EBAii Assemblage & Annotation 2024

Construction and analysis of a prokaryotic genomic dataset

V. Loux, H. Chiapello, G. Gautreau

INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Construction and analysis of prokaryotic genomic dataset

Outline

> Constructing a genome dataset from public ressources

> Analyzing the genome dataset : intrinsec metrics & distances

> 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

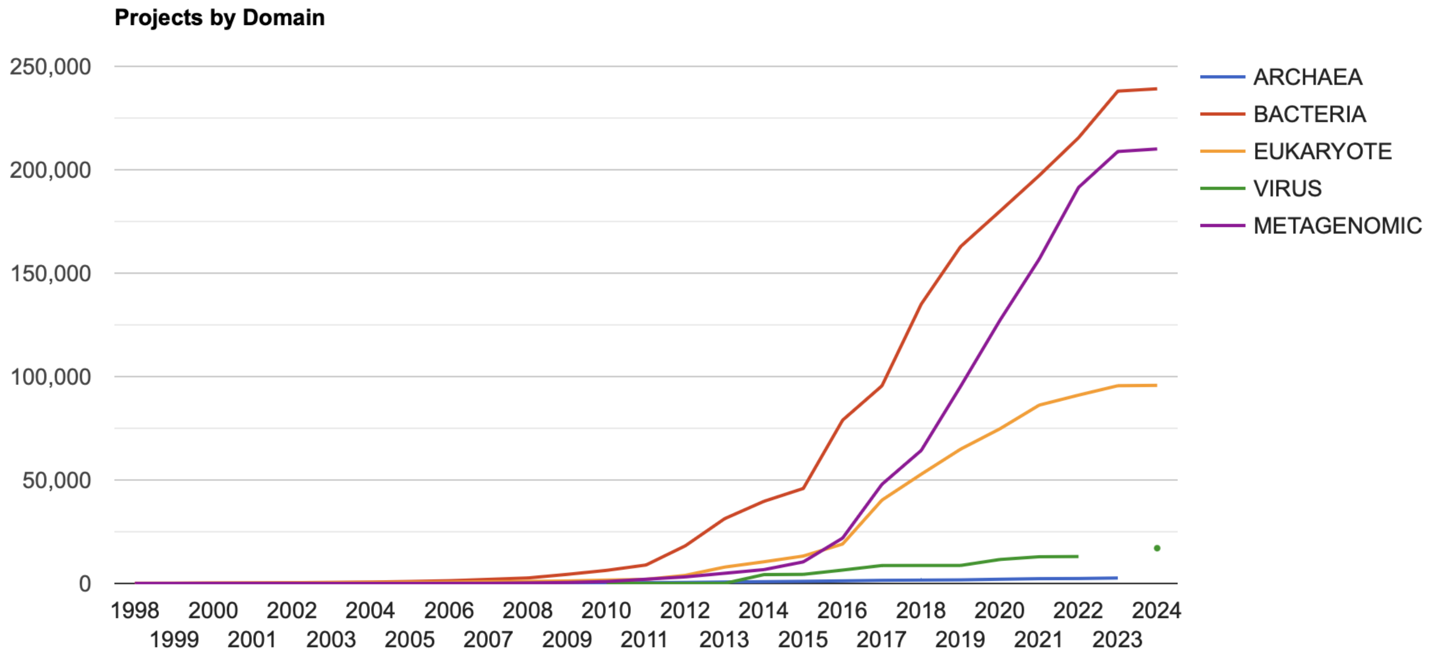
https://documents.migale.inrae.fr/#training-materials

Hélène Chiapello

Valentin Loux

Sequenced genomes by kingdom

Project Totals in GOLD (by year and Domain Group)



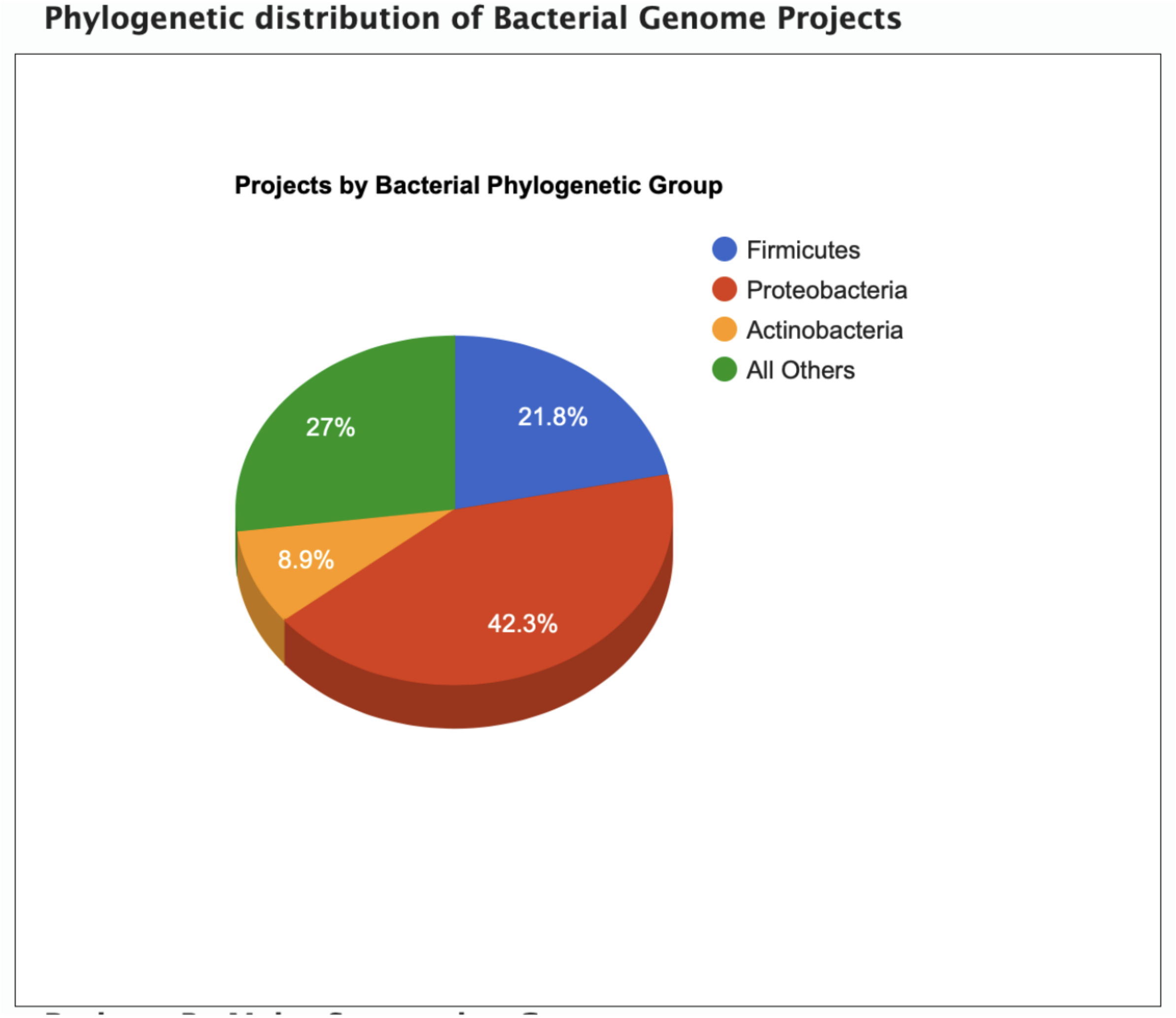
June 24

- Archea : 2 629
- Eukaryote : 95 593
- Virus : 13 009
- Metagenomics : 210 060
- Bacteria : 239 146

⚠ genomes in GOLD ≠ assembly in Genbank (> 2 M bacteria)

Source : Gold (https://gold.jgi.doe.gov/statistics)

Sequenced genomes by kingdom



Phylogenetic distribution of Bacterial Genome Projects

Source : Gold (https://gold.jgi.doe.gov/statistics)

Why using public data and do comparative genomics ?

- Answer to (not so simple) questions like :
  - What is the genomic diversity into a species / genus ?
  - Is the genome structure conserved into a species / genus ?
  - How does the gene repertory evolves into a species / genus ?
  - Is the gene repertory corelated to a given habitat ?
- Does this diversity could explain a given phenotype :
  - metabolism
  - probiotics (anti-inflamatory)
  - pathogenicity
- …

Dataset building

Genomes of interest could be

- already **published** and **available** at public databanks (ENA, NCBI, …)
- private, not yet published.

At least, **we need** :

- [As much as possible] **complete genome assemblies** (contigs / scaffolds in Fasta format)
- Syntactic and functional annotation :
  - Genbank or GFF format
  - For private genomes, you could/should use Bakta [remember what GG told you yesterday afternoon ?]

It's always better, if not mandatory, if [syntaxic] **annotation is homogeneous**

Quick reminder on formats !

FASTA format

The **FASTA format** is used to represent **sequence information**.

The format is very simple:

A > symbol on the FASTA header line indicates a fasta record start.

A string of letters called the sequence id may follow the > symbol.

The header line may contain an arbitrary amount of text (including spaces) on the same line.

Subsequent lines contain the sequence.

```
>foo
ATGCC
>bar other optional text could go here
CCGTA
>bidou
ACTGCAGT
TTCGN
>repeatmasker
ATGTGTcggggggATTTT
>prot2; my_favourite_prot
MTSRRSVKSGPREVPRDEYEDLYYTP
SSGMASP
```

Genbank Format

The Genbank format is used to represent sequence **and** annotation information together.

The start of the annotation section is marked by a line beginning with the word “LOCUS”.

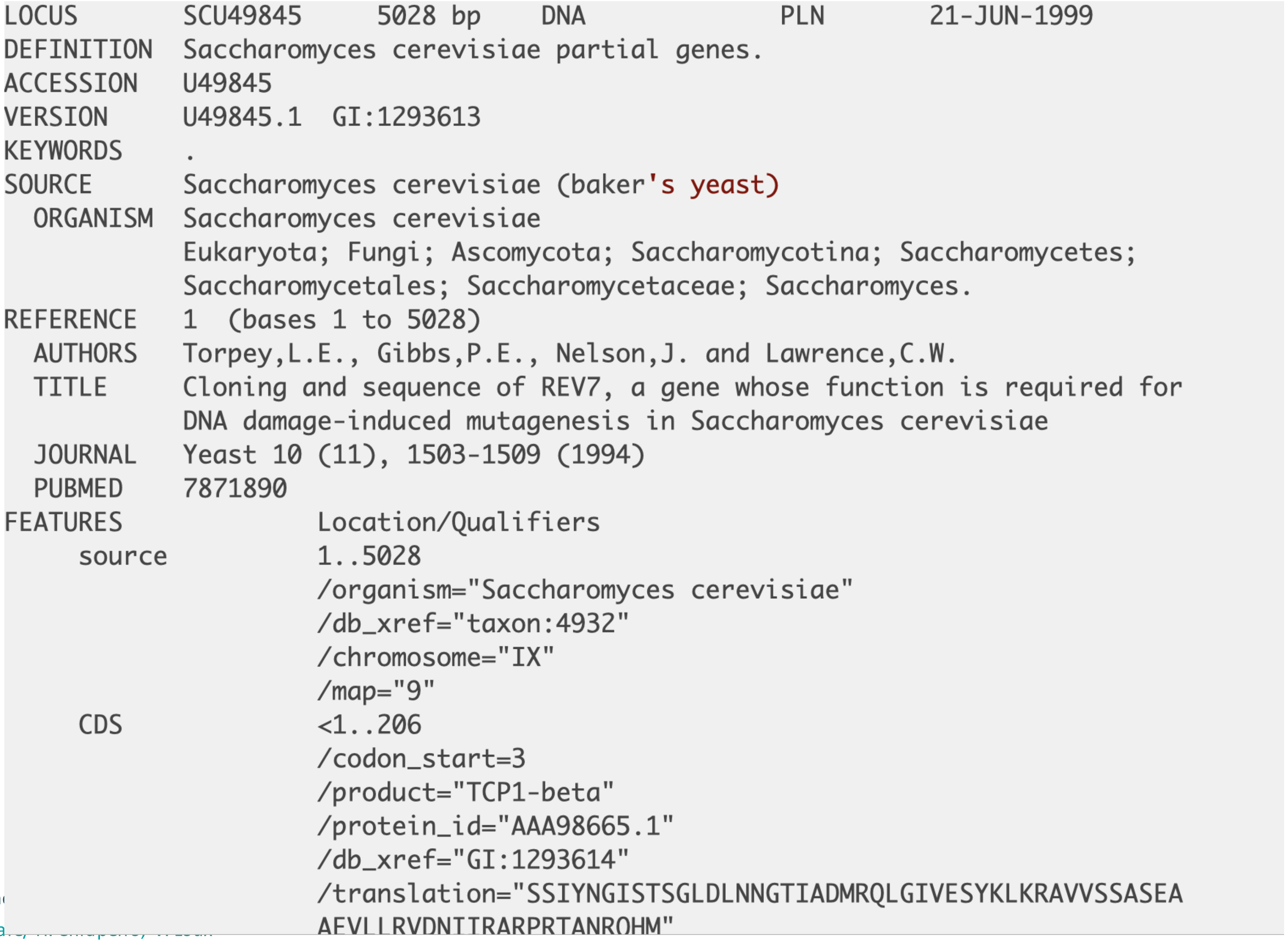
Features (CDS, genes) are annotated with their position , strand, and qualifiers that contain the annotation.

The **start of sequence** section is marked by a line beginning with the word “ORIGIN” and the end of the section is marked by a line with only “//”.

Those three bank agree on the list of **feature / qualifier** that one can use to annotate sequence. (Cf https://www.ebi.ac.uk/ena/WebFeat/)

NCBI, ENA (European Nucleotide Archive) et DDBJ (Japan) entries are synchronized each day.

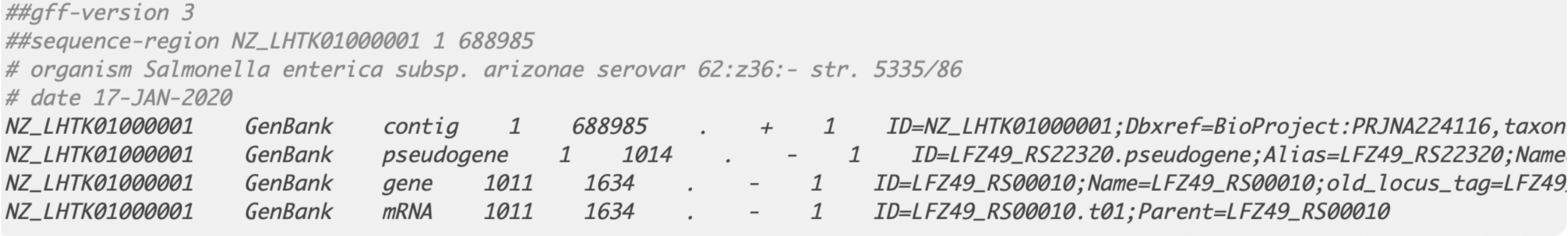
Genbank entry example



```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae partial genes.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
```

GFF Format

The **General Feature Format** contains **annotation** and (optionally) **sequence**. It consists of one line per feature, each containing 9 columns of data, plus optional track definition line.



```
##gff-version 3
##sequence-region NZ_LHTK01000001 1 688985
# organism Salmonella enterica subsp. arizonae serovar 62:z36:- str. 5335/86
# date 17-JAN-2020
NZ_LHTK01000001    GenBank    contig     1     688985     .     +     1     ID=NZ_LHTK01000001;Dbxref=BioProject:PRJNA224116,taxon
NZ_LHTK01000001    GenBank    pseudogene     1     1014     .     -     1     ID=LFZ49_RS22320.pseudogene;Alias=LFZ49_RS22320;Name
NZ_LHTK01000001    GenBank    gene     1011     1634     .     -     1     ID=LFZ49_RS00010;Name=LFZ49_RS00010;old_locus_tag=LFZ49
NZ_LHTK01000001    GenBank    mRNA     1011     1634     .     -     1     ID=LFZ49_RS00010.t01;Parent=LFZ49_RS00010
```

Practical : public genomes

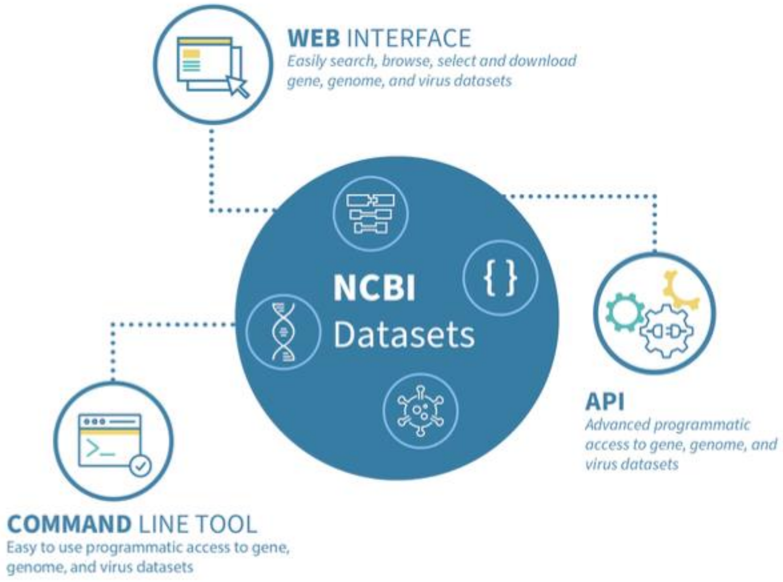
How to gather a list of public genomes of interest ?

