Microbial genome annotation and comparison

Cycle de formation à la bioinformatique par la pratique

Hélène Chiapello - Valentin Loux
(helene.chiapello|valentin.loux)@inrae.fr

19 et 20 mai 2025

Practical informations

- 9h30 - 17h00
- 2 breaks morning and afternoon
- Lunch at INRAE restaurant (not mandatory)
- Questions allowed
- Everyone has something to learn from each other

A quick round table presentation

- 🌍 **Who are you** ?
  - Institution, laboratory, position …
- Are you (somewhat) familiar with **Galaxy** ?
- 🦠 What are your needs in **microbial genomes annotation** ?
- 🦠 🦠 What are your needs in **microbial genomes comparison** ?
- Have you already dealt with **microbial genomics data** ?
  - Aim of the study ?
  - Species studied
  - Number of genomes
  - Difficulties ?
- How do you feel today ? 👍 or 👎 ?

Migale team

- Migale website
- INRAE infrastructure dedicated to provide
  - Calculation & storage infrastructure
  - Trainings
  - Data analysis service (collaboration or accompaniement)
  - Bioinformatics tool development
- Member of the Institut Français de Bioinformatique

Objectives

After this training, you will:

- Be able to evaluate the quality of a private or public genome assembly
- Be able to automatically annotate a bacterial genome with **Bakta** and visualize it with **Jbrowse**
- Be able to construct a genomic dataset from public ressources and evaluate its quality and diversity
- Know the outlines, advantages and limits of main microbial genome comparison approaches
- Be able to use several tools like **dRep**, **PPanGGOLiN** and **FastTree** under Galaxy or using a graphical interface on the training dataset
- Have some keys to interpret results

Program

Day 1 : Genome annotation

- **Morning**:
  - Introduction:
  - Sequencers types, errors
  - Genome assembly quality
  - Practical : genome quality evaluation
- **Afternoon**:
  - key points about genome annotation
  - Practical : Annotate your own Genome
  - Practical : Visualize your annotated genomes
  - Annotation : specialized tools
  - Annotation : questions and wrap-up

Program

Day 2 : Comparative genomics

- **Morning**:
  - Introduction to comparative genomics
  - Dataset construction
  - Dataset quality evaluation
  - Dataset diversity analysis and dereplication
- **Afternoon**:
  - pangenome construction and interpretation
  - First steps in phylogenomics
  - Data visualization and interpretation
  - Comparison : questions and wrap-up

Hands on : dataset presentation

- Training dataset: *Streptococcus salivarius* genomes
- A species of *Bacilotta* found in human microbiomes (oral, pharyngeal and gut) and contributes to the maintenance of oral, pharyngeal and gut health
  - Some strains described as opportunistic pathogens (meningitis, endocartis, bacteremia,…)
  - Genomes exhibit high diversity, caused mainly by Mobile Genetic Elements
- Annotation of one genome:
  - Genome assembly ASM1102908v1 (GCF_011029085.1), our "**private genome**"
- Comparison with a set of *Streptococcus salivarius* genomes:
  - Compare ASM1102908v1 with a dataset of 49 public genomes

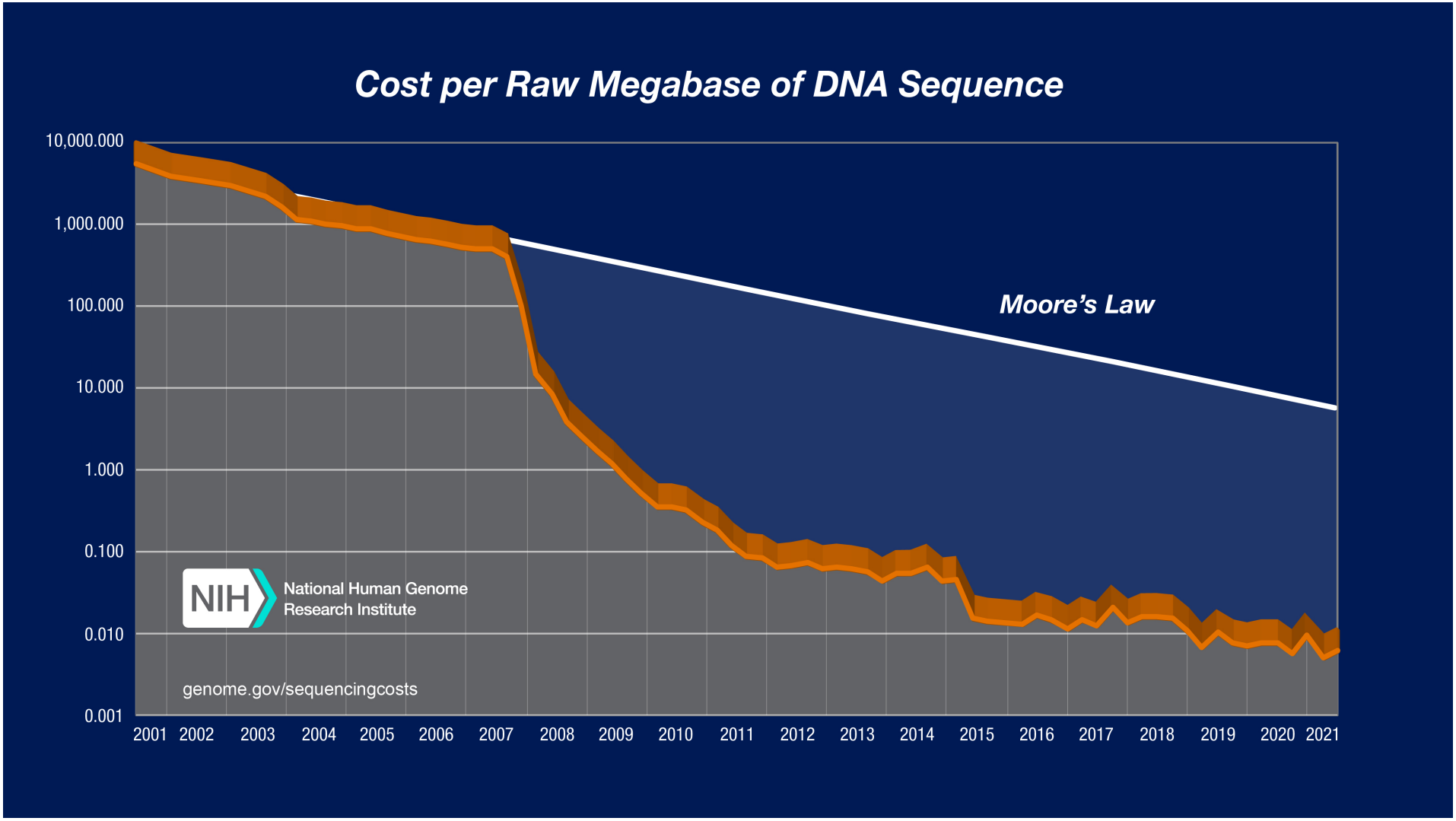
Hands on: questions

- Do you have an idea of the number of public bacterial genomes ?
- Do you knwow what are the top more sequenced bacterial species ?
- How many genomes for the more sequenced bacterial species ?
- How many public genomes for Streptococcus salivarius ?
- What are the main issues regarding
  - Genome annotation ?
  - Genome comparison ?

Sequencing technologies

Next generation Sequencing in a few slides

Sequencing Cost per Megabase (Source)



Genome Sequencing, why ?

Interest in a genome that has not yet been sequenced

- Assembly and annotation
  - de novo sequencing
  - chromosomal rearrangements
  - metagenomics

Genome Sequencing, why ?

A reference genome is available

- Alignment (mapping) of reads on the genome
  - Detection of genomic variants (SNPs)
  - RNA-seq (gene expression)
  - ChIP-seq (regulation of gene expression)
  - Chromosomal rearrangements, variation in gene copy number
  - Detection of small non-coding RNAs
  - metagenomics

Sequencing challenges

- Smallest known (non viral) genome:
  - 🦠 *Nasuia deltocephalinicola* = 112 kb
  - 🦠 *Candidatus Hodgkinia cicadicola* = 144 kb
  - 🦠 *Carsonella ruddii* = 160 kb
- Largest known genome:
  - 🪴 *Paris japonica* = 150 Gb
  - 🪴 *Tmesipteris oblanceolata* ~ 147 Gb
  - 🐟 *Protopterus aethiopicus* = 130 Gb

Sequencing challenges

- Maximum Reads Size :
  - 1st generation (Sanger): up to 900 bp
  - 2nd generation: up to 500 bp
  - 3rd generation: up to 100 - 1000 Kbp

Need to cut the genome into millions of fragments (**shotgun sequencing**) from the 2 DNA strands.

The operation to reconstruct the genetic elements from the raw reads is called **assembly**.

Sequencing technologies

- First generation :
  - Sanger sequencing
    - First step : fragment cloning
    - Reads up to 900 bp
    - Expensive
    - low throughput

Next generation Sequencing technologies

Second generation (since 2007)

- **454** - Sequencing by Synthesis - PCR Amplification
- **SOLiD**~~ : Sequencing by Ligation - PCR Amplification
- **Ion Torrent** : Sequencing by Synthesis - PCR Amplification
- **Illumina** : Sequencing by Synthesis - PCR Amplification

454 discontinuited in 2013, SOLiD no longer actively maintained.

Illumina : principles

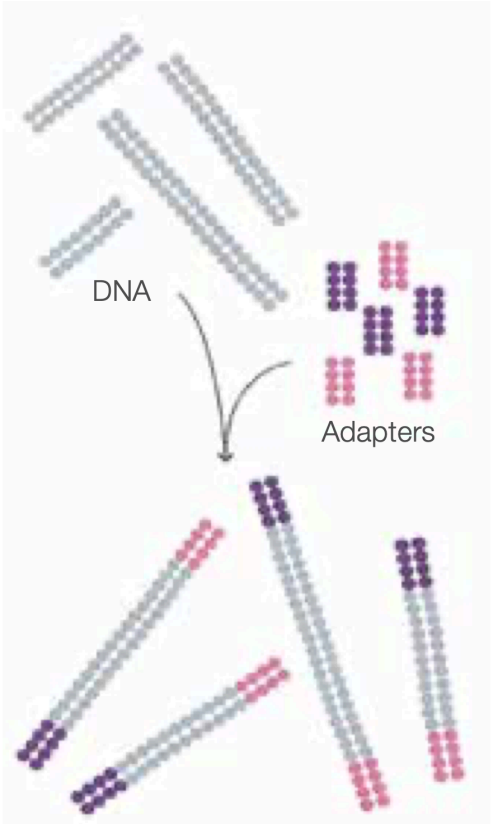
- Based on "reversible terminated chemistry" : reversible terminators that enable the identification of single nucleotides as they are washed over DNA strands.

Three steps :

- Amplification of DNA fragments
- Sequencing
- Analysis

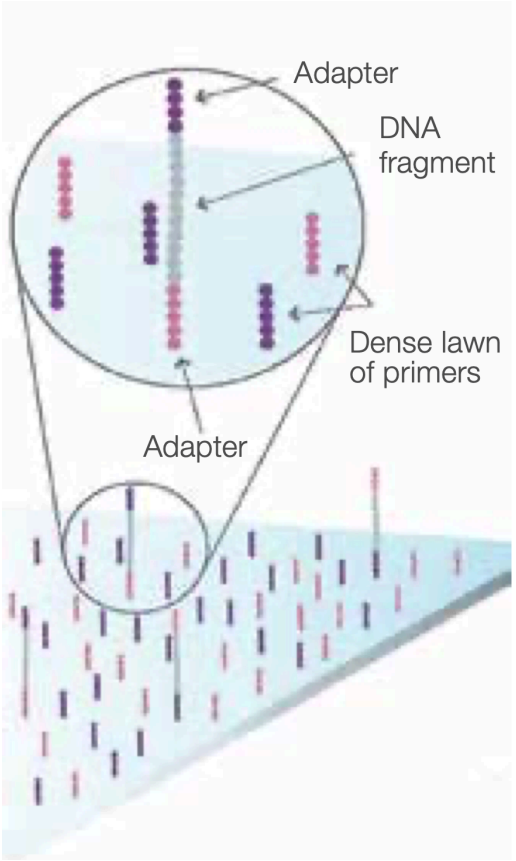
Reference : Technology Spotlight: Illumina Sequencing

Prepare genomic DNA samples



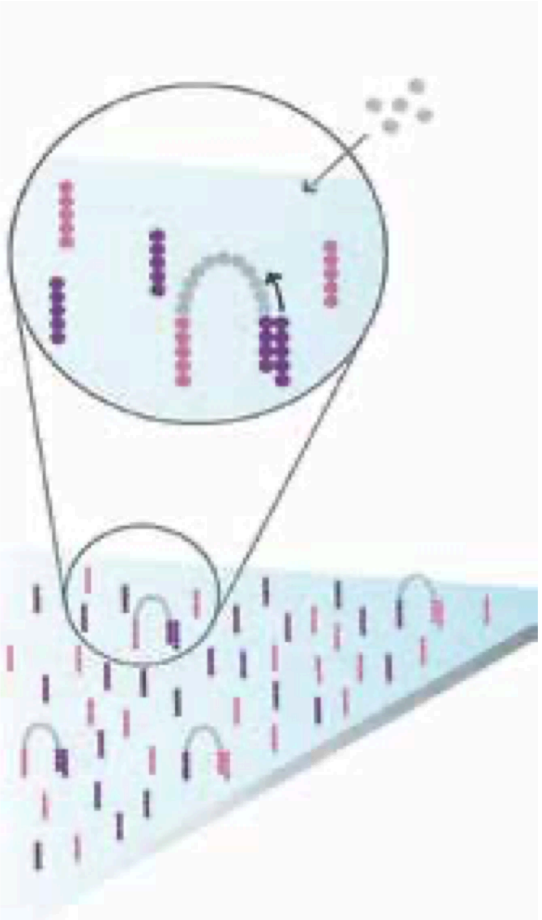
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments

Attach DNA to Flow Cell Surface



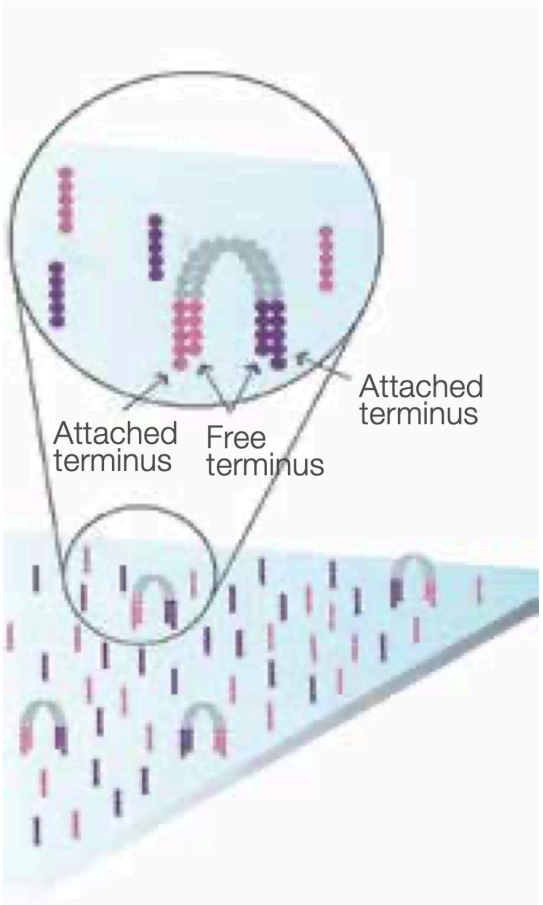
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Bridge Amplification



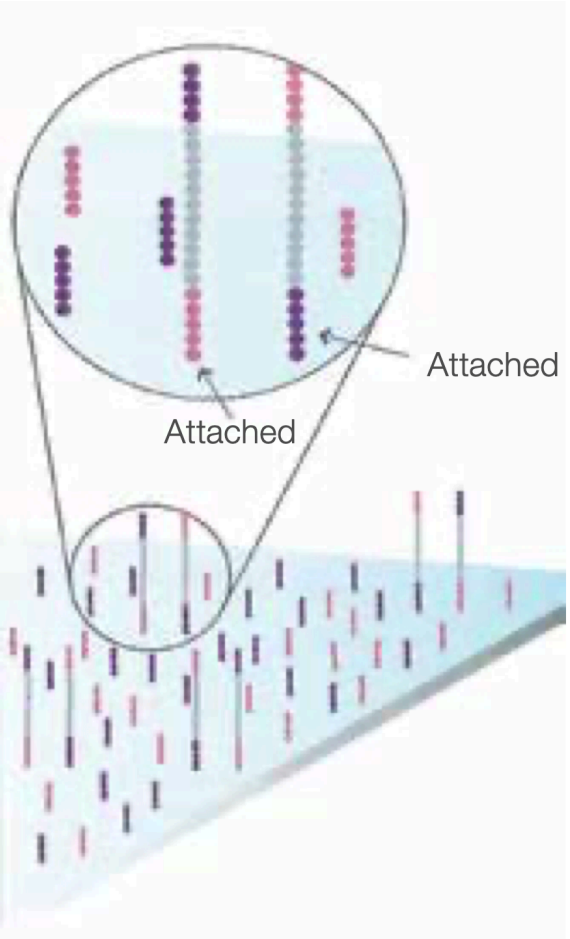
Add **unlabelled** nucleotides and enzyme to initiate solid-phase bridge amplification.

