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

Dataset construction

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 Prokka [*See module 9*]
- It's always better if annotation is homogeneous

Quick reminder on format

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.

Example

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

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 annotaed with thier position , strand and qualifiers that contains the n 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 **“//”**.

- NCBI, ENA (European Nucleotide Archive) et DDBJ (Japan) entries are synchronized each day.
- Those three bank agree on the list of feature / qualifier that one can use to annotate sequence.

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
```

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.

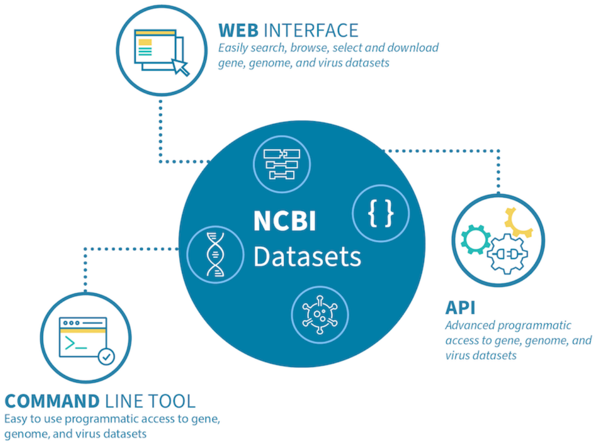
```
##gff-version 3
##sequence-region NZ_LHTK01000001 1 688985
# organism Salmonella enterica subsp. arizonae serovar 62:z36:- str. 5335/86
# date 17-JAN-2020
NZ_LHTK01000001    GenBank    contig    1    688985    .    +    1    ID=NZ_LHTK01000001;Dbxref=BioP
NZ_LHTK01000001    GenBank    pseudogene    1    1014    .    -    1    ID=LFZ49_RS22320.pseudogene;
NZ_LHTK01000001    GenBank    gene    1011    1634    .    -    1    ID=LFZ49_RS00010;Name=LFZ49_RS0
