A workflow to build a relevant and reduced bacterial dataset from NCBI

Sam Ah-Lone1, Sandra Dérozier2, Valentin Loux3, Hélène Chiapello3
1 MalAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France

Rapid progress in High-Throughput Sequencing (HTS) technologies has made more than 190,000 bacterial assemblies available in public databases [1]. Among those assemblies, there are redundancies due to very close or even identical genomes. Yet, most genome comparison tools are not scalable to large datasets. In order to overcome this issue, we set up rules to build a representative subdataset by taking into account assembly quality and genomic diversity of the original one. We implemented this approach in a Snakemake [2] workflow that allows to rapidly analyze and filter large sets of closely related bacterial genomes. This procedure has been first tested on two datasets of 300 assemblies from *S. enterica* and *B. subtilis*, then on 9,520 *S. enterica* chromosome assemblies.

**Workflow**

**Genomic distance estimation**
- Dataset: assembly data from RefSeq
- Genomic distance computation with Mash [3]
- Exploration through a Neighbour Joining tree

**Data quality assessment**
- Quality metrics computation with Quast [4]
- Quality metrics exploration with MultiQC [5]

**Closely related genomes clustering**
- Most reliable representatives selection out of each cluster

**Subdataset construction**

**Selection of representative genomes**

- **Principle:** From the Neighbour Joining tree built with the Mash distance matrix, closely-related genomes are grouped into clusters. Each cluster is built according to a user-defined threshold so that the maximum genomic distance between two assemblies of the same cluster is lower than this threshold. This method determines two types of representative data: clusters and singletons.

**Application example on *S. enterica***

**Initial dataset**
- Organism: *Salmonella enterica*
- Number of assemblies: 300
- Assembly quality distribution:
  - Complete: 150
  - Chromosome: 50
  - Scaffold: 50
  - Contig: 50
- Max. genomic distance in NJ tree: 0.05

**Reduced datasets**

- **Threshold: 5e-04**
  - Number of assemblies: 170
  - Assembly quality distribution:
    - Complete: 91
    - Chromosome: 8
    - Scaffold: 30
    - Contig: 41

- **Threshold: 5e-06**
  - Number of assemblies: 274
  - Assembly quality distribution:
    - Complete: 137
    - Chromosome: 42
    - Scaffold: 47
    - Contig: 48

**Distribution of the number of assemblies constituting the reduced dataset for each threshold**

**Conclusion**

- Representative selection according to assembly quality.
- Handle large number of genomes: scale-up on a set of ~10k *S. enterica* assemblies.

**Perspectives**

- Automation of the genomic distance choice as threshold, according to the dataset.
- Availability through a web application.
- Improvement of outputs according to users' feedbacks.

**Bibliography**