Fragments Become Double Stranded



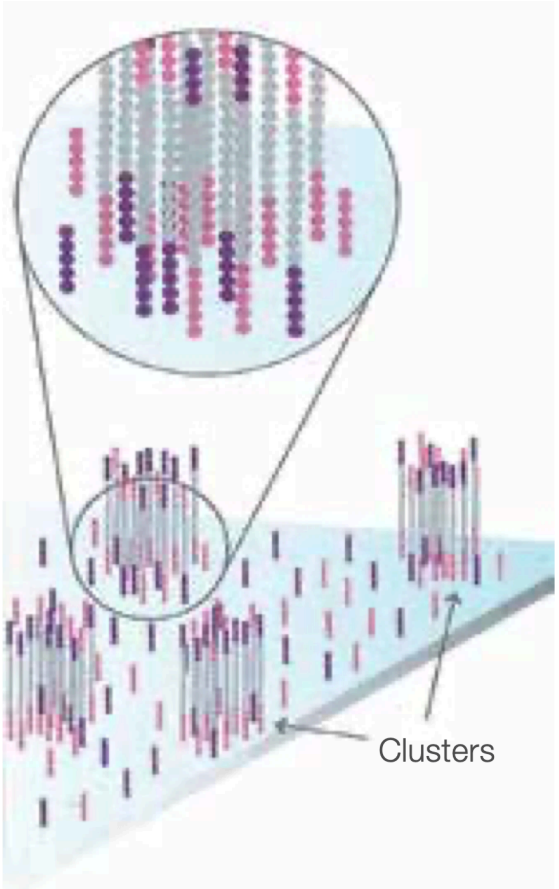
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

Denature the Double-Stranded Molecule



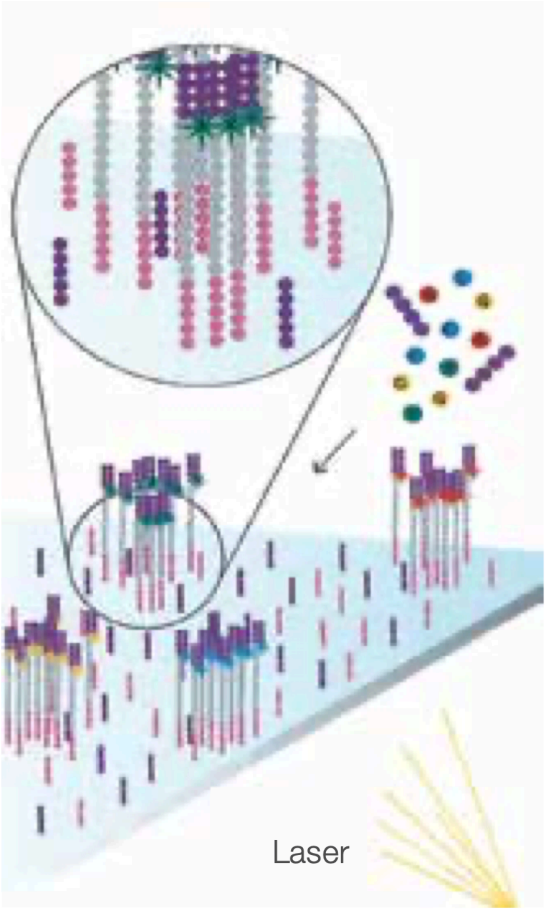
Denaturation leaves single-stranded templates anchored to the substrate.

Complete Amplification



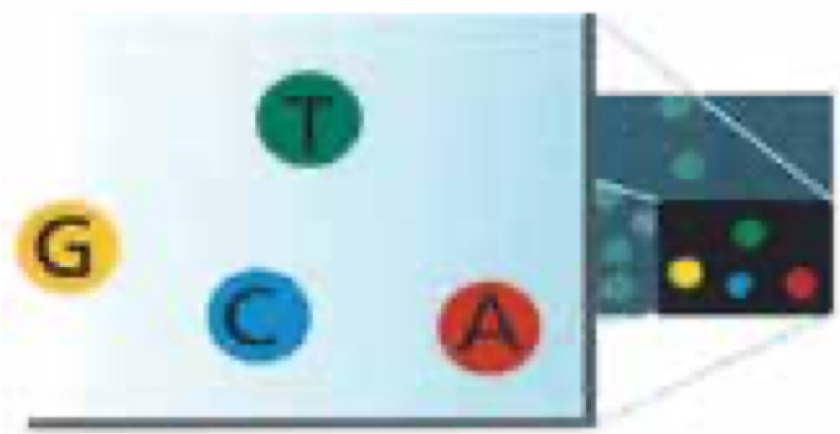
Several millions dense clusters of double-stranded DNA are grated in in channel of the flow cell.

Determine First Base



The first sequencing cycle begins by adding four labelled reversible terminators, primers, and DNA polymerase.

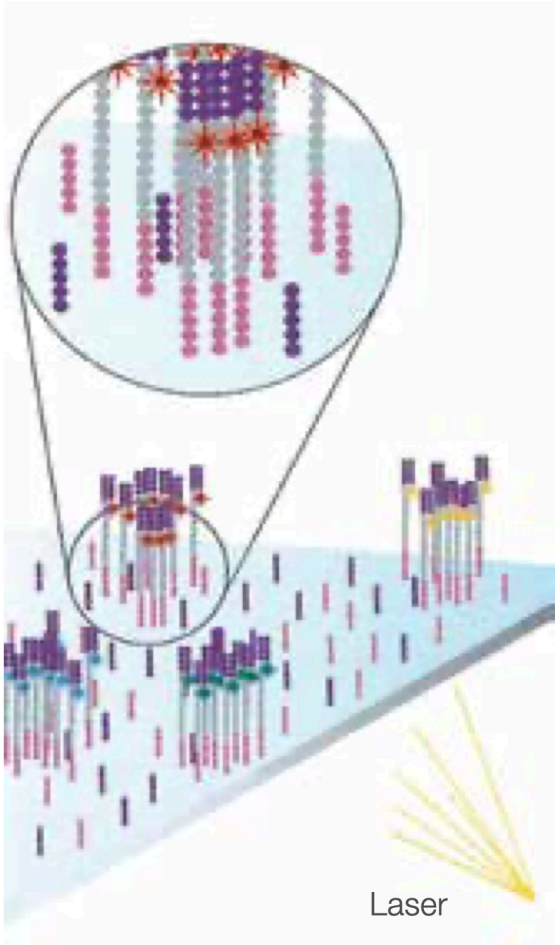
Image First Base



After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

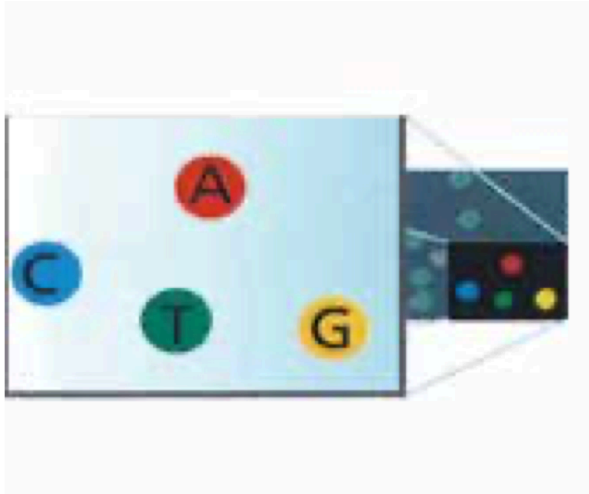
The blocked 3' terminus and florphore are removed,flow cell washed, leaving the terminator free for a second cycle.

Determine Second Base



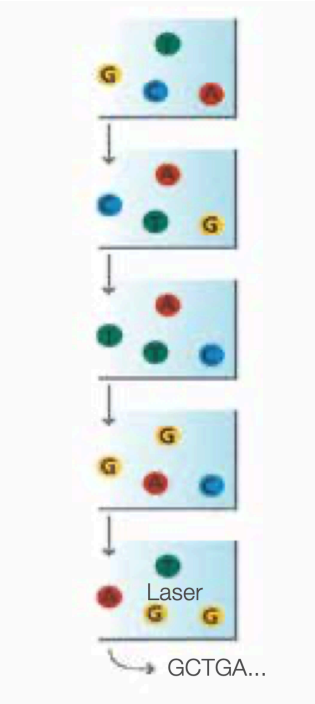
The next cycle repeats the incorporation of four labelled reversible terminators, primers, and DNA polymerase.

Image Second Chemistry Cycle



After laser excitation, the image is captured as before, and the identity of the second base is recorded.

Sequencing Over Multiple Chemistry Cycles



The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

Millions of clusters are processed in parallel, allowing high-throughput sequencing.

Illumina : summary

- High precision >99.5% (main type or errors : substitutions Reference)
- Short reads (maximum 2 x 300)
- Huge throughput (up to 6 Tbp per run on NovaSeq)
- Some under-representation of rich AT- and GC- regions.

Sumary video about Illumina sequencing

Sequencing - Vocabulary

- **Read**: piece of sequenced DNA
- **DNA fragment**: 1 or more reads depending on whether the sequencing is single- or paired-end
- **Insert**: Fragment size
- **Depth**: $\frac{N*L}{G}$\ $N =$ number of reads\ $L =$ reads size\ $G =$ genome size
- **Coverage**: % of genome covered

3d generation

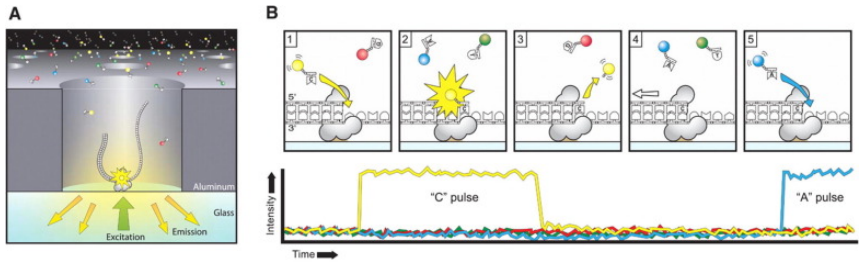
Target the **weaknesses** of the 2nd generation :

- PCR amplification
- Short reads

Two main competitors (in production) :

- Pacific Bioscience (PacBio)
- Oxford Nanopore Technologies (ONT)

PacBio



> A polymerase is immobilized at the bottom of a sequencing unit called zero-mode waveguide (ZMW) .Four fluorescent-labelled nucleotides, which generate distinct emission spectrums, are added to the SMRT cell. As a base is held by the polymerase, a light pulse is produced that identifies the base. The replication processes in all ZMWs of a SMRT cell are recorded by a "movie" of light pulses, and the pulses corresponding to each ZMW can be interpreted to be a sequence of bases.

Reference

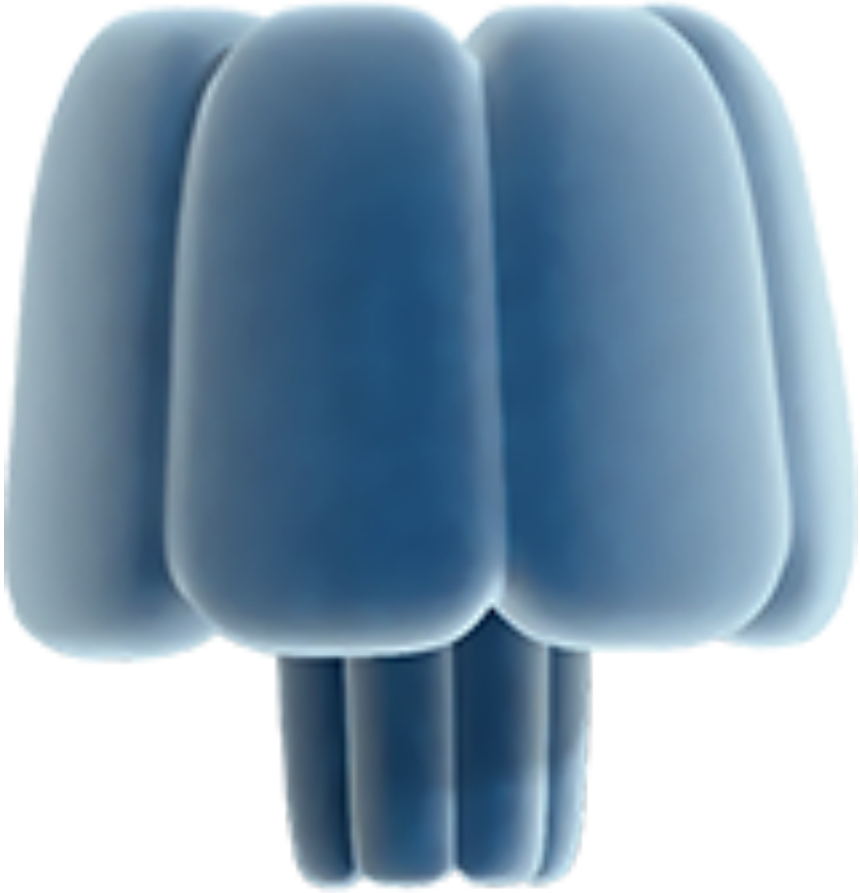
PacBio : summary

- Long reads (up to Kbs with SequelII)
- Depends on DNA quality
- High error rate. Tend to lower with depth
- Medium throughput

Applications :

- IsoSeq (RNA Isoform full length sequencing)
- Detection of DNA modification
- Assembly

Oxford Nanopore



MinION, GridION, PromethION

Sequencing on The ISS

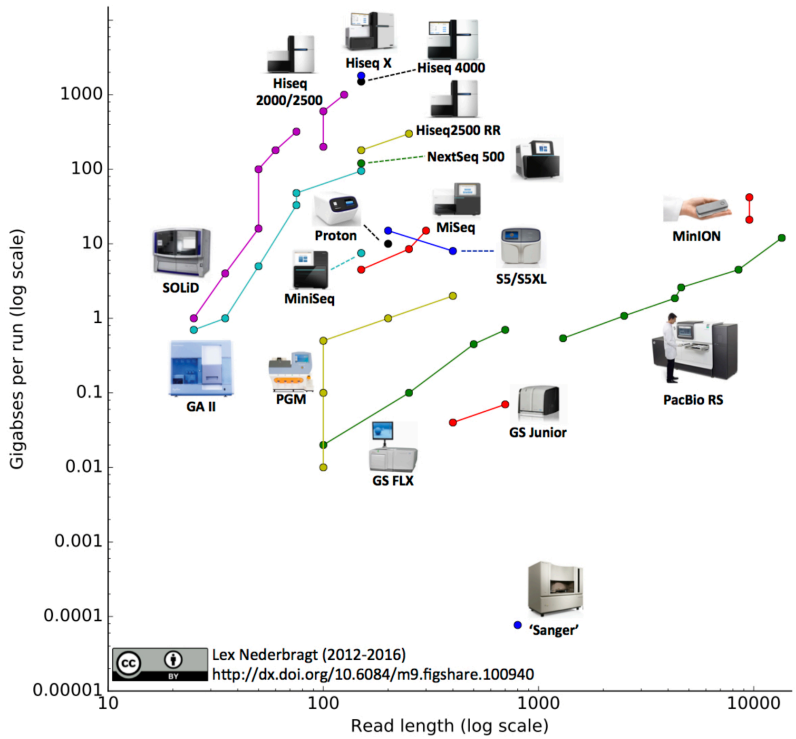
ONT Summary

- Ultra long reads (up to 1 Mb (!))
- Length of the reads depends on DNA quality
- Low to high throughput
- "On field" sequencing
- Direct RNA sequencing, peptide sequencing
- High error rate (5-10%), tends to lower with new chemistry, base calling algorithms and depth

ONT Applications

- Full length isofrom sequencing, direct RNA sequencing
- Detection of DNA modification
- Assembly

An other view on sequencing technologies (probably out of date)



source

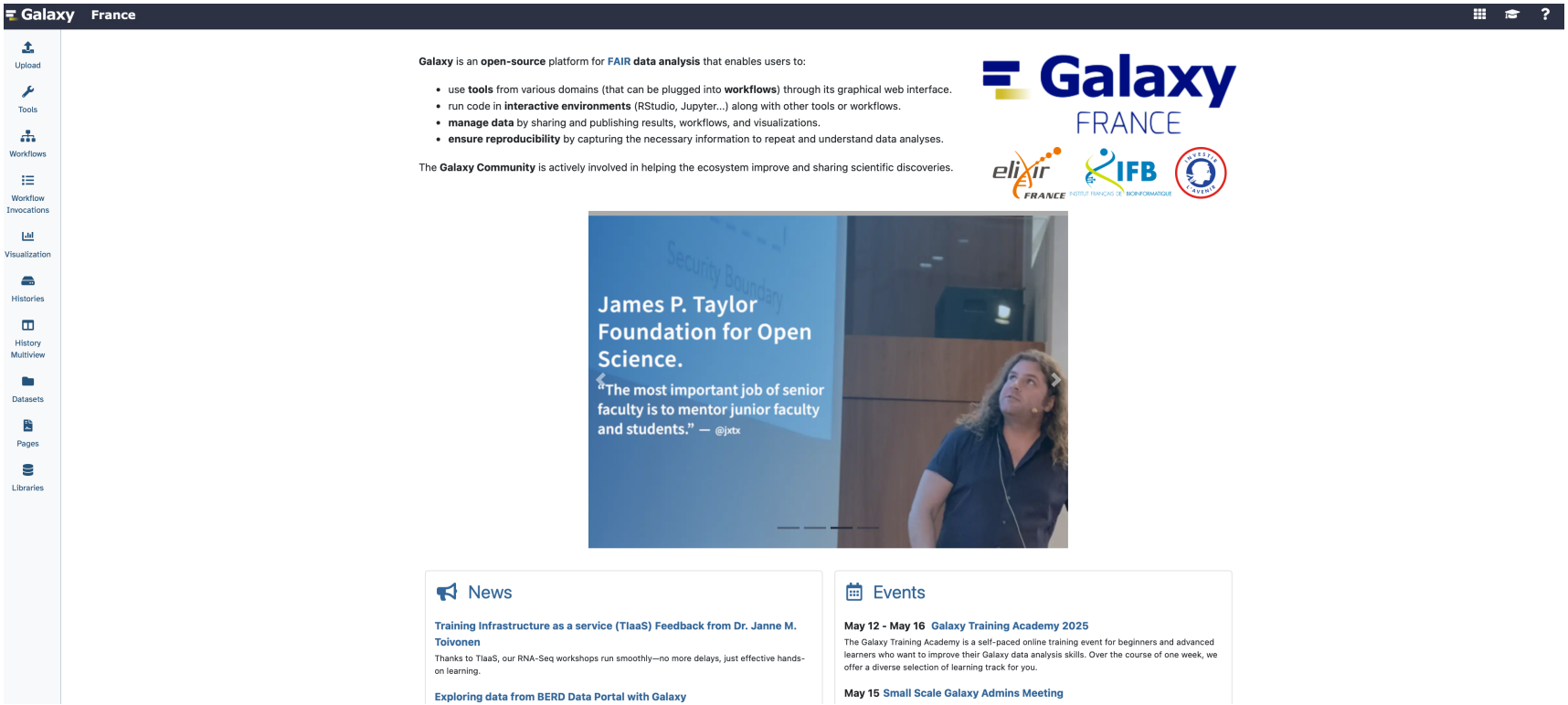
Global Summary

Platform	Read length in bp	Throughput per run	# of reads per run	Runtime	Error profile	Cost/Gbp (US$)
Roche 454 GS FLX titanium XL+	Up to 1000	700 Mb	~1 M	1d	1%, indels	$9500
Illumina MiSeq v3	300 (PE)	15 Gb	50 M	2d	0.1% substitutions	$110
Illumina NextSeq 500/550	150 (PE)	120 Gb	800 M	1d	<1%, substitutions	$33
Illumina HiSeq 3000/4000	150 (PE)	700 Gb	2.5 B (SE)	3d	0.1% substitutions	$22
Illumina HiSeq X	150 (PE)	850 Gb x 10	3 B (PE)	<3d	0.1% substitutions	$7
Illumina NovaSeq	150 (PE)	6 Tbp	20 B (PE)	4d	0.1% substitutions	$7
Ion Torrent PGM	200 (SE)	600 Mb – 1 Gb	5 M	4h	1%, indels	$600
Ion Torrent Proton	200 (SE)	10 Gb	70 M	3h	1%, indels	$80
Pacific Biosciences sequel	Up to 60 Kb	5-10 Gb	<100 K	4h	10-15%, indels	$800
ONT MK1 MinION	Up to 1Mb!	Up to 1 Gb	>100 K	2d	15%, indels	$750
Illumina synthetic long reads	~100 Kb	500 Gb	4B (PE)	6d	0.1%, substitutions	$33 + $500 per sample