NZ_LHTK01000001    GenBank    mRNA    1011    1634    .    -    1    ID=LFZ49_RS00010.t01;Parent=LFZ
```

Practical : public genomes

How to filter and download publicly available genomes ?

- list all publicly available genomes with
  - metadata
  - quality metrics (size, completeness,…)
- filter according to above criteria
- download genomes in various formats

A solution : NCBI 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:

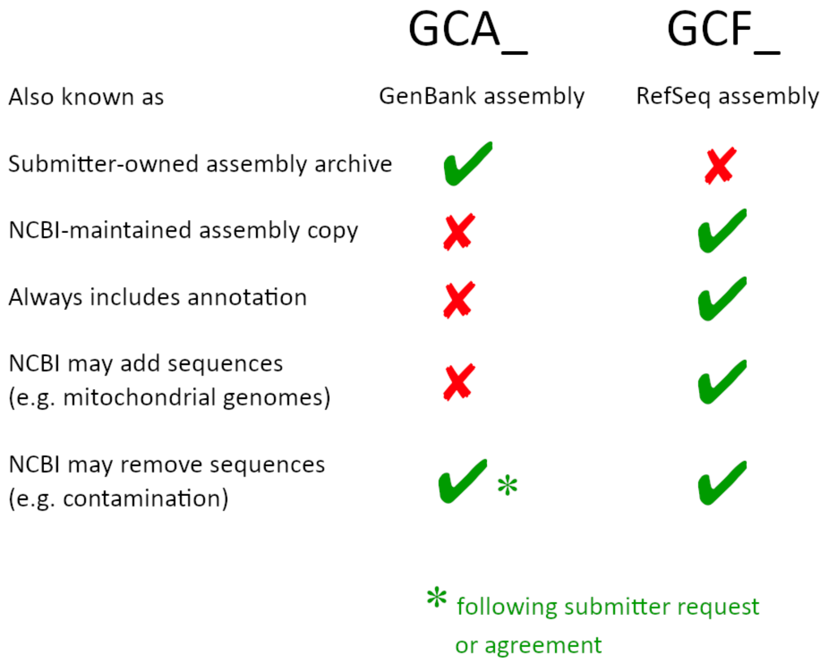
- NCBI Datasets website
- Command-line tools
- API (Application programming Interface)

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



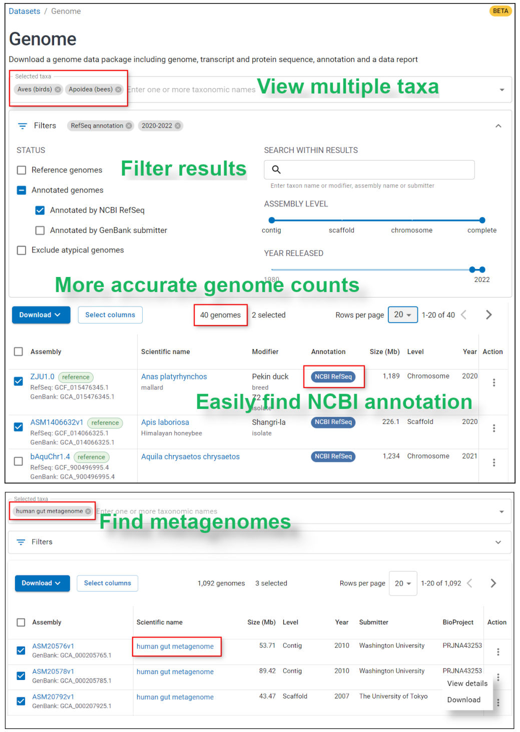
	GCA_	GCF_
Also known as	GenBank assembly	RefSeq assembly
Submitter-owned assembly archive	✓	✗
NCBI-maintained assembly copy	✗	✓
Always includes annotation	✗	✓
NCBI may add sequences (e.g. mitochondrial genomes)	✗	✓
NCBI may remove sequences (e.g. contamination)	✓ *	✓

* 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: 1980 – 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 Z2 isolate	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		⋮

View multiple taxa

Filter results

More accurate genome counts

Easily find NCBI annotation

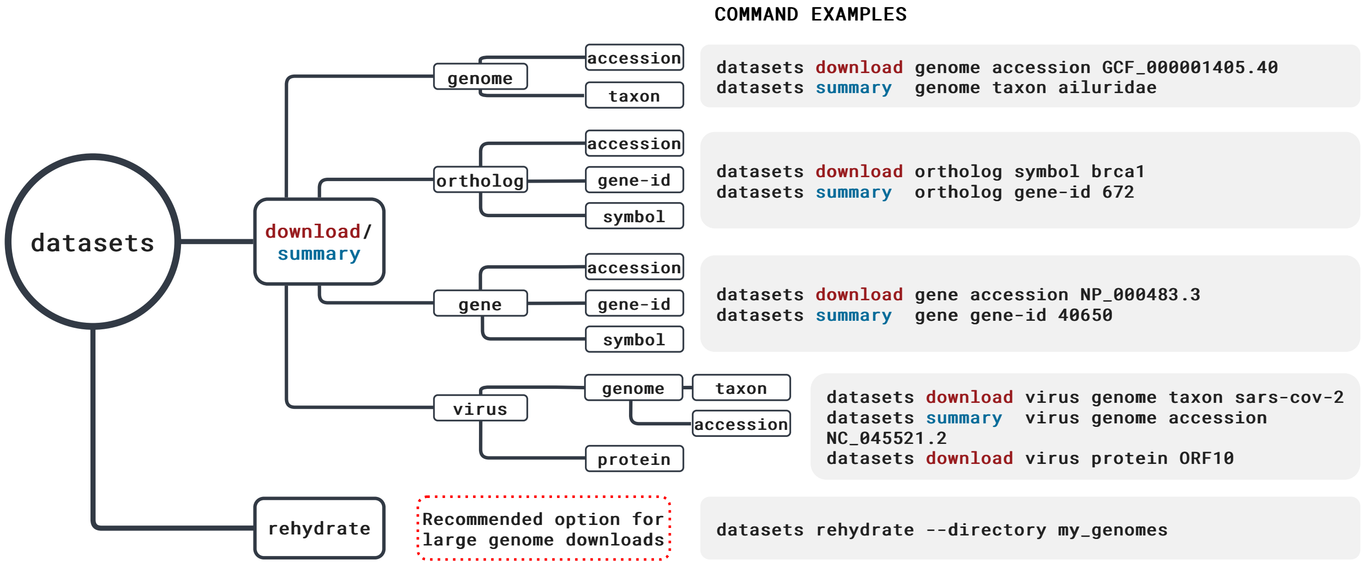
Find metagenomes

NCBI Datasets Genome Page

Genome Table | | 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
```

