



Comparaison de génomes microbiens

Cycle de formation à la bioinformatique par la pratique

Hélène Chiapello - Valentin Loux

(helene.chiapello | valentin.loux)@inrae.fr

2022/05/10

Practical informations

- 9h30 - 17h00
- 2 breaks in the morning and in the afternoon
- Lunck break of 1 hour

These supports, by [INRAE-Migale Bioinformatics Facility](#) are licensed under
CC BY-SA 4.0 

A quick round table presentation

- Who are you ?
 - Institution, laboratory, position ...
- Are you (somewhat) familiar with Galaxy ?
- 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 ? Ok or Ko ?

Migale team



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

Objectives

After this training, you will:

- Be able to construct a genomic dataset from public resources 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**, **MAUVE** and **ROARY** under Galaxy or using a graphical interface on the training data set
- Have some keys to interpret results

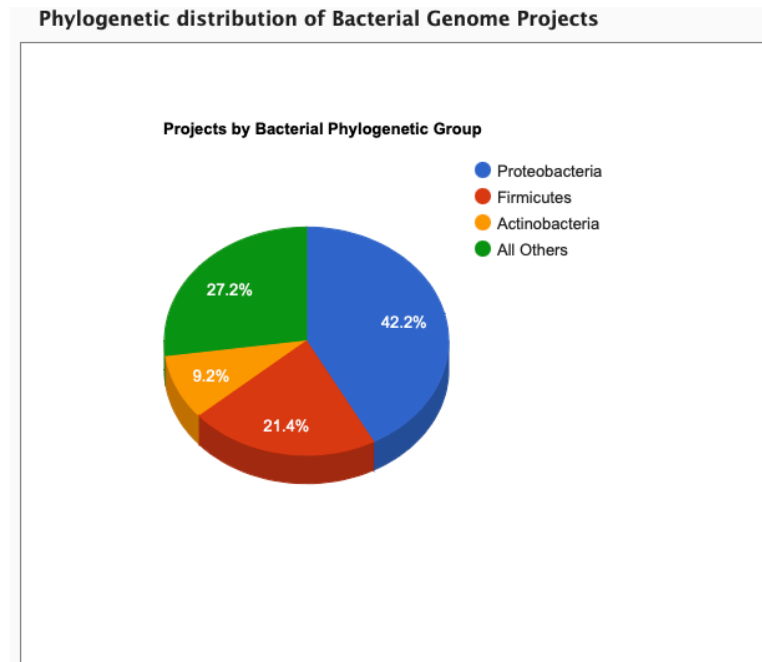
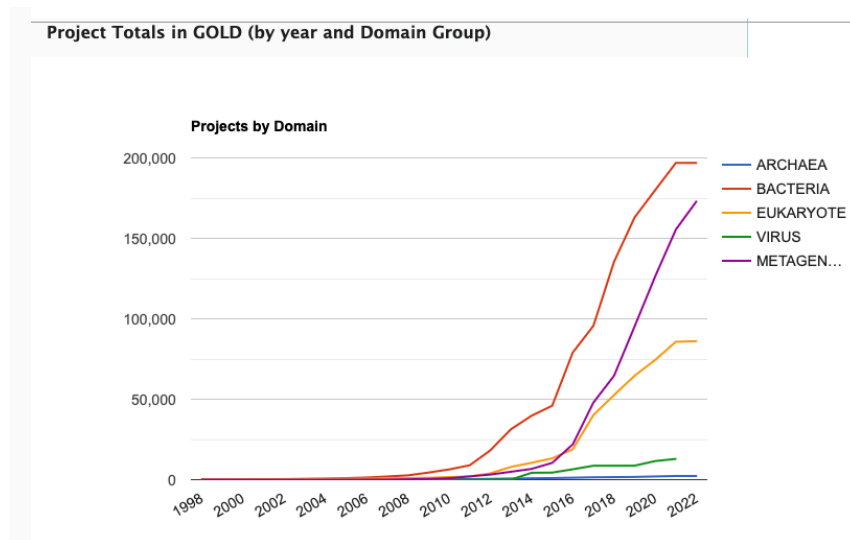
Program

- Morning:
 - Dataset construction
 - Dataset quality evaluation
 - Dataset diversity analysis
 - Genome alignment
- Afternoon:
 - Pan-Genome construction
 - First steps in phylogenomics
 - Data visualization and interpretation

Microbial comparative genomics

A huge number of microbial genomes

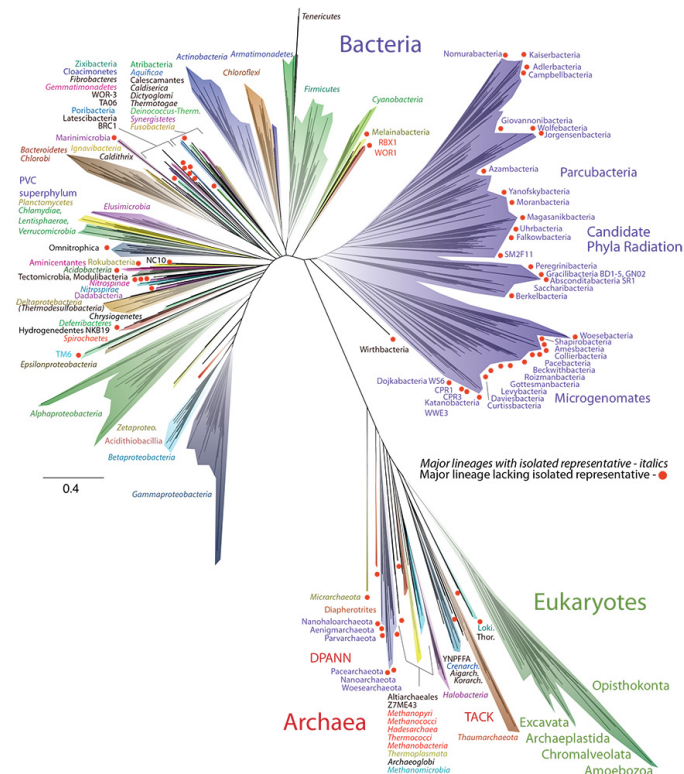
Bacterial and metagenomic genome projects: the top of the sequencing projects



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 were recovered from diverse metagenomic samples
- tree reconstructed from an alignment 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

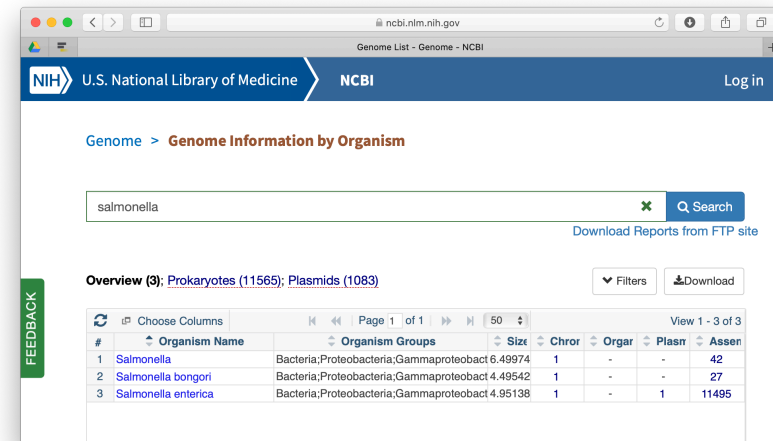
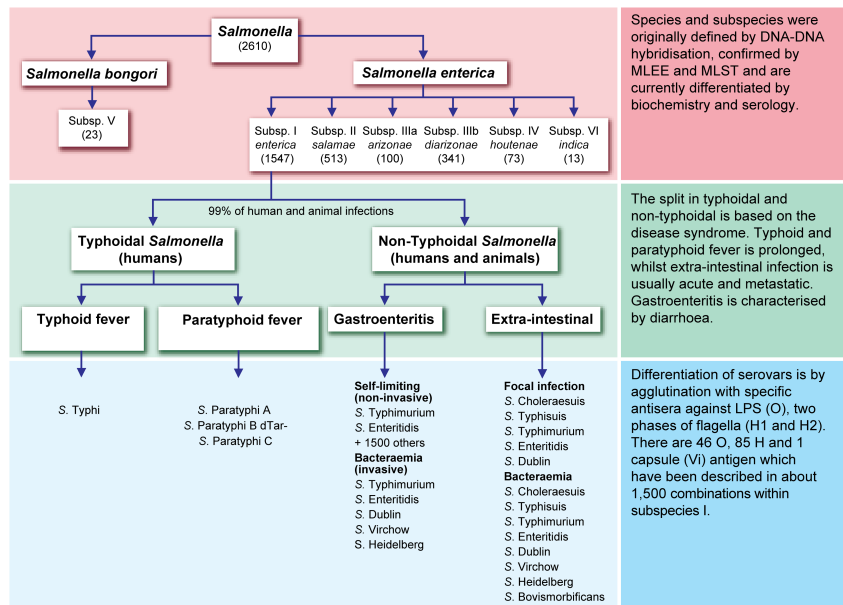
- Heterogenous quality of sequencing and assembly
- Presence of huge number of public genomes OR absence of any close genomes of the same species in public databases
- Difficulties regarding microbial taxonomy (classification) and nomenclature (naming of genus, species and strain naming) for many non-model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

Why comparative genomics

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

The training dataset

We will work on a reduced dataset of public *Salmonella* genomes



13.327 salmonella enterica public assemblies at NCBI!