- New competitors relaunching the game : PacBio, ONT but also AVITI, BGI, Roche SBX (soon)
- Up-to-date figures : NGS spec tables
- An interesting review
- Nature review : Milestones in Genomic Sequencing

Hands-On : Galaxy

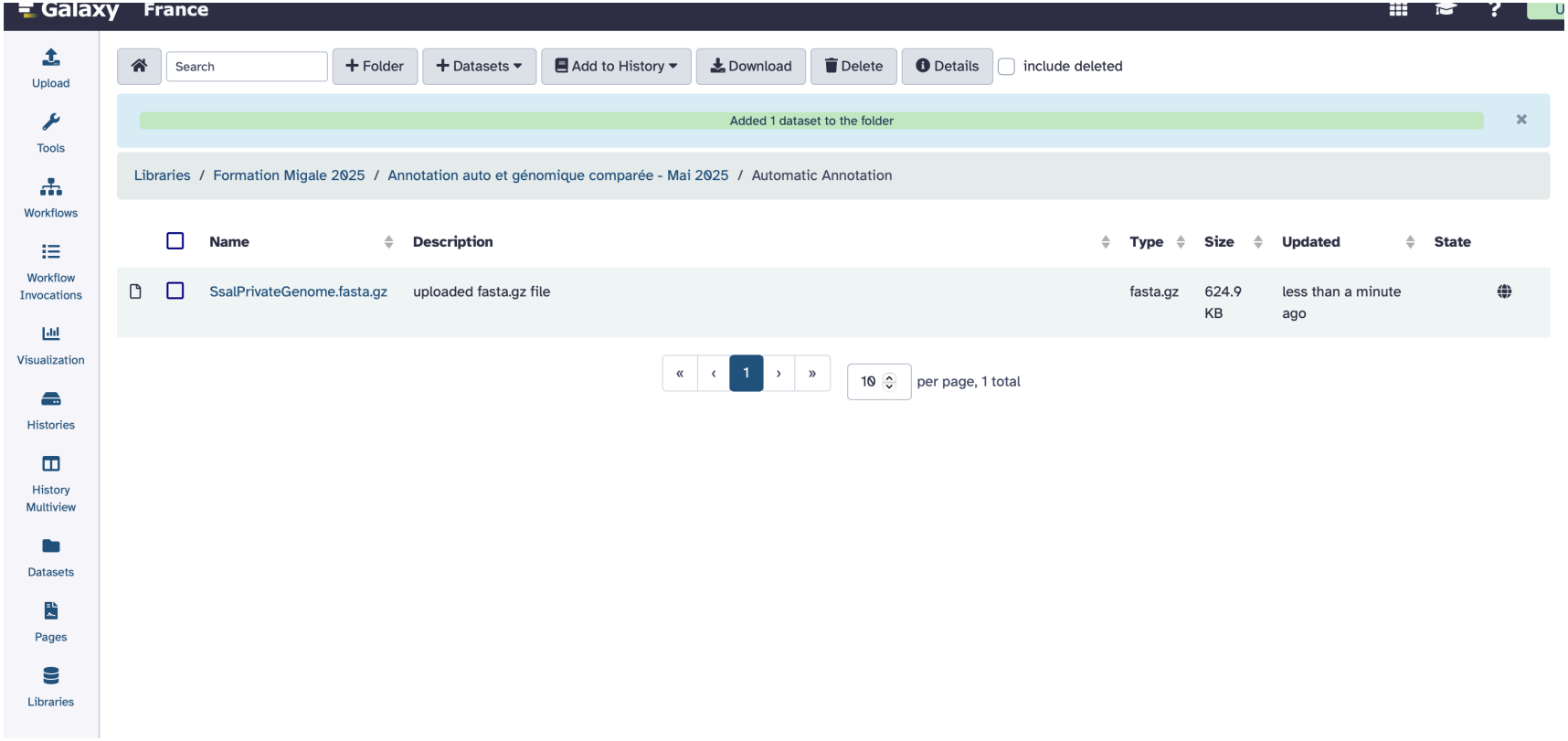
Connect to Usegalaxy.fr



usegalaxy.fr

Connect also to the Training Session

Import your "private" genome assembly



`Library / Formation Migale 2025 / Annotation auto et génomique comparée - Mai 2025 / Automatic Annotation / SsalPrivateGenome.fasta.gz`

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
>bar
ACTGCAGT
TTCGN
```

Hands-On : count the number of "sequences" in the fasta file

Hands-On : count the number of "sequences" in the fasta file (correction)

- Multiple ways of doing that. On example :
  - **Extract the number of line starting with** `>`
  - 🔧 `Select lines that match an expression`
  - **Select lines from** 📂 `SsalPrivateGenome.fasta.gz`
  - **that** `matching`
  - **the pattern** `^>`
  - **Run tool**
  - Edit result dataset name (optionnal, for clarity) : `` `Select on data 1 `` -> ✏️ `Header lines of SsalPrivateGenome.fasta.gz`
  - 🔧`Line/Word/Character count of a dataset`
    - **Select** `Line count` - 📂 `Header lines of SsalPrivateGenome.fasta.gz`
  - **Run tool**

ℹ️ Answer was also in the "deployed view" of the `SsalPrivateGenome.fasta.gz` dataset

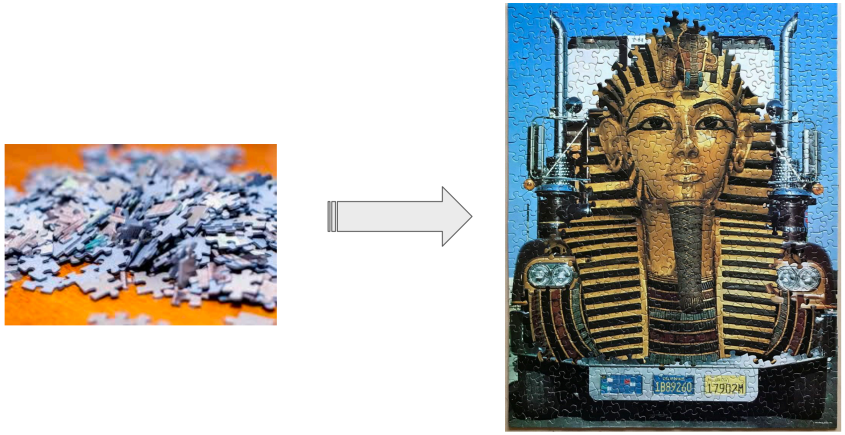
Assembly

Assembly : principles

From raw reads to complete replicons.

Similar to a puzzle :

- millions of pieces
- without the original image
- with pieces in both sense
- the pieces do not necessarily fit together (sequencing errors)
- parts of the puzzle are missing (cover + sequencing bias)



For training about genome assembly, see module 8bis or GTN

Vocabulary

- **Contigs**:Contigs are continuous stretches of sequence containing only A, C, G, or T bases without gaps.
- **Scaffold**: Scaffolds are created by chaining contigs together using additional information about the relative position and orientation of the contigs in the genome.
- **Assembly**: a set of contigs or scaffolds

Assembly Quality control

Why QC'ing your genomes ?

Try to answer to (not always) simple questions :

- What is the "quality" of an assembly [compared to what we expect] ? Is the assembly fragmented or **complete** and **continuous**?
  - Length
  - Number of contigs
  - Number of scaffolds
  - GC%
- What is the "quality" of an annotation [compared to what we expect]? Are ther more of fewere genes than expected. Are those genes correct compared to a refernce (SNPs…)
  - Number of (pseudo)genes
  - number of rRNA genes
  - number of tRNA genes

3 **C** : **C**ontiguity, **C**ompleteness & **C**orectness

Tools to QC your dataset : Quast

Quast (Quality Assessment Tool for Genome Assemblies, (Gurevich, Saveliev, Vyahhi, and Tesler, 2013)) is an easy to use software to evaluate genome assemblies.

It gives you, in one single report different metrics about one or more assemblies.

Without reference :

- Number of contigs / scaffolds (>0, >500bp, > 1kb)
- Largest contig
- N50 : the sequence length of the **shortest contig** at 50% of the total genome length (equivalent to a median of contig lengths)
- Number of Ns in the consensus sequence.

Additional metrics **with a reference** genome :

- NG50 (N50 for reference genome size)
- number of "misassemblies"

De novo metrics

Evaluation of the assembly based on:

- Number of contigs greater than a given threshold (0, 1kb, ...)
- Total / thresholded assembly size
- Largest contig size
- N50 : the sequence length of the shortest contig at 50% of the total assembly length, equivalent to a median of contig lengths. (N75 idem, for 75%)
- L50 : the number of contigs at 50% of the total assembly length. (L75 idem, for 75%)

Reference-based metrics

- Metrics based on based on an alignment of all contigs on a reference genome. :
  - duplication rate
  - percent genome complete
  - NGA50 : equivalent of N50 but with the aligned block of the contigs
  - "Misassemblies" : breakpoint of alignment in a contigs. "
  - Visualisation available

Hands-On : Quast on your genome (without reference)

- Quast 📂 `SsalPrivateGenome.fasta.gz` without reference
- ❓ How many contigs ?
- ❓ Are there scaffolds ?
- ❓ N50, L50 ?

Hands-On : Quast on your genome (without reference)

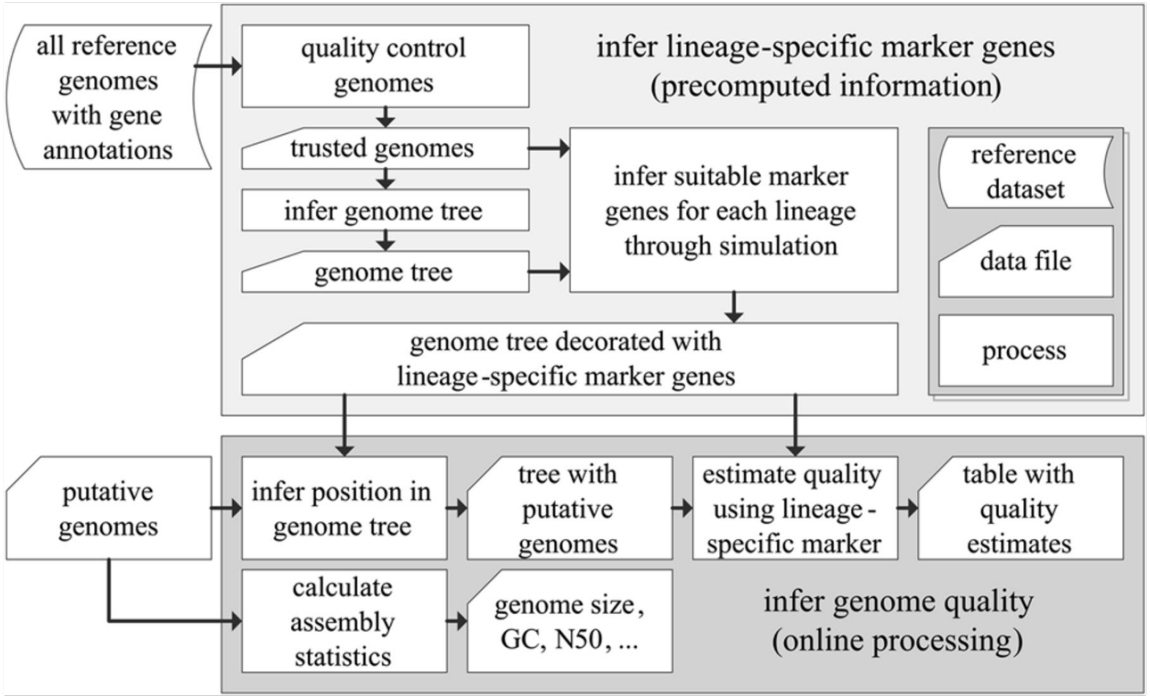
🔧 `Quast Genome assembly Quality`

- **Assembly mode?**
  - `Individual assembly`
  - **Contigs/scaffolds file** 📂 `SsalPrivateGenome.fasta.gz`
  - **Run tool**

Tools to QC your dataset : 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

CheckM workflow



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)

Source : (Parks, Imelfort, Skennerton, Hugenholtz, and Tyson, 2015)

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: genomes 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".

Source : (Parks, Imelfort, Skennerton et al., 2015)

Those quality metrics are somewhat outdated, nowdays for isolates it is more Completeness >= 95% and Contamination <= 5%

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!

Hands-On : CheckM on our private Genome

- CheckM 📂 `SsalPrivateGenome.fasta.gz` (lineage workflow)
- ❓ Lineage
- ❓ Completeness, Contamination, Heterogeneity

Hands-On : CheckM on our private Genome (correction)

🔧 `CheckM lineage_wf`

- **Data structure for bins?**
  - `In Individual datasets`
  - **Contigs/scaffolds file** 📂 `SsalPrivateGenome.fasta.gz`
  - **Run tool**

LUNCH

Genome annotation

What is Annotation ?

Adding biological info to sequences

ribosome binding site

delta toxin
PubMed: 15353161