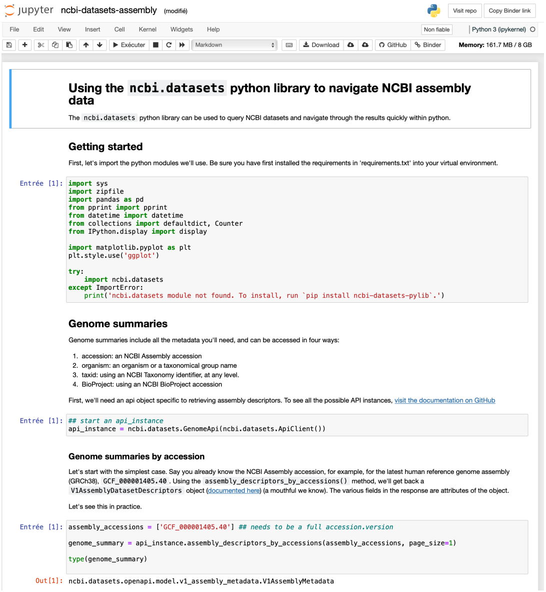
```
datasets rehydrate --directory my_genomes
```

NCBI Datasets Command Line

genome options :

- summary according to *accession* or *taxid*
- filter according to quality criteria & metadata
- donwload 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.

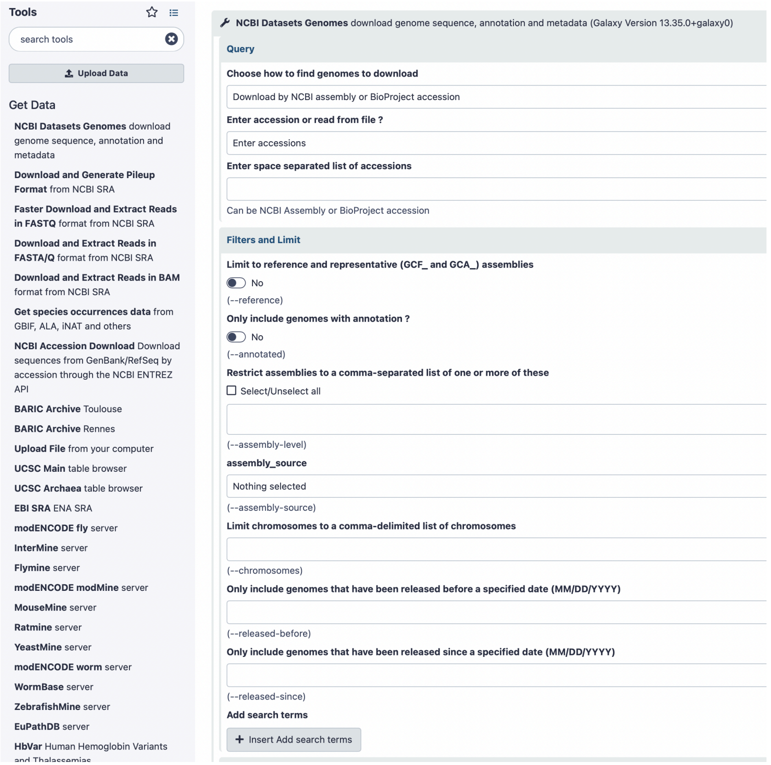
Let's see this in practice.

```
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
```