The training dataset: a list of 16 salmonella 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

The training dataset: a list of 16 salmonella 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

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 annotated with their 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 banks 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="SSIIYNGISTSGLDLNNGTIADMRQLGIVESYKLGKRAVVSSASEA
```

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

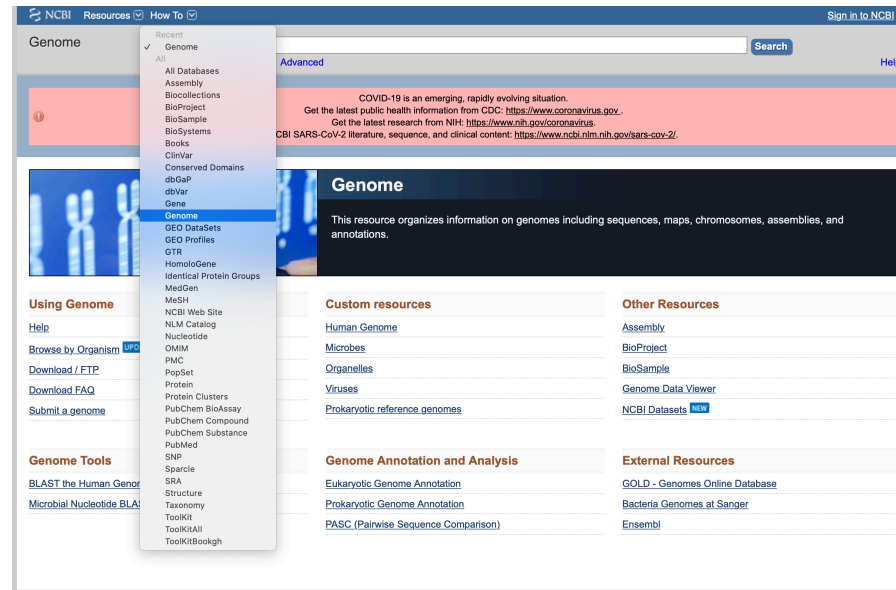
1 How to gather a list of public genomes of interest ?

- Work from the [prokaryotic public genomes available at NCBI](#)
- Use the interface to filter, then download this table
- From this list of **accession** you will have to download a list of files.

Demonstration : download genbank and nct fasta file from NCBI

Practical : Public genomes - NCBI web site

- Go to the NCBI web site
- <https://www.ncbi.nlm.nih.gov/>
- browse to the "Genomes" section



NCBI web site

Practical : Public genomes list

- You will obtain a list of *complete* genomes with different informations :
 - accession (unique id) number
 - species
 - strain
 - completeness
 - a link to download the genome file(s) (Refseq or Deposited)**

Overview (1) Prokaryotes (11648) | Strains (11055)

Page 1 of 233

#	Organism Name	Organism Group	Strain	Bioproject	Assembly	Size	GC%	RefSeqs	WGI	Scatter	COG	Release Dt.	FTP
1	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.L12	SAMN020470	PRJNA241	GCA_00009940.2	4,91	52.22	Chromosome: NZ_CP014051.1 CP014051.1 Plasmid pSL1: NC_025777.2 AB294112.2	2	4048	28-Oct-2001	IR G
2	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.L12	SAMN020644	PRJNA21221	GCA_00155830.2	4,91	52.22	Chromosome: NZ_CP014051.1 CP014051.1 Plasmid pSL1: NC_025777.2 AB294112.2	2	4056	16-Feb-2016	IR G
3	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.TDARR05_748	SAMN115540	PRJNA21221	GCA_00383326.1	4,91	52.20	Chromosome: NZ_CP014051.1 CP014051.1 Plasmid pSL1: NC_025777.2 AB294112.2	3	4059	19-Jun-2016	IR G
4	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.TDCC_13011	SAMN020487	PRJNA35847	GCA_00014306.1	4,83	52.11	Chromosome: NZ_CP014051.1 CP014051.1 Plasmid pSL1: NC_025777.2 AB294112.2	2	4469	29-Aug-2014	IR G
5	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.Ty2	SAMN020496	PRJNA3171	GCA_00007940.1	4,79	52.10	Chromosome: NZ_CP014051.1 CP014051.1	1	4341	21-Mar-2003	IR G
6	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.S42021406	SAMN020487	PRJNA34244	GCA_00323216.1	5,70	51.27	Chromosome: NZ_CP014051.1 CP014051.1 Plasmid pSL1: NC_025777.2 AB294112.2	4	5204	17-Jun-2016	IR G
7	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.S420212091	SAMN0207544	PRJNA34244	GCA_00323470.1	5,48	51.20	Chromosome: NZ_CP014051.1 CP014051.1 Plasmid pSL1: NC_025777.2 AB294112.2	2	4917	17-Jun-2016	IR G
8	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.NCTC11081	SAMEA33044	PRJEB3463	GCA_00474105.1	5,22	51.20	Chromosome: NZ_CP014051.1 CP014051.1	1	4660	17-Jun-2016	IR G
9	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.S420192001	SAMN020496	PRJNA34244	GCA_00383326.1	5,20	51.20	Chromosome: NZ_CP014051.1 CP014051.1	1	4564	17-Jun-2016	IR G
10	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.AIR_8487	SAMN1204897	PRJNA48271	GCA_00078706.1	5,33	51.88	Chromosome: NZ_CP014051.1 CP014051.1 Plasmid pSL1: NC_025777.2 AB294112.2	4	5040	29-Sep-2016	IR G
11	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.TDARR05_739	SAMN115540	PRJNA21221	GCA_00078706.1	5,14	52.00	Chromosome: NZ_CP014051.1 CP014051.1	1	4702	05-Oct-2016	IR G
12	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.S1118T01542413	SAMEA130441	PRJEB1287	GCA_001449106.1	5,23	52.14	Chromosome: NZ_CP014051.1 CP014051.1 Plasmid pSL1: NC_025777.2 AB294112.2	2	4827	29-Oct-2015	IR G
13	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.T17_008	SAMN020644	PRJNA42316	GCA_00355175.1	5,15	52.11	Chromosome: NZ_CP014051.1 CP014051.1 Plasmid pSL1: NC_025777.2 AB294112.2	2	4277	04-Feb-2016	IR G
14	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.GTA-FD-2016-M-02553.2	SAMN1001887	PRJNA41786	GCA_00479805.1	5,15	52.12	Chromosome: NZ_CP014051.1 CP014051.1 Plasmid pSL1: NC_025777.2 AB294112.2	4	4779	19-Apr-2016	IR G
15	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.GTA-FD-2016-M-02553.3	SAMN1001888	PRJNA41786	GCA_00479805.1	5,15	52.12	Chromosome: NZ_CP014051.1 CP014051.1 Plasmid pSL1: NC_025777.2 AB294112.2	4	4778	17-Apr-2016	IR G
16	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.GTA-FD-2016-M-02553.3	SAMN1001888	PRJNA41786	GCA_00479805.1	5,15	52.12	Chromosome: NZ_CP014051.1 CP014051.1 Plasmid pSL1: NC_025777.2 AB294112.2	4	4777	15-Apr-2016	IR G
17	<i>Stenotrophomonas maltophilia</i> strain ATCC 13637	Bacteria	Proteobacteria.Gam.T1701939P	SAMN1001888	PRJNA41786	GCA_00479805.1	5,15	51.91	Chromosome: NZ_CP014051.1 CP014051.1	1	4676	17-Jun-2016	IR G

NCBI web site public genome list

<https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/>

Practical : Public genomes - filter and download

- The list can be
 - filtered with the *filter* button
 - downloaded (csv file) with the "download" button

Overview (1) Prokaryotes (1646) (Phaemids (106))

Filters

Kingdom Bacteria (1,646)

Group Proteobacteria (1,646)

Subgroup Gammaproteobacteria (1,646)

Assembly level Chromosome (198) Complete (656) Contig (7,371) Scaffold (5,148)

Partial All (1,646) Exclude partial (1,645) Include partial only (1)

Annotated All (1,646) Exclude unannotated (11,617) Include unannotated only (25)

Host Human (1,646)

RefSeq category reference (1)

Organism name

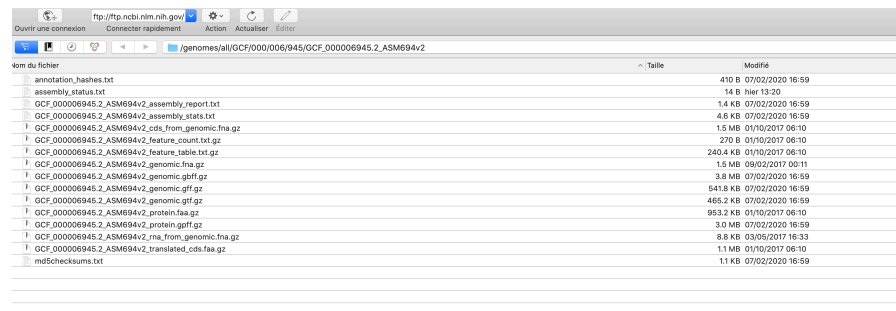
Download

#	Organism Name	Organism Group	Strain	BioSample	BioProject	Assembly	Level	Size	GC%	Replicons	WGS	Scaffold	CDS	Release Date	FTP
1	<i>Salmonella enterica subsp. enterica</i> serovar LT2 Typhimurium str. LT2	Bacteria:Proteobacteria:Gam LT2	SARR0304315	PRJNA411	GCA_00006646.2	●	4.95	52.20	chromosome: NZ_CP003972.1:AB004649.2 plasmid pSLT2: NZ_CP003972.1:AB004649.1.2	2	4548	26-Oct-2001	R	G	
2	<i>Salmonella enterica</i>	Bacteria:Proteobacteria:Gam LT2	SARR0399249	PRJNA431201	GCA_001566203.2	●	4.95	52.20	chromosome: NZ_CP041005.1:CP041005.1 plasmid unannotated: NZ_CP041005.1:CP041005.1	2	4596	11-Feb-2016	R	G	
3	<i>Salmonella enterica</i>	Bacteria:Proteobacteria:Gam FDMARGOS_708	SARR11059483	PRJNA431201	GCA_003630205.1	●	4.95	52.20	chromosome: NZ_CP041005.1:CP041005.1 plasmid unannotated: NZ_CP041005.1:CP041005.1	2	4619	16-Jun-2019	R	G	
4	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Typhimurium	Bacteria:Proteobacteria:Gam ATCC 13311	SARR0294517	PRJNA356476	GCA_000743005.1	●	4.93	52.11	chromosome: NZ_CP009103.1:CP009103.1 plasmid pSLT1: NZ_CP009103.1:CP009103.1	2	4469	29-Aug-2014	R	G	
