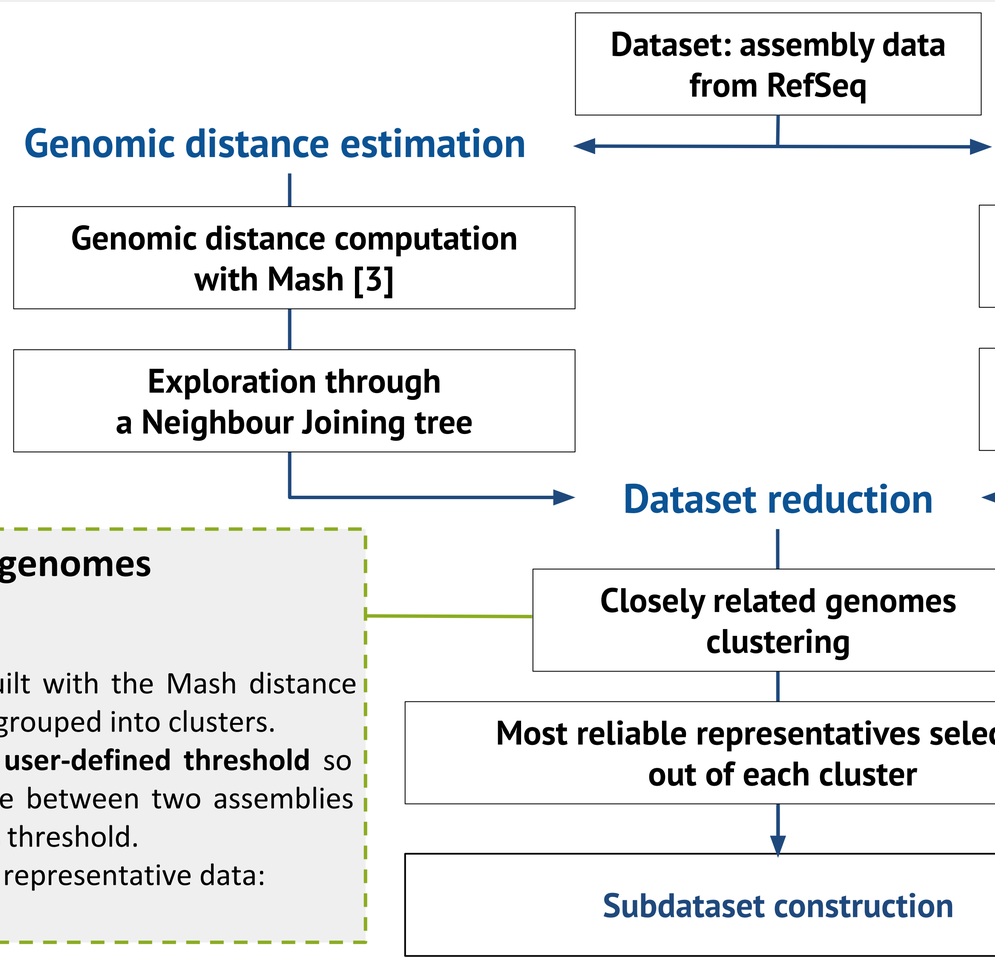
A workflow to build a relevant and reduced bacterial dataset from NCBI

Sam Ah-Lone[1], Sandra Dérozier[1], Valentin Loux[1], Hélène Chiapello[1]

[1] MaIAGE, 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



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 determine two types of representative data: **clusters** and **singletons**.

Quality assessment rules

rule 1: **Highest assembly level**

rule 2: **Exclusion of assemblies with too many contigs** (default = 80)

rule 3: **Greatest N50 value**

Application example on *S. enterica*

Initial dataset

Organism: ***Salmonella enterica***

Number of assemblies: **300**

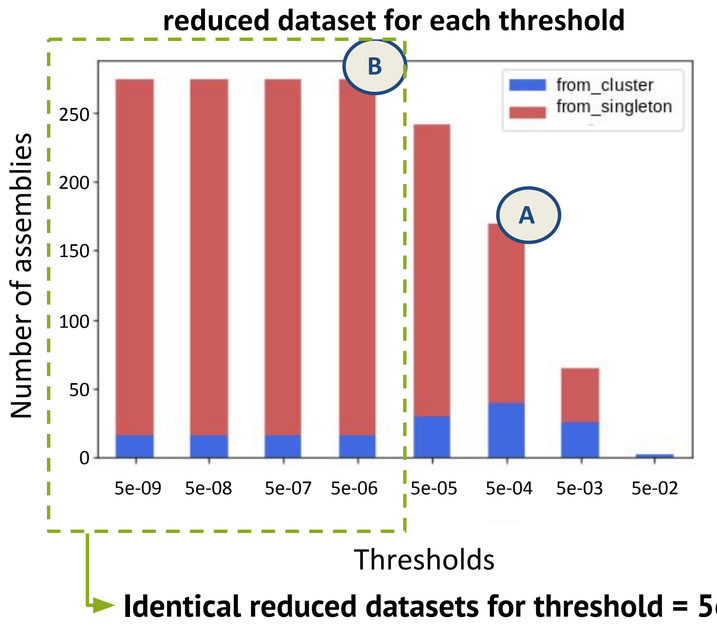
Assembly quality distribution:

- Complete: 150
- Chromosome: 50
- Scaffold: 50
- Contig: 50

Max. genomic distance in NJ tree: **0.05**

Workflow application on 8 different thresholds

Distribution of the number of data constituting the reduced dataset for each threshold



Identical reduced datasets for threshold = 5e-06 and less.