NCBI Datasets Python API

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)

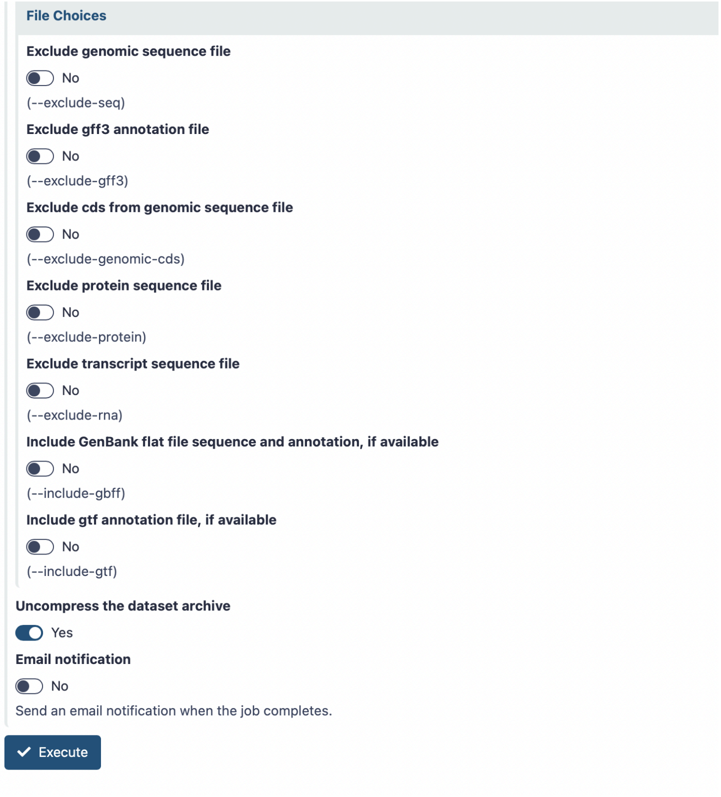
Only include genomes that have been released since a specified date (MM/DD/YYYY)

(--released-since)