```
ACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGA
AAAGCAGCCTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTC
CCAGGCCAGTGCCGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTG
GCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGG
ACAGAATGCCCTGCAGGAACTTCTTCTAGAAGACCTTCTCCTCCTG
CAAATAAAACCTCACCCATGAATGCTCACGCAAGTTTAATTACAGA
CCTGAAACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCT
CTCCGTCCGTCCGTGGGCCACGGCCACCGCTTTTTTTTTTGCC
```

transfer RNA
Leu-(UUR)

tandem repeat
CCGT x 3

homopolymer
10 x T

Annotation means having at least :

- a genomic sequence location (described using coordinates on the sequence)
- a biological meaning for this sequence,ex ::
  - Is it a gene ?
  - What is its function ?
  - Is it a coding gene or a non coding gene ? -Is it in an operon?
  - Is it regulated by a common factor ? -Is it a RBS, a repet element, a tRNA ?
- Is it a integration hotspot ?
- Is it a replication origin ?
- …

Three levels of annotation :

- Syntaxic annotation : *Where are the genes and other biological features*
- Functional annotation : *What are there functions*
- Relational annotation : *How they interact in a biological process*

The automatic annotation of bacterial genomes

Software for bacterial genome annotation

Usually, runs numerous sepcilized sofware and integrate results.

Example of software :

- PGAP (NCBI)
- dFAST (DDBJ)
- MicroScope

Trade-off between Specificty / Completeness / Ease of use and Speed

Bakta

- **Bakta** rapid & standardized annotation of bacterial genomes and plasmids from both isolates and MAGs.

Main Features :

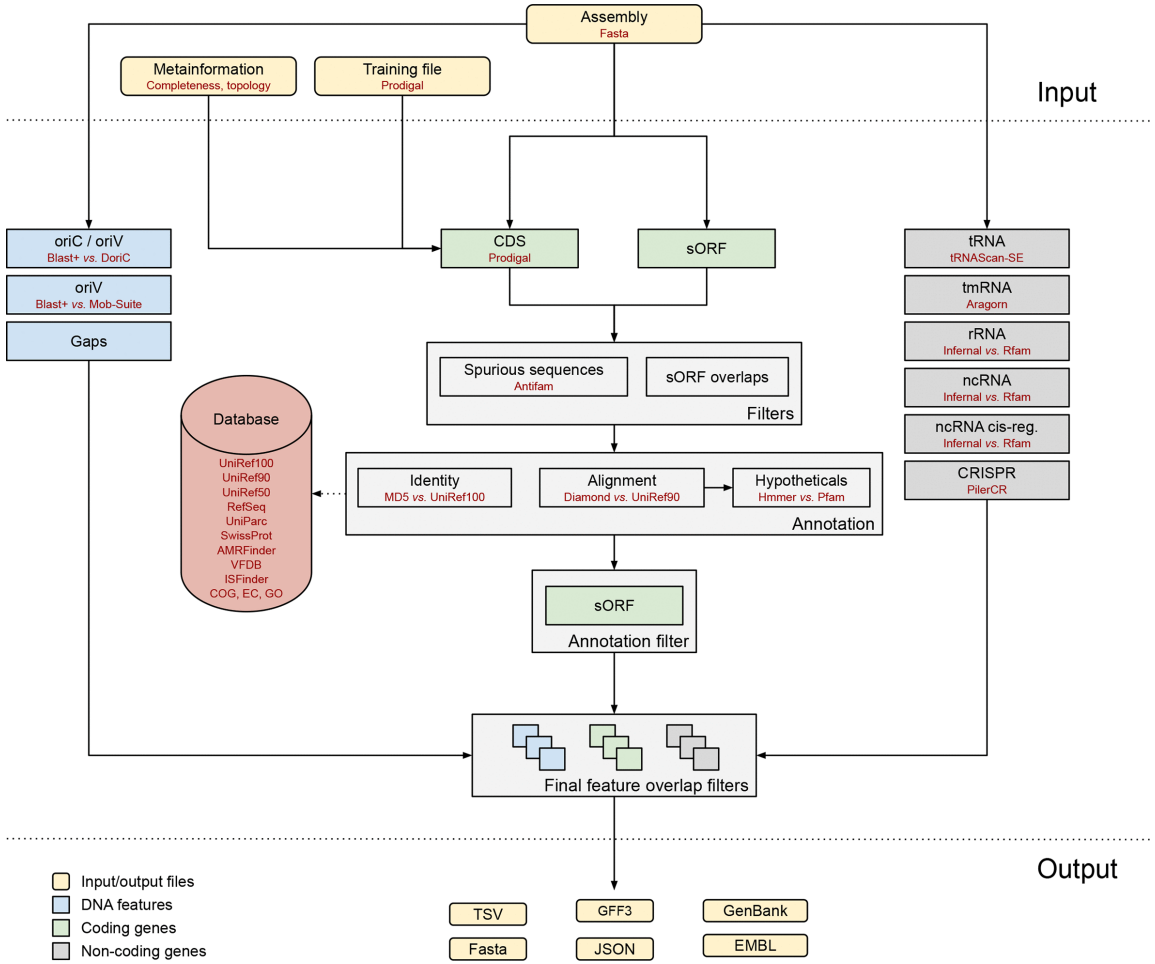
- **Fast** : Bakta can annotate a typical bacterial genome in 10 ±5 min
- Comprehensive & taxonomy-independent database ng from well-studied species to unknown genomes from MAGs.
- Alignement free approach for protein comparison
- Database cross-references & FAIR annotations
- Small proteins / short open reading frames detection

- Not only CDS : ncRNA cis-regulatory regions, oriC/oriV/oriT and assembly gaps, tRNA, tmRNA, rRNA, ncRNA genes, CRISPR, pseudogenes.

- user-provided trusted protein sequences `--protein`

- Ready-to-submid to INSDC member databases `--compliant` mode!!!!

Bakta


Assembly
Fasta
Metainformation
Completeness, topology
Training file
Prodigal
Input
oriC / oriV
Blast+ vs. DoriC
oriV
Blast+ vs. Mob-Suite
Gaps
CDS
Prodigal
sORF
tRNA
tRNAScan-SE
tmRNA
Aragorn
rRNA
Infernal vs. Rfam
ncRNA
Infernal vs. Rfam
ncRNA cis-reg.
Infernal vs. Rfam
CRISPR
PilerCR
Spurious sequences
Antifam
sORF overlaps
Filters
Database
UniRef100
UniRef90
UniRef50
RefSeq
UniParc
SwissProt
AMRFinder
VFDB
ISFinder
COG, EC, GO
Identity
MD5 vs. UniRef100
Alignment
Diamond vs. UniRef90
Hypotheticals
Hmmer vs. Pfam
Annotation
sORF
Annotation filter
Final feature overlap filters
Output
Input/output files
DNA features
Coding genes
Non-coding genes
TSV
Fasta
GFF3
JSON
GenBank
EMBL

Hands-on : Annotate your own Genome with Bakta

Explore the parameters !

Hands-on : Annotate your own Genome with Bakta (correction)

🔧 `Bakta`

- **Select genome in fasta format**
  - 📂 `SsalPrivateGenome.fasta.gz`
- **Optionnal organism options*
  - Specify genus name `Streptococcus`
  - Specify species name `salivarius`
  - Specify strain name `migalicus`
- **Selection of the output files*
  - Select all outputs
    - **Run tool**

Bakta outputs

- *.tsv: annotations as simple human readble TSV
- *.gff3: annotations & sequences in GFF3 format
- *.gbff: annotations & sequences in (multi) GenBank format
- *.embl: annotations & sequences in (multi) EMBL format
- *.fna: replicon/contig DNA sequences as FASTA
- *.ffn: feature nucleotide sequences as FASTA
- *.faa: CDS/sORF amino acid sequences as FASTA
- *.inference.tsv: inference metrics (score, evalue, coverage, identity) for annotated accessions as TSV
- *.hypotheticals.tsv: further information on hypothetical protein CDS as simple human readble tab separated values
- *.hypotheticals.faa: hypothetical protein CDS amino acid sequences as FASTA
- *.txt: summary as TXT
- *.png: circular genome annotation plot as PNG
- *.svg: circular genome annotation plot as SVG
- *.json: all (internal) annotation & sequence information as JSON

Quick reminder on format

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 their position , strand and qualifiers that contains 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 **“//”**.

- 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 .
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"
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
```

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=BioProjec
NZ_LHTK01000001    GenBank    pseudogene    1    1014    .    -    1    ID=LFZ49_RS22320.pseudogene;Alias=
NZ_LHTK01000001    GenBank    gene    1011    1634    .    -    1    ID=LFZ49_RS00010;Name=LFZ49_RS00010;
NZ_LHTK01000001    GenBank    mRNA    1011    1634    .    -    1    ID=LFZ49_RS00010.t01;Parent=LFZ49_RS
```

Bakta results exploration

- How many protein coding genes ?
- Hown many rRNA, tRNA ?
- Explore the dbXrefs for gene `pepF` (Uniprot, Uniref…)

Advice for private genomes:

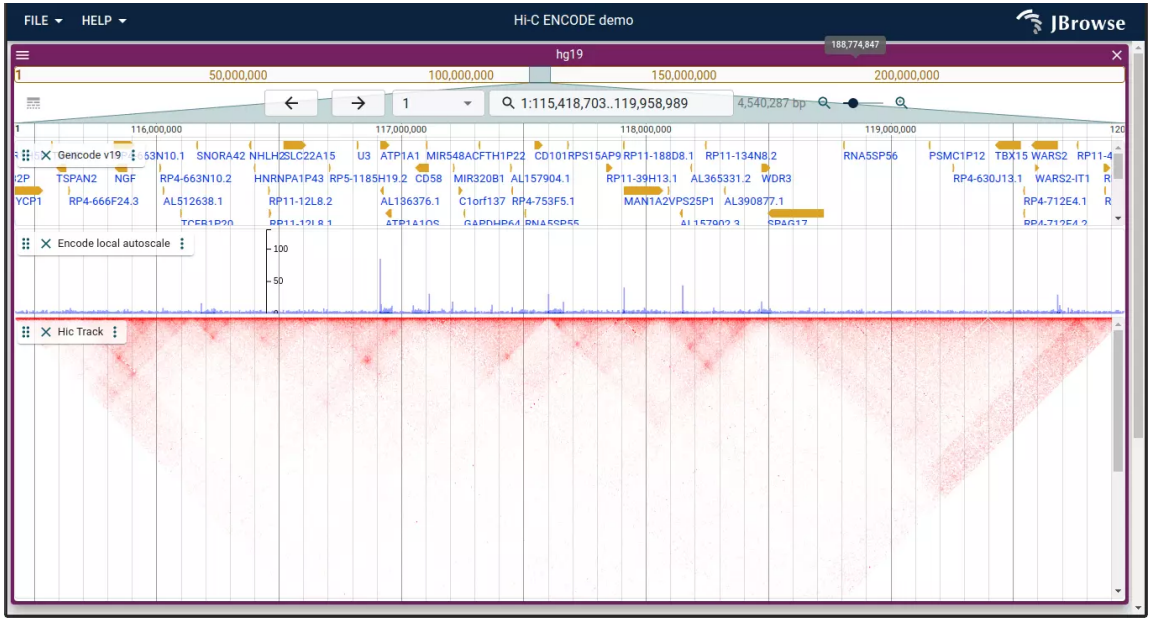
Annotate your genome directly with the "good" meta-data : strain name, project id, locus tags …

Frome the documentation :

> Most genomes annotated with Bakta should be ready-to-submid to INSDC member databases GenBank and ENA. As a first step, please register your BioProject (e.g. PRJNA123456) and your locus_tag prefix (e.g. ESAKAI).

- First declare your study in EMBL
  - Documentation
  - You will obtain a Bioproject number and a locus_tag prefix
  - Submit the raw reads as soon a possible (embargo possible)
  - Submit the genome annotation (embargo possible)

Hands-On : Visualize your annotated genomes with Jbrowse



Jbrowse : Open Source Genome Browser

- Right in you browser
- Alternatives : igv webapp (not yet in Galaxy)

Hands-On : Visualize your annotated genomes with Jbrowse (correction)

🔧 `JBrowse genome browser`

- **Use a genome from history**
  - 📂 `Bakta on data 1 : Replicon/contig DNA sequences`
- **Genetic code*
  - `11. The bacterial, Archeal and Plant Plastid Code`
- **Annotation Track**
  - **Track type**
    - `GFF/GFF3/BED features`
    - **GFF/GFF3/BED Track Data**
    - 📂 `Bakta on data 1: Annotation and sequence (GFF3)`
  - **Run tool**

Specialized Annotation

- Often relies on specialized and curated databases
- Results in tables with homology/identity and overlap/coverage informations
- What specialized databases do you know?

Specialized Annotation : Mobile Genetic Elements

- Some popular tools available on **Galaxy**
  - Conjscan: detect both conjugative plasmids and integrated conjugative elements [https://doi.org/10.1007/978-1-4939-9877-7_19]
  - ICEscreen: detect integrated conjugative elements in Bacillota genomes [https://doi.org/10.1093/nargab/lqac079]
  - VirSorter: DNA and RNA virus identification [https://doi.org/10.1186/s40168-020-00990-y]

Specialized annotation : antibioresistance & beyond

- Some popular tools available on **Galaxy**
  - Abricate: antimicrobial resistance or virulence genes, include many databases (NCBi, CARD, ARG-ANNOT, Resfinder, PlasmidFinder,...) [https://github.com/tseemann/abricate]
  - StarAMR: antimicrobial resistance genes, Scans genome assemblies against the ResFinder, PlasmidFinder, and PointFinder [https://doi.org/10.3390/microorganisms10020292]

Hands on

🔧 `ABRicate`

- **Select the GCF_903908965 genome in gbff format**
  - 📂 `GCF_903908965.1-genomic.gbff`
  - default parameters

🔧 `ABRicate Summary`

- **Combine ABRicate results into a simple matrix of gene** presence/absence
- default parameters

Look at result files ant try to interpret

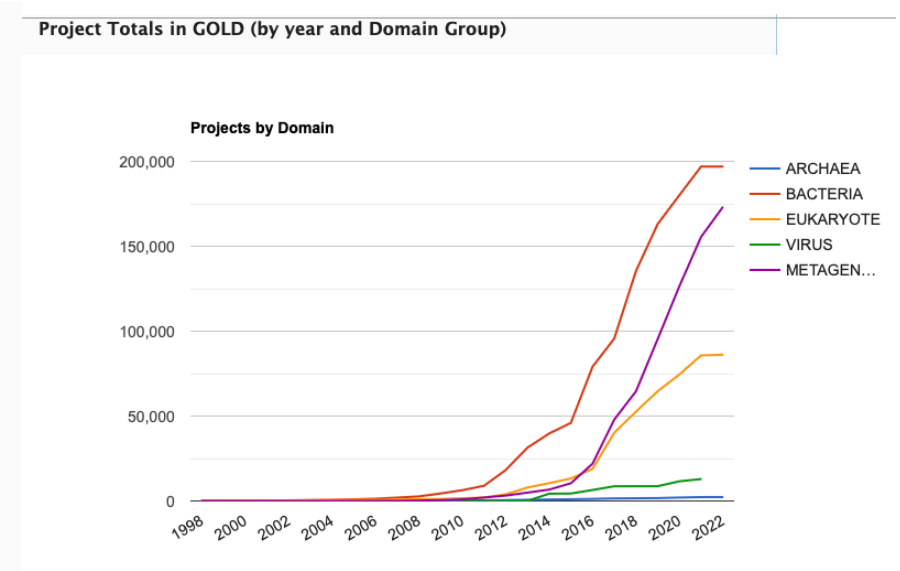
Day 1 wrap-up

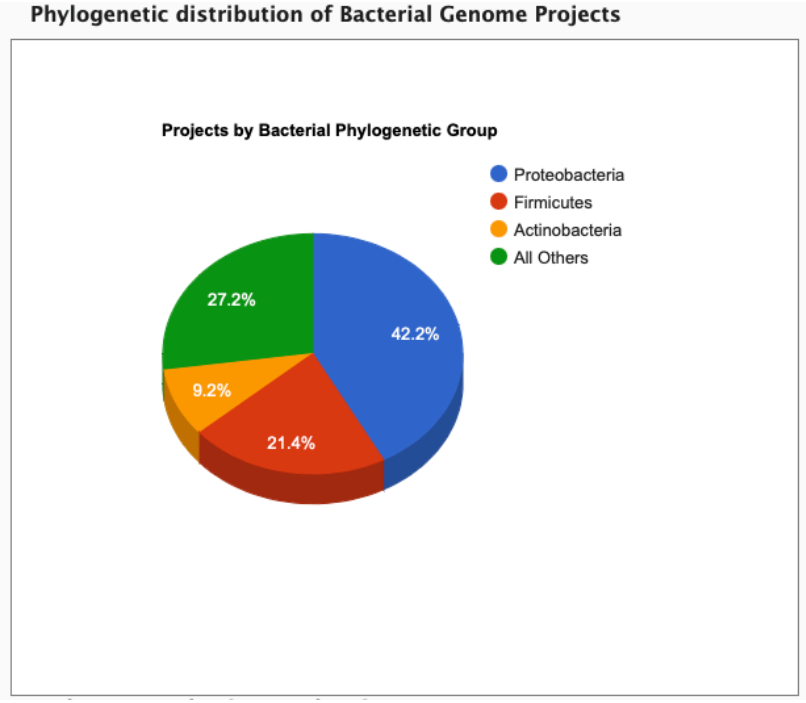
- What have you learned today ?
- Any questions ?

Microbial comparative genomics

A huge number of microbial genomes

Bacterial and metagenomic genome projects: the top of the sequencing projects

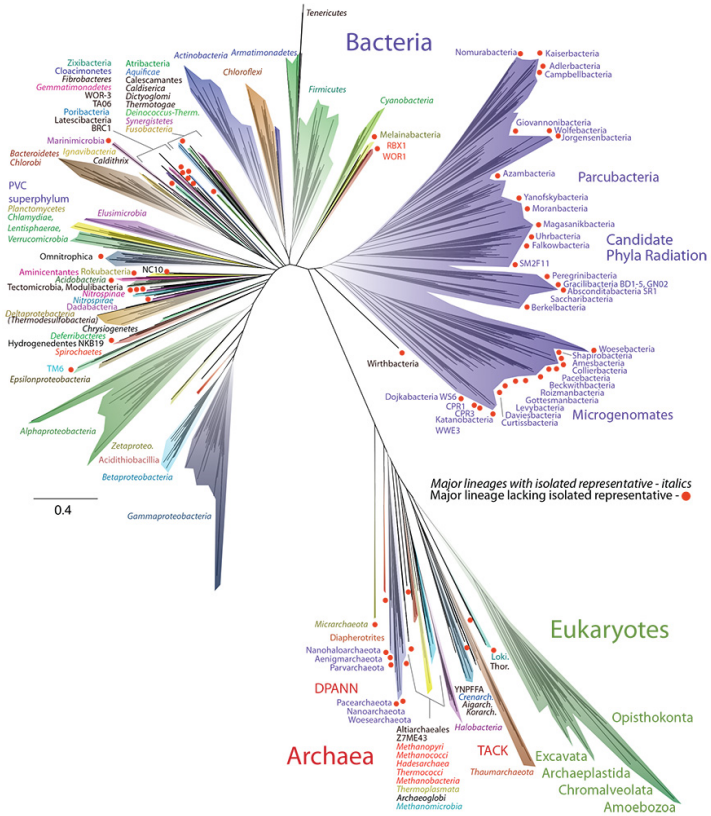




Proteobacteria and Firmicutes: the two most sequenced group of genomes

Source: GOLD statistics

And there is still a lot more to explore, especially for microbes



- genomic data where recovered from diverse metagenomic samples
- tree reconstructed from an alignemnt of 16 ribosomal proteins
- red dots indicate lineages lacking an isolated representative
- there are a large number of major lineages without isolated representatives

Source : Hug, L., Baker, B., Anantharaman, K. et al. A new view of the tree of life. Nat Microbiol 1, 16048 (2016). https://doi.org/10.1038/nmicrobiol.2016.48

Frequent problems for microbial genome analysis and comparison

- Heterogenous quality of sequencing and assembly
- Contaminations
- 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

Why comparative genomics

- Answer to (not so simple) questions like :
  - What is the genomic diversity into a microbial species / genus ? Is the taxonomy of my strains consistent ?
  - How does the gene repertory evolves into a species / genus ?
  - Does this diversity could explain a given phenotype :
    - metabolism
    - probiotics (anti-inflamatory)
    - pathogenicity
  - …

Dataset construction

Dataset building

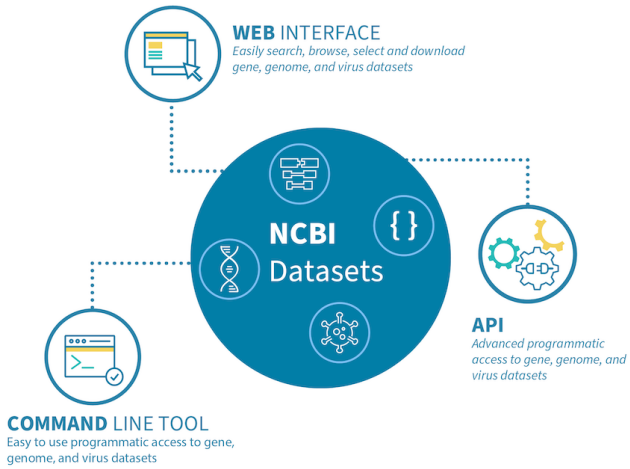
- Genomes of interest could be
  - already published and available at public databanks (ENA, NCBI, …)
  - **private**, not yet published.
- At least, we need :
  - [as much as possible] complete genome assemblies (contigs / scaffolds in fasta format)
  - Syntactic and functional annotation :
  - Genbank or GFF format
- For private genomes, think about what we have learned yesterday
- It's always better if annotation is homogeneous

Practical : public genomes

How to **list**, **filter** and **download** publicly available genomes ?

- **list** all publicly available genomes
- **select** a subset of them according to
  - metadata
  - quality metrics (size, completeness,…)
- **download** genomes in various formats

A solution : NCBI 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

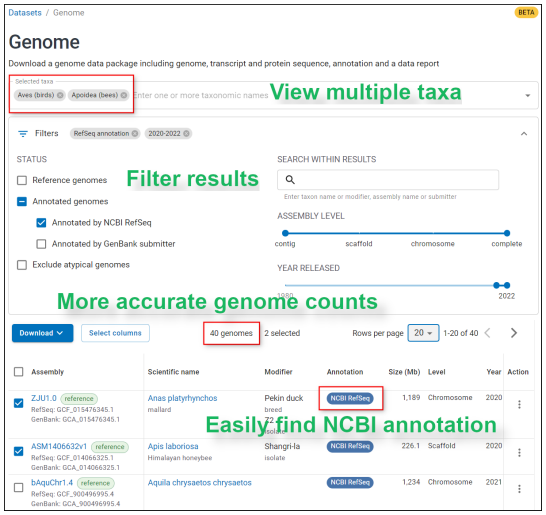
	GCA_	GCF_
Also known as	GenBank assembly	RefSeq assembly
Submitter-owned assembly archive	✔	✘
NCBI-maintained assembly copy	✘	✔
Always includes annotation	✘	✔
NCBI may add sequences (e.g. mitochondrial genomes)	✘	✔
NCBI may remove sequences (e.g. contamination)	✔ *	✔

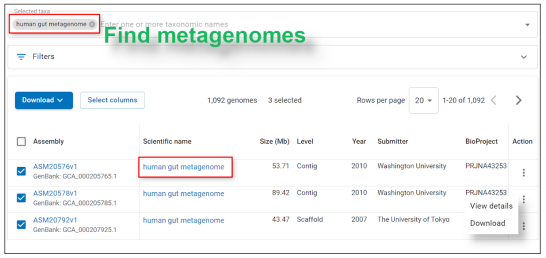
* following submitter request or agreement

NCBI Datasets website genome sources

Source : Dataset documentation

NCBI Datasets : Datasets Genome Table



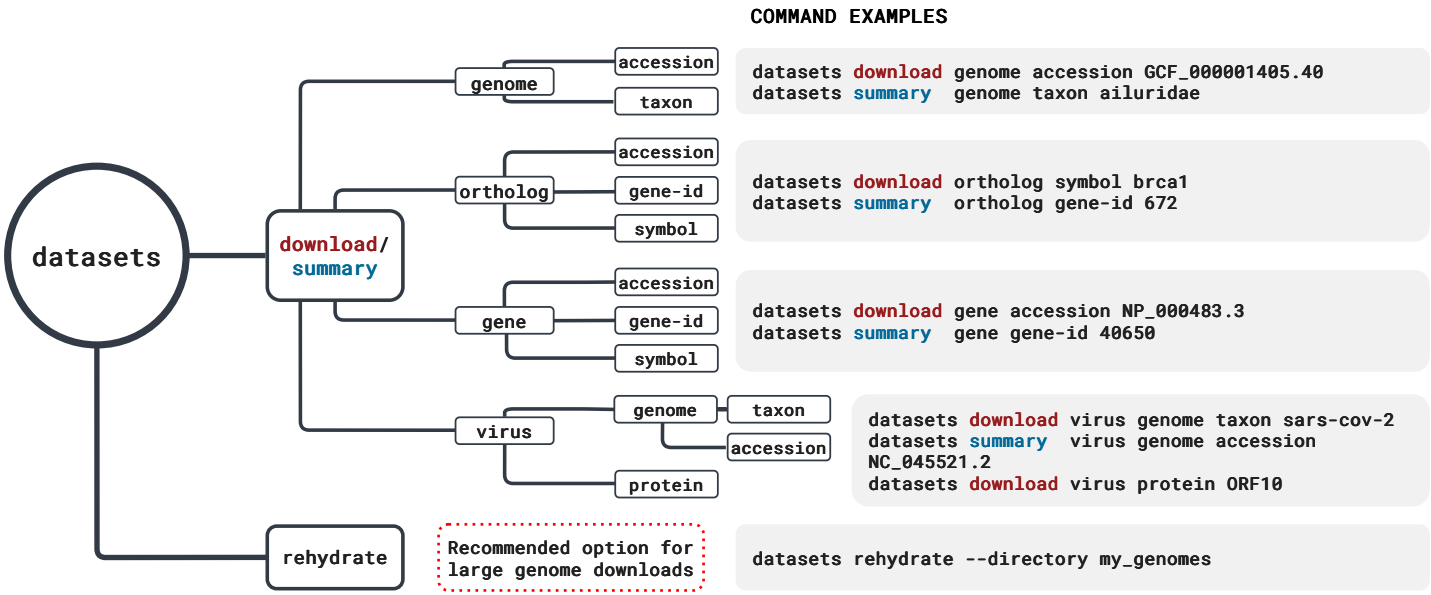


NCBI Datasets Genome Page

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



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 : **A**plication **P**rogrammatic **I**nterface

jupyter ncbi-datasets-assembly (modifié)

Visit repo | Copy Binder link

File Edit View Insert Cell Kernel Widgets Help | Non fiable | Python 3 (ipykernel)

Exécuter | Markdown | Download | GitHub | Binder | **Memory:** 161.7 MB / 8 GB

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.

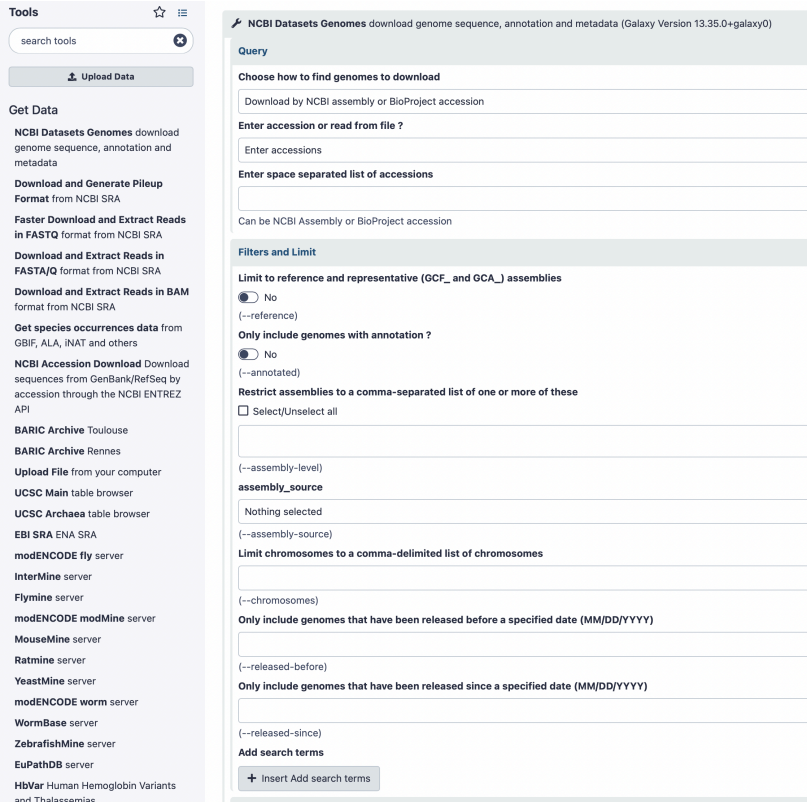
Entrée [1]:
```
assembly_accessions = ['GCF_000001405.40'] ## needs to be a full accession.version