List **all** available public genomes of interest with :

- Associated metadata
- Metrics quality (size, N50, completnesse…)
- Filter according to above criteria
- Download genomes in **various formats**

Work from the complete list of prokaryotic public genomes available at NCBI [https://www.ncbi.nlm.nih.gov/datasets/genome/]

A solution : NSBI datasets


WEB INTERFACE
Easily search, browse, select and download gene, genome, and virus datasets
NCBI Datasets
{ }
API
Advanced programmatic access to gene, genome, and virus datasets
COMMAND LINE TOOL
Easy to use programmatic access to gene, genome, and virus datasets

NCBI Datasets components

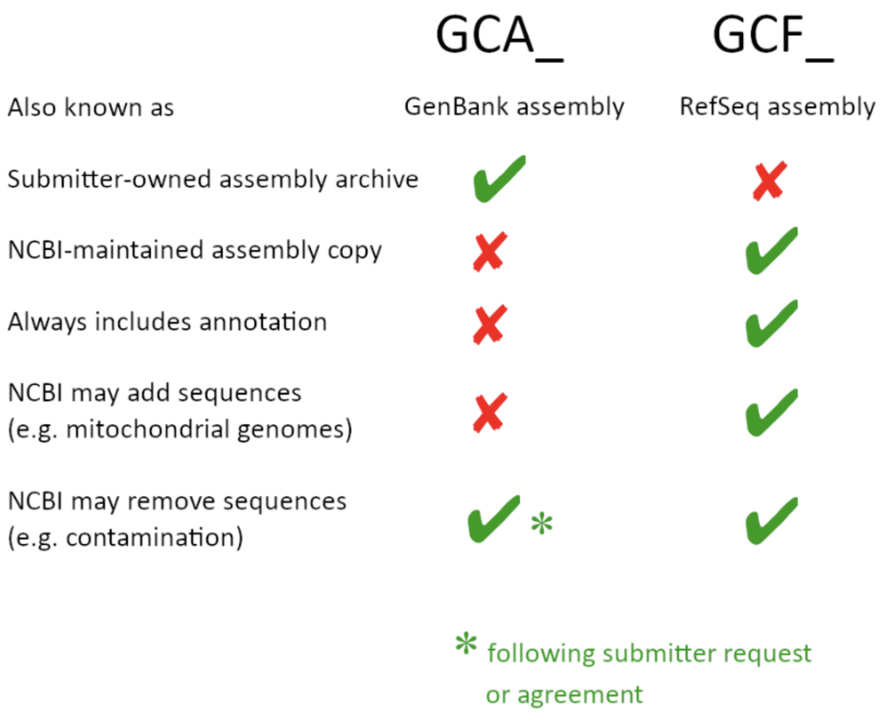
NCBI Datasets is a new resource that lets you easily gather data from across NCBI databases. You have the choice of getting the data through three interfaces:

- NCBI Datasets website Command-line tools
- API (Application programming Interface)

NCBI Datasets delivers data and metadata as a **cohesive data package** contained in a zip archive. i.e., for an **assembly : sequences, annotation (CDS, transcripts, genome...) and metadata.**

Source for genome assemblies

- **A GenBank (GCA)** genome assembly contains assembled genome sequences submitted by investigators to GenBank or another member of the International Nucleotide Sequence Database Collaboration (INSDC)
- A **RefSeq (GCF)** genome assembly represents an NCBI-derived copy of a submitted GenBank (GCA) assembly. In the majority of cases, the annotation is generated by the NCBI prokaryotic or eukaryotic genome annotation pipelines



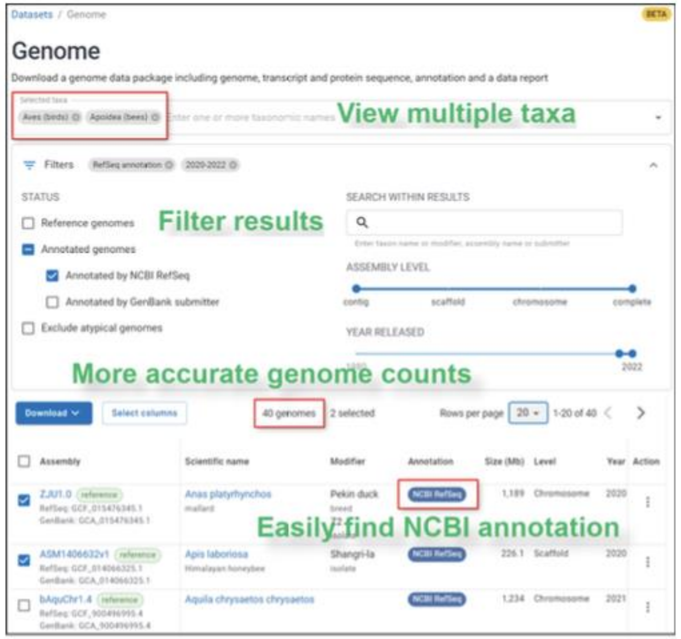
	GCA_	GCF_
Also known as	GenBank assembly	RefSeq assembly
Submitter-owned assembly archive	✓	✗
NCBI-maintained assembly copy	✗	✓
Always includes annotation	✗	✓
NCBI may add sequences (e.g. mitochondrial genomes)	✗	✓
NCBI may remove sequences (e.g. contamination)	✓ *	✓

* following submitter request or agreement

NCBI Datasets website genome sources

Source : Dataset documentation

NCBI Datasets : Datasets Genome Table



Datasets / Genome BETA

Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa: Aves (birds) Apoidea (bees) Enter one or more taxonomic names

Filters: RefSeq annotation 2020-2022

STATUS
□ Reference genomes
▣ Annotated genomes
☑ Annotated by NCBI RefSeq
□ Annotated by GenBank submitter
□ Exclude atypical genomes

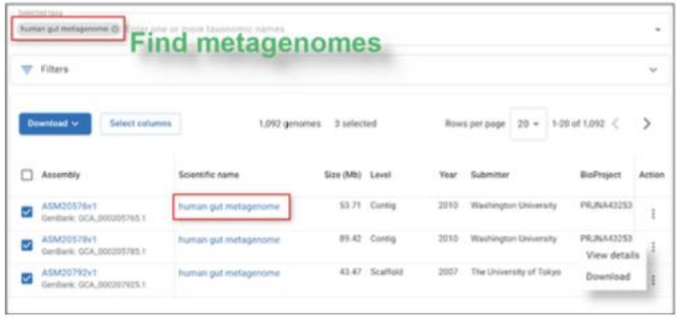
SEARCH WITHIN RESULTS
Enter taxon name or modifier, assembly name or submitter

ASSEMBLY LEVEL: contig scaffold chromosome complete

YEAR RELEASED: 2022

Download Select columns 40 genomes 2 selected Rows per page 20 1-20 of 40

☐	Assembly	Scientific name	Modifier	Annotation	Size (Mb)	Level	Year	Action
☑	ZJU1.0 reference RefSeq: GCF_015476345.1 GenBank: GCA_015476345.1	Anas platyrhynchos mallard	Pekin duck breed	NCBI RefSeq	1,189	Chromosome	2020	⋮
☑	ASM1406632v1 reference RefSeq: GCF_014066325.1 GenBank: GCA_014066325.1	Apis laboriosa Himalayan honeybee	Shangri-la isolate	NCBI RefSeq	226.1	Scaffold	2020	⋮
☐	bAquChr1.4 reference RefSeq: GCF_900496995.4 GenBank: GCA_900496995.4	Aquila chrysaetos chrysaetos		NCBI RefSeq	1,234	Chromosome	2021	⋮



Selected taxa: human gut metagenome Enter one or more taxonomic names

Filters

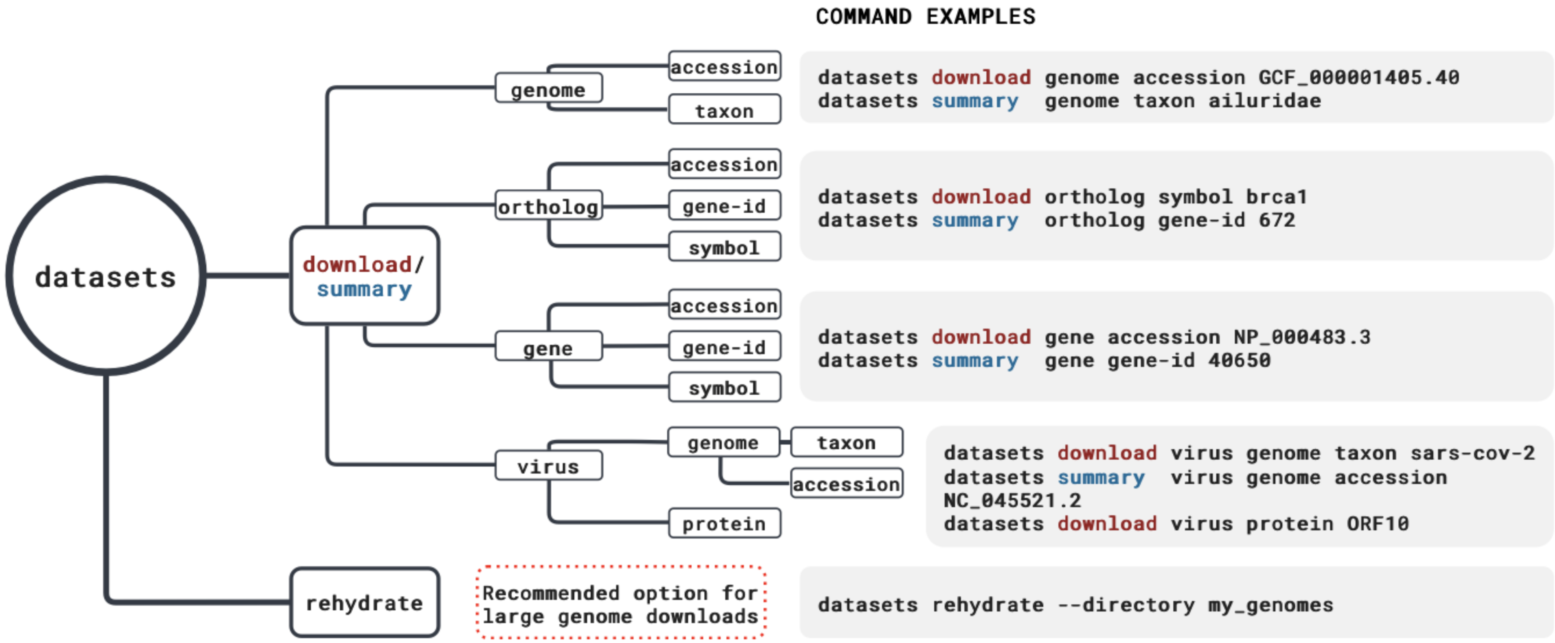
Download Select columns 1,092 genomes 3 selected Rows per page 20 1-20 of 1,092

☐	Assembly	Scientific name	Size (Mb)	Level	Year	Submitter	BioProject	Action
☑	ASM20576v1 GenBank: GCA_000205765.1	human gut metagenome	53.71	Contig	2010	Washington University	PRJNA43253	⋮
☑	ASM20578v1 GenBank: GCA_000205785.1	human gut metagenome	89.42	Contig	2010	Washington University	PRJNA43253	⋮ View details Download
☑	ASM20792v1 GenBank: GCA_000207925.1	human gut metagenome	43.47	Scaffold	2007	The University of Tokyo		⋮

Figure Source

- Find **all current genomes**, including metagenomes
  View **multiple taxa** such as birds and bees, or polyphyletic groups like fish
- Easily find genomes with **NCBI RefSeq** annotations
  Get more accurate genome counts, since **each row now represents a single genome with GenBank and RefSeq accessions** for that genome in the same row
- **Customize your downloads** to include either GenBank or RefSeq files, or both
  Download **tables** or **data packages**

NCBI Datasets : Command Line



COMMAND EXAMPLES

```
datasets download genome accession GCF_000001405.40
datasets summary  genome taxon ailuridae
```

```
datasets download ortholog symbol brca1
datasets summary  ortholog gene-id 672
```

```
datasets download gene accession NP_000483.3
datasets summary  gene gene-id 40650
```

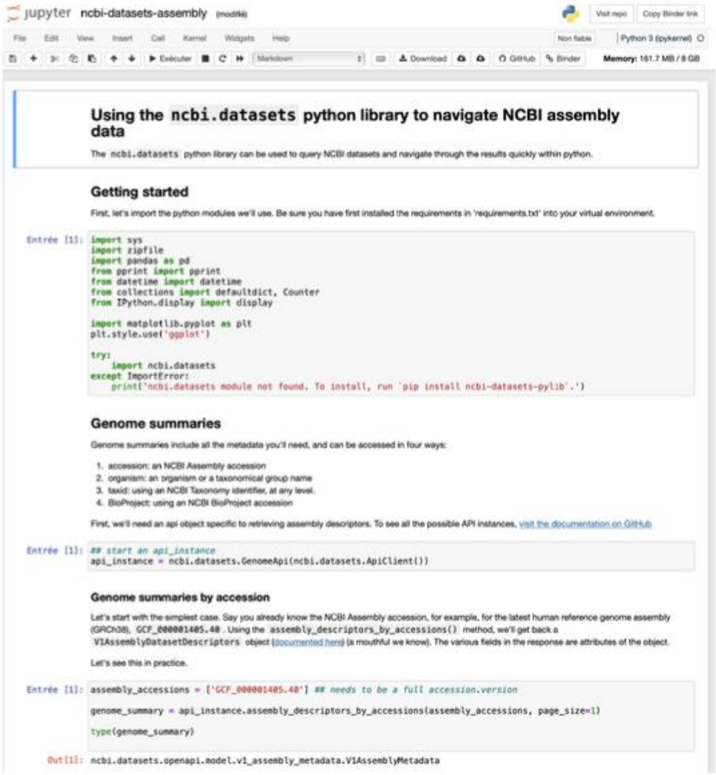
```
datasets download virus genome taxon sars-cov-2
datasets summary  virus genome accession
NC_045521.2
datasets download virus protein ORF10
```

```
datasets rehydrate --directory my_genomes
```

genome options :

- summary according to accession or taxid
- filter according to quality criteria & metadata
- download packages (or rehydrate) in various formats

NCBI Datasets : Aplication Programmatic Interface



Using the `ncbi.datasets` python library to navigate NCBI assembly data

The `ncbi.datasets` python library can be used to query NCBI datasets and navigate through the results quickly within python.