Add search terms

+ Insert Add search terms

A wrapper of the command line tool



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

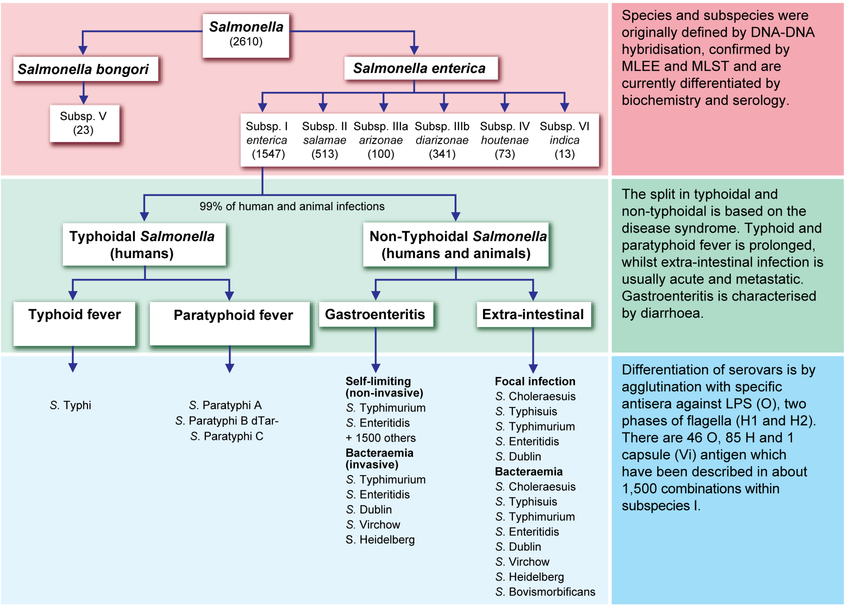
Parameters to define packages files

NCBI Datasets : Galaxy Integration

- A few caveats of the wrapper :
    - Some (not so easy) errors when select / filter fails
    - Impossible to just download a list of genomes as a file and "rehydrate" it after
- What recomend to :
    - use the NCBI dataset genome page to browse / filter a list of genomes of interest
    - download the list as a `tsv` file
    - feed NCBI datset with the list to donwload the genomes in diverse formats

The training datasets

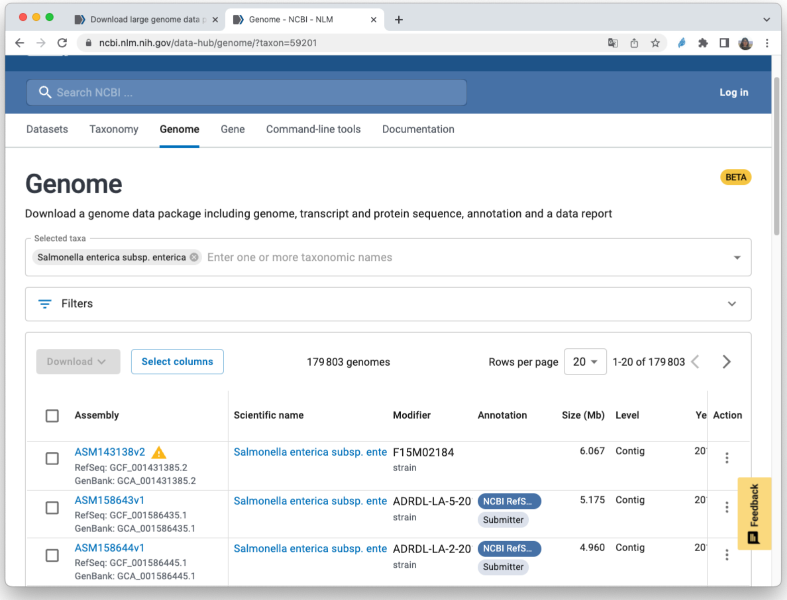
We will work on 3 datasets of public *Salmonella* genomes



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

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.

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

179 803 genomes — Rows per page 20 — 1-20 of 179 803

	Assembly	Scientific name	Modifier	Annotation	Size (Mb)	Level	Ye	Action
□	ASM143138v2 ⚠ RefSeq: GCF_001431385.2 GenBank: GCA_001431385.2	Salmonella enterica subsp. ente	F15M02184 strain		6.067	Contig	20	⋮
□	ASM158643v1 RefSeq: GCF_001586435.1 GenBank: GCA_001586435.1	Salmonella enterica subsp. ente	ADRDL-LA-5-20 strain	NCBI RefS... Submitter	5.175	Contig	20	⋮
□	ASM158644v1 RefSeq: GCF_001586445.1 GenBank: GCA_001586445.1	Salmonella enterica subsp. ente	ADRDL-LA-2-20 strain	NCBI RefS... Submitter	4.960	Contig	20	⋮

179.803 salmonella enterica enterica public assemblies at NCBI!!