Zoom on two thresholds

Reduced datasets

A **Threshold: 5e-04**

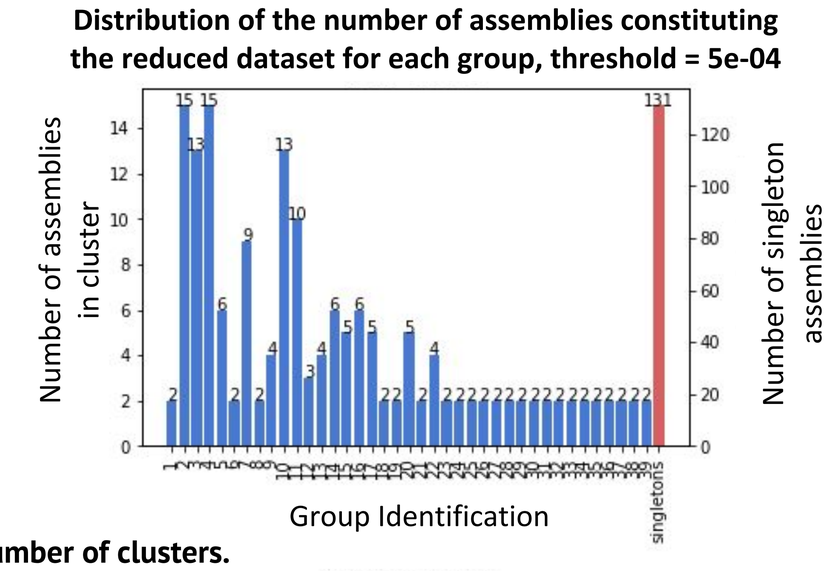
Number of assemblies: **170**

Assembly quality distribution:

- Complete: 91
- Chromosome: 8
- Scaffold: 30
- Contig: 41

Dataset reduction with a maximal number of clusters.

Distribution of the number of assemblies constituting the reduced dataset for each group, threshold = 5e-04



B **Threshold: 5e-06**

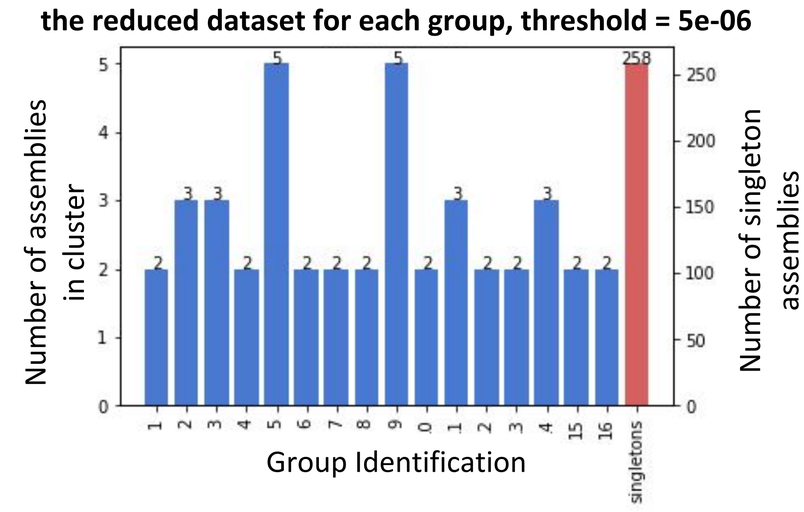
Number of assemblies: **274**

Assembly quality distribution:

- Complete: 137
- Chromosome: 42
- Scaffold: 47
- Contig: 48

Clusters of identical assemblies according to Mash distance.

Distribution of the number of assemblies constituting the reduced dataset for each group, threshold = 5e-06



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

[1] NCBI Genbank FTP site. 'prokaryotes.txt'. [online] https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt. Accessed 14/03/2019. [2] Köster J., *et al.* **Snakemake - A scalable bioinformatics workflow engine.** Bioinformatics, 2012. [3] Ondov BD, *et al.* **Mash: fast genome and metagenome distance estimation using MinHash.** Genome Biol. (2016). [4] Gurevich A., *et al.* **QUAST: quality assessment tool for genome assemblies.** Bioinformatics, 2013. [5] Ewels P., *et al.* **MultiQC: summarize analysis results for multiple tools and samples in a single report.** Bioinformatics, 2016.