Getting started

First, let's import the python modules we'll use. Be sure you have first installed the requirements in 'requirements.txt' into your virtual environment.

```
Entrée [1]: import sys
import zipfile
import pandas as pd
from pprint import pprint
from datetime import datetime
from collections import defaultdict, Counter
from IPython.display import display

import matplotlib.pyplot as plt
plt.style.use('ggplot')

try:
    import ncbi.datasets
except ImportError:
    print('ncbi.datasets module not found. To install, run 'pip install ncbi-datasets-pylib'.')
```

Genome summaries

Genome summaries include all the metadata you'll need, and can be accessed in four ways:

1. accession: an NCBI Assembly accession
2. organism: an organism or a taxonomical group name
3. taxid: using an NCBI Taxonomy identifier, at any level.
4. BioProject: using an NCBI BioProject accession

First, we'll need an api object specific to retrieving assembly descriptors. To see all the possible API instances, visit the documentation on GitHub

```
Entrée [1]: ## start an api_instance
api_instance = ncbi.datasets.GenomeApi(ncbi.datasets.ApiClient())
```

Genome summaries by accession

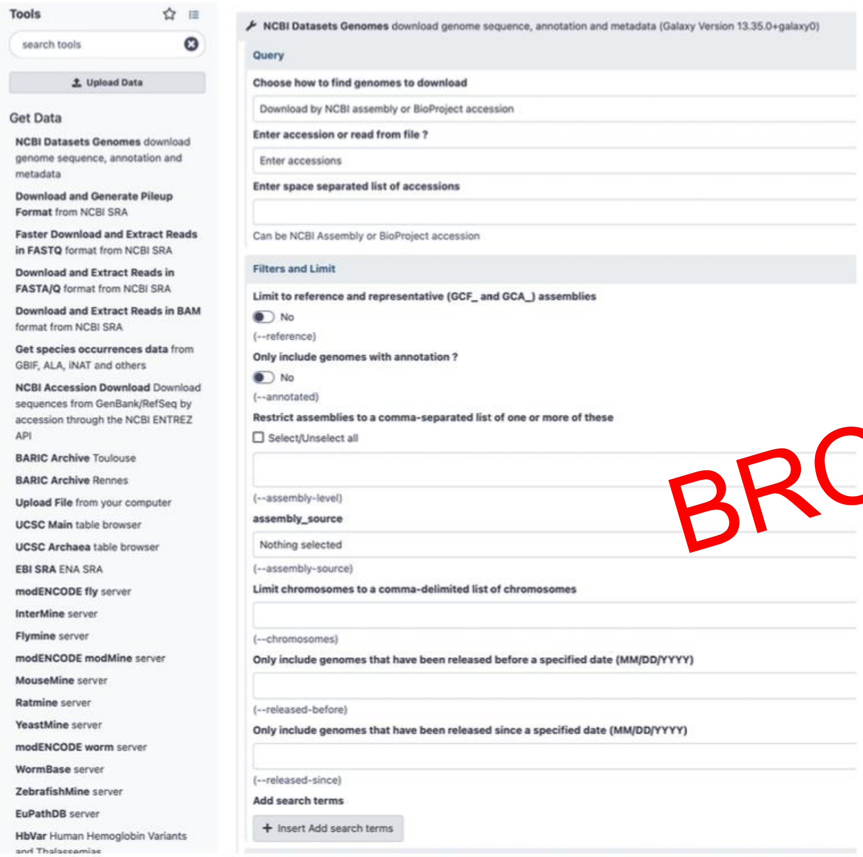
Let's start with the simplest case. Say you already know the NCBI Assembly accession, for example, for the latest human reference genome assembly (GRCh38), `GCF_000001405.40` . Using the `assembly_descriptors_by_accessions()` method, we'll get back a `V1AssemblyDatasetDescriptors` object (documented here) (a mouthful we know). The various fields in the response are attributes of the object.

Let's see this in practice.

```
Entrée [1]: assembly_accessions = ['GCF_000001405.40'] ## needs to be a full accession.version

genome_summary = api_instance.assembly_descriptors_by_accessions(assembly_accessions, page_size=1)

type(genome_summary)

Out[1]: ncbi.datasets.openapi.model.v1_assembly_metadata.V1AssemblyMetadata
```

Jupyter Notebook

NCBI Datasets : Galaxy Integration


Tools

search tools

Upload Data

Get Data

- **NCBI Datasets Genomes** download genome sequence, annotation and metadata
- **Download and Generate Pileup Format** from NCBI SRA
- **Faster Download and Extract Reads in FASTQ** format from NCBI SRA
- **Download and Extract Reads in FASTA/Q** format from NCBI SRA
- **Download and Extract Reads in BAM** format from NCBI SRA
- **Get species occurrences data** from GBIF, ALA, iNAT and others
- **NCBI Accession Download** Download sequences from GenBank/RefSeq by accession through the NCBI ENTREZ API
- **BARIC Archive** Toulouse
- **BARIC Archive** Rennes
- **Upload File** from your computer
- **UCSC Main** table browser
- **UCSC Archaea** table browser
- **EBI SRA** ENA SRA
- **modENCODE fly** server
- **InterMine** server
- **Flymine** server
- **modENCODE modMine** server
- **MouseMine** server
- **Ratmine** server
- **YeastMine** server
- **modENCODE worm** server
- **WormBase** server
- **ZebrafishMine** server
- **EuPathDB** server
- **HbVar** Human Hemoglobin Variants and Thalassemias

NCBI Datasets Genomes download genome sequence, annotation and metadata (Galaxy Version 13.35.0+galaxy0)

Query

Choose how to find genomes to download
Download by NCBI assembly or BioProject accession

Enter accession or read from file ?
Enter accessions

Enter space separated list of accessions

Can be NCBI Assembly or BioProject accession

Filters and Limit

Limit to reference and representative (GCF_ and GCA_) assemblies
No
(--reference)

Only include genomes with annotation ?
No
(--annotated)

Restrict assemblies to a comma-separated list of one or more of these
□ Select/Unselect all

(--assembly-level)

assembly_source
Nothing selected
(--assembly-source)

Limit chromosomes to a comma-delimited list of chromosomes

(--chromosomes)

Only include genomes that have been released before a specified date (MM/DD/YYYY)

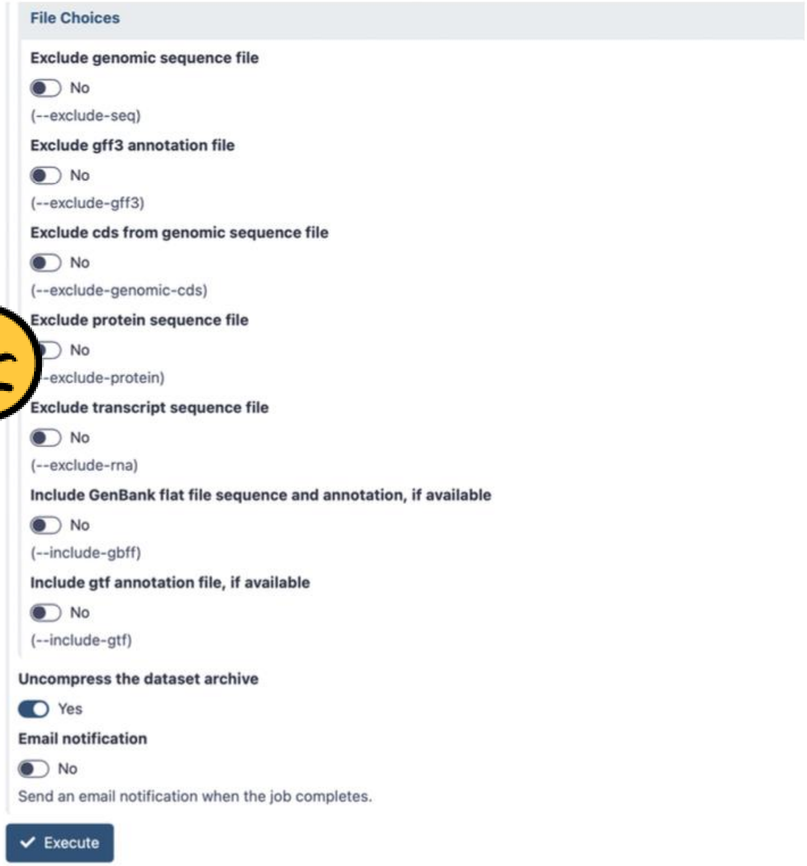
(--released-before)

Only include genomes that have been released since a specified date (MM/DD/YYYY)

(--released-since)

Add search terms
+ Insert Add search terms

BROKEN


File Choices

Exclude genomic sequence file
No
(--exclude-seq)

Exclude gff3 annotation file
No
(--exclude-gff3)

Exclude cds from genomic sequence file
No
(--exclude-genomic-cds)

Exclude protein sequence file
No
(--exclude-protein)

Exclude transcript sequence file
No
(--exclude-rna)

Include GenBank flat file sequence and annotation, if available
No
(--include-gbff)

Include gtf annotation file, if available
No
(--include-gtf)

Uncompress the dataset archive
Yes

Email notification
No
Send an email notification when the job completes.

✓ Execute

A wrapper of the command line tool

Parameters to define packages files

NCBI Datasets : Galaxy Integration

A few caveats of the wrapper :

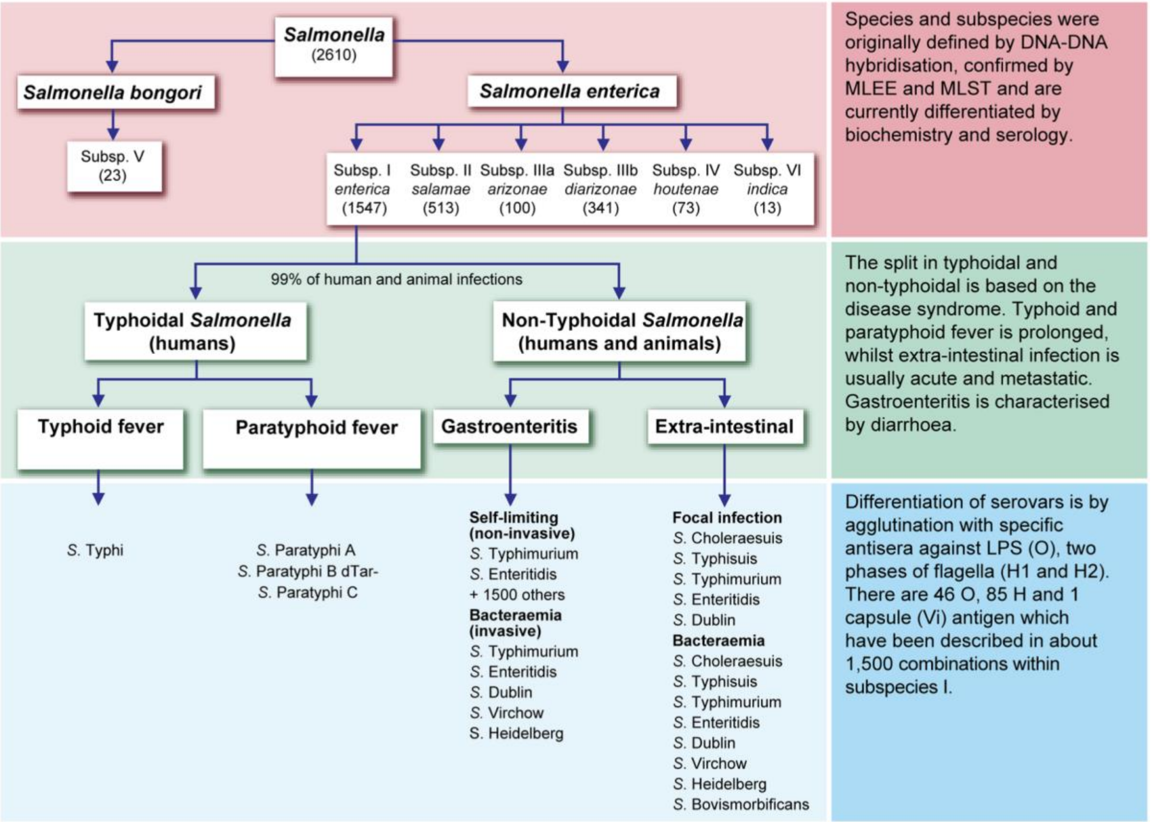
- Some (not so easy) errors when select / filter fails
- Often broken 😔
- Impossible to just download a list of **genomes as a file** and "**rehydrate**" it after

What we recomend to do [for now] :

- use the NCBI dataset genome page to **browse / filter a list of genomes** of interest, and choose one of these solution
  - **A few genomes** :
    - Download the genomes as package un re-upload into Galaxy 😔
  - **A lot of genomes :**
    - Extract the URL of interest of the files from the TSV, feed it into Galaxy

The training datasets

We will work on 3 datasets of public Salmonella genomes



Salmonella (2610)

Salmonella bongori

Subsp. V (23)

Salmonella enterica

Subsp. I *enterica* (1547)

Subsp. II *salamae* (513)

Subsp. IIIa *arizonae* (100)

Subsp. IIIb *diarizonae* (341)

Subsp. IV *houtenae* (73)

Subsp. VI *indica* (13)

Species and subspecies were originally defined by DNA-DNA hybridisation, confirmed by MLEE and MLST and are currently differentiated by biochemistry and serology.

99% of human and animal infections

Typhoidal *Salmonella* (humans)

Non-Typhoidal *Salmonella* (humans and animals)

Typhoid fever

Paratyphoid fever

Gastroenteritis

Extra-intestinal

The split in typhoidal and non-typhoidal is based on the disease syndrome. Typhoid and paratyphoid fever is prolonged, whilst extra-intestinal infection is usually acute and metastatic. Gastroenteritis is characterised by diarrhoea.

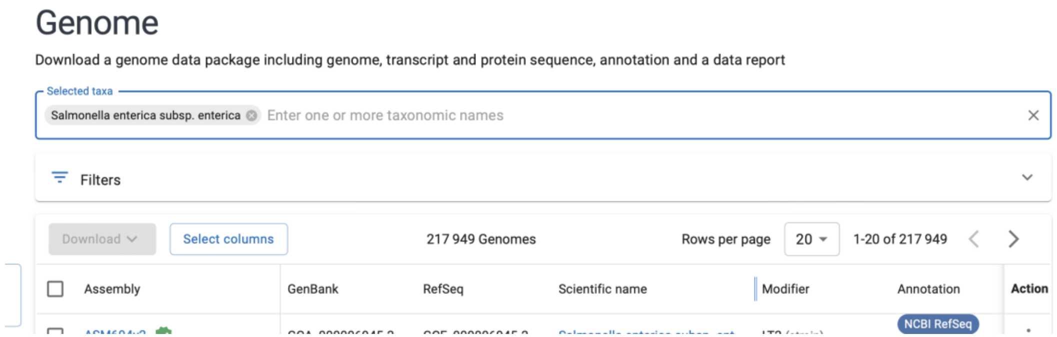
S. Typhi

S. Paratyphi A
S. Paratyphi B dTar-
S. Paratyphi C

Self-limiting (non-invasive)
S. Typhimurium
S. Enteritidis
+ 1500 others
Bacteraemia (invasive)
S. Typhimurium
S. Enteritidis
S. Dublin
S. Virchow
S. Heidelberg

Focal infection
S. Choleraesuis
S. Typhisuis
S. Typhimurium
S. Enteritidis
S. Dublin
Bacteraemia
S. Choleraesuis
S. Typhisuis
S. Typhimurium
S. Enteritidis
S. Dublin
S. Virchow
S. Heidelberg
S. Bovismorbificans

Differentiation of serovars is by agglutination with specific antisera against LPS (O), two phases of flagella (H1 and H2). There are 46 O, 85 H and 1 capsule (Vi) antigen which have been described in about 1,500 combinations within subspecies I.



Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa: Salmonella enterica subsp. enterica ⊗ Enter one or more taxonomic names ✕

217 949 Genomes

Rows per page 20 ▾ 1-20 of 217 949

□ Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action

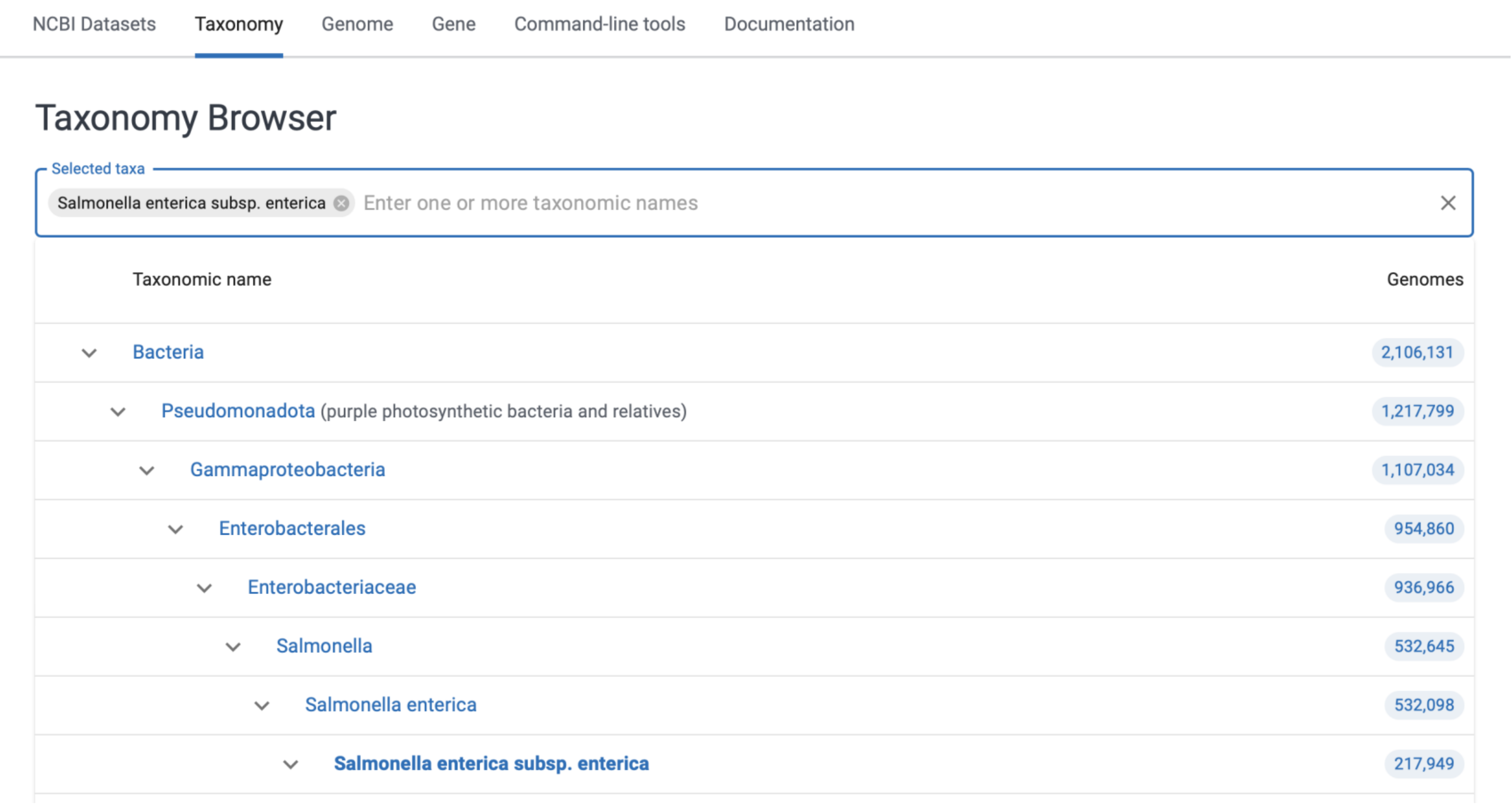
- *217 949 Salmonella enterica enterica* public assemblies at NCBI!!

Training

We will construct 3 datasets of public *Salmonella* genomes

- **Dataset 1**: list all Salmonella enterica subsp. enterica assemblies using their taxon id and assembly level (Chromosome)
- **Dataset 2**: list all the **Salmonella bongori** assemblies to choose and download the best outgroup of a salmonella enterica dataset
- **Dataset 3**: download 16 Salmonella enterica public assemblies (2 sub-species, 4 serotypes) from their accession numbers

Data set 1 : Taxonomy browser



Taxonomy Browser

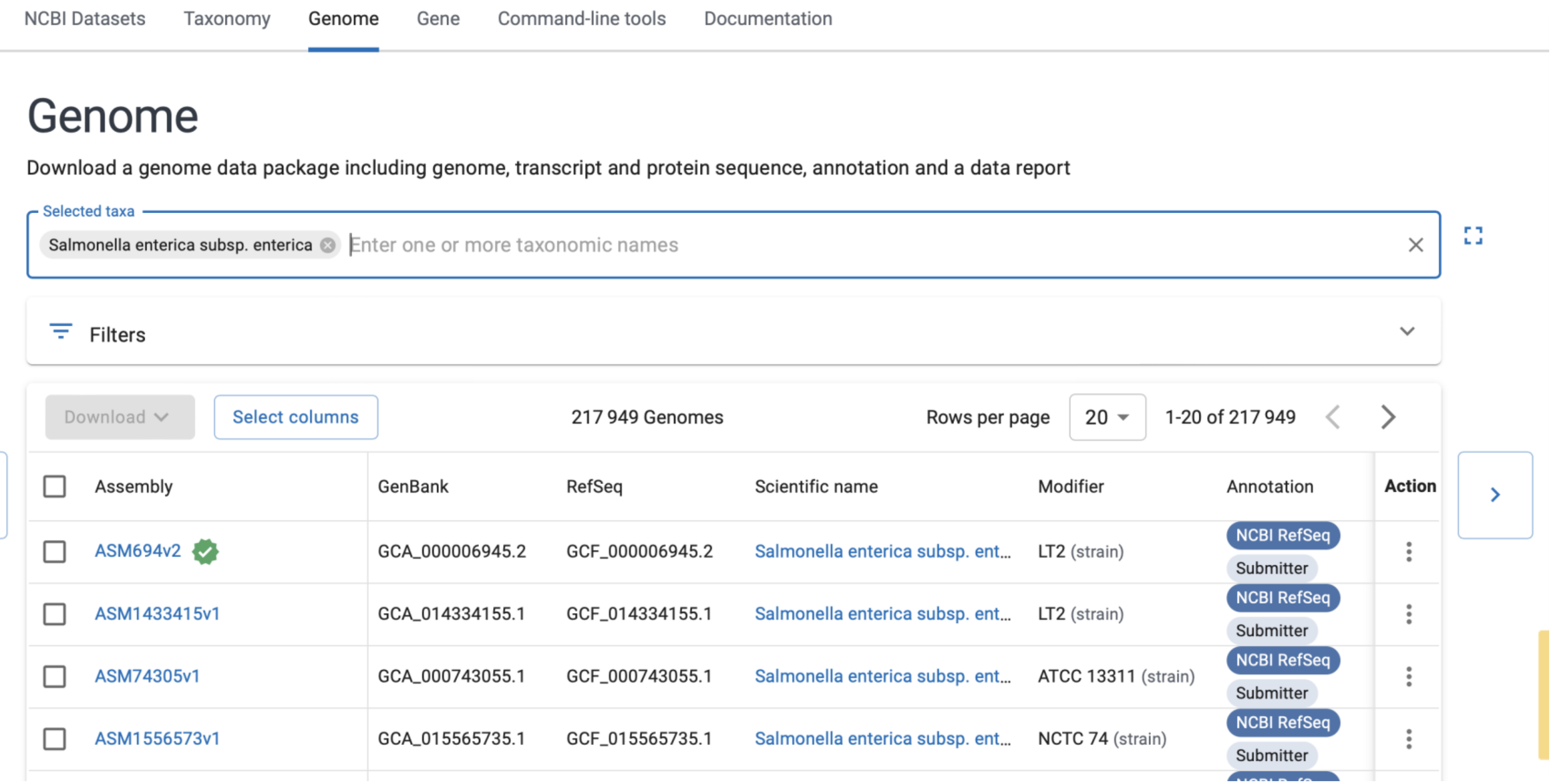
Selected taxa

Salmonella enterica subsp. enterica ⊗ Enter one or more taxonomic names ✕

Taxonomic name	Genomes
Bacteria	2,106,131
Pseudomonadota (purple photosynthetic bacteria and relatives)	1,217,799
Gammaproteobacteria	1,107,034
Enterobacterales	954,860
Enterobacteriaceae	936,966
Salmonella	532,645
Salmonella enterica	532,098
Salmonella enterica subsp. enterica	217,949

List all *Salmonella enterica subsp. enterica* assemblies using their *taxon id* and *assembly level (Chromosome)*

Data set 1 : genome table



Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

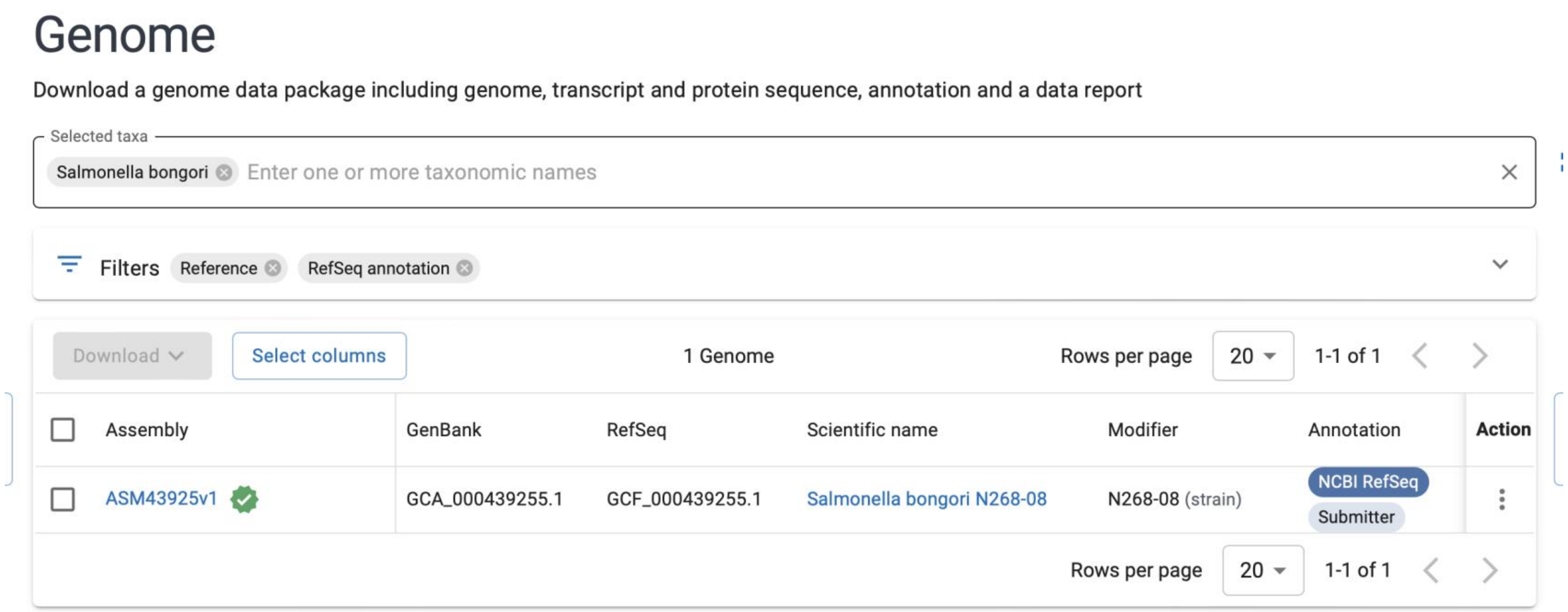
Selected taxa: Salmonella enterica subsp. enterica | Enter one or more taxonomic names

217 949 Genomes

□	Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action
□	ASM694v2 ✓	GCA_000006945.2	GCF_000006945.2	Salmonella enterica subsp. ent...	LT2 (strain)	NCBI RefSeq Submitter	⋮
□	ASM1433415v1	GCA_014334155.1	GCF_014334155.1	Salmonella enterica subsp. ent...	LT2 (strain)	NCBI RefSeq Submitter	⋮
□	ASM74305v1	GCA_000743055.1	GCF_000743055.1	Salmonella enterica subsp. ent...	ATCC 13311 (strain)	NCBI RefSeq Submitter	⋮
□	ASM1556573v1	GCA_015565735.1	GCF_015565735.1	Salmonella enterica subsp. ent...	NCTC 74 (strain)	NCBI RefSeq Submitter	⋮

- Notice :
  - Filter parameters
  - Select columns button
  - Download button (table or package)

•Data set 2: download a genome



Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa: Salmonella bongori ⊗ Enter one or more taxonomic names ✕

Filters: Reference ⊗ RefSeq annotation ⊗

Download ⌄ | Select columns | 1 Genome | Rows per page 20 ▾ | 1-1 of 1 ‹ ›

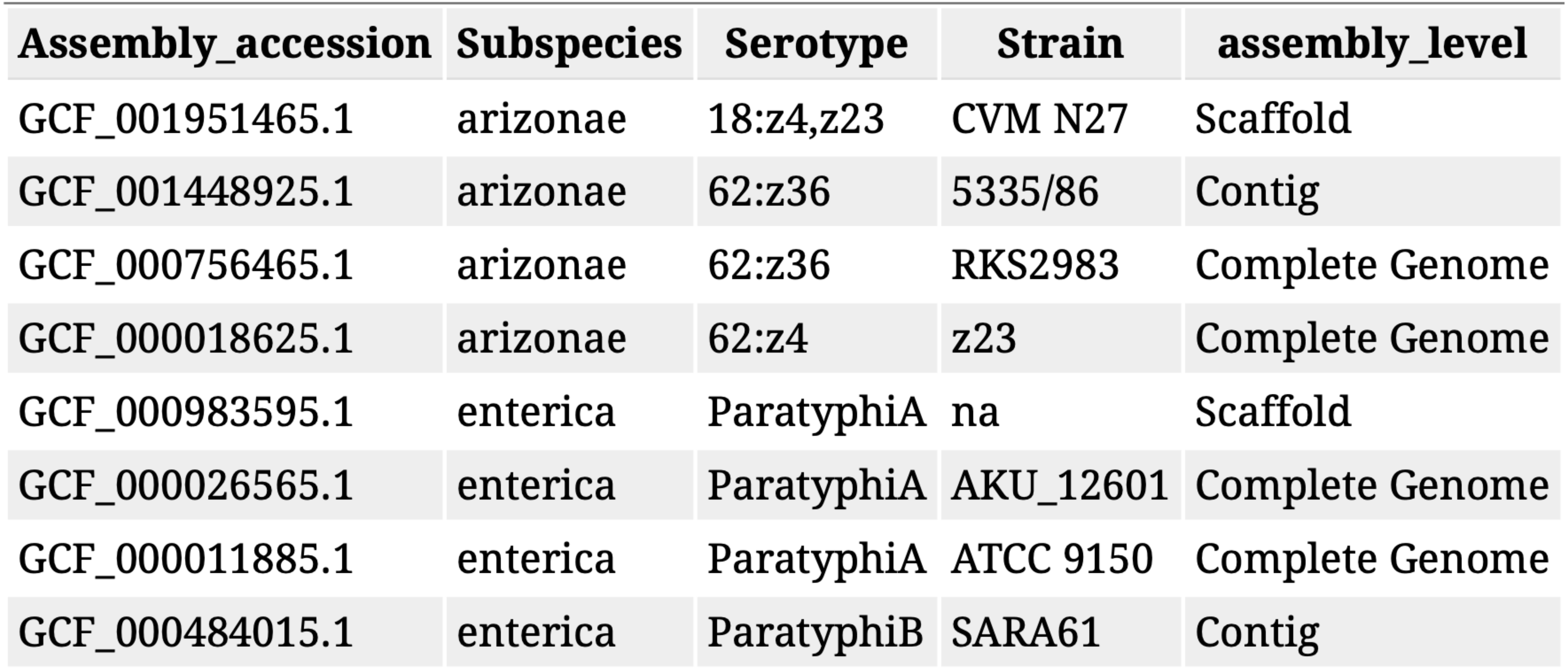
□	Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action
□	ASM43925v1 ✓	GCA_000439255.1	GCF_000439255.1	Salmonella bongori N268-08	N268-08 (strain)	NCBI RefSeq Submitter	⋮

Rows per page 20 ▾ | 1-1 of 1 ‹ ›

List all the *Salmonella bongori assemblies* to **choose the best outgroup** of a *Salmonella enterica* dataset