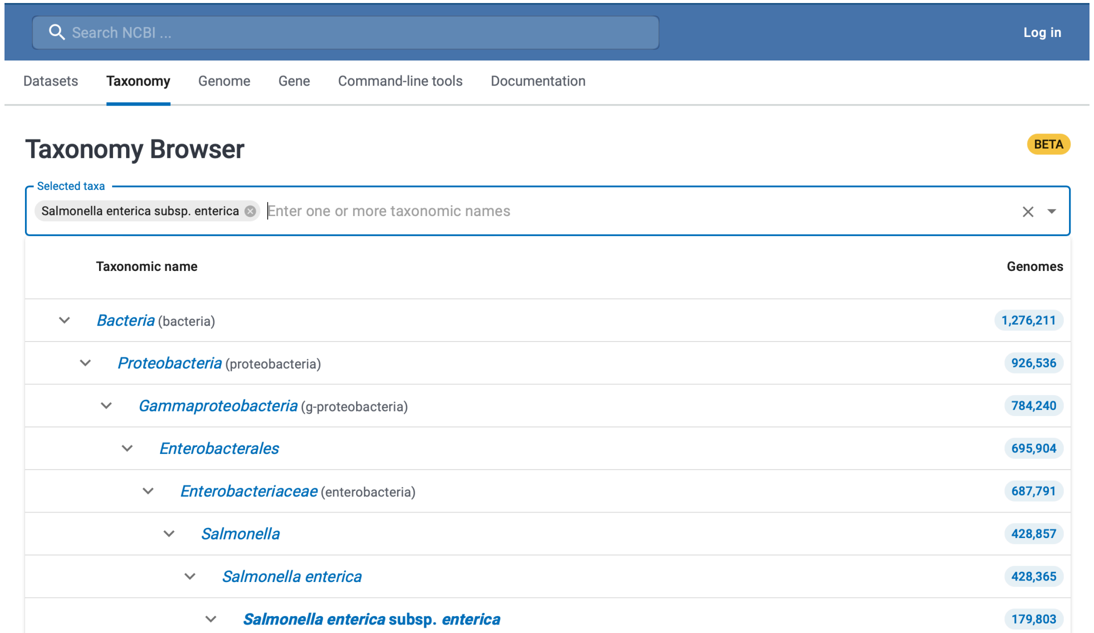
The training datasets

We will work on 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 from their *taxon id*
- dataset 3: download 16 Salmonella enterica public assemblies (2 sub-species, 4 serotypes) from their *accession numbers*

Use case 1 : Taxonomy browser

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



Taxonomy Browser

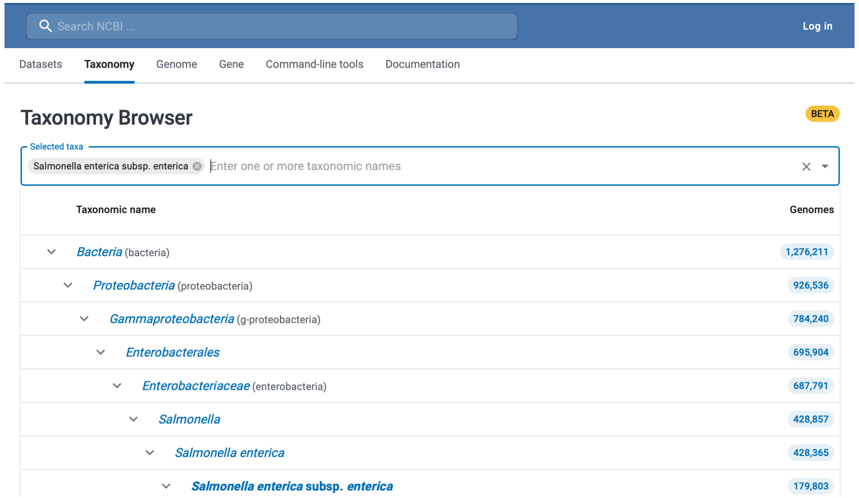
Selected taxa: Salmonella enterica subsp. enterica