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



A wrapper of the command line tool

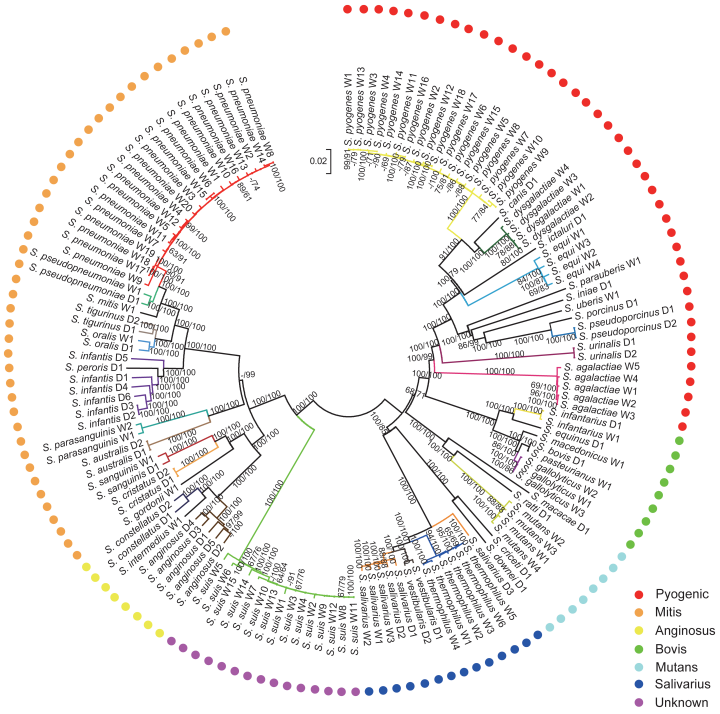
Parameters to define packages files

NCBI Datasets: Galaxy Integration

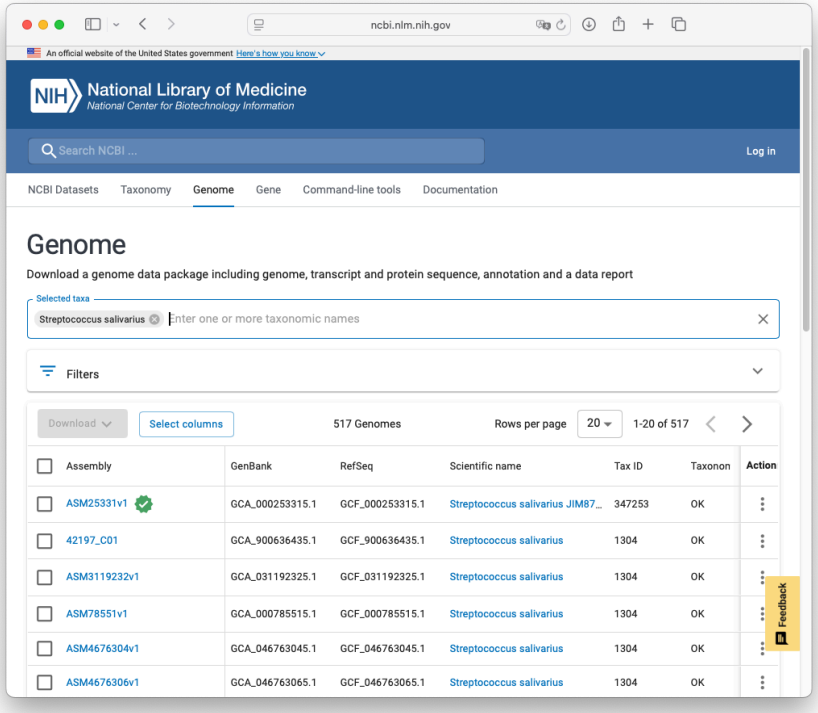
- A few caveats of the wrapper :
  - Some (not so easy) errors when select / filter fails
  - Impossible to just download a list of genomes as a file and "rehydrate" it after
- What recomend to :
  - use the NCBI dataset genome page to browse / filter a list of genomes of interest
  - download the list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

The training datasets

We will work on a dataset of public *Streptococcus salivarius* genomes



Gao et al. PLoS One.2014(https://doi.org/10.1371/journal.pone.0101229)



517 S. salivarius public assemblies at NCBI

The training dataset

We will build a dataset including

- The genome ASM1102908v1 (GCF_011029085.1), our "private genome"
- A list of 49 public genomes of Streptococcus salivarius
- We will download a dataset of 50 genomes from their *accession numbers* using the tool **NCBI datasets**

Practical: connections to the tools

Two tools needed : **Galaxy** and **NCBI datasets**

- Connect to Galaxy (https://usegalaxy.fr) with your account.
- Do not forget to login (upper right …)
- Create a new history
- Connect to NCBI Datasets in a separate tab (https://www.ncbi.nlm.nih.gov/datasets/)

Hands on: retrieve genomes from a tabular file in Galaxy with NCBI datasets (2)

Download 50 Streptococcus salivarous public assemblies from their *accession numbers*.

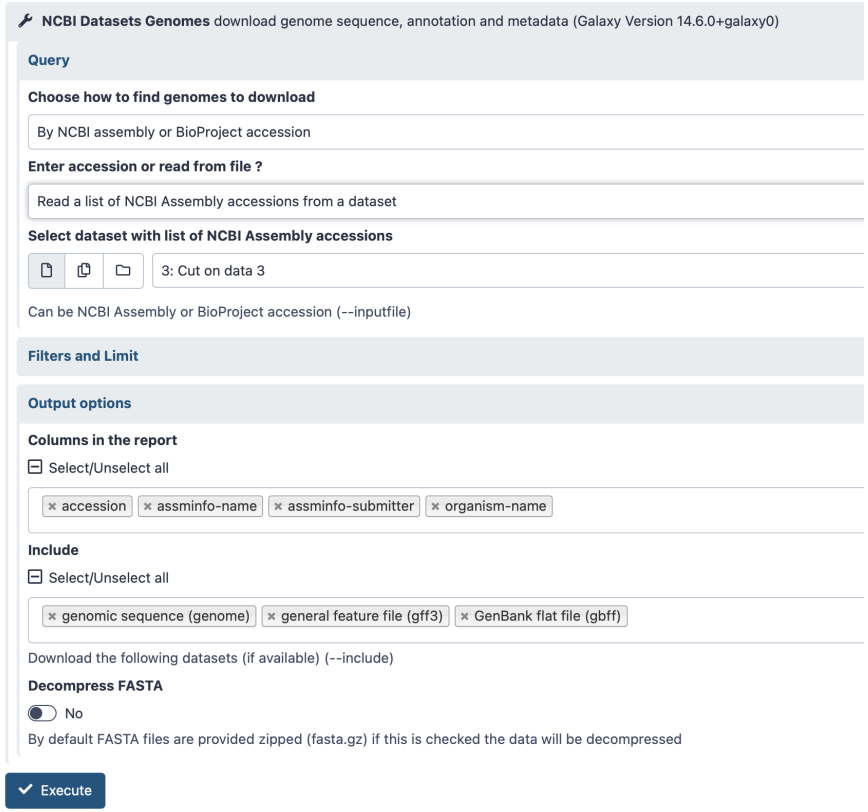
- List of assembly accession in a tabular file downloaded from Dataset genome Table
- Import `Ssal_50G_dataset.tsv` from `Libraries /Formation Migale 2025 / Annotation auto et génomique comparée - Mai 2025 / Comparative Genomics / Dataset`l
- Select the first column of the file (Assembly Accession) using `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

Use case: from a tabular file in Galaxy with NCBI datasets (correction)

- Select the first column of the file (Assembly Accession) using `Cut columns from a table`

Use case: from a tabular file in Galaxy with NCBI datasets (correction)

- 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



Practical : Galaxy - Transform list into flat datatsets

- Results from NCBI Datasets Genome Download are stored in Lists of Lists (List of Lists of datasets)
- We must "flatten the collection" , ie, do only list of datsets, not list of list
- List or dataset Collection (see Galaxy documentation) allow you to group together related datasets into collections that can be processed alltogether.

-> **Do this for : Genbank flat file, Fasta files and GFF3 files**

Practical : Galaxy - Transform list into flat datatsets (2)

🔧 `Flatten collection`

- **Input collectiont**
  - 📂 `NCBI Genome datasets : : Genbank flat file`
  - **Run tool**

Practical : Quast your dataset !

Apply quast to the 50 assemblies of you dataset.

🔧 `Quast Genome assembly Quality`

- **Assembly mode?**
  - `Co-assembly`
  - **Contigs/scaffolds file** Dataset Collection 📂 `NCBI Genome datasest : fasta`
  - **Run tool**


COFFEE
BREAK

Dataset diversity analysis

Genome diversity evaluation

Why ?

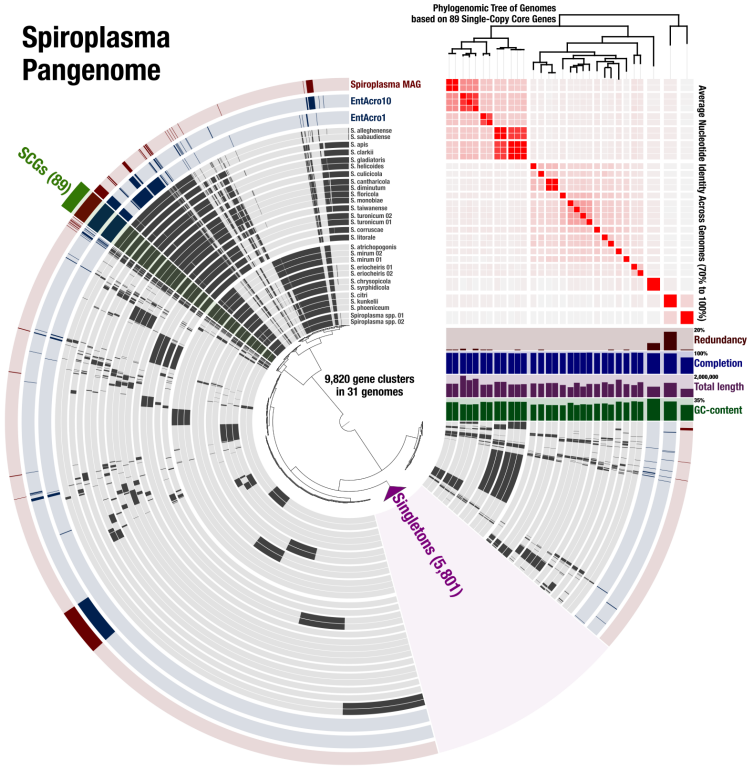
- Identify outlier genomes
- Identify groups of (very) similar genomes and de-replicate datasets
- Estimate genome similarity in a dataset and design an adapted comparative strategy

How ?