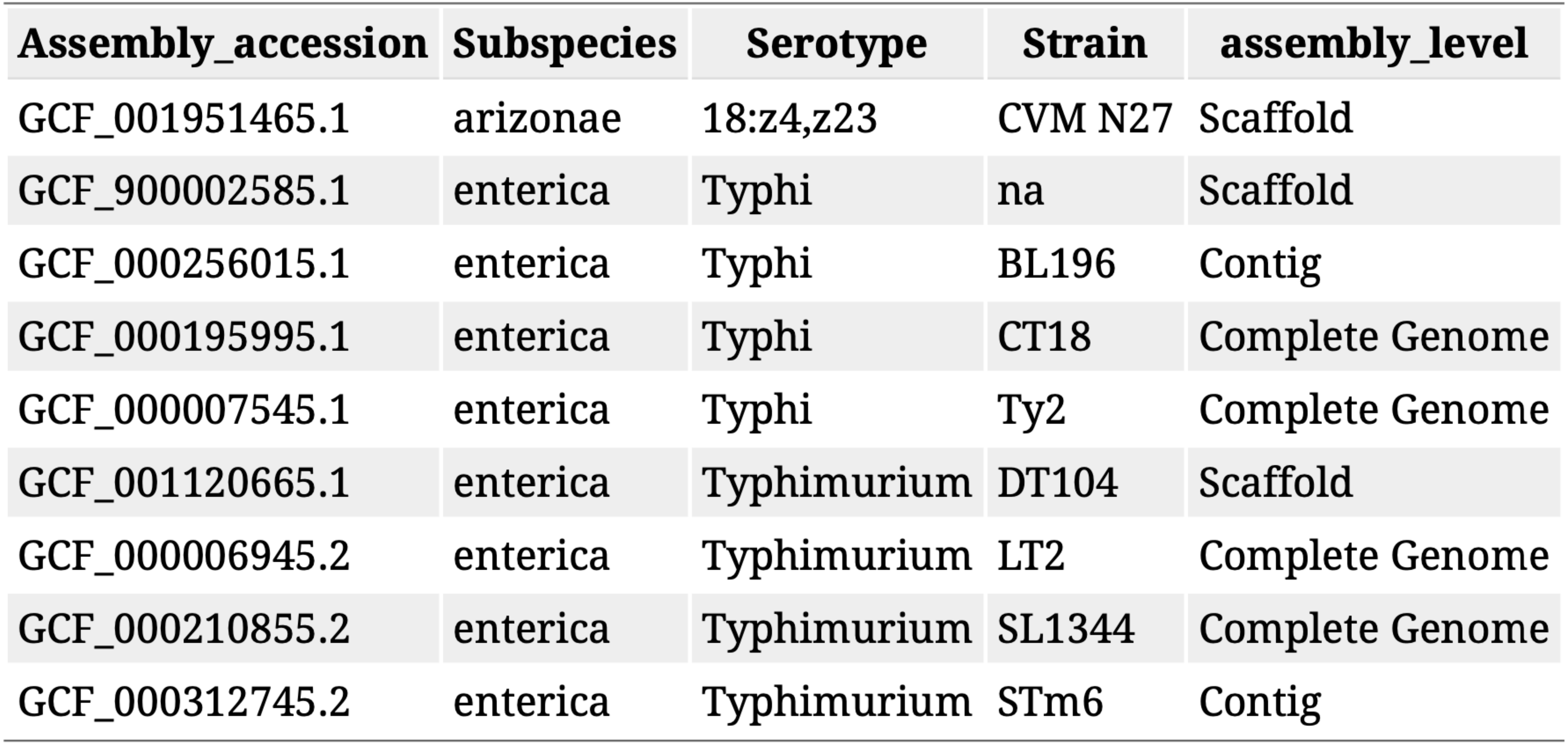
Download the genome in Genbank, nuclotide and fasta and GFF format

Data set 3 : 16 S. enterica public genomes (part 1)



Assembly_accession	Subspecies	Serotype	Strain	assembly_level
GCF_001951465.1	arizonae	18:z4,z23	CVM N27	Scaffold
GCF_001448925.1	arizonae	62:z36	5335/86	Contig
GCF_000756465.1	arizonae	62:z36	RKS2983	Complete Genome
GCF_000018625.1	arizonae	62:z4	z23	Complete Genome
GCF_000983595.1	enterica	ParatyphiA	na	Scaffold
GCF_000026565.1	enterica	ParatyphiA	AKU_12601	Complete Genome
GCF_000011885.1	enterica	ParatyphiA	ATCC 9150	Complete Genome
GCF_000484015.1	enterica	ParatyphiB	SARA61	Contig

Data set 3 : 16 S. enterica public genomes (part 1)



Assembly_accession	Subspecies	Serotype	Strain	assembly_level
GCF_001951465.1	arizonae	18:z4,z23	CVM N27	Scaffold
GCF_900002585.1	enterica	Typhi	na	Scaffold
GCF_000256015.1	enterica	Typhi	BL196	Contig
GCF_000195995.1	enterica	Typhi	CT18	Complete Genome
GCF_000007545.1	enterica	Typhi	Ty2	Complete Genome
GCF_001120665.1	enterica	Typhimurium	DT104	Scaffold
GCF_000006945.2	enterica	Typhimurium	LT2	Complete Genome
GCF_000210855.2	enterica	Typhimurium	SL1344	Complete Genome
GCF_000312745.2	enterica	Typhimurium	STm6	Contig

Data set 3 : from a tabular file

Download 16 Salmonella enterica public assemblies (2 sub-species, 4 serotypes) from their *accession numbers*.

Input : [Filtered] List of assembly accession in a tabular file downloaded from **Dataset genome Table**

- Import ncbi_dataset_salmonella_genome_table.tsv from Shared Data / Data Library / EBAII A&A 2022 / Prokaryotic Annotation / NCBI Dataset
- **Filter lines concerning Refseq assemblies** (starts with "GCF_") using **Select lines that match an expression** tool
- Select the first column of the file (Assembly Accession) using Cut columns from a table
- Feed **NCBI Datasets Genomes** download genome sequence, annotation and metadata with the list of accession

BROKEN 😓

- Retrieve all file format of interest **including** genbank annotated files
  - [also in Shared Data / Data Library/ Libraries /EBAII A&A 2022 Prokaryotic annotation/ Salmonella dataset]

BROKEN 😓

Construction and analysis of prokaryotic genomic dataset

Outline

> Constructing a genome dataset

> Analyzing the genome dataset

> Comparing and dereplicating the dataset

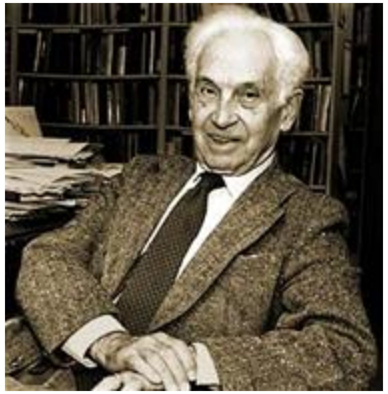
Analyzing a genome dataset

Why?

- Frequent problems in genome analysis and comparison
  - Heterogenous quality of sequencing and assembly
  - Presence of huge number or public genomes OR absence of any close genomes of the same species in public databases
  - Difficulties regarding microbial taxonomy (classification) and nomenclature (naming of genus, species and strain naming) for many non-model organisms

Introduction

- What is a species?



Ernst Mayr (1942) :

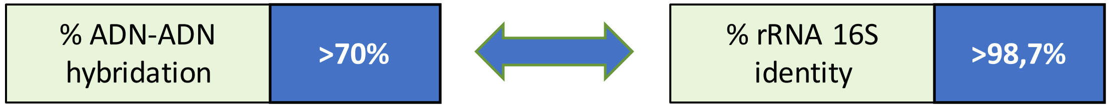
"Species are groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups"

⇨ ***Not relevant for bacteria***

What is a bacterial species?

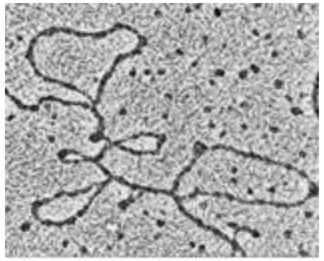
No consensual definition for procaryotes

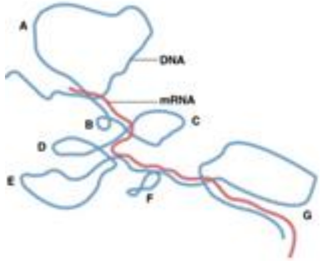
- No universal criteria
- Several approaches used to classify bacterial
    - Phenotypes and morphological criteria
    - DNA-DNA hybridization
- Universal markers
    - 16S rRNA
    - MLST (Multi Locus Sequence Typing)
- Genomic-based taxonomy are now becoming a gold-standard

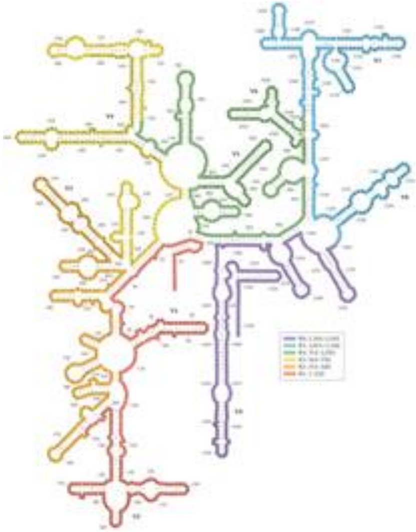


% ADN-ADN hybridation | >70%

% rRNA 16S identity | >98,7%



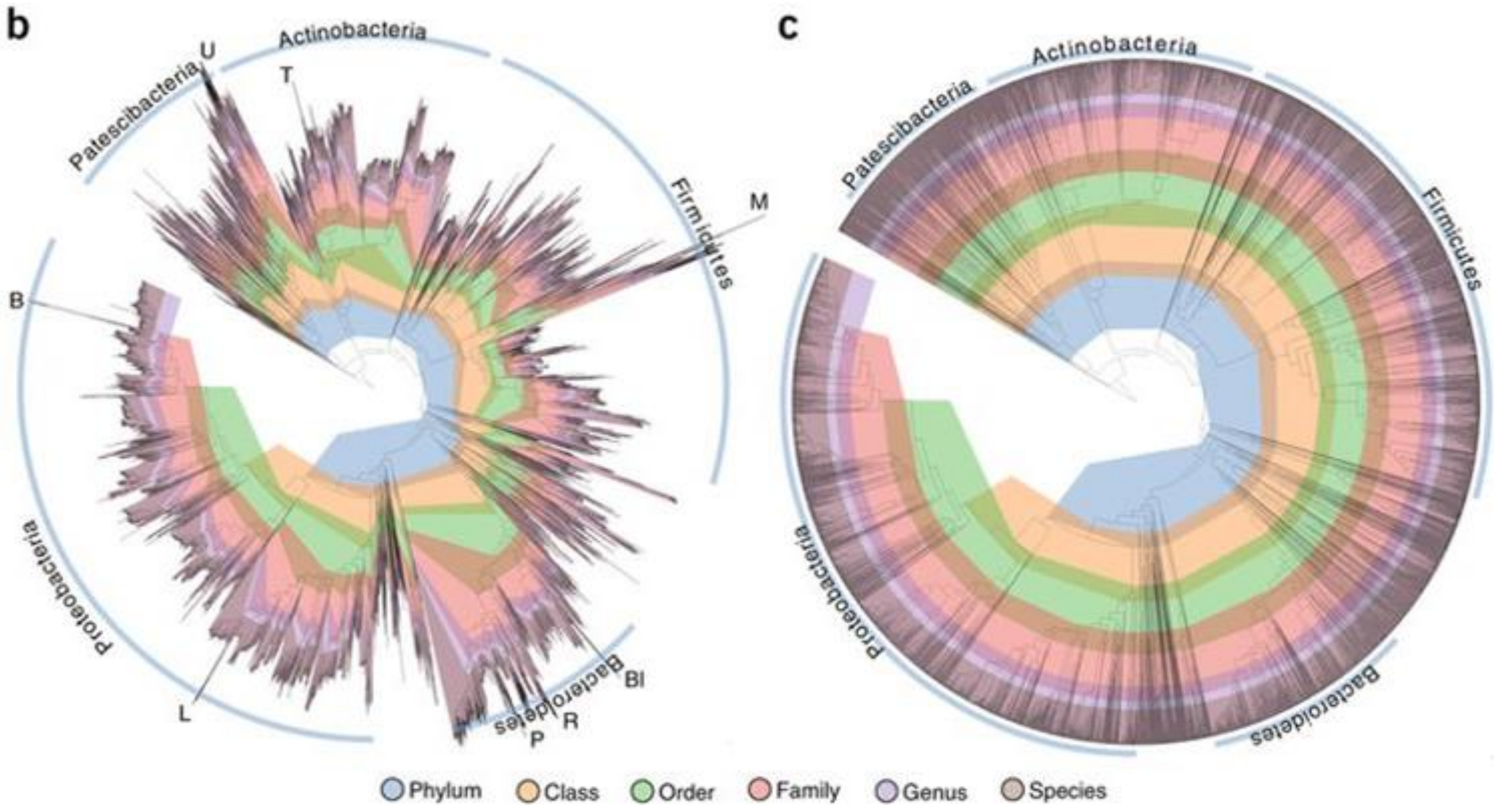




Example: the Genome-based taxonomy for prokaryotic genomes

- Objective: a standardized microbial taxonomy based on genome phylogeny

- Taxonomy inferred from concatenated single copy marker proteins



Parks et al. 2018, 2021

https://gtdb.ecogenomic.org/

Evaluating genome diversity in a dataset

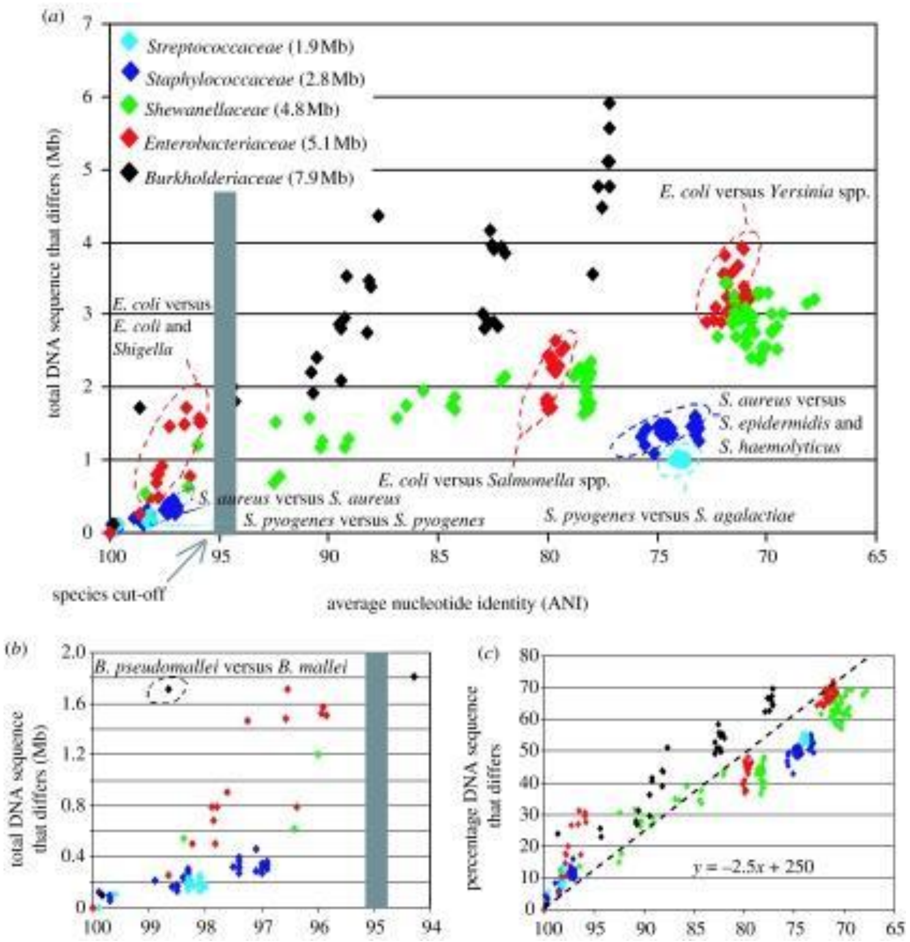
- Why ?
  - Identify outlier genomes
  - Identify groups of (very) similar genomes and de-replicate datasets
  - Estimate genome similarity in a dataset and design an adapted comparative strategy

How ?