5	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Typhimurium	Bacteria:Proteobacteria:Gam Tyf	SARR0294506	PRJNA371	GCA_00007546.1	●	4.79	52.10	chromosome: NC_004831.1:ME14813.1	1	4341	21-Mar-2003	R	G	
6	<i>Salmonella enterica</i>	Bacteria:Proteobacteria:Gam SA20021456	SARR0064497	PRJNA342444	GCA_003329215.1	●	5.70	51.27	chromosome: NZ_CP003218.1:CP003218.1 plasmid pSA20021456.1: NZ_CP003218.1:CP003218.1 plasmid pSA20021456.2: NZ_CP003218.1:CP003218.1 Show all 4 replicons	4	5324	17-Jun-2018	R	G	
7	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar 4812	Bacteria:Proteobacteria:Gam SA20121591	SARR0037364	PRJNA342444	GCA_003324765.1	●	5.48	51.20	chromosome: NZ_CP033948.1:CP033948.1 plasmid pSA20121591.1: NZ_CP033948.1:CP033948.1	2	4917	17-Jun-2018	R	G	
8	<i>Salmonella enterica</i> subsp. <i>enterica</i>	Bacteria:Proteobacteria:Gam NCTC10281	SARR0428944	PRJEB3643	GCA_000476185.1	●	5.22	51.20	chromosome: 1_NZ_LS48474.1:LS48474.1	1	4692	17-Jun-2018	R	G	
9	<i>Salmonella enterica</i>	Bacteria:Proteobacteria:Gam SA20102001	SARR0045182	PRJNA342444	GCA_003325005.1	●	5.20	51.20	chromosome: NZ_CP003180.1:CP003180.1	1	4544	17-Jun-2018	R	G	
10	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Concord	Bacteria:Proteobacteria:Gam AB-0407	SARR11044907	PRJNA69271	GCA_008727503.1	●	5.33	51.88	chromosome: NZ_CP041478.1:CP041478.1 plasmid pAB-0407-1: NZ_CP041478.1:CP041478.1 plasmid pAB-0407-2: NZ_CP041478.1:CP041478.1 Show all 4 replicons	4	8040	28-Sep-2019	R	G	
11	<i>Salmonella enterica</i>	Bacteria:Proteobacteria:Gam FDMARGOS_708	SARR11059483	PRJNA431201	GCA_003739915.1	●	5.14	52.00	chromosome: NZ_CP041005.1:CP041005.1	1	4792	16-Jun-2019	R	G	
12	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Weltevreden	Bacteria:Proteobacteria:Gam 25118T01Y942413	SARR11044901	PRJEB1397	GCA_001409185.1	●	5.23	52.14	chromosome: 1_NZ_LM89520.1:LM89520.1 plasmid p11-001: NZ_LM89520.1:LM89520.1	2	4827	20-Oct-2019	R	G	
13	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Senftenberg	Bacteria:Proteobacteria:Gam N17-009	SARR0368946	PRJNA421162	GCA_002953175.1	●	5.15	52.11	chromosome: NZ_CP003076.1:CP003076.1 plasmid pN17-009: NZ_CP003076.1:CP003076.1	2	4577	04-Feb-2016	R	G	
14	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Senftenberg	Bacteria:Proteobacteria:Gam GTA-FD-2016-MI-02533-2	SARR11051887	PRJNA417862	GCA_004176805.1	●	5.15	52.12	chromosome: NZ_CP003800.1:CP003800.1 plasmid pGTA-FD-2016-MI-02533-2: NZ_CP003800.1:CP003800.1 plasmid pGTA-FD-2016-MI-02533-2: NZ_CP003800.1:CP003800.1 Show all 4 replicons	4	4779	10-Apr-2019	R	G	
15	<i>Salmonella enterica</i>	Bacteria:Proteobacteria:Gam GTA-FD-2016-MI-02533-1	SARR11051886	PRJNA417862	GCA_004176785.1	●	5.15	52.12	chromosome: NZ_CP003800.1:CP003800.1 plasmid pGTA-FD-2016-MI-02533-1: NZ_CP003800.1:CP003800.1 plasmid pGTA-FD-2016-MI-02533-2: NZ_CP003800.1:CP003800.1 Show all 4 replicons	4	4778	17-Apr-2019	R	G	

NCBI web site public genome list- filter

Practical : Public genomes - Remote Web Site Structure Exploration

- Explore the remote web site.
- Example :
 - accession **GCA_003181115.1_ASM318111v1**
 - **FTP directory :**
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/003/181/115/GCA_003181115.1_ASM318111v1
- Different file format, including :
 - *accession_genomic_gbff.gz* : compressed **Genbank file**
 - *accession_genomic_fna.gz* : compressed **genomic Fasta file**
 - Full description : <ftp://ftp.ncbi.nlm.nih.gov/genomes/all/README.txt>



nom du fichier	Taille	Modifié
annotation_hashes.txt	410 B	07/02/2020 16:59
assembly_status.txt	14 B	hier 13:20
GCF_000006945.2_ASM694v2_assembly_report.txt	14 KB	07/02/2020 16:59
GCF_000006945.2_ASM694v2_assembly_stats.txt	4,6 KB	07/02/2020 16:59
GCF_000006945.2_ASM694v2_cds_from_genomic.fna.gz	1,5 MB	01/10/2017 06:10
GCF_000006945.2_ASM694v2_feature_count.txt.gz	270 B	01/10/2017 06:10
GCF_000006945.2_ASM694v2_feature_table.txt.gz	240,4 KB	01/10/2017 06:10
GCF_000006945.2_ASM694v2_genomic.fna.gz	1,5 MB	09/02/2017 00:11
GCF_000006945.2_ASM694v2_genomic.gbff.gz	3,8 MB	07/02/2020 16:59
GCF_000006945.2_ASM694v2_genomic.gff.gz	641,6 KB	07/02/2020 16:59
GCF_000006945.2_ASM694v2_genomic.gff.gz	452,2 KB	07/02/2020 16:59
GCF_000006945.2_ASM694v2_protein.faa.gz	953,2 KB	01/10/2017 06:10
GCF_000006945.2_ASM694v2_protein.gpff.gz	3,0 MB	07/02/2020 16:59
GCF_000006945.2_ASM694v2_rna_from_genomic.fna.gz	8,6 KB	03/05/2017 16:33
GCF_000006945.2_ASM694v2_translated_cds.faa.gz	1,1 MB	01/10/2017 06:10
md5checksums.txt	1,1 KB	07/02/2020 16:59

How to download a list of genomes files in Galaxy ?

- **Galaxy** can handle list of files to download.
- Needs only a **list of URLs** (http, ftp protocols)
- But, no simple way to have a direct download link to a (Genbank | GFF | Fasta) file.
 - We will have to manipulate the tabular file to reconstruct the URL with a concatenation of
 - FTP site (column **FTP**)
 - accession number (end of URL in column **FTP**)
 - file suffix (ex: *genomic_fna.gz*)

From :

ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/003/181/115/GCA_003181115.1_ASM318111v1

to

ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/003/181/115/GCA_003181115.1_ASM318111v1\\GCA_003181115.1_ASM318111v1_genomic_fna.gz

- **Two ways** of doing this :
 - In your favorite spreadsheet software (Excel, LibreOffice)
 - Directly in **Galaxy** with Rule-based upload.

Practical : Public genomes - Connect to galaxy

Galaxy / Migale

Analyze Data Workflow Visualize Shared Data Admin Help User

Tools ☆ ⬇

search tools ✕

Get Data

BASIC TOOLS

Collection Operations

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Statistics

Fasta manipulation

NGS TOOLS

Quality control

Nanopore

FASTQ cleaning

Mapping

SAM/BAM manipulation

RNAseq

migale

Welcome to the Migale Galaxy instance!

20 %

<https://galaxy.migale.inrae.fr>

Practical : Public genomes list - Upload

- Select upper-left upload button
 - Upload the csv file, convert it to tabular (pen icon)
 - Upload button (again)
- Rule-based tab
- Load tabular from history
- Build

You will then be able to apply a list of **rules and transformation** to this tabular file.

Practical : Public genomes list - Remove first row

Build Rules for Uploading Datasets

Use this form to describe rules for import datasets. At least one column

Rules

One or more column definitions must be specified. These are required to specify how to build collections and datasets from rows and columns of the table. [Click here to manage column definitions.](#)

#Organism Name
Salmonella ente
Salmonella ente
Salmonella ente
Salmonella ente
Salmonella ente
Salmonella ente
Salmonella ente
Salmonella ente

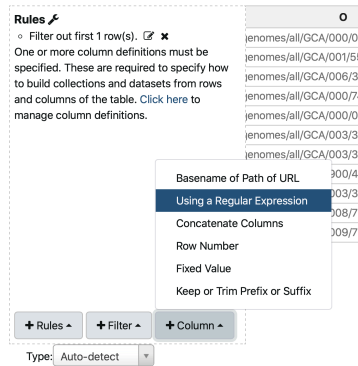
- Using a Regular Expression