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

Average Nucleotide Identity (ANI)

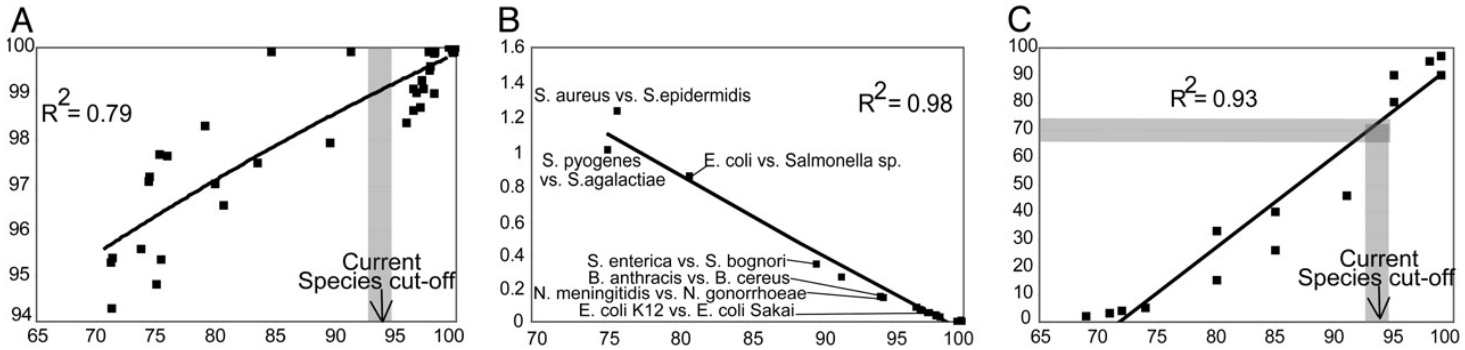
- Meet the need for a robust measure of genomic reladness and a systematic and scalable species assignation technique
- Mean identity percent of aligned regions of a pair of genomes
- Rely on pairwise alignments that may come either from aligned core genes or from genomic alignements
- Can easily be used to build phylogenetics tree using distance methods
- Is implemented in several bioinformatics tools (gANI, fastANI)



Pangenomics, phylogenomics, and ANI of 31 Spiroplasma genomes.

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



Relationships between ANI, 16S rRNA, mutation rate, and DNA–DNA reassociation

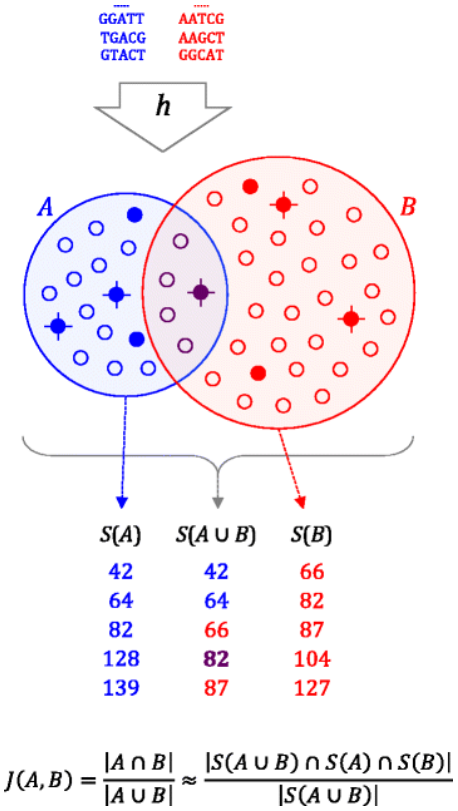
Source : (Konstantinidis and Tiedje, 2005)

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, which estimates the rate of sequence mutation under a simple evolutionary model



Overview of the MinHash bottom sketch strategy for estimating the Jaccard index.

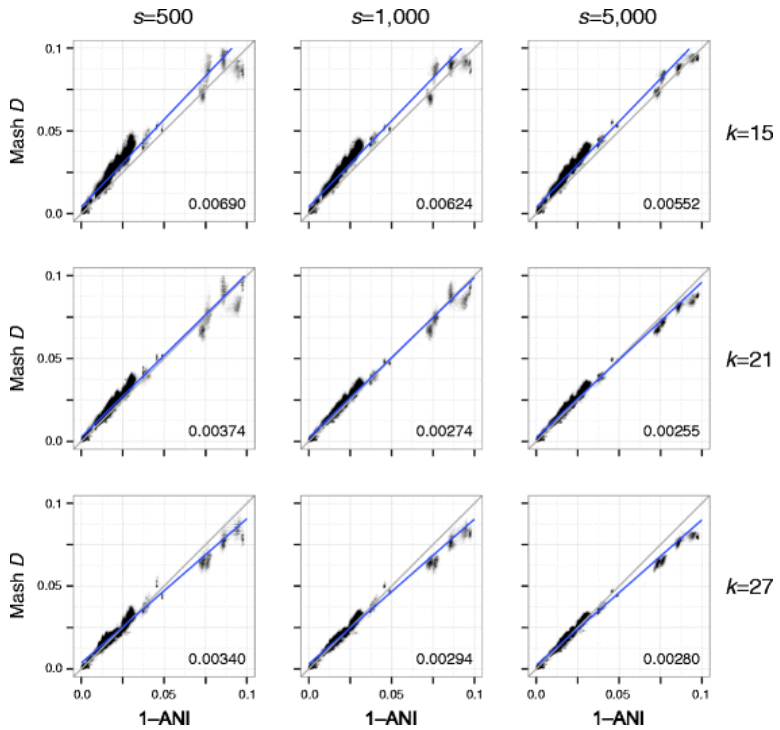
Source : (Ondov, Treangen, Melsted, Mallonee, Bergman, Koren, and Phillippy, 2016)

MASH distances correlate well with ANI

- Dataset: 500 complete E. coli genomes
- Gray lines: model relationship D = 1–ANI

- Each plot column shows a different sketch size
- Each plot row a different k-mer size k.
  .

- 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)



Scatterplots illustrating the relationship between ANI and Mash distance for a collection of Escherichia genomes.

Source : (Ondov, Treangen, Melsted et al., 2016)

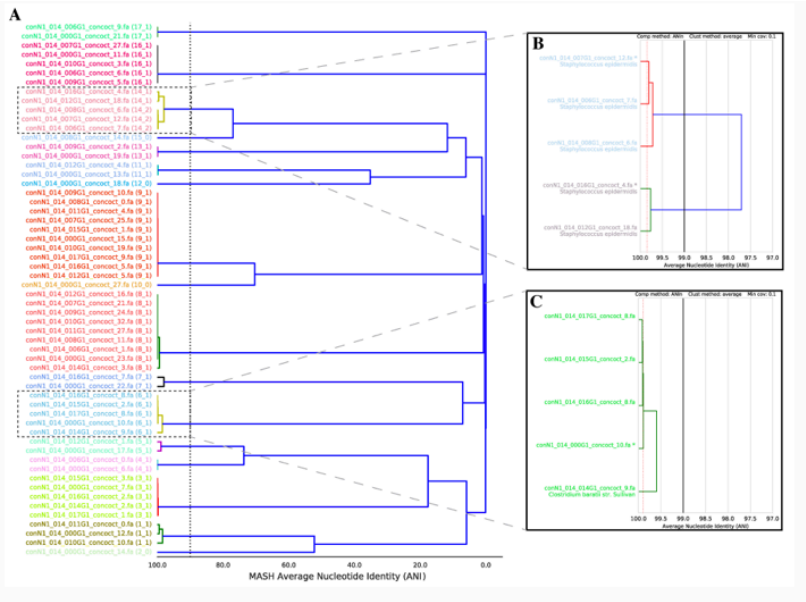
dREP: comparison and dereplication

- 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

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

Source : (Olm, Brown, Brooks, and Banfield, 2017)



Assembly and de-replication with dRep

dREP important concepts and parameters

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 completness:** 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. **The score 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 produce many results 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.