Enter one or more taxonomic names

Taxonomic name	Genomes
Bacteria (bacteria)	1,276,211
Proteobacteria (proteobacteria)	926,536
Gammaproteobacteria (g-proteobacteria)	784,240
Enterobacterales	695,904
Enterobacteriaceae (enterobacteria)	687,791
Salmonella	428,857
Salmonella enterica	428,365
Salmonella enterica* subsp. *enterica	179,803

Use case 1 : genome table

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



Taxonomy Browser

Selected taxa: Salmonella enterica subsp. enterica

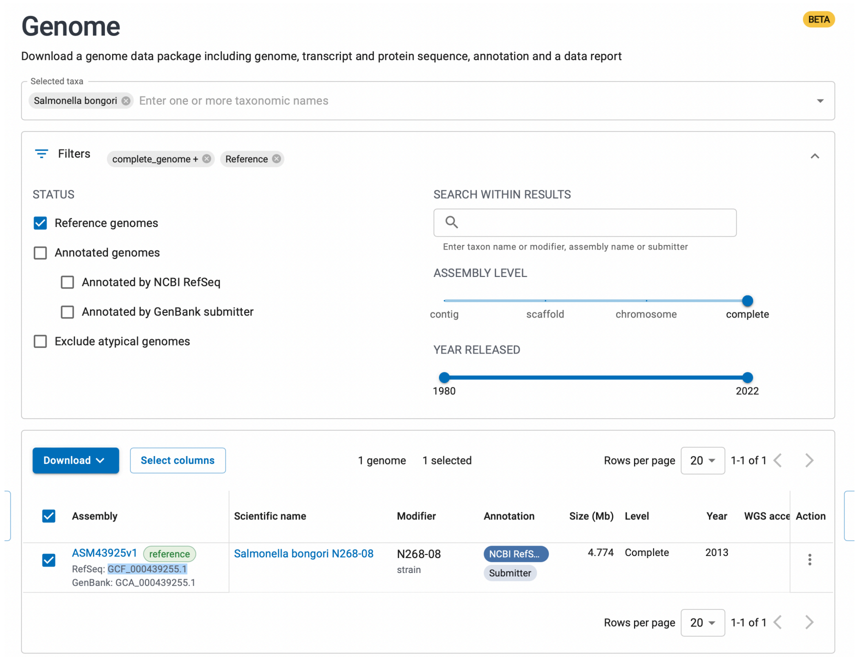
Taxonomic name	Genomes
Bacteria (bacteria)	1,276,211
Proteobacteria (proteobacteria)	926,536
Gammaproteobacteria (g-proteobacteria)	784,240
Enterobacterales	695,904
Enterobacteriaceae (enterobacteria)	687,791
Salmonella	428,857
Salmonella enterica	428,365
Salmonella enterica* subsp. *enterica	179,803

Notice :

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

Use case 2 : genome table

List all the *Salmonella bongori assemblies* to choose and download the best outgroup of a salmonella enterica dataset from their *taxon id*



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: complete_genome + ⊗ Reference ⊗

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: ● 1980 — 2022 ●

1 genome 1 selected Rows per page 20 1-1 of 1

☑	Assembly	Scientific name	Modifier	Annotation	Size (Mb)	Level	Year	WGS acce	Action
☑	ASM43925v1 reference RefSeq: GCF_000439255.1 GenBank: GCA_000439255.1	Salmonella bongori N268-08	N268-08 strain	NCBI RefS... Submitter	4.774	Complete	2013		⋮

Rows per page 20 1-1 of 1

Notice :

- Filter parameters :
  - "reference genome"
  - annotated by RefSeq
- Refseq accession for Galaxy

Use case 2 : Galaxy dataset

- **List all** the *Salmonella bongori* assemblies using their species name or taxon id
- **Choose and Download** the best genome for being an outgroup of a *Salmonella enterica* dataset

Training dataset 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

Training dataset 3: 16 *S. enterica* public genomes (part 2)

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

Use case 3 : from a tabular file

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

- 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
  - Retrieve all file format of interest **including** genbank annotated files