- Matching a Supplied Value
- By Comparing to a Numeric Value
- On Emptiness
- First or Last N Rows**

+ Rules - + Filter - + Column -

Type: Auto-detect

Practical : Public genomes list- Extract id(1)

Use a *regular expressions* to extract the id



Practical : Public genomes list- Extract id(2)

Use a *regular expressions* to extract the id :

- Applied a column P
- Create column matching expression groups (between brackets) :
 - ftp://.*/(.*)
 - ".*" means any character
 - This expression means, capture all the character you found after the last /
 - It will create a new column with what they have captured on each line

Use this form to describe rules for import datasets. At least one column should be defined to a source to fetch data from (URLs, FTP files, etc...).






From Column	O	P
<input type="radio"/> Create column matching expression.	genomes GCA 000 006 945 GCA_000006945.2_ASM684v2	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/006/945/GCF_000006945.2_ASM684v2
<input checked="" type="radio"/> Create columns matching expression groups.	genomes GCA 001 1588 355 GCA_001588355.2_ASM158835v2	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/1588/355/GCF_001588355.2_ASM158835v2
<input type="radio"/> Create column from expression replacement.	genomes GCA 006 365 335 GCA_006365335.1_ASM636533v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/006/365/335/GCF_006365335.1_ASM636533v1
	genomes GCA 007 13355 GCA_000713355.1_ASM71335v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/007/13355/GCF_000713355.1_ASM71335v1
	genomes GCA 000 007 545 GCA_000007545.1_ASM754v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/007/545/GCF_000007545.1_ASM754v1
	genomes GCA 003 325 215 GCA_003325215.1_ASM332521v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/325/215/GCF_003325215.1_ASM332521v1
	genomes GCA 003 324 755 GCA_003324755.1_ASM332475v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/324/755/GCF_003324755.1_ASM332475v1
	genomes GCA 900 478 155 GCA_900478155.1_47328_D01	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/900/478/155/GCF_900478155.1_47328_D01
	genomes GCA 003 325 035 GCA_003325035.1_ASM332503v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/325/035/GCF_003325035.1_ASM332503v1
	genomes GCA 008 727 535 GCA_008727535.1_ASM872753v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/008/727/535/GCF_008727535.1_ASM872753v1
	genomes GCA 009 729 915 GCA_009729915.1_ASM972991v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/009/729/915/GCF_009729915.1_ASM972991v1

Regular Expression ?
ftp://.*/(.*)
Number of Groups
1

Cancel Apply

Practical : Public genomes list - Identify Column with ID

- Column Q is now filled with the ID

Rules 		P	Q
<ul style="list-style-type: none"> Filter out first 1 row(s).   Add new column using ftp:\V.*\V(*) applied to column P   <p>One or more column definitions must be specified. These are required to specify how to build collections and datasets from rows and columns of the table. Click here to manage column definitions.</p>	.00006945.2_ASM694v2	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/006/945/GCF_000006945.2_ASM694v2	GCF_000006945.2_ASM694v2
	001558355.2_ASM155835v2	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/558/355/GCF_001558355.2_ASM155835v2	GCF_001558355.2_ASM155835v2
	.006365335.1_ASM636533v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/006/365/335/GCF_006365335.1_ASM636533v1	GCF_006365335.1_ASM636533v1
	000743055.1_ASM74305v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/743/055/GCF_000743055.1_ASM74305v1	GCF_000743055.1_ASM74305v1
	.000007545.1_ASM754v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/007/545/GCF_000007545.1_ASM754v1	GCF_000007545.1_ASM754v1
	003325215.1_ASM332521v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/325/215/GCF_003325215.1_ASM332521v1	GCF_003325215.1_ASM332521v1
	.003324755.1_ASM332475v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/324/755/GCF_003324755.1_ASM332475v1	GCF_003324755.1_ASM332475v1
	900478155.1_47328_D01	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/900/478/155/GCF_900478155.1_47328_D01	GCF_900478155.1_47328_D01
	.003325035.1_ASM332503v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/325/035/GCF_003325035.1_ASM332503v1	GCF_003325035.1_ASM332503v1
	008727535.1_ASM872753v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/008/727/535/GCF_008727535.1_ASM872753v1	GCF_008727535.1_ASM872753v1
	009729915.1_ASM972991v1	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/009/729/915/GCF_009729915.1_ASM972991v1	GCF_009729915.1_ASM972991v1

Practical : Public genomes list - Add a column with fixed value

- Add a column with "/"
- Add a column with "suffix" (ie genomic_fna.gz)

The screenshot shows a 'Rules' panel on the left with the following content:

- Filter out first 1 row(s).
- Add new column using ftp:\V:*V(*) applied to column P

One or more column definitions must be specified. These are required to specify how to build collections and datasets from rows and columns of the table. [Click here](#) to manage column definitions.

On the right, a table displays the following values:

.000006945.2_A5
001558355.2_ASI
.006365335.1_AS
.000743055.1_ASI
.000007545.1_ASI
003325215.1_ASM
003324755.1_ASI

A dropdown menu is open, showing the following options:

- Basename of Path of URL
- Using a Regular Expression
- Concatenate Columns
- Row Number
- Fixed Value**
- Keep or Trim Prefix or Suffix

Practical : Public genomes list - Concatenate columns

- Concatenate column "URL" and "fixed value with /"
- Concatenate preceding column and accession
- Concatenate preceding column and suffix

Build Rules for Uploading Datasets

Use this form to describe rules for import datasets. At least one column must be specified.

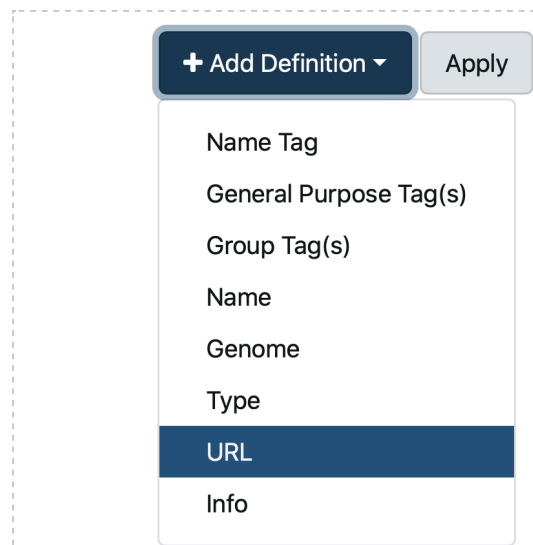
Rules	Q
Filter out first 1 row(s).	GCF_000006945.2_A
Add new column using ftp:\V:(*) applied to column P	GCF_001558355.2_A
Add column for the constant value of /	GCF_006365335.1_A
Add column for the constant value of _genomic.gbff.gz.	GCF_000743055.1_A
Concatenate column P and column R	GCF_000007545.1_A
	GCF_003325215.1_A
	GCF_003324755.1_A
	55.1_47
	35.1_A
	35.1_A
	15.1_A

One or more column definitions specified. These are required to build collections and datasets and columns of the table. Click to manage column definitions.

- Basename of Path of URL
- Using a Regular Expression
- Concatenate Columns
- Row Number
- Fixed Value
- Keep or Trim Prefix or Suffix

Practical : Public genomes list - Define columns

- define the last column (with the URL to the file you have constructed) as an URL
- It will tell Galaxy where to look for the files to download



Practical : Public genomes list - Define columns(2)