Output files of dRep

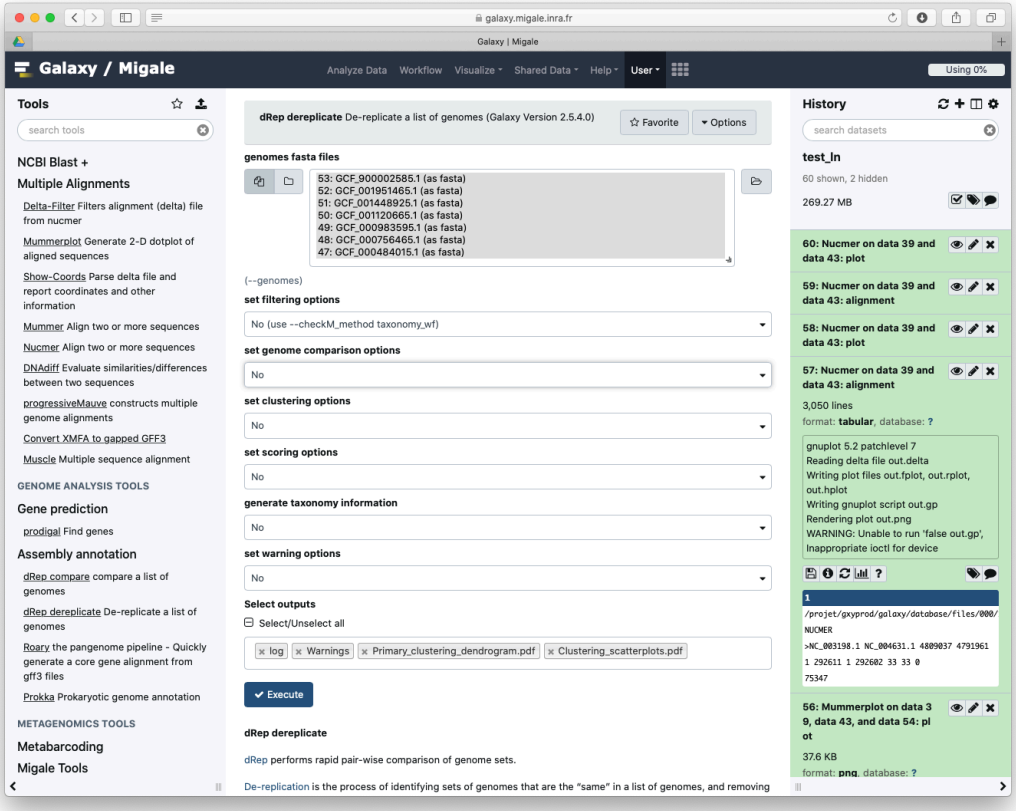
```
workDirectory
./data
...../checkM/
...../Clustering_files/
...../gANI_files/
...../MASH_files/
...../ANIn_files/
...../prodigal/
./data_tables
...../Bdb.csv  # Sequence locations and filenames
...../Cdb.csv  # Genomes and cluster designations
...../Chdb.csv # CheckM results for Bdb
...../Mdb.csv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

dRep results

Source : (Olm, Brown, Brooks et al., 2017)

Practical : dreplicate your dataset !

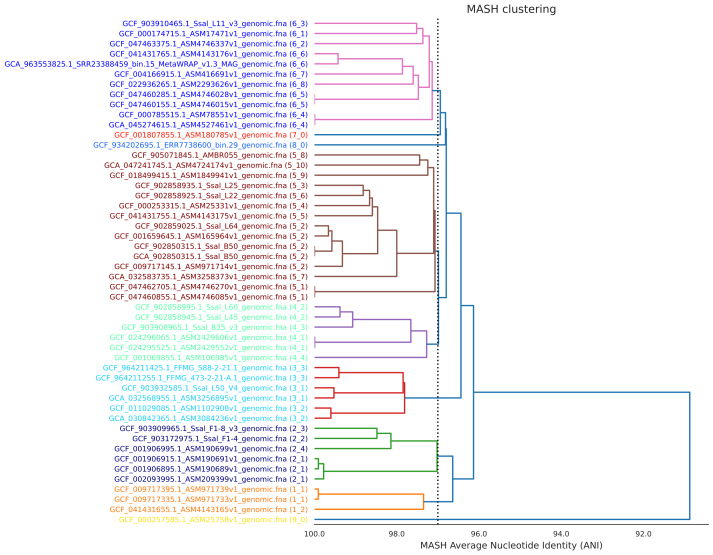
- use 🔧 `dREP-dreplicate` to explore the Streptococcus salivarius genome dataset diversity and completenes and dereplicate the dataset
- explore and interpret results
- input: 50 genome fasta files
- Change `ANI threshold to form primary clusters` to `0.97`



dRep on mMigale Galaxy server

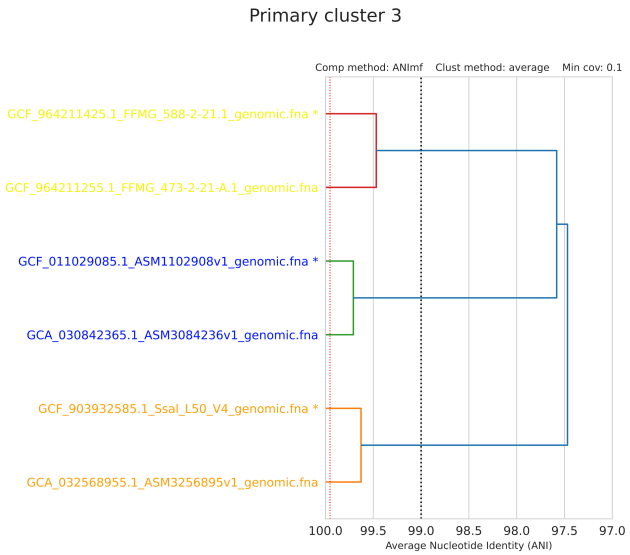
dRep results interpretation

Important outputs of dRep



Primary_clustering_dendrogram.pdf

The "Primary_clustering_dendrogram.pdf" output file



Secondary_clustering_dendrogram.pdf

The "Secondary_clustering_dendrograms.pdf" output file and the deReplicated genomes list

Practical : construct dereplicate collection !

- `dREP-dreplicate` outputs a list of dereplicated genomes in `fasta` format.
- We need to construct a collection of dereplicated genomes in the other formats available (Genbank, GFF). Those collections will be used as ipnput for susbsequent analysis.
- *How to do that ?* :
    - extract the list of dereplicated genomes from drep csv output
    - select in the collection, the correponding files

Practical : construct dereplicate collection - 1 - Extract list of dereplicated genomes

🔧 `**Cut** columns from a table`

- **Cut columns**
  - `c1`
- **Delimited by**
  - `comma`
- **From**
  - 📂 `drep dreplicate : Widb.csv`
    - **Run tool**

Practical : construct dereplicate collection - 2 - Clean list

dreP tends to rewrite assembly names, we have to extract, in each line, the original assembly accession :

- From `GCF_009717395.1_GCF_009717395.1_ASM971739v1.fasta` to `GCF_009717395.1`
- We will use Regular Expression to do a super powerfull search and replace for each line (See here for a complete explanation)

🔧 `**Regex Replace**`

- **From**
  - 📂 `cut on data XXX`
- **Search String**
  - `(.+\d)_(GC.*)`
- **Replace String**
  - `\1`
  - **Run tool**

Practical : construct dereplicate collection - 2 - Clean list

Remove the first ligne, non informative (genome)

🔧 `**Remove beginning of a file**`

- **Remove first**
  - `1`
- **From**
  - 📂 `regex replace on data XXX`
  - **Run tool**

Practical : construct dereplicate collection - 3 - Filter collection

Use the clen list to filter collection

🔧 `**Filter collection**`

- **Input Collection**
  - 📂 `Genbank datsets (flattened)`
- **How should the elements to remove be determined**
  - `Remove if identifiers are ABSENT from file`
- **Filter out identifiers absent from**
  - 📂 `Remove of begining of data xx`

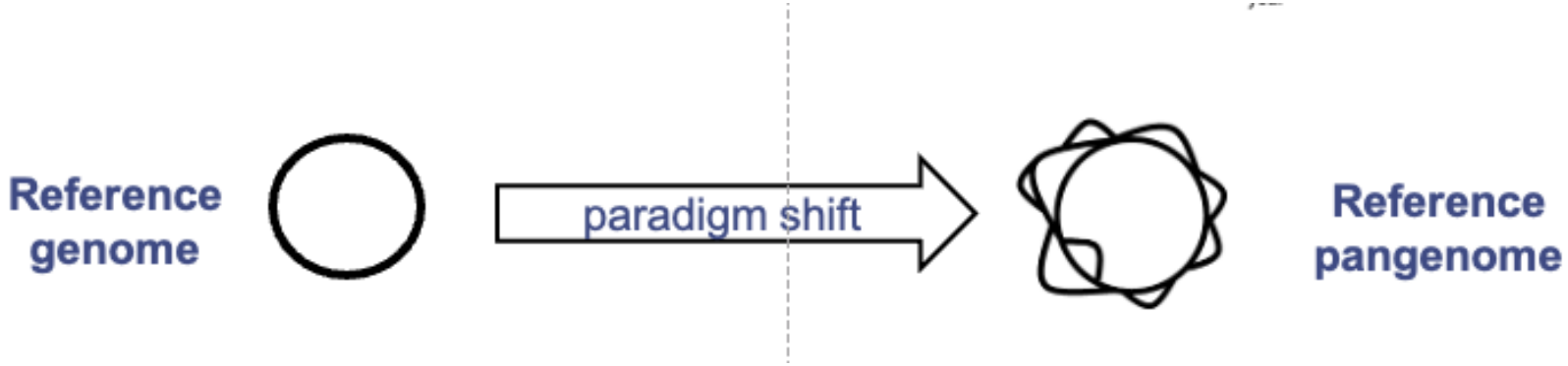
This tool will produce two collection, `filtered` that contains dereplicated genomes, `discarded` that contains remaining genomes. rename the `filtered` collection and dlete the `discarded`.

LUNCH

The microbial pangenome

From genomes to pangenomes

- With the evolution of sequencing technology, there is an explosion of prokaryotic genomes available in databases
- Prokaryotic genomics studies now rely on the comparison of thousands of genomes from the same species
  - High diversity of gene content from horizontal gene transfer (5% to 40% of variable genes)
  - high level of polymorphism



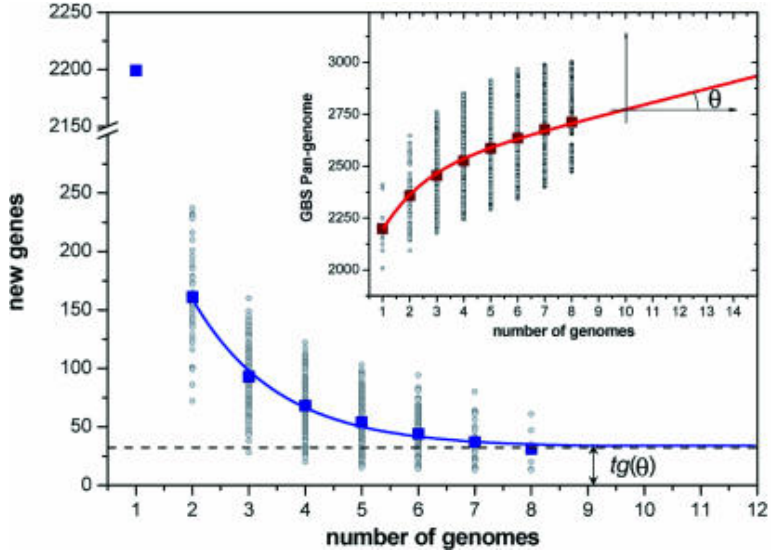
From reference genoems to pangenomes

The microbial pangenome

First term apparition in 2005 in two publications

- Tettelin et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pangenome” Proc Natl Acad Sci U S A.
- Medini et al. "The microbial pangenome" Curr Opin Genet Dev.

*A bacterial species can be described by its **pangenome** composed of a **core genome** containing genes present in all strains, and a **dispensable genome** containing genes present in two or more strains and genes unique to single strains.*

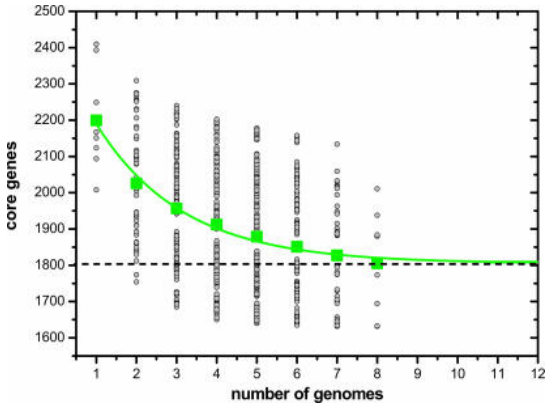


Streptococcus group B pan genome

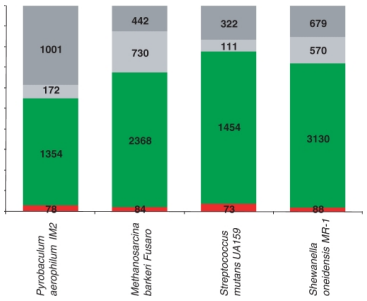
References: (Tettelin, Masignani, and Cieslewicz MJ, 2005) and (Medini, Donati, Tettelin, Masignani, and Rappuoli, 2005)

The microbial pangenome

- Definition refinment by Koonin (2008) and Collins (2012): the 3 classes of prokaryotic genes
  - **core (or persitent) genes**: a small fraction of highly conserved genes
  - **shell genes**: a larger set of moderately conserved genes
  - **cloud genes**: (nearly) unique genes



Streptococcus group B core genome

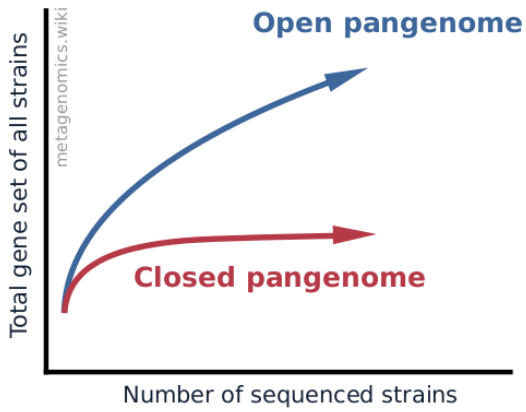


A Common and rare genes in selected archaeal and bacterial genomes. Red, core; green, shell; light gray, cloud; dark gray, ORFans.

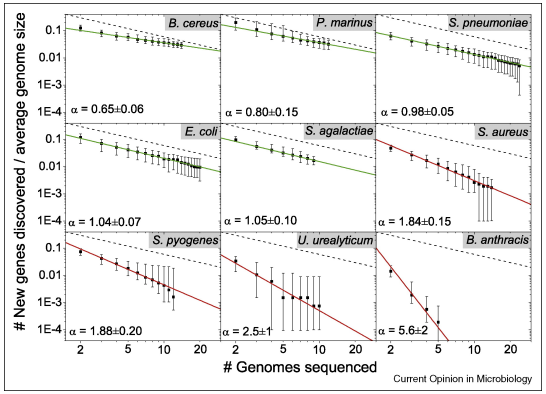
Source : (Koonin and Wolf, 2008)
Source : (Collins and Higgs, 2012)

Open or closed pangenome

- Some bacterial species are considered to have an unlimited large gene repertoire => **open pangenome**
- Other species seem to be limited by a maximum number of genes in their gene pool=> **closed pangenome**
- Authors use **Power or Heaps law** to fit of the overall number of genes (pangenome) obtained according to the number of sequenced genomes



Open and closed pangenomes



Power law regression for species with open and closed pangenomes.Red curves indicate closed pangenomes, green curves indicate open ones.

Source : (Tettelin, Riley, Cattuto, and Medini, 2008)

Roary: the first rapid large-scale prokaryote pangenome analysis

Roary, the pan genome pipeline, takes *closely related* annotated genomes in GFF3 file format and calculates the pan genome.

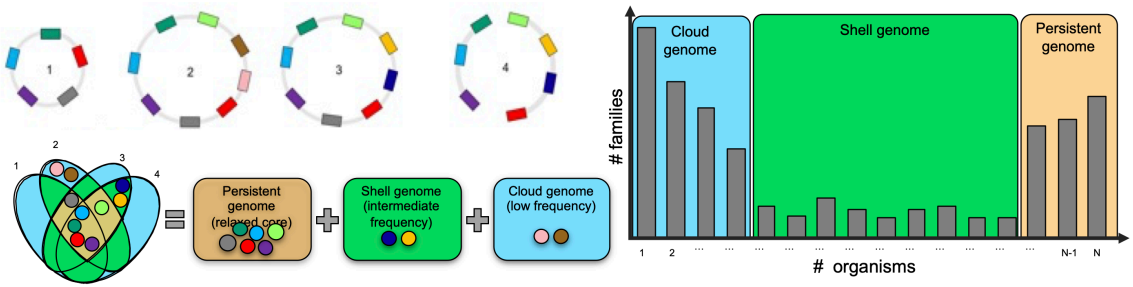
- Takes and input annotated genomes in **GFF3** format
  - *Very* sensitive to the validity of the format
  - GFFs generated by **Prokka** are valid
  - Locus tags must be uniques
  - GFF from NCBI are **invalid** (sequence is missing)
    - Must be converted from Genbank using "Genbank to GFF3" converter. You can use "From Genbank (NCBI datasets genome) to gff3" workflow

- What does Roary do ?
  - converts annotated coding sequences (CDS) into protein sequences
  - cluster these protein sequences iteratively by several methods (cd-hit, all vs all blastp)
  - further refines clusters into orthologous genes
  - for each sample, determines if a gene is present/absent
  - uses this information to build a tree, using FastTree
  - overall, calculates the number of genes that are shared, and unique

Andrew et al. Bioinformatics 2015

PanGGOLiN: depicting microbial diversity via a partitioned pangenome graph

- Gautreau et al. 2020 (https://doi.org/10.1371/journal.pcbi.1007732)
- builds pangenomes for large sets of prokaryotic genomes (i.e. several thousands)
- classify gene families into three classes: persistent, cloud, and one or several shell partitions
- relies on a statistical model that makes a more robust estimation of the persistent genome in comparison to classical approaches based on gene family frequencies in isolate genomes and also in MAGs



Apply PanGGOLiN to the 34 dereplicated genomes of your dataset

- Tool : 🔧 `PPanGGOLiN all`
- Input dataset:
  - the set of 34 dereplicated genomes **with annotation in Genbank format**, ie the 34 S. salivarius "gbff" files - 📂 `Ssalivarius_*.gbff` files
- Parameters :
  - Select all the output files
  - Set the **Translation table** option to `11- Bacterial and plant plastid`
  - **Run tool**

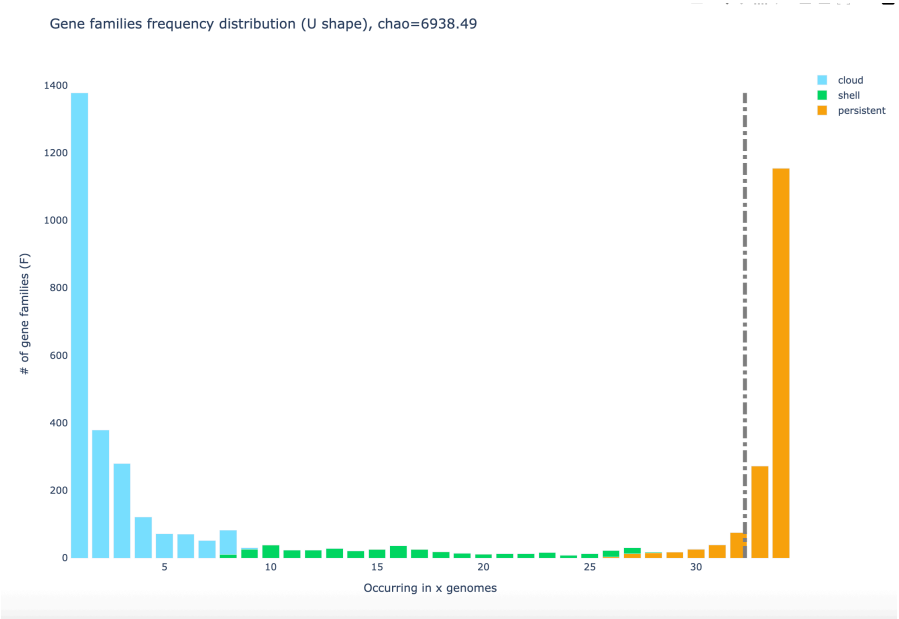
PanGGOLiN results interpretation

- PanGGOLiN outputs

- **genomes_statistics.tsv**: Statistics about genomes a tab-separated file summarizing the content of each of the genomes used for building the pangenome. Look at the online documentation
- **matrix.tsv**: Informations about gene families. A tab separated presence absence matrix of genomes and gene families. Similar to **gene_presence_absence.Rtab**
- **Ushaped_plot.html**: U-shaped plot is a figure presenting the number of families (y-axis) per number of genomes (x-axis)
- **tile_plot.html**: a heatmap representing the gene families (y-axis) in the genomes (x-axis) making up your pangenome. Useful to detect pangenome structure and outlier

PanGGOLiN results

- Number of families: 4437
- persistent: 1614
- shell: 399
- cloud: 2424



Falaxy-Fasttree

Apply PanGGOLiN to align the persistent genes

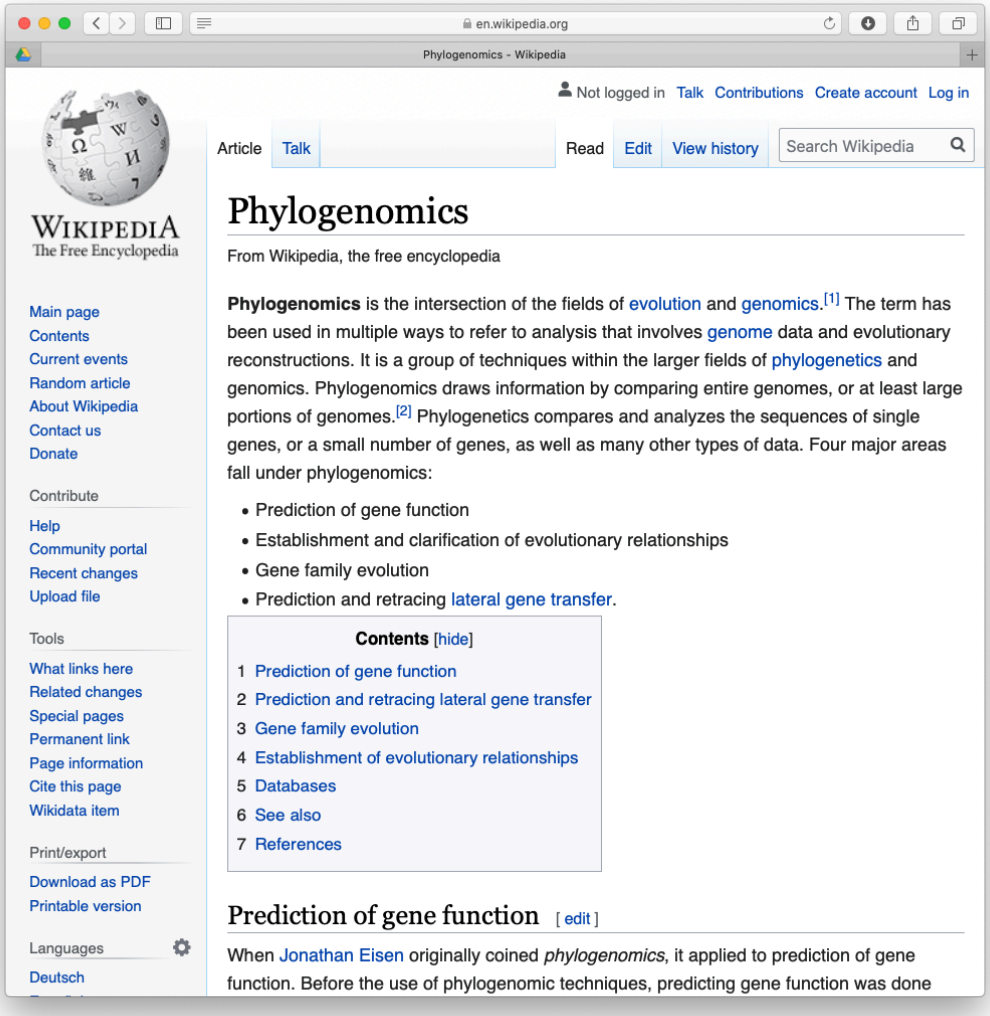
- Tool : 🔧 `PPanGGOLiN MSA`
- Input dataset: - 📂 `Pangenome HDF5 file` files
- Parameters:
  - All the output files selected
  - Set the **Partition** option: `persistent`
  - Set the **Source** option to `DNA`
  - Set the **Translation table** option to `11- Bacterial and plant plastid`
  - **Run tool**


COFFEE
BREAK

Phylogenomics basics

A few concepts on phylogenomics

- Phylogenomics definition



en.wikipedia.org

Phylogenomics - Wikipedia

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Languages
Deutsch

Phylogenomics

From Wikipedia, the free encyclopedia

Phylogenomics is the intersection of the fields of evolution and genomics.[1] The term has been used in multiple ways to refer to analysis that involves genome data and evolutionary reconstructions. It is a group of techniques within the larger fields of phylogenetics and genomics. Phylogenomics draws information by comparing entire genomes, or at least large portions of genomes.[2] Phylogenetics compares and analyzes the sequences of single genes, or a small number of genes, as well as many other types of data. Four major areas fall under phylogenomics:

- Prediction of gene function
