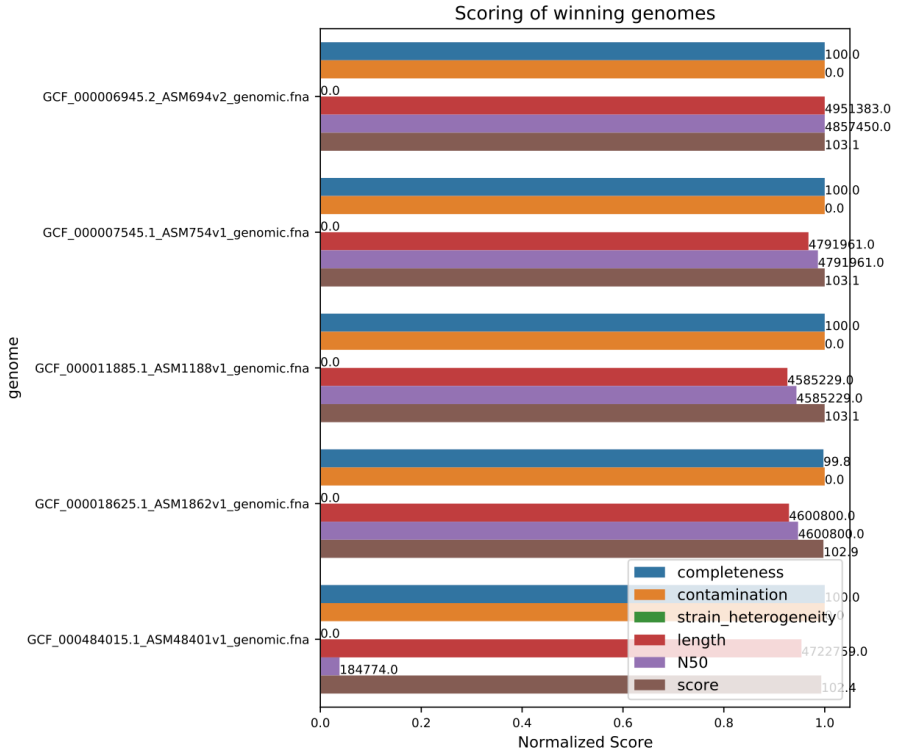
Choose « browse datasets » and select the 16 fasta files of Salmonella dataset 3



dRep results interpretation



Secondary_clustering_dendrograms.pdf



Winning_genomes.pdf

Questions ?