- [Optional] define the accession column as a "name"
- It will tell Galaxy where to look for the name to give to the files downloaded (otherwise it gives the URL as the name)

Build Rules for Uploading Datasets

Use this form to describe rules for import datasets. At least one column should be defined for a source to fetch data from (URLs, FTP files, etc.).

Rules	URL
Filter out first 1 row(s) of #	493412
Add new column using the following formula	3475183362
Add column for the constant value of #	3465323241
Add column for the constant value of #	41320161
Add column for the constant value of #	5411
Concatenate column P and column R of #	403231741
Concatenate column T and column R of #	403247511
Concatenate column U and column S of #	401
Set column Q as Name of #	3465323241
Set column V as URL of #	487272311
	48729911

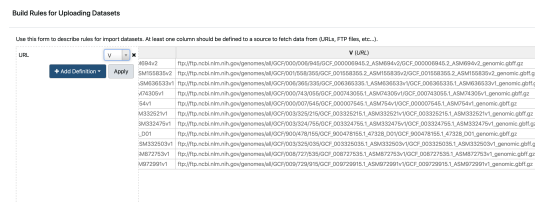
Rules: 10 Filter Columns: 10

Genome: Additional Species Are Below

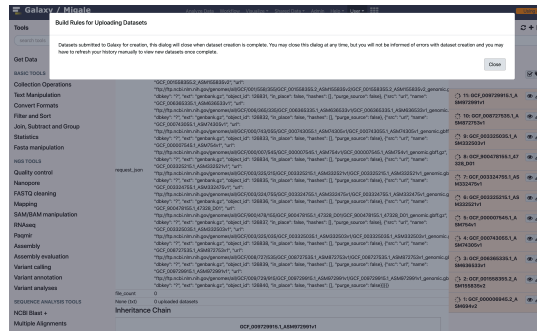
Cancel Reset Upload

Practical : Public genomes list - Upload files form built list

- Check the rules
- Save it (wrench icon) for later
- click on upload



Practical : Public genomes list - Launch Upload



- The tabular genome description file is in "Shared Data/ Data Library/ Formation Génomique Comparée/DataSet/DataSalmonella.tabular"
- The backup of the rules file is in "Shared Data/ Data Library/ Formation Génomique Comparée/ Correction/rule_based_ipload.json".
- Rules should be adapted to your tabular file

Practical : create your dataset in galaxy

- Connect to Galaxy(<https://galaxy.migale.inrae.fr>) with your (or stage) account.
- Do not forget to login (upper right ...)
- Create a new history
- Copy all the genomes fasta & GFF from "Shared Data / Data Libraries/ Formation Génomique Comparée/ Dataset/Fasta" and "Shared Data / Data Libraries/ Formation Génomique Comparée/ Dataset/GFF"

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 ?
 - Length
 - Number of contigs
 - Number of scaffolds
 - GC%
- What is the "quality" of an annotation [compared to what we expect]?
- Number of (pseudo)genes
- number of rRNA genes
- number of tRNA genes

Tools to QC your dataset :

Quast (Quality Assessment Tool for Genome Assemblies, (Gurevich, Saveliev, Vyahhi, et al., 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"

Practical : Quast your dataset !

Apply quast to the 16 assemblies of you dataset.



COFFEE
BREAK

Dataset diversity analysis

Genome diversity evaluation

Why ?

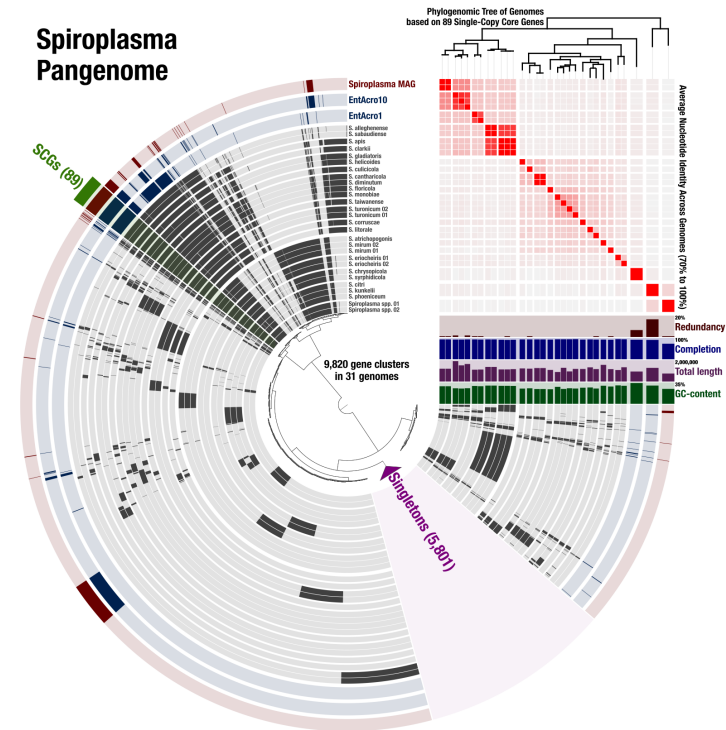
- Build and de-replicate genome datasets
- Estimate genome similarity in a dataset and design an adapted comparative strategy

How ?

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

Average Nucleotide Identity (ANI)

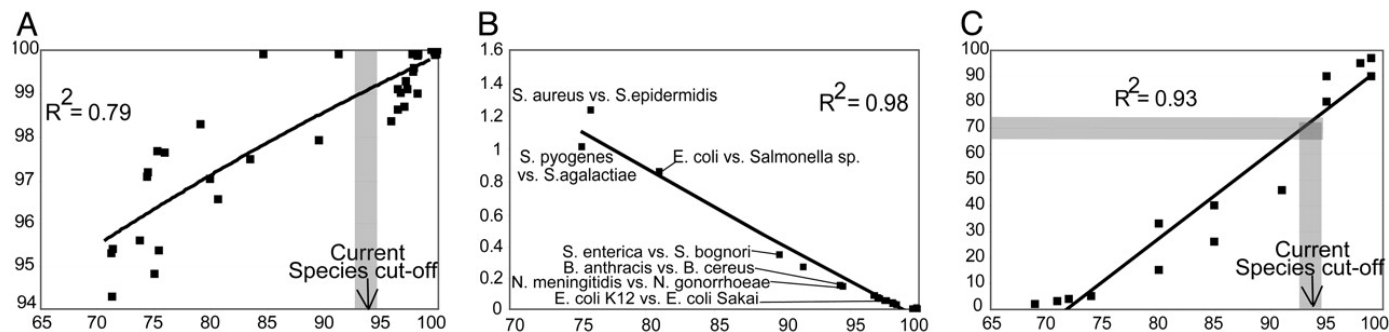
- Meet the need for a robust measure of genomic relatedness and a systematic and scalable species assignment technique
- Mean identity percent of aligned regions of a pair of genomes
- Rely on pairwise alignments that may come either from aligned core genes or from genomic alignments
- Can easily be 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

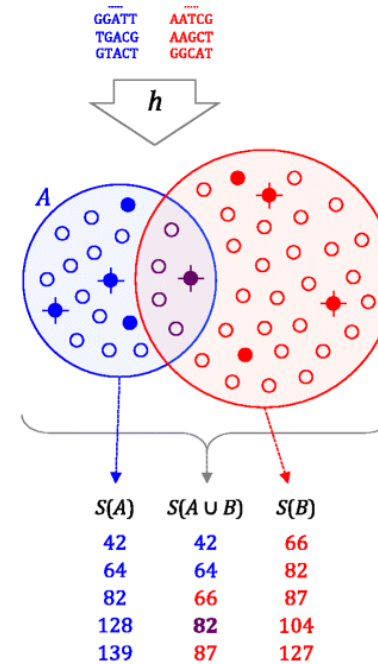
Source : (Konstantinidis and Tiedje, 2005)

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

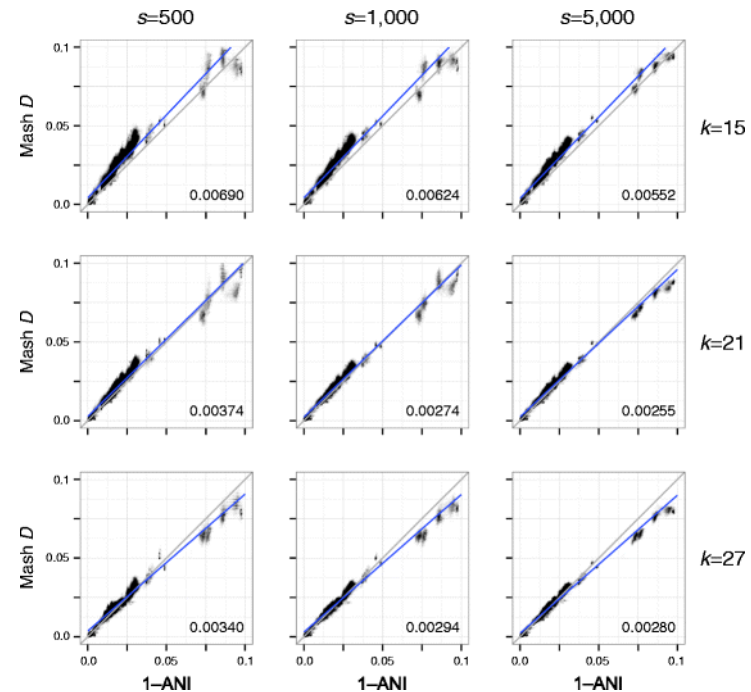


$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

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

MASH distances correlate well with ANI

- Dataset: 500 complete *E. coli* genomes
- Gray lines: model relationship $D = 1 - \text{ANI}$
- 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.

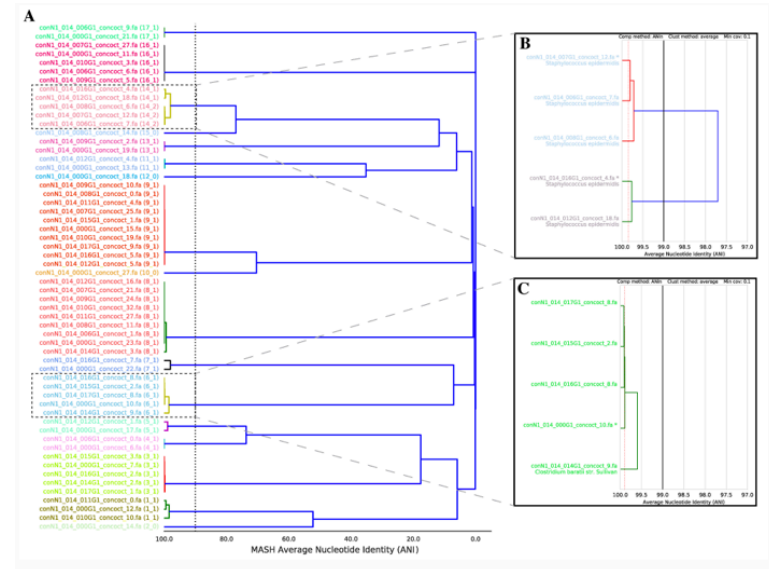
Source : (Ondov, Treangen, Melsted, et al., 2016)

dREP: comparison and de-replication

- 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



Assembly and de-replication with dRep

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

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 completeness**: MASH is very sensitive to genome completeness. 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. <<<<<<< HEAD

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/
.... /ANI_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 ANI comparisons
.... /Sdb.csv # Scoring information
.... /Wdb.csv # Winning genomes
.... /Wfdb.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)

Practice

- use **dREP-duplicate** to explore the Salmonella genome dataset diversity and completeness and dereplicate the dataset
- explore and interpret results
- input : 16 genome fasta files

The screenshot displays the Galaxy web interface for the dRep-duplicate tool. The main panel shows the tool configuration with the following details:

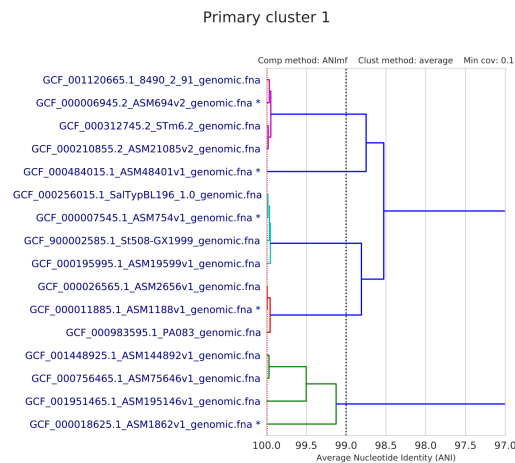
- Tool Name:** dRep-duplicate (De-replicate a list of genomes) (Galaxy Version 2.5.4.0)
- Input:** A list of 16 genome FASTA files, including GCF_900002585.1, GCF_001951465.1, GCF_001448925.1, GCF_001120665.1, GCF_000983595.1, GCF_000756465.1, and GCF_000484015.1.
- Options:** The tool is configured with default settings for filtering, comparison, clustering, and scoring options, all set to "No".
- Output:** The workflow includes a "Select outputs" section with checkboxes for "log", "Warnings", "Primary_clustering_dendrogram.pdf", and "Clustering_scatterplots.pdf".

The right-hand panel shows the workflow history, listing several jobs such as "Nucmer on data 39 and data 43: plot" and "Mummerplot on data 39, data 43, and data 54: plot".

dRep results interpretation

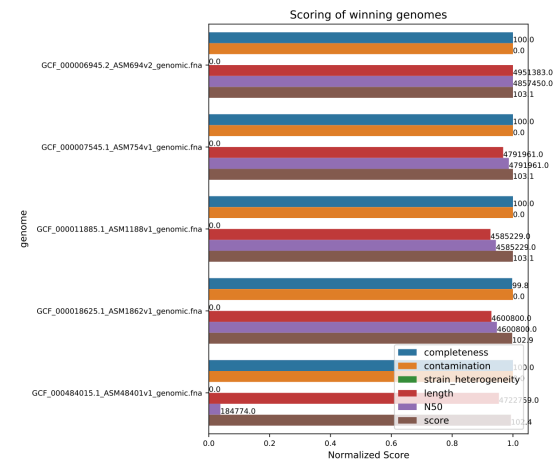
Important outputs of dRep

The "Secondary_clustering_dendrograms.pdf" output file



Secondary_clustering_dendrograms.pdf

the "Winning_genomes.pdf" output file and the deReplicated genomes list

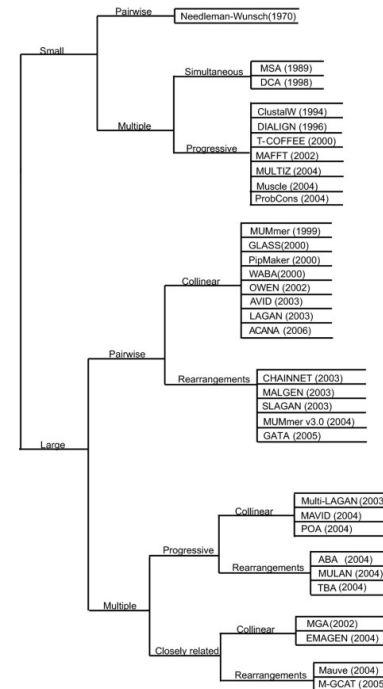


Winning_genomes.pdf

Genome alignment

Genome alignment

- Mostly targeted to **close genome comparisons** (generally at the intra-species level)
 - A variety of applications:
 - help for genome assembly, scaffolding and annotation
 - genome architecture comparison
 - genome micro-evolution analysis
 - discovery of DNA motifs or elements in conserved non-coding regions
 -
- Aligning whole genome sequences is a challenge:
 - computational intensive
 - heterogenous quality of assemblies
 - broad variety of mutational and evolutionary events (including rearrangements)



An approximate phylogeny of genome comparison tools over the past 30 years

Source : (Treangen and Messeguer, 2006)

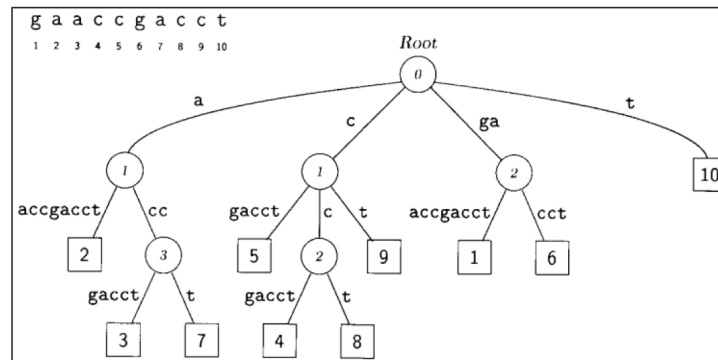
Mummer: pairwise alignment with rearrangements

Based on three main steps:

Step 1: Perform a maximal unique match (MUM) decomposition of the two genomes using suffix trees