- Establishment and clarification of evolutionary relationships
- Gene family evolution
- Prediction and retracing lateral gene transfer.

Contents [hide]

1 Prediction of gene function
2 Prediction and retracing lateral gene transfer
3 Gene family evolution
4 Establishment of evolutionary relationships
5 Databases
6 See also
7 References

Prediction of gene function [edit]

When Jonathan Eisen originally coined *phylogenomics*, it applied to prediction of gene function. Before the use of phylogenomic techniques, predicting gene function was done

A few concepts on phylogenomics

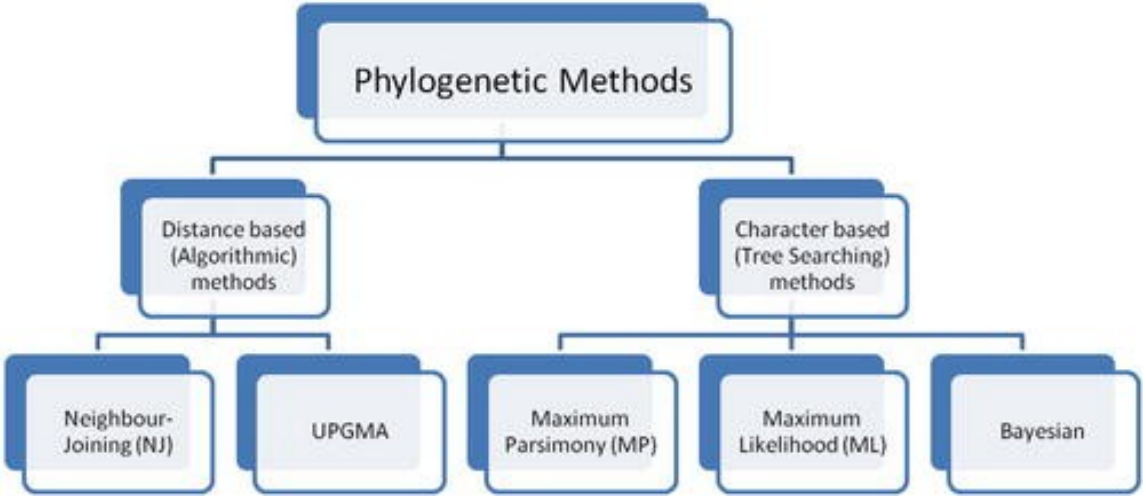
- Original definition
  - The application of phylogenetic methods for gene function analysis (Eisen, 1996)
  - Organism evolution based on whole genome analyses
- Recent usage: Various types of studies mixing genomics and phylogenetics, such as:
  - Global patterns of synteny (conserved gene order) across species
  - Global patterns of gene presence and absence studies across species
  - Genome rearrangments analyses
  - DNA substitution patterns seen in noncoding regions analyses
  - Genomic epidemiological studies
  - ...
- These analyses can be used to understand metabolism, pathogenicity, physiology, and behavior, speciation...

Reference: (Eisen and Fraser, 2003)

Some basics about phylogenetic tree reconstruction methods

3 main methods:

- Neighbor-Joining (distance matrix)
- Parsimony (presence/absence patterns)
- Maximum likehood method (alignment)



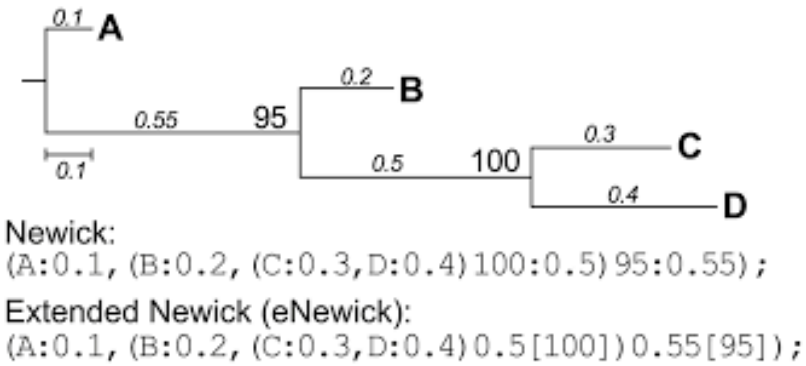
Phylogenetics main methods

Reference: (Sleator, 2015)

The tree Newick format

Newick is a text-based format for representing trees in computer-readable form using (nested) parentheses and commas

- The tree ends with a semicolon
- Interior nodes are represented by a pair of matched parentheses, separated by commas
- Branch lengths are incorporated by putting a real number after a node and preceded by a colon



Phylogenetics main methods

Reference: (Stephens, Bhattacharya, Ragan, and Chan, 2016)

FastTree: Approximately Maximum-Likelihood Trees for Large Alignments

FastTree 2 allows the inference of maximum-likelihood phylogenies for huge alignments

- Can deal with core-gene or core-genome alignments
- Can deal with hundred of thousands of sequences
- Relies on robust Maximum-Likehood statistical models
- Compute local support values with the Shimodaira-Hasegawa test to estimate the reliability of each split in the tree

FastTree in practice:

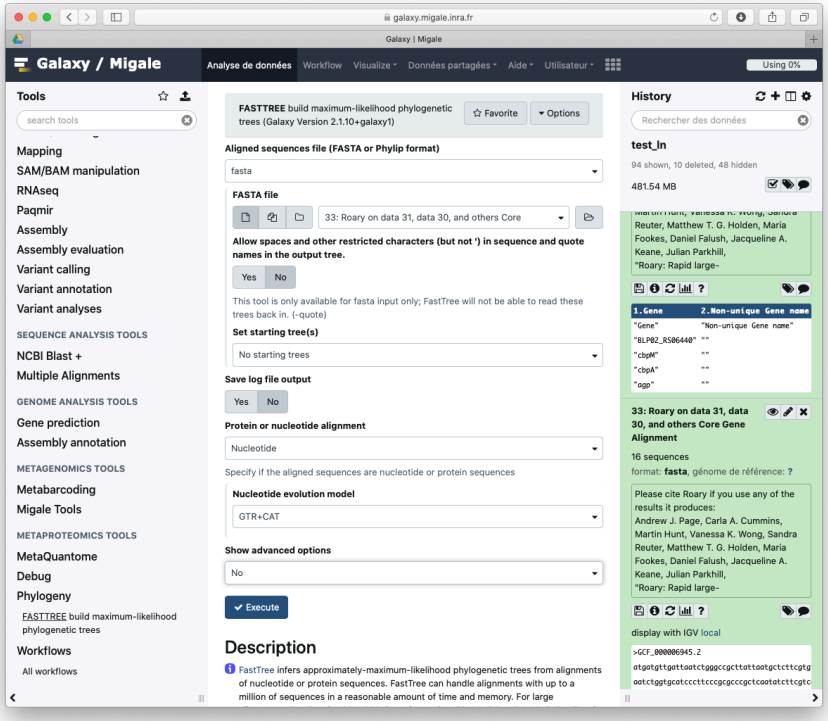
- takes as input an alignment file (Fasta or Phylip interleaved format)
- needs an evolution model: JTT or WAG or LG for protein, JC or GTR for nucleotide
- produces a tree in Newick format with SH support values [0-1] given as names for the internal nodes

http://www.microbesonline.org/fasttree/

FastTree: practice

Use the Tool : 🔧 `FASTTREE` to build a Maximum likehood tree on the aligned core-genes

- Input dataset:
  - the Panggolin multiple fasta alignment of the persistent genome - 📂 `PPanGGOLin msa ...` file
- Paramaters:
  - All the output files selected
  - Set the **Protein or nucleotide alignment** option to `nucleotide`
  - Set the **Nucleotide evolution model** option to `GTR+CAT`



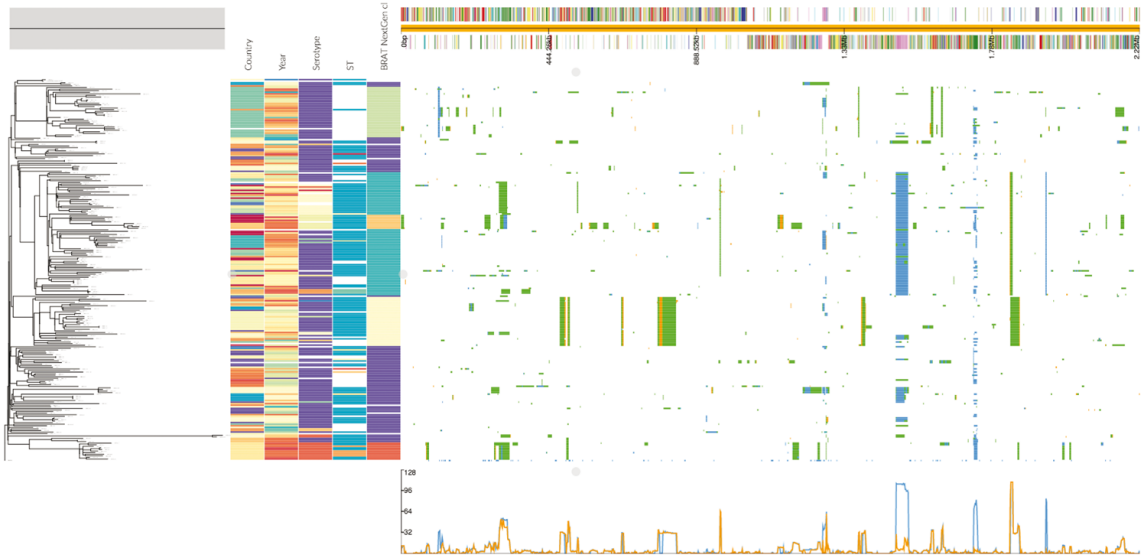
Falaxy-Fasttree

How can I add metadata to my tree and view results ?

The Phandango viewer

Phandango: an interactive viewer for bacterial population genomics

- run directly in a web browser (drag files to upload data)
- many possible inputs like: a phylogenetic tree (Newick format), pangenome data (from Roary for instance), genome annotations (GFF3 format) or any metadata (in simple CSV format)
- a valuable ressource for results interpretation



Phandango

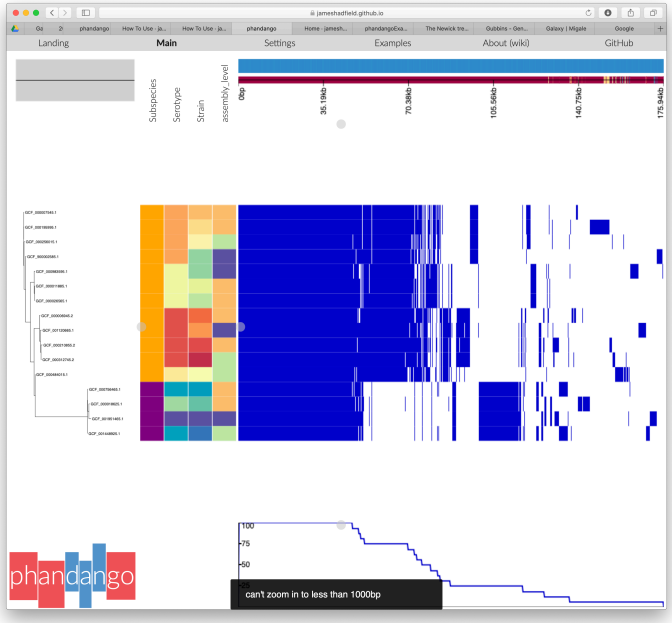
https://jameshadfield.github.io/phandango/#/

Phandango: practice

Open https://jameshadfield.github.io/phandango/#/ in a web browse of your local computer

Upload 3 datafiles just by draging them:

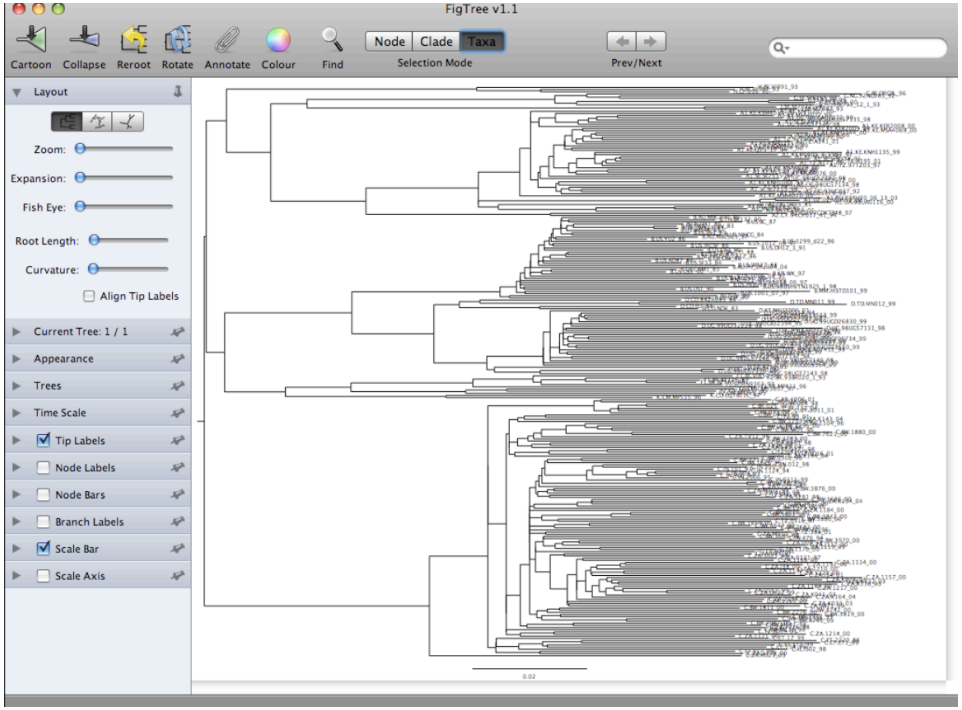
- the **Panggolin** gene presence-absence file: `matrix.csv`
- the **FastTree** phylogenetic tree (change the extension file in *.tree*): `fasttree_persistent_genome.tree`
- The metadata csv file: `Ssal_34G_metadata.csv` Look at results



Phandango results on the Salmonella dataset

Tree visualization with FigTree

- FigTree (https://tree.bio.ed.ac.uk/software/figtree/
- A graphical viewer of phylogenetic trees useful for producing publication-ready figures
- Compiled binaries for Mac, Windows and Linux
- A good tutorial here (https://beast.community/workshop_figtree)



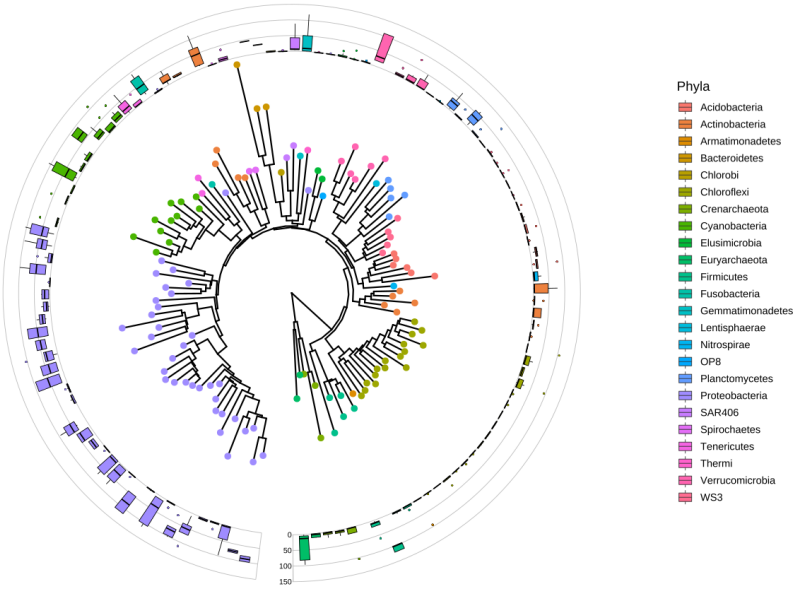
Tree visualization with iTol

> iTol (interactive tree of life) Letunic 2024

Tree visualization with ggtree

ggtree (https://doi.org/10.1002/cpbi.96)

- an R package that extends **ggplot2** for visualizating, manipulating and annotating phylogenetic trees
- available from **Bioconductor**
- very powerful but need **R** and **ggplot** expertise



Example of ggtree plot

Take home message

- Dataset construction, quality and diversity evaluation is a **mandatory** first-step and may be time-consuming
- Dataset dereplication may be helpful for some well-studied organisms
- Comparative strategy depends on the addressed question and on the genome diversity level
- Genome comparison is still an ongoing active bioinformatics research field, recent tools often produce better results
- Phylogenomics approaches are powerful and promising

THANK

YOU

References

Collins, R. E. and P. G. Higgs (2012). "Testing the Infinitely Many Genes Model for the Evolution of the Bacterial Core Genome and Pangenome". In: *Molecular Biology and Evolution* 29.11, pp. 3413-3425. ISSN: 0737-4038. DOI: 10.1093/molbev/mss163. eprint: https://academic.oup.com/mbe/article-pdf/29/11/3413/13648372/mss163.pdf. URL: https://doi.org/10.1093/molbev/mss163.

Eisen, J. A. and C. M. Fraser (2003). "Phylogenomics: Intersection of Evolution and Genomics". In: *Science* 300.5626, pp. 1706-1707. ISSN: 0036-8075. DOI: 10.1126/science.1086292. eprint: https://science.sciencemag.org/content/300/5626/1706.full.pdf. URL: https://science.sciencemag.org/content/300/5626/1706.

Gurevich, A., V. Saveliev, N. Vyahhi, et al. (2013). "QUAST: quality assessment tool for genome assemblies". In: *Bioinformatics* 29.8, pp. 1072-1075. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt086. eprint: https://academic.oup.com/bioinformatics/article-pdf/29/8/1072/17106244/btt086.pdf. URL: https://doi.org/10.1093/bioinformatics/btt086.

Hadfield, J., N. J. Croucher, R. J. Goater, et al. (2017). "Phandango: an interactive viewer for bacterial population genomics". In: *Bioinformatics* 34.2, pp. 292-293. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx610. URL: https://doi.org/10.1093/bioinformatics/btx610.

Konstantinidis, K. T. and J. M. Tiedje (2005). "Genomic insights that advance the species definition for prokaryotes". In: *Proceedings of the National Academy of Sciences* 102.7, pp. 2567-2572. ISSN: 0027-8424. DOI: 10.1073/pnas.0409727102. eprint: https://www.pnas.org/content/102/7/2567.full.pdf. URL: https://www.pnas.org/content/102/7/2567.

References(2)

Konstantinidis, K. T. and J. M. Tiedje (2005). "Genomic insights that advance the species definition for prokaryotes". In: *Proceedings of the National Academy of Sciences* 102.7, pp. 2567-2572. ISSN: 0027-8424. DOI: 10.1073/pnas.0409727102. eprint: https://www.pnas.org/content/102/7/2567.full.pdf. URL: https://www.pnas.org/content/102/7/2567.

Koonin, E. and Y. Wolf (2008). "Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world". In: *Nucleic Acids Res* 36(21), pp. 6688-6719. DOI: 10.1093/nar/gkn668.

Medini, D., C. Donati, H. Tettelin, et al. (2005). "The microbial pan-genome". In: *Current Opinion in Genetics & Development* 15.6. Genomes and evolution, pp. 589

- 1. DOI: https://doi.org/10.1016/j.gde.2005.09.006. URL: http://www.sciencedirect.com/science/article/pii/S0959437X05001759.

Olm, M. R., C. T. Brown, B. Brooks, et al. (2017). "dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication". In: *The ISME Journal* 11.12, pp. 2864-2868. DOI: 10.1038/ismej.2017.126. URL: https://doi.org/10.1038/ismej.2017.126.

Ondov, B. D., T. J. Treangen, P. Melsted, et al. (2016). "Mash: fast genome and metagenome distance estimation using MinHash". In: *Genome Biology* 17.1, p. 132. DOI: 10.1186/s13059-016-0997-x. URL: https://doi.org/10.1186/s13059-016-0997-x.

References(3)

Parks, D. H., M. Imelfort, C. T. Skennerton, et al. (2015). "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes". In: *Genome Research* 25.7, pp. 1043-1055. DOI: 10.1101/gr.186072.114. URL: https://doi.org/10.1101/gr.186072.114.

Sleator, R. (2015). "Phylogenetics, Overview". In: *Encyclopedia of Metagenomics: Genes, Genomes and Metagenomes: Basics, Methods, Databases and Tools*. Ed. by K. E. Nelson. Boston, MA: Springer US, pp. 577-582. ISBN: 978-1-4899-7478-5. DOI: 10.1007/978-1-4899-7478-5_708. URL: https://doi.org/10.1007/978-1-4899-7478-5_708.

Stephens, T. G., D. Bhattacharya, M. A. Ragan, et al. (2016). "PhySortR: a fast, flexible tool for sorting phylogenetic trees in R". In: *PeerJ* 4, p. e2038. ISSN: 2167-8359. DOI: 10.7717/peerj.2038. URL: https://doi.org/10.7717/peerj.2038.

Tettelin, H., V. Masignani, and e. a. Cieslewicz MJ (2005). In: *Proc Natl Acad Sci U S A* 102(39), pp. 13950-13955. DOI: 10.1073/pnas.0506758102.

Tettelin, H., D. Riley, C. Cattuto, et al. (2008). "Comparative genomics: the bacterial pan-genome". In: *Current Opinion in Microbiology* 11.5. Antimicrobials/Genomics, pp. 472 - 477. ISSN: 1369-5274. DOI: https://doi.org/10.1016/j.mib.2008.09.006. URL: http://www.sciencedirect.com/science/article/pii/S1369527408001239.