- Alignment based approaches (ANI)
- k-mer based approaches (MASH)

Average Nucleotide Identity (ANI)

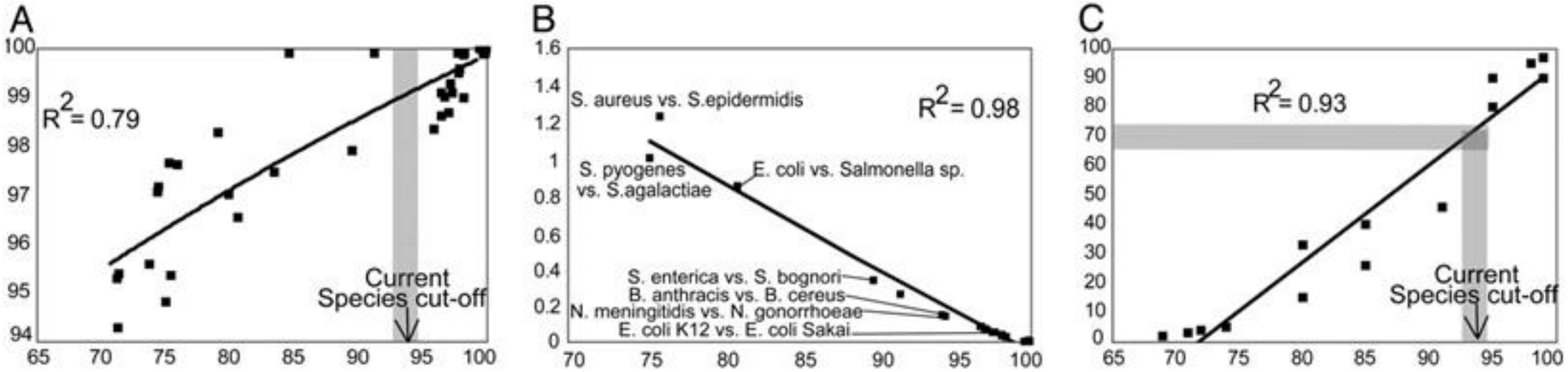
- Meet the need for a robust measure of genomic relatedness and a systematic and scalable species assignation technique
- Mean identity percent of aligned regions of a pair of genomes
- Rely on pairwise alignments from
  - aligned core genes
  - genomic alignments
- Can easily be used to build phylogenetics tree using distance methods
- Is implemented in several bioinformatics tools: ANIn (nucmer based, Richter 2009) gANI (coding regions, Varghese 2015), fastANI (computing efficiency, Jain 2018), …



Genetic diversity within five important bacterial groups. Konstantidinis et al. 2006. The bacterial species definition in the genomic era
DOI: 10.1098/rstb.2006.1920

Average Nucleotide Identity (ANI)

- ANI strongly correlates (R = 0.79 for logarithmic correlation) with the 16S rRNA gene sequence identity and can resolve areas where the 16S rRNA gene is inadequate (intra-species level)
- The average rate of synonymous substitutions shows a tight correspondence to ANI, suggesting that ANI may also be a useful descriptor of the evolutionary distance
- ANI shows a strong linear correlation to DNA–DNA reassociation values, and the 70% DNA–DNA reassociation standard corresponds to ≈93–94% ANI i.e. strains that show >94% ANI should belong to the same species



ANI vs 16 rRNA %id

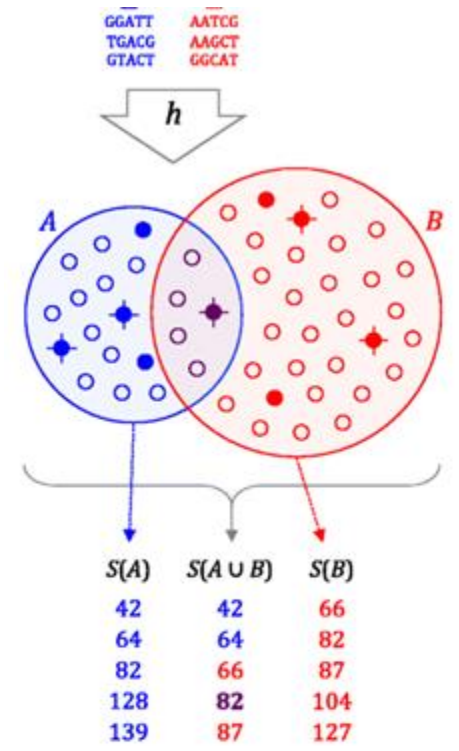
ANI vs average syn. Mutation rate

ANI vs DNA-DNA reassociation value

Konstantidinis et al. 2005. Genomic insights that advance the species definition for prokaryotes
https://doi.org/10.1073/pnas.0409727102

MASH: fast (meta)genome distance estimation using MinHash

- Mash allows to compute a pairwise mutation distance without alignment using k-mer counts
- Mash provides two basic functions for sequence comparisons:
  - **sketch**: converts a sequence or collection of sequences into a MinHash sketch
  - **dist**: compares two sketches and returns an estimate of the Jaccard index (i.e. the fraction of shared k- mers), a P value, and the Mash distance



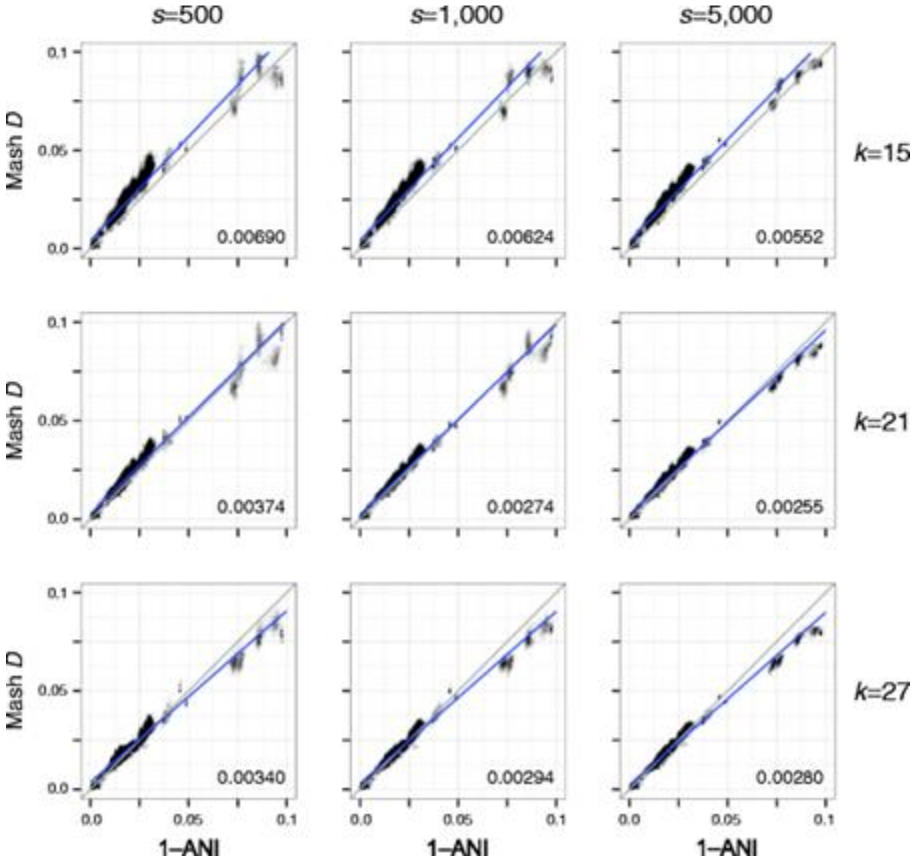
$S(A)$	$S(A \cup B)$	$S(B)$
42	42	66
64	64	82
82	66	87
128	**82**	104
139	**87**	127

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

Ondov, B.D., Treangen, T.J., Melsted, P. et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17, 132 (2016). https://doi.org/10.1186/s13059-016-0997-x

MASH distances correlate well with ANI

- Dataset: 500 complete E. coli genomes
  - Each plot column shows a different sketch size
  - Each plot row a different k-mer size k.
  - Gray lines: model relationship D = 1–ANI
- Increasing the sketch size improves the accuracy of the MASH distance, especially for more divergent sequences.
- Limit on how well the MASH distance can approximate ANI, especially for more divergent genomes (e.g. ANI considers only the core genome)



Ondov, B.D., Treangen, T.J., Melsted, P. et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17, 132 (2016). https://doi.org/10.1186/s13059-016-0997-x

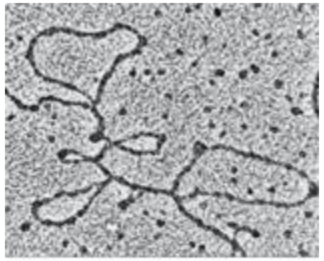
Back to procaryote taxonomy

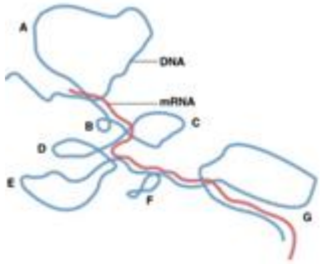
% ADN-ADN hybridation >70%

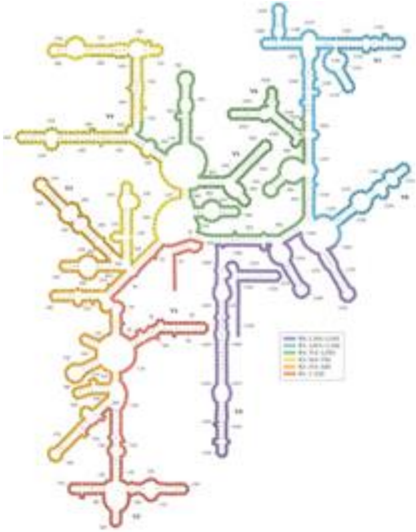
% rRNA 16S identity >98,7%

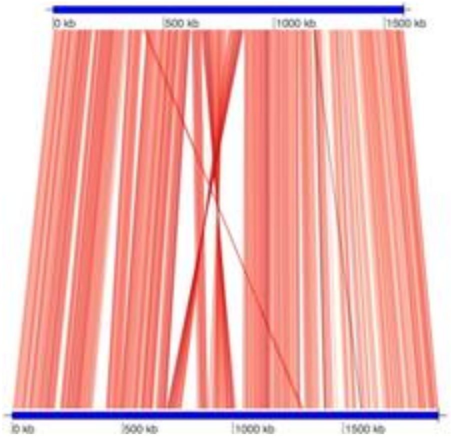
% genome identity (ANI) >94%

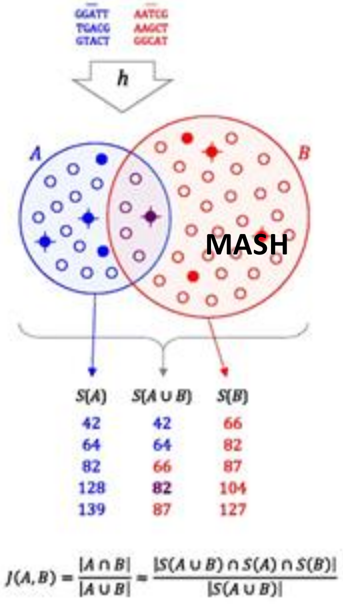
K-mer distance <0,06











Quality Control & filter

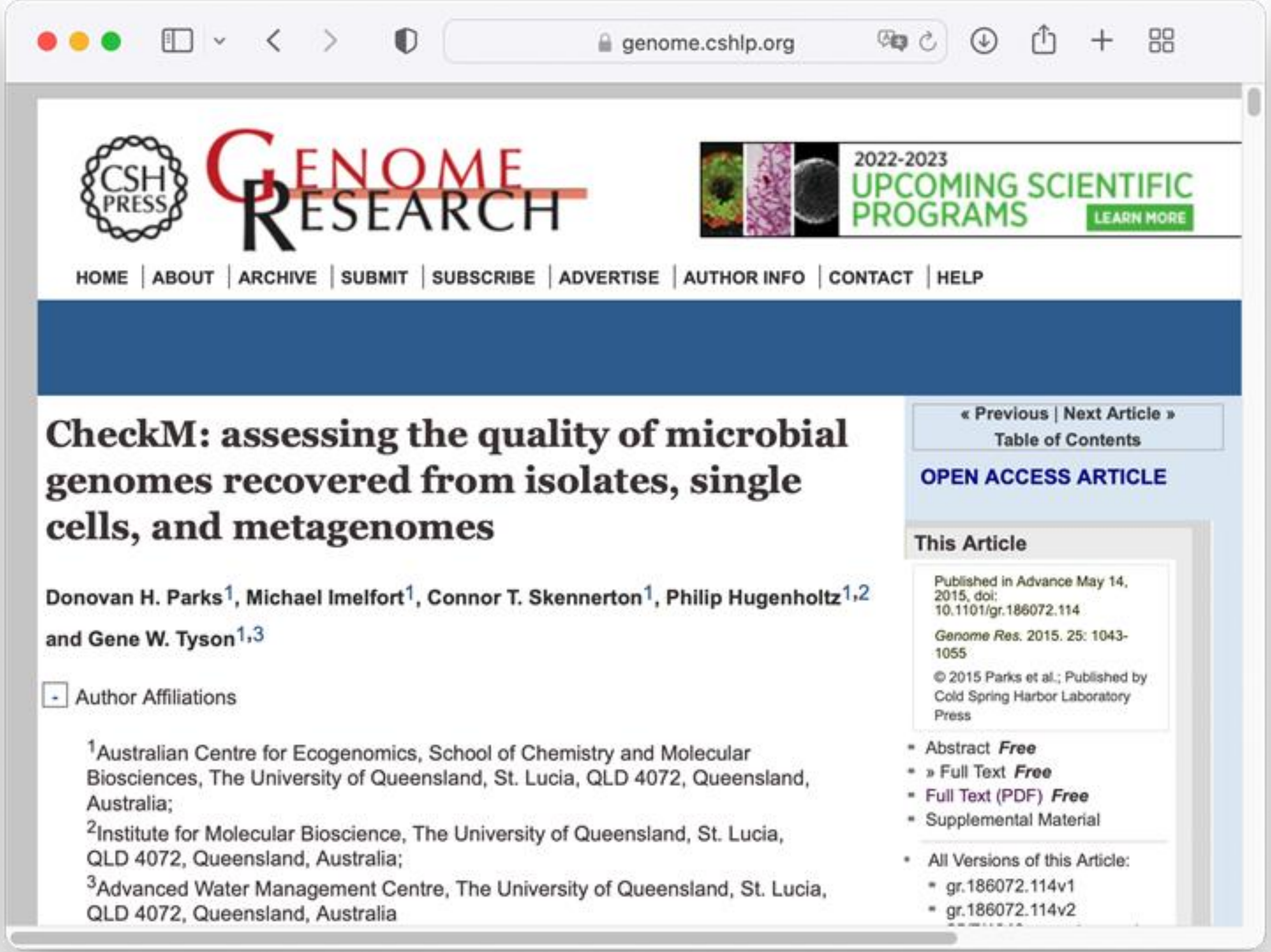
- Already presented: evaluate Quality using the 3Cs

1. **Contiguity**. Produce the longest possible contigs.
2. **Correctness**. Assemble contigs with few/no errors.
3. **Completeness**. Cover the entire original sequence and minimize missing regions

- An additional key point for microbes: evaluate **Contamination**
  - From genomic fragments of divergent taxa
  - From genomic fragments of multiple strains (i.e. strain **heterogeneity**)
  - Be sure that the "announced" taxa is good

CheckM

- a set of tools for assessing the quality of genomes recovered from isolates, single cells, or metagenomes
- provides robust estimates of genome **completeness** and **contamination**
    - use collocated sets of genes that are ubiquitous and single-copy within a phylogenetic lineage
    - propose a fixed vocabulary for defining genome quality based on estimates of completeness and contamination
- Evaluate by simulations the accuracy of quality estimates


genome.cshlp.org
CSH PRESS
GENOME RESEARCH
2022-2023
UPCOMING SCIENTIFIC PROGRAMS
LEARN MORE
HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP
« Previous | Next Article »
Table of Contents
OPEN ACCESS ARTICLE

CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes

Donovan H. Parks[1], Michael Imelfort[1], Connor T. Skennerton[1], Philip Hugenholtz[1,2] and Gene W. Tyson[1,3]

□ Author Affiliations

[1] Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia, QLD 4072, Queensland, Australia;
[2] Institute for Molecular Bioscience, The University of Queensland, St. Lucia, QLD 4072, Queensland, Australia;
[3] Advanced Water Management Centre, The University of Queensland, St. Lucia, QLD 4072, Queensland, Australia