```
Genome A: tcgatcGACGATCGCGGCCGTAGATCGAATAACGAGAGAGCATAAcgactta  
Genome B: gcattaGACGATCGCGGCCGTAGATCGAATAACGAGAGAGCATAAtccagag
```

A maximal unique matching subsequence (MUM) of 39 nt (shown in uppercase) shared by Genome A and Genome B

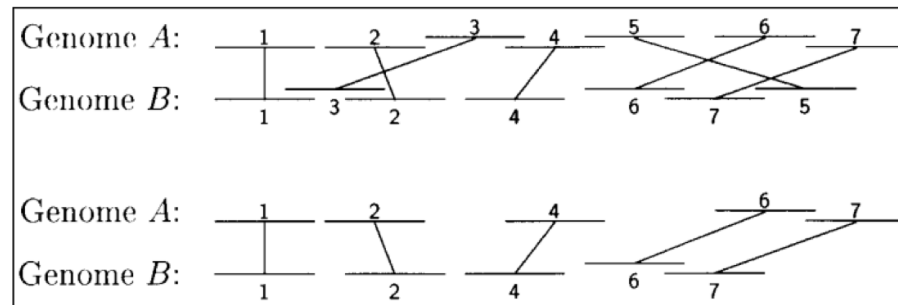


A Suffix tree for the sequence gaaccgacct

Source : (Delcher, Kasif, Fleischmann, et al., 1999)

Mummer: pairwise alignment with rearrangements

Step 2: Sort the matches found in the MUM alignment, and extract the longest possible set of matches that occur in the same order in both genomes



LIS algorithm to find the longest set of MUMs whose sequences occur in ascending order in both Genome A and Genome B

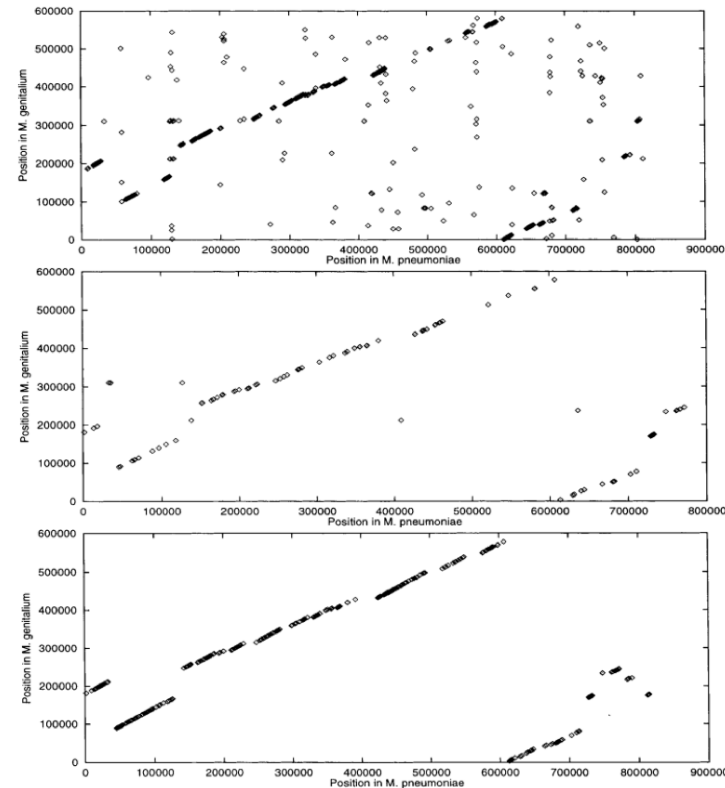
Step 3: Close the gaps (regions between the MUMs) by

- detecting SNPs between MUMs
- identifying large inserts (transpositions or insertions) and repeats (overlapping MUMs)
- aligning small polymorphic regions using a standard dynamic programming algorithm approach

Mummer: pairwise alignment with rearrangements

Example of Nucmer results

- Alignment of M.genitalium (580 074 nt) x M.pneumoniae (816 394 nt)
- The MUM alignment clearly shows five translocations of M.genitalium sequence with respect to M.pneumoniae, in agreement with the analysis of Himmelreich et al. 1997 x Source : (Delcher, Kasif, Fleischmann, et al., 1999)



Alignment of M.genitalium and M.pneumoniae using FASTA (top), 25mers (middle) and MUMs (bottom)

Practice

- Use **Galaxy-Nucmer** to align the two Salmonella typhi CT18 (Refseq accession:GCF_000195995.1) and Ty2 (Refseq accession:GCF_000007545.1) complete genomes
- Look at result files
- What do you conclude accorging their genome structure?
- Generate a list of coordinates of aligned regions using the **Show-Coords** program

Nucmer result interpretation

The Galaxy-nucmer outputs

- The *dotplot* output

Nucmer result interpretation

The Galaxy-nucmer outputs

- The *alignment* output

Nucmer result interpretation

The Galaxy-nucmer outputs

- The *show-coords* output

Mauve: multiple alignment with rearrangements

<http://darlinglab.org/mauve/mauve.html>

- One of the first multiple genome aligner that can deal with rearrangements
- Well suited to bacterial genome alignment
- Success largely due to its Graphical User Interface

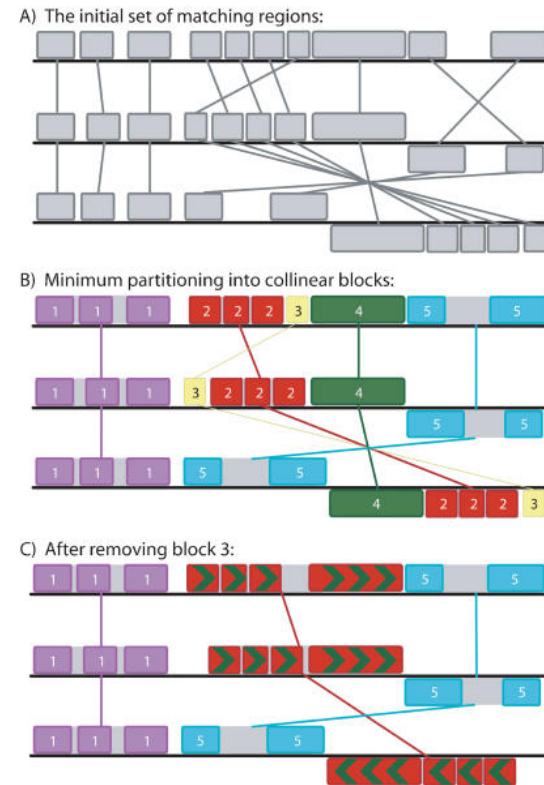
Source : (Darling, Mau, Blattner, et al., 2004)

Mauve: how it works?

Mauve alignment algorithm main steps:

- Find local alignments (multi-MUMs).
- Use the multi-MUMs to calculate a phylogenetic guide tree.
- Select a subset of the multi-MUMs to use as anchors—these anchors are partitioned into collinear groups called LCBs.
- Perform recursive anchoring to identify additional alignment anchors within and outside each LCB.
- Perform a progressive alignment of each LCB using the guide tree.

Source : (Darling, Mau, Blattner, et al., 2004)

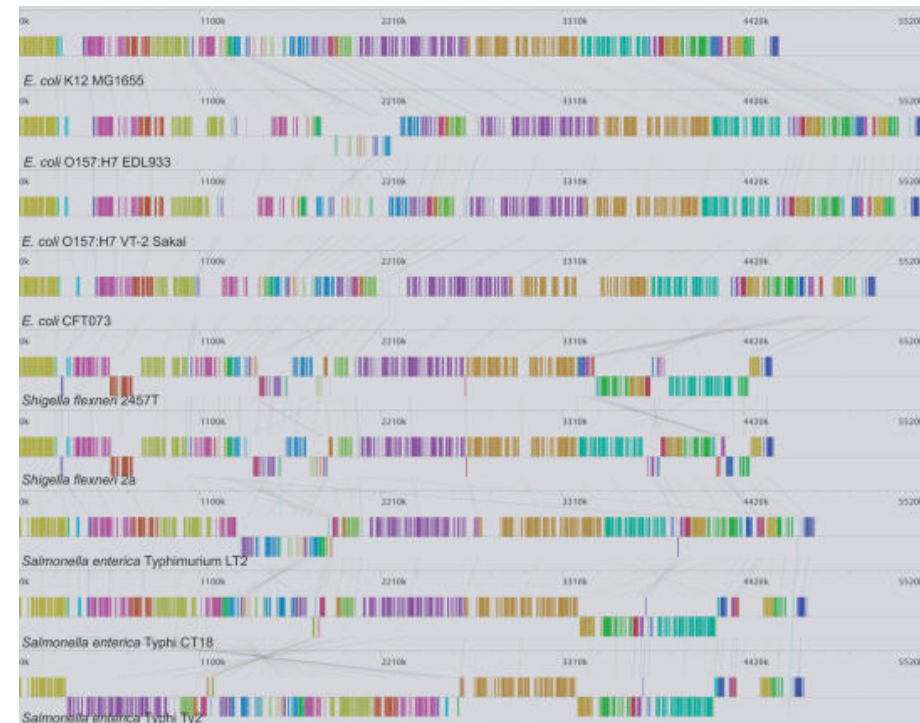


A pictorial representation of greedy breakpoint elimination in three genomes

Mauve: Alignment of Nine Enterobacterial Genomes

Genome alignment features

- Each contiguously colored region is a locally collinear block (LCB)
- LCB can be in reverse complement orientation relatively to reference genome (K12)
- 45 LCB with minimum weight of 69 consisting of 2.86 Mb of conserved backbone sequence broken into 1252 segments
- Several known inversions are confirmed such as the O157:H7 EDL933 inversion relative to K12 and the large inversion about the origin of replication among the *S. enterica* serovars Typhi CT18 and Ty2



Locally collinear blocks identified among the nine enterobacterial genomes

Mauve companion tools

Mauve Contig Mover (Rissman et al. 2009)

- Can order contigs of a draft genome relative to a related reference genome
- Based on iterative genome alignment using Mauve and requires anchors at both ends of contigs
- The reference used may be draft quality itself, or may have divergent genetic content

ProgressiveMauve (Darling & Perna 2010)

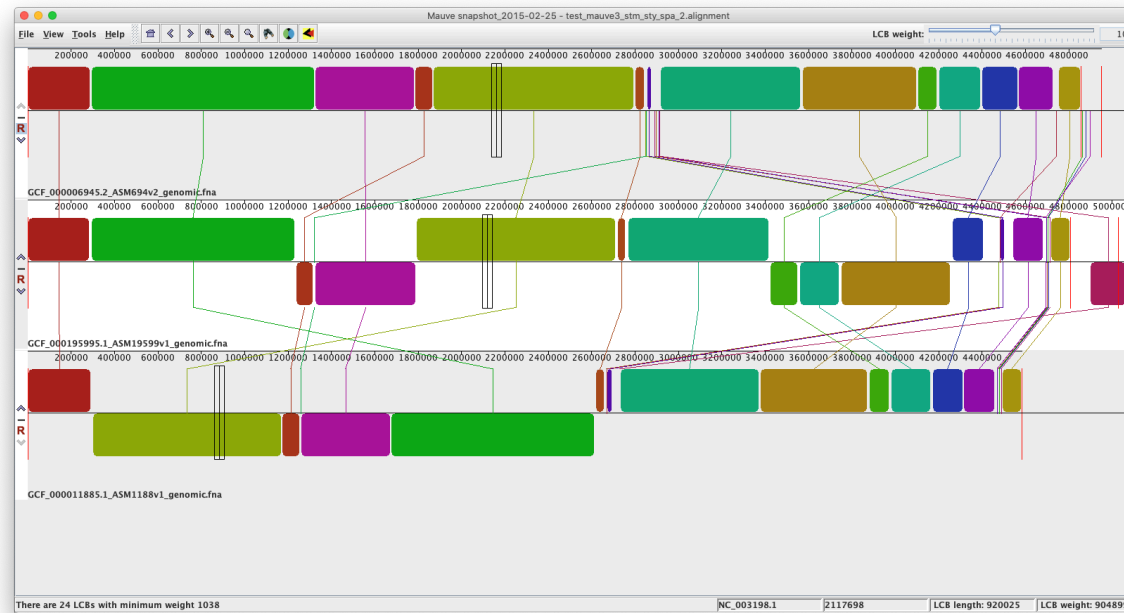
- Can align regions conserved only in subsets of the genomes
- Set up an anchor scoring function that penalizes alignment anchoring in repetitive regions of the genome and penalizes genomic rearrangement
- Use a probabilistic scoring strategy (HMM) to reject erroneous alignments of unrelated sequence produced by Mauve
- In summary: can align faster and more accurately than Mauve more distant and big dataset of genomes

Practice Mauve

- Use *Mauve* **on your local computer** to align the 3 complete genomes of serotypes typhi (CT18, Refseq accession:GCF_000195995.1), typhimurium (LT2, Refseq accession:GCF_000006945.2) and Paratyphi A (ATCC 9150, Refseq accession: GCF_000011885.1)
- Mauve input: fasta (or Genbank) files
- Choose *Mauve* and **not** *ProgressiveMauve* algorithm

Mauve results interpretation

- Genome alignment of serotypes typhi (CT18), typhimurium (LT2) and Paratyphi A
- Look at the LCB output (other output files description : <http://darlinglab.org/mauve/user-guide/files.html>)
- What do you conclude regarding genome structure ?



Mauve on ly local computer

LUNCH

The microbial pan-genome

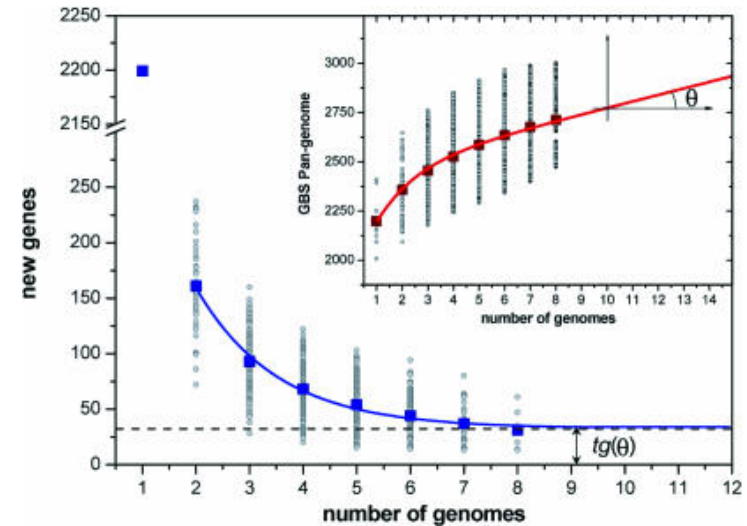
The microbial pan-genome

First term apparition in 2005 in two publications

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

*A bacterial species can be described by its **pan-genome** 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.*

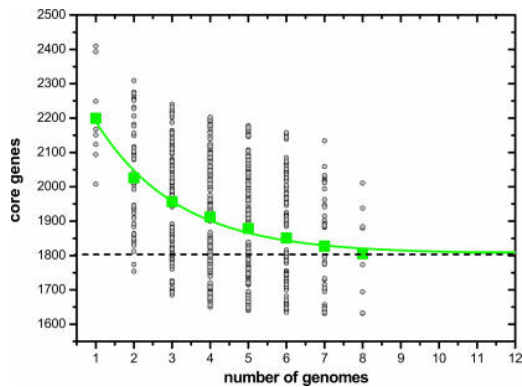