This Article

Published in Advance May 14, 2015, doi: 10.1101/gr.186072.114

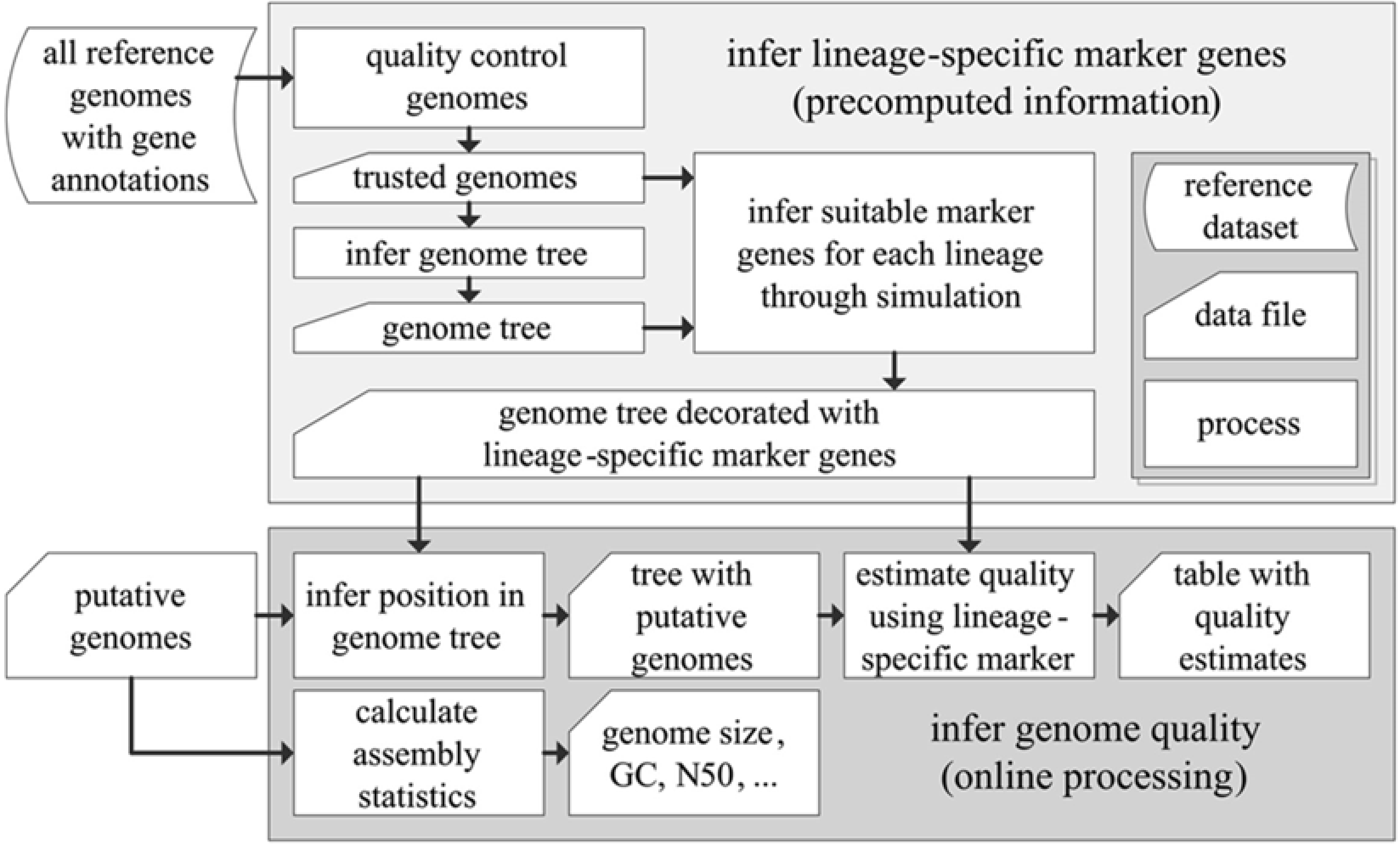
Genome Res. 2015. 25: 1043-1055

© 2015 Parks et al.; Published by Cold Spring Harbor Laboratory Press

- Abstract ***Free***
- » Full Text ***Free***
- Full Text (PDF) ***Free***
- Supplemental Material

- All Versions of this Article:
    - gr.186072.114v1
    - gr.186072.114v2

CheckM consists of a workflow for precomputing lineage-specific marker genes for each branch within a reference genome tree (top box) and an online workflow for inferring the quality of putative genomes (bottom box).


all reference genomes with gene annotations
quality control genomes
infer lineage-specific marker genes (precomputed information)
trusted genomes
infer suitable marker genes for each lineage through simulation
reference dataset
infer genome tree
data file
genome tree
genome tree decorated with lineage-specific marker genes
process
putative genomes
infer position in genome tree
tree with putative genomes
estimate quality using lineage-specific marker
table with quality estimates
calculate assembly statistics
genome size, GC, N50, ...
infer genome quality (online processing)

Donovan H. Parks et al. Genome Res. 2015;25:1043-1055

CheckM relies on several other tools and data

- *prodigal* to predict genes
- A reference genome tree based on 43 phylogenetically informative marker genes and 5656 trusted reference genomes
  - Marker genes are identified in assemblies using HMMER
  - The resulting genes are used to placed the genome into the tree using *pplacer*
- Lineage-specific marker sets determined for all nodes within the reference genome tree by identifying single-copy genes present in ≥97% of all descendant genomes.

CheckM report

Provides classic quality metrics and plots, including:

- Results of binning

  >Marker lineage, #genomes, #markers, #marker sets

- CheckM metrics

  > Completeness, Contamination, Strain heterogeneity

- Classical Quality metrics

> #ambiguous bases, #scaffolds, #contigs, N50 (scaffolds), N50 (contigs), Mean scaffold length (bp),Mean contig length (bp), Longest scaffold (bp),Longest contig (bp), GC, GC std (scaffolds > 1kbp)

CheckM report – binning part

Marker lineage: indicates the taxonomic rank of the lineage-specific marker set used to estimated genome completeness, contamination, and strain heterogeneity.

#genomes: number of reference genomes used to infer the lineage-specific marker set

#markers: number of marker genes within the inferred lineage-specific marker set

#marker sets: number of co-located marker sets within the inferred lineage-specific marker set

0-5+: number of times each marker gene is identified

CheckM report

- Completeness: estimated completeness of genome as determined from the presence/absence of marker genes and the expected colocalization of these genes
- Contamination: estimated contamination of genome as determined by the presence of multi-copy marker
- Strain heterogeneity: % determined from the number of multi-copy marker pairs which exceed a specified **amino acid identity threshold** (default = 90%).
  - High strain heterogeneity suggests the majority of reported contamination is from one or more closely related organisms (i.e. potentially the same species),
  - Low strain heterogeneity suggests the majority of contamination is from more phylogenetically diverse sources

CheckM: proposed genome quality classification scheme

- ***Finished genomes***: genomes assembled into a single contiguous sequence containing no gaps or ambiguities and where extensive efforts have been made to identify errors
- ***Noncontiguous finished:*** *g*enomes assembled into multiple sequences as a result of repetitive regions, but otherwise of a finished quality
- ***Draft genomes:*** all other genomes

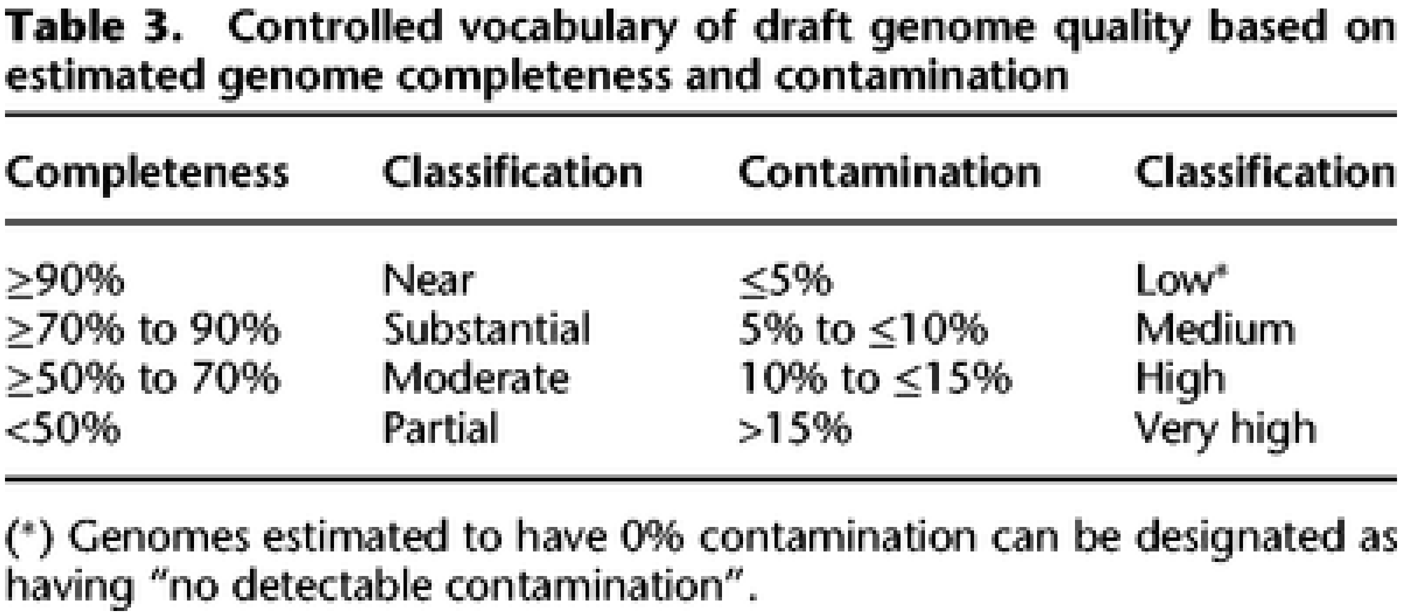


Table 3. Controlled vocabulary of draft genome quality based on estimated genome completeness and contamination

Completeness	Classification	Contamination	Classification
≥90%	Near	≤5%	Low*
≥70% to 90%	Substantial	5% to ≤10%	Medium
≥50% to 70%	Moderate	10% to ≤15%	High
<50%	Partial	>15%	Very high

(*) Genomes estimated to have 0% contamination can be designated as having "no detectable contamination".

Donovan H. Parks et al. Genome Res. 2015;25:1043-1055

⚠ **Evolution of the thresholds with the evolution of sequencing technologies** :

- Those threshold are now more for **MAGs** (see Pasolli et al. 2019,Bowers et al., 2017)
- For **isolates genome**, for instance , Refseq defines high quality as >98

CheckM result interpretation limits

- CheckM is dedicated to eubacterial and archeal genomes
  - Eukaryotic or phage genomes will be reported as highly incomplete
  - The quality of plasmids must also be assessed independently of CheckM
- The novelty of a genome will also influence the accuracy of CheckM estimates
  - Estimates for bacterial and archaeal genomes from deep basal lineages with few reference genomes are generally based on domain-level marker sets
  - Quality estimates may be not reliable for genomes of novel lineages
  - Gene loss or duplication may be an issue

Conclusion : use CheckM as a tool to detect outliers and further investigate!

Construction and analysis of prokaryotic genomic dataset

Outline

> Constructing a genome dataset

> Analyzing the genome dataset

> **Comparing and dereplicating the dataset**

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

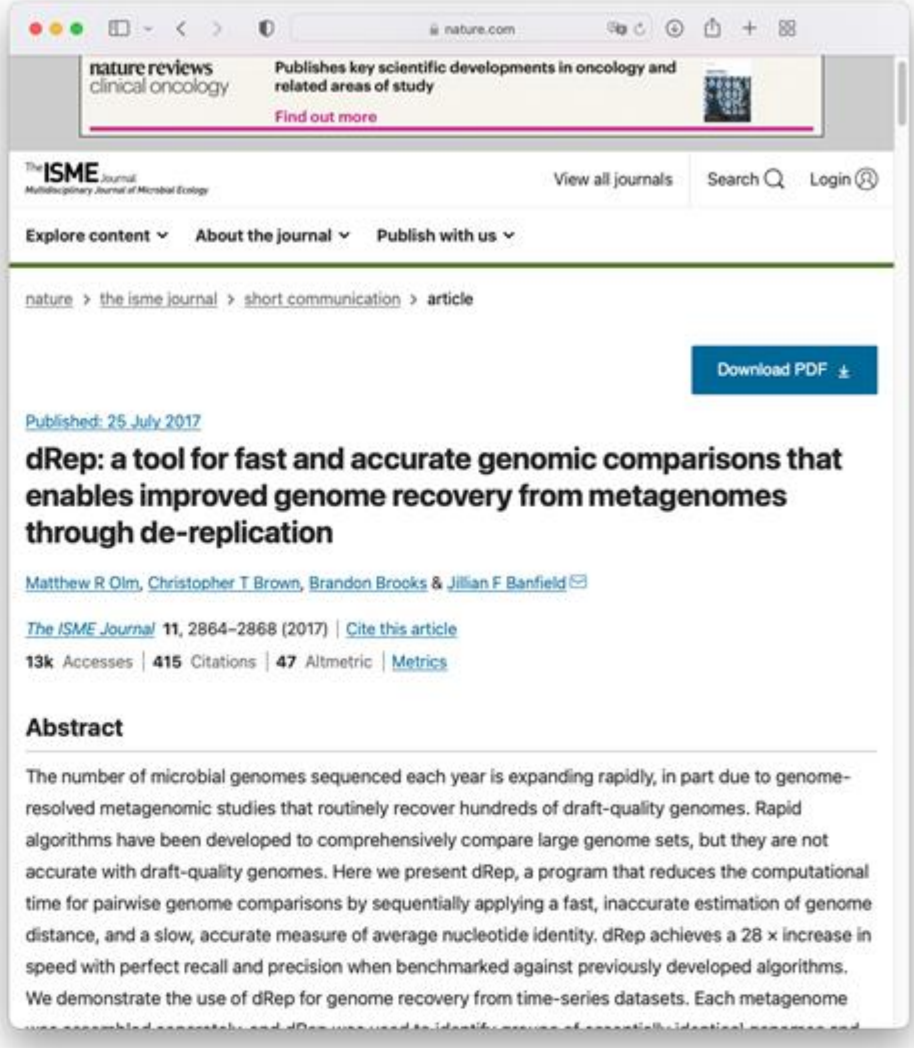
• Comparing and dereplicating a genome dataset

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/



nature.com
nature reviews clinical oncology
Publishes key scientific developments in oncology and related areas of study
Find out more
The ISME Journal
Multidisciplinary Journal of Microbial Ecology
View all journals
Search
Login
Explore content
About the journal
Publish with us
nature › the isme journal › short communication › article
Download PDF

Published: 25 July 2017

dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication

Matthew R Olm, Christopher T Brown, Brandon Brooks & Jillian F Banfield

The ISME Journal **11**, 2864–2868 (2017) | Cite this article

13k Accesses | **415** Citations | **47** Altmetric | Metrics

Abstract

The number of microbial genomes sequenced each year is expanding rapidly, in part due to genome-resolved metagenomic studies that routinely recover hundreds of draft-quality genomes. Rapid algorithms have been developed to comprehensively compare large genome sets, but they are not accurate with draft-quality genomes. Here we present dRep, a program that reduces the computational time for pairwise genome comparisons by sequentially applying a fast, inaccurate estimation of genome distance, and a slow, accurate measure of average nucleotide identity. dRep achieves a 28 × increase in speed with perfect recall and precision when benchmarked against previously developed algorithms. We demonstrate the use of dRep for genome recovery from time-series datasets. Each metagenome was assembled separately and dRep was used to identify groups of essentially identical genomes and

• Comparing and dereplicating a genome dataset

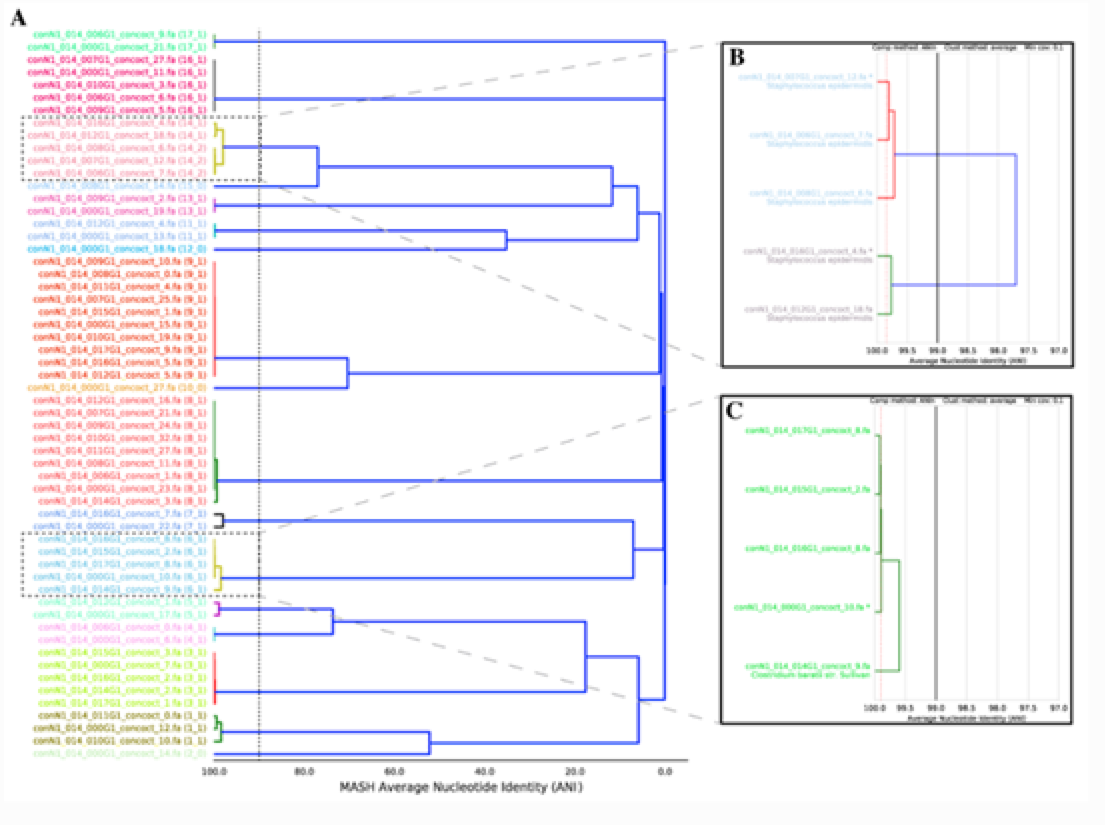
The dRep tool

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)
```

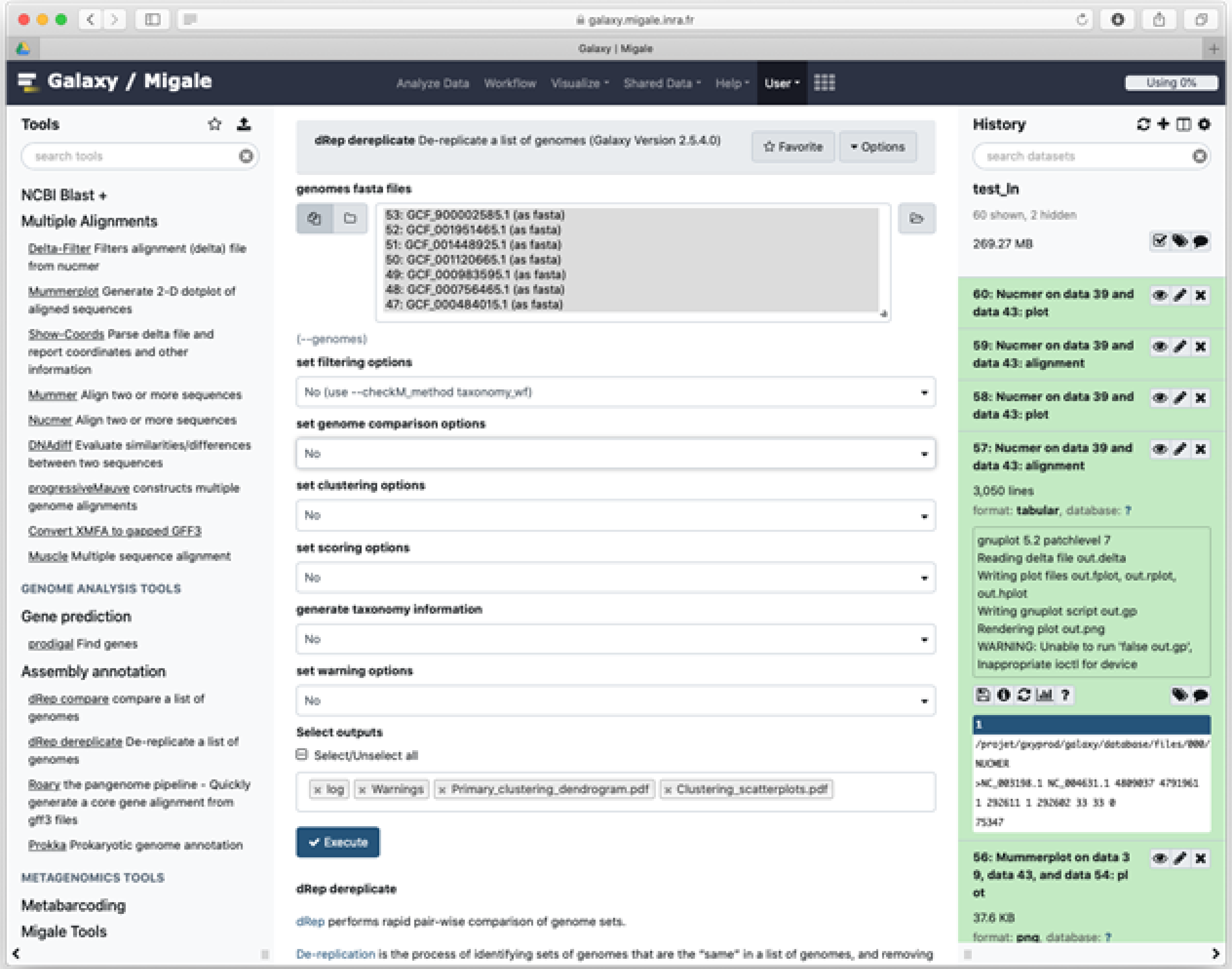
dRep practice

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

> input : 16 Salmonella genome fasta files

- ➢ Default parameters
- ➢ **All outputs**

explore and interpret results



galaxy.migale.inra.fr

Galaxy | Migale

Galaxy / Migale

Tools

NCBI Blast +

Multiple Alignments

Delta-Filter Filters alignment (delta) file from nucmer

Mummerplot Generate 2-D dotplot of aligned sequences

Show-Coords Parse delta file and report coordinates and other information

Mummer Align two or more sequences

Nucmer Align two or more sequences

DNAdiff Evaluate similarities/differences between two sequences

progressiveMauve constructs multiple genome alignments

Convert XMFA to gapped GFF3

Muscle Multiple sequence alignment

GENOME ANALYSIS TOOLS

Gene prediction

prodigal Find genes

Assembly annotation

dRep compare compare a list of genomes

dRep dereplicate De-replicate a list of genomes

Roary the pangenome pipeline - Quickly generate a core gene alignment from gff3 files

Prokka Prokaryotic genome annotation

METAGENOMICS TOOLS

Metabarcoding

Migale Tools

dRep dereplicate De-replicate a list of genomes (Galaxy Version 2.5.4.0)

genomes fasta files

53: GCF_900002585.1 (as fasta)
52: GCF_001951465.1 (as fasta)
51: GCF_001448925.1 (as fasta)
50: GCF_001120665.1 (as fasta)
49: GCF_000983595.1 (as fasta)
48: GCF_000756465.1 (as fasta)
47: GCF_000484015.1 (as fasta)

(--genomes)

set filtering options

No (use --checkM_method taxonomy_wf)

set genome comparison options

No

set clustering options

No

set scoring options

No

generate taxonomy information

No

set warning options

No

Select outputs

⊟ Select/Unselect all

× log × Warnings × Primary_clustering_dendrogram.pdf × Clustering_scatterplots.pdf

dRep dereplicate

dRep performs rapid pair-wise comparison of genome sets.

De-replication is the process of identifying sets of genomes that are the "same" in a list of genomes, and removing

History

test_ln

60 shown, 2 hidden

269.27 MB

60: Nucmer on data 39 and data 43: plot

59: Nucmer on data 39 and data 43: alignment

58: Nucmer on data 39 and data 43: plot

57: Nucmer on data 39 and data 43: alignment

3,050 lines

format: **tabular**, database: ?

```
gnuplot 5.2 patchlevel 7
Reading delta file out.delta
Writing plot files out.fplot, out.rplot,
out.hplot
Writing gnuplot script out.gp
Rendering plot out.png
WARNING: Unable to run 'false out.gp',
Inappropriate ioctl for device
```

```
1
/projet/gxyprod/galaxy/database/files/000/
NUCMER
>NC_003198.1 NC_004631.1 4809037 4791961
1 292611 1 292602 33 33 0
75347
```

56: Mummerplot on data 39, data 43, and data 54: plot

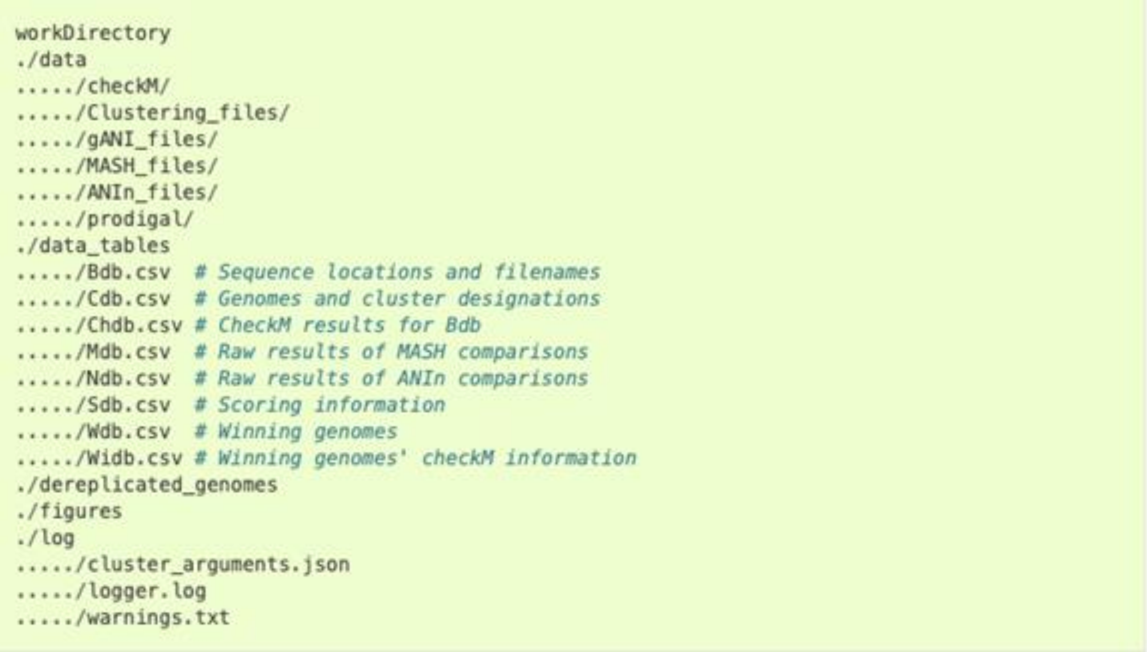
37.6 KB

format: **png**, database: ?

dRep tools and result files

dRep rely on several other programs:

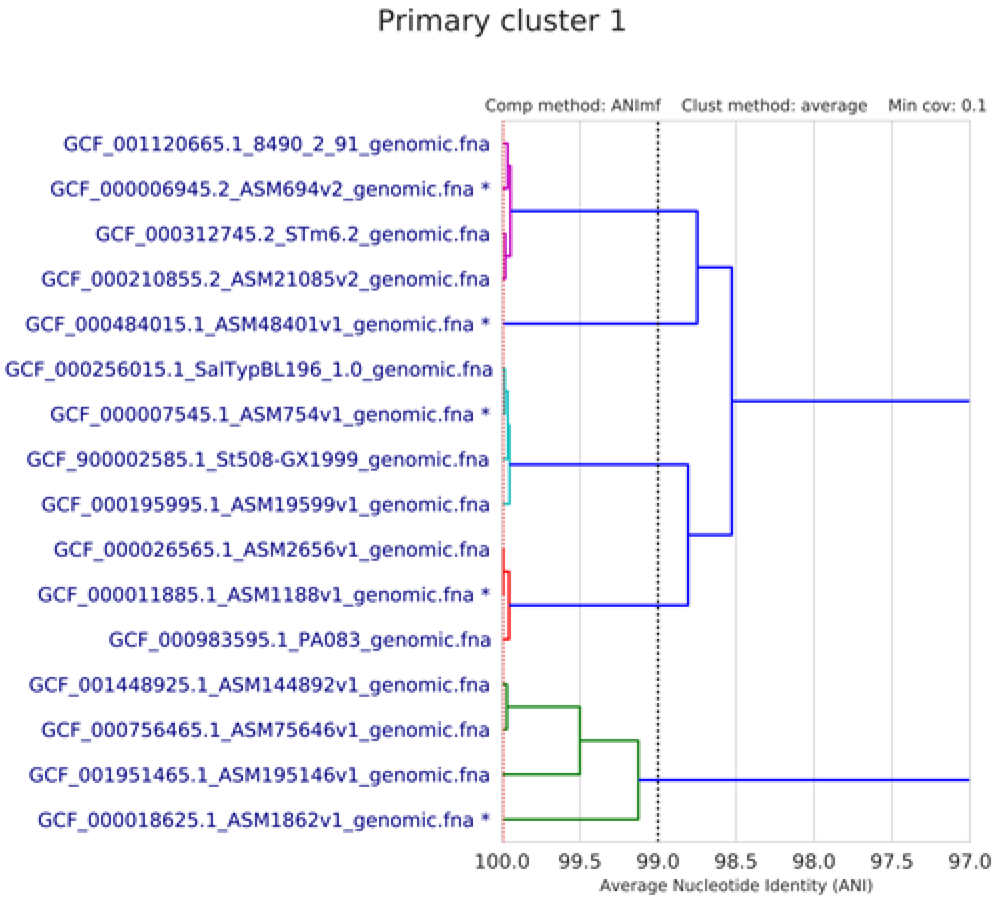
1. **Mash**: to build the primary clusters
2. **Mummer**: to perform the ANI computation on pairwise genome alignements (used by default but **fastANI** or **gANI** may also be used)
3. **checkM** (Parks et al. 2015) to determine contamination and completeness of genomes
4. **Prodigal** (Hyatte et al.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



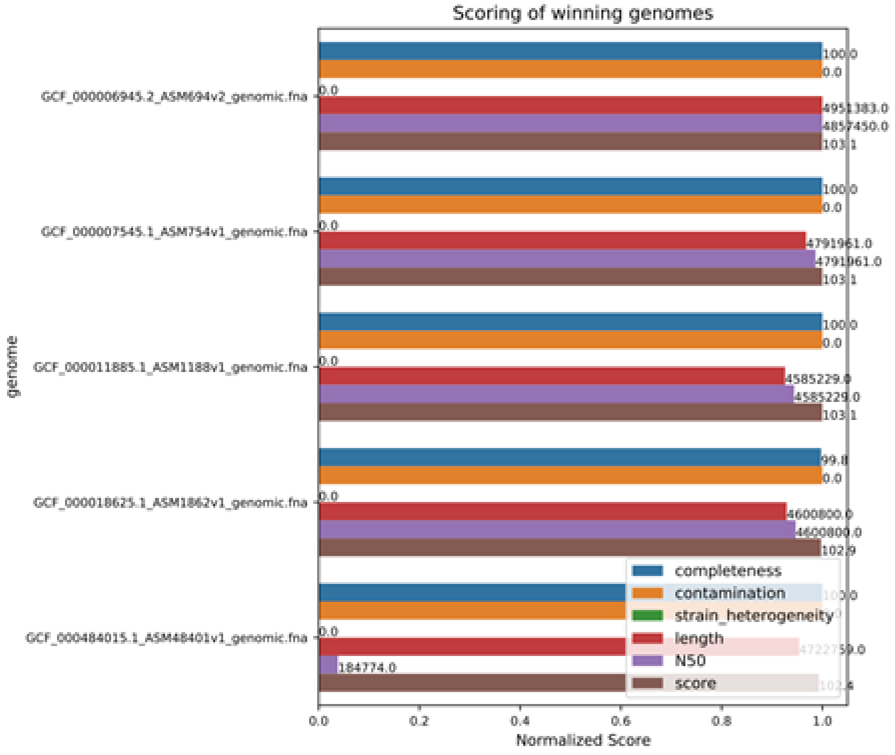
```
workDirectory
./data
...../checkM/
...../Clustering_files/
...../gANI_files/
...../MASH_files/
...../ANIn_files/
...../prodigal/
./data_tables
...../Bdb.csv  # Sequence locations and filenames
...../Cdb.csv  # Genomes and cluster designations
...../Chdb.csv # CheckM results for Bdb
...../Mdb.csv  # 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 results interpretation



Secondary_clustering_dendrograms.pdf



Winning_genomes.pdf

•Next steps

- Linking Genomes with meta-data :
  - Meta-data extracted from « Genome » csv file
  - **Omnicrobe** : a database of habitats, phenotypes and uses of microorganisms. (Litterature , genbank, DSMZ, CIRM…)

  https://omnicrobe.migale.inrae.fr

- Pangenome analysis [Cf next talk]
  - **Roary**
    - ⚠ File format. GFF from Bakta OR GFF convetered from Gbk with Genbank to GFF3 converter (Cf https://sanger-pathogens.github.io/Roary/)
  - **PPanGGOLiN** : Depicting microbial species diversity via a Partitioned PanGenome Graph Of Linked Neighbors