References: (Tettelin, Masignani, and Cieslewicz MJ, 2005) and (Medini, Donati, Tettelin, et al., 2005)



Streptococcus group B pan genome

The microbial pan-genome

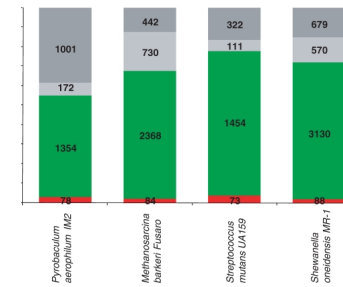
- Definition refinement by Koonin (2008) and Collins (2012): the 3 classes of prokaryotic genes
 - **core (or persitent) genes**: a small fraction of highly conserved genes
 - **shell genes**: a larger set of moderately conserved genes
 - **cloud genes**: (nearly) unique genes



Streptococcus group B core genome

Source : (Koonin and Wolf, 2008)

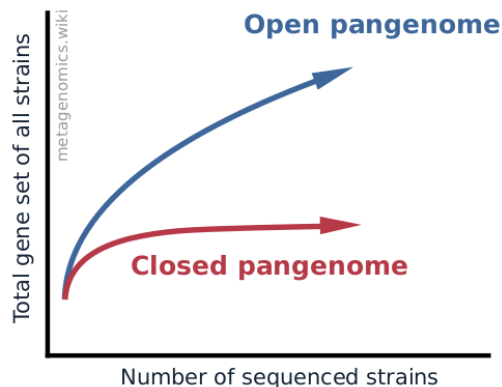
Source : (Collins and Higgs, 2012)



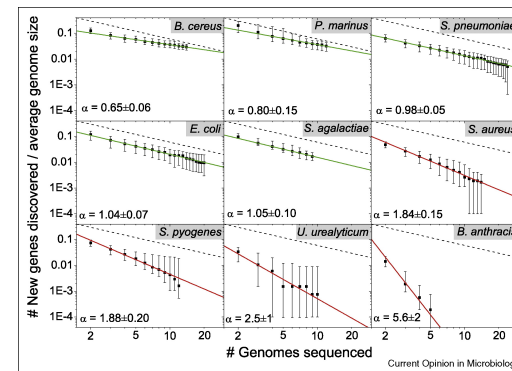
A Common and rare genes in selected archaeal and bacterial genomes. Red, core; green, shell; light gray, cloud; dark gray, ORFans.

Open or closed pan-genome

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



Open and closed pangnomes



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

Source : (Tettelin, Riley, Cattuto, et al., 2008)

Roary: rapid large-scale prokaryote pan genome analysis

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

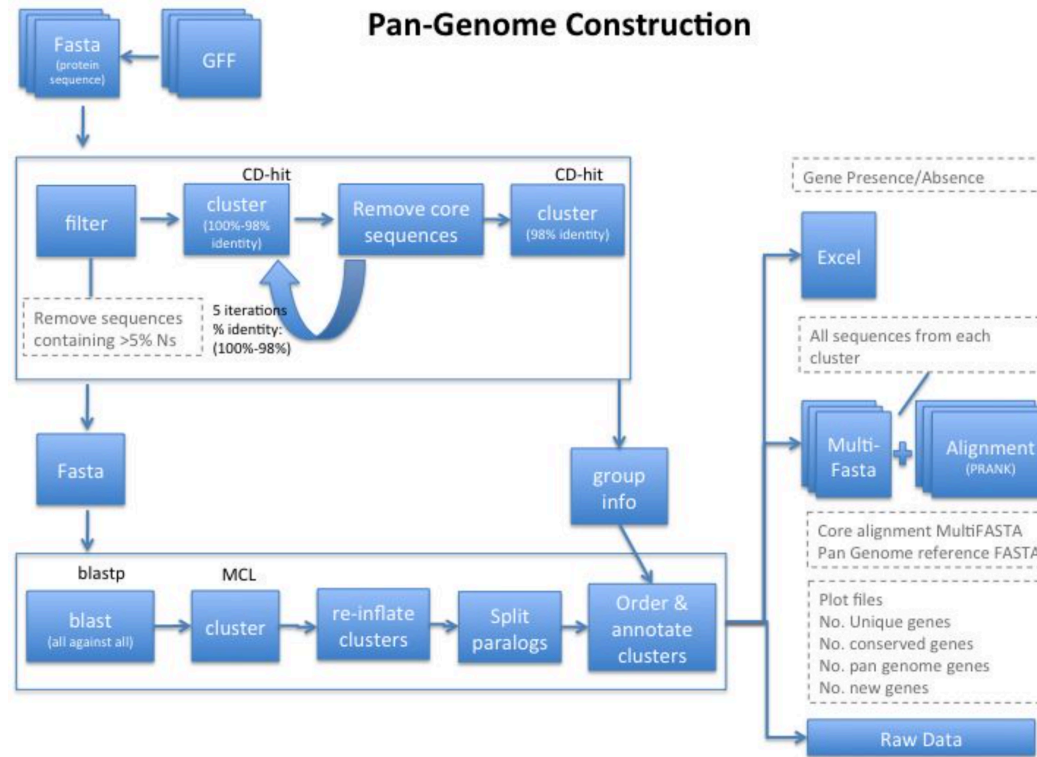
Input :

- annotated genomes in **GFF3** format
 - Roary is *very* sensitive to the validity of the GFF format
 - GFFs generated by **Prokka** are valid
 - Locus tags must be unique across datasets.
 - GFF from NCBI are **invalid** (sequence is missing)
 - Must be converted from Genbank using "Genbank to

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

Roary workflow

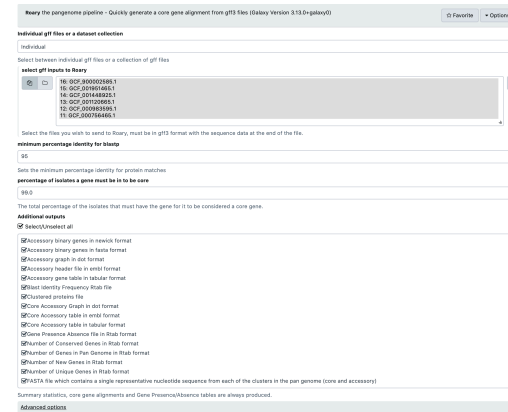


Sub. Fig. 13: A flowchart of the steps in the application.

Practical : Roary your dataset !

Apply roary to the 16 assemblies of you dataset.

- Input :
 - the 16 gff files
- Paramaters :
 - All the output files selected
 - No specific parameter (-split-paralog to "yes")



Roary outputs

- *Summary statistics* about number of gene in the core/pan/accessory genomes
- *Gene Presence Absence* : lists each cluster of gene, the most common annotation within the cluster and which genomes it is present in.
- *Core gene alignment* : a multiple alignment file of the core genes created using PRANK
- *Clustered Proteins* : a file that gives for each cluster id the list of locus tags it is made of
- *pan-genome reference* : this fasta file contains a single nucleotide sequence (representative) from each of the clusters in the pan genome
- Other various files in R of CSF formats.

Gene	Non-unique Gene name	Annotation	No. isolates	No. sequences	Avg sequences per isolate
Gene	*Non-unique Gene name*	*Annotation*	*No. isolates*	*No. sequences*	*Avg sequences per isolate*
EFZ9_RS07075	**	"molecular chaperone"	"16"	"16"	"1"
yJ2F	**	"GUP100 domain-containing protein"	"16"	"16"	"1"
yJ2B	**	"O-acetyl-ADP-ribose deacetylase"	"16"	"16"	"1"
mdcC	**	"glucane biosynthesis protein MdcC"	"16"	"16"	"1"
yJ2M	**	"N-methyl-L-tryptophan oxidase"	"16"	"16"	"1"
wblB_RS06790	**	"oxae family protein"	"16"	"16"	"1"
STY_RS05615	**	"Glycyl-MecA family oxidoreductase"	"16"	"16"	"1"
muuJ	**	"muonin biosynthesis integral membrane protein MuuJ"	"16"	"16"	"1"
yJ2H	**	"tagellar biosynthesis chaperone FlgH"	"16"	"16"	"1"
yJ2I	**	"anti-sigma-28 factor FlgI"	"16"	"16"	"1"
yJ2A	**	"tagellar basal body P-ring formation protein FlgA"	"16"	"16"	"1"
yJ2J	**	"tagellar basal body rod protein FlgJ"	"16"	"16"	"1"
yJ2G	**	"tagellar hook protein FlgG"	"16"	"16"	"1"
yJ2K	**	"tagellar basal body rod protein FlgK"	"16"	"16"	"1"
yJ2L	**	"tagellar hook-associated protein FlgL"	"16"	"16"	"1"
yJ2D	**	"tagellar hook-flament junction protein FlgD"	"16"	"16"	"1"
yJ2C	**	"ZSS (RNA accumulation protein) YcdC"	"16"	"16"	"1"
yJ2B	**	"ACP S-malonyltransferase"	"16"	"16"	"1"
yJ2P	**	"beta-keetoacyl-ACP synthase II"	"16"	"16"	"1"



COFFEE
BREAK

Phylogenomics basics

A few concepts on phylogenomics

- Phylogenomics definition



The image shows a screenshot of the Wikipedia article for "Phylogenomics". The browser address bar shows "en.wikipedia.org". The page title is "Phylogenomics - Wikipedia". The article content includes a definition: "Phylogenomics is the intersection of the fields of evolution and genomics." It lists four major areas: Prediction of gene function, Establishment and clarification of evolutionary relationships, Gene family evolution, and Prediction and retracing lateral gene transfer. A table of contents is visible, listing sections 1 through 7. The "Prediction of gene function" section is highlighted, with a sub-heading "Prediction of gene function" and a paragraph starting with "When Jonathan Eisen originally coined phylogenomics, it applied to prediction of gene function. Before the use of phylogenomic techniques, predicting gene function was done".

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

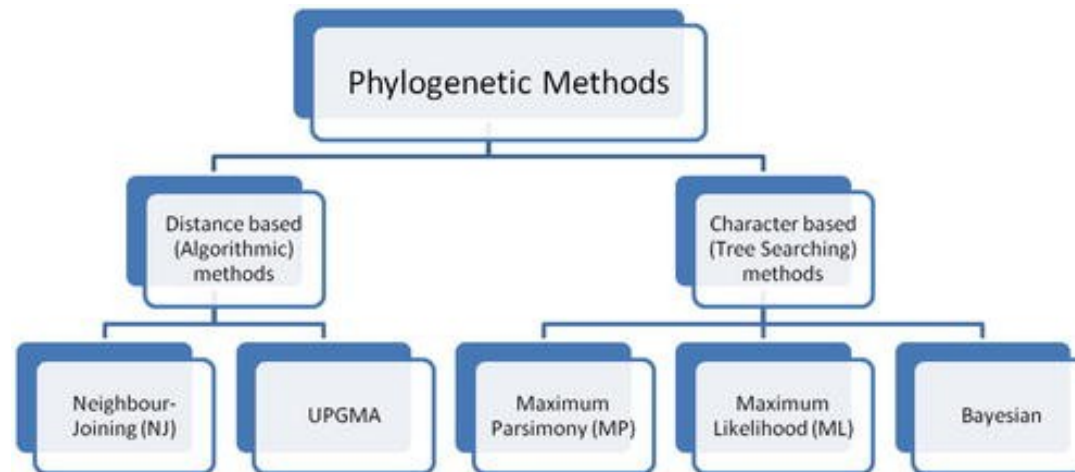
- Original definition
 - The application of phylogenetic methods for gene function analysis (Eisen, 1996)
 - Organism evolution based on whole genome analyses
- Recent usage: Various types of studies mixing genomics and phylogenetics, such as:
 - Global patterns of synteny (conserved gene order) across species
 - Global patterns of gene presence and absence studies across species
 - Genome rearrangements 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 likelihood method (alignment)

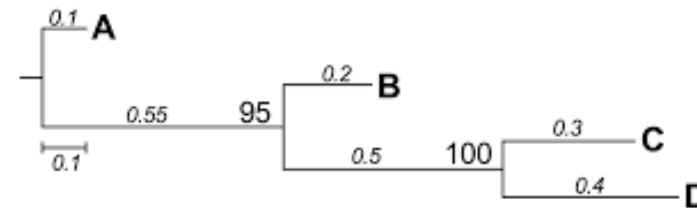


Phylogenetics main methods

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



Newick:

```
(A:0.1, (B:0.2, (C:0.3, D:0.4) 100:0.5) 95:0.55);
```

Extended Newick (eNewick):

```
(A:0.1, (B:0.2, (C:0.3, D:0.4) 0.5 [100]) 0.55 [95]);
```

Phylogenetics main methods

Reference: (Stephens, Bhattacharya, Ragan, et al., 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-Likelihood statistical models
- Compute local support values with the Shimodaira-Hasegawa test to estimate the reliability of each split in the tree

FastTree in practice:

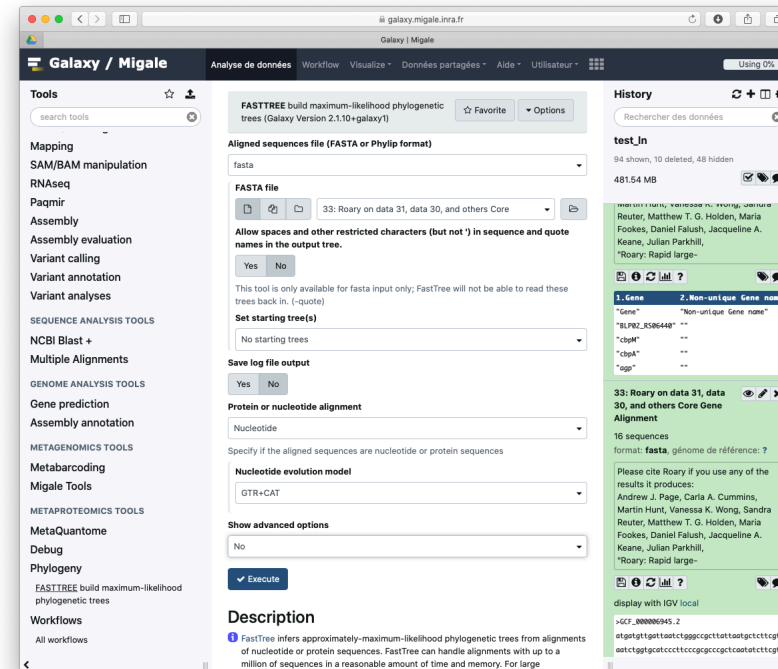
- takes as input an alignment file (Fasta or Phylip interleaved format)
- needs an evolution model: JTT or WAG or LG for protein, JC or GTR for nucleotide
- produces a tree in Newick format with SH support values [0-1] given as names for the internal nodes

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

FastTree: practice

Use **Galaxy-FastTree** to build a Maximum likelihood tree on the aligned core-genes

- input: the *Roary core genome alignment* file in fasta format
- Choose *Nucleotide alignment*
- Choose *GTR+CAT nucleotide evolution model*



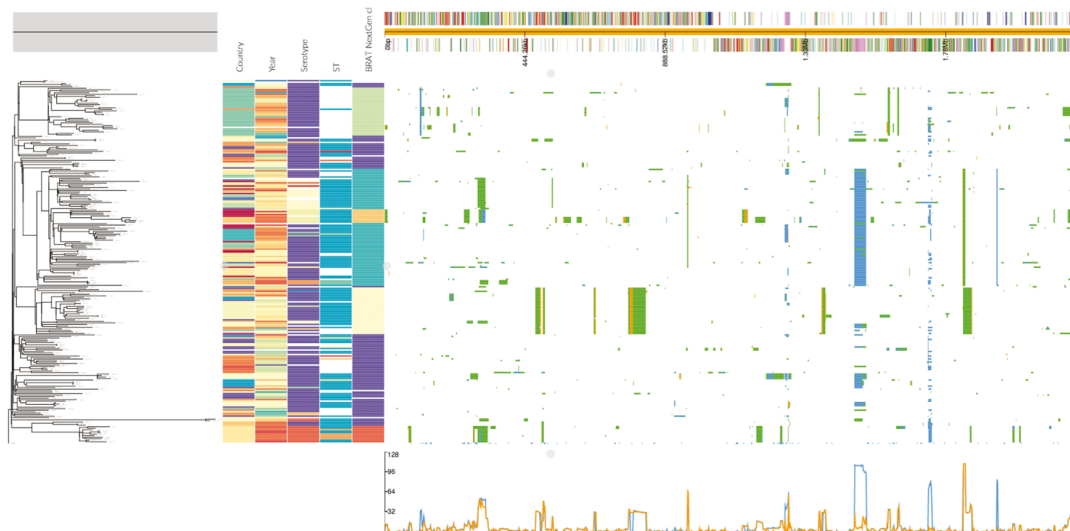
Galaxy-Fasttree

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), pan-genome data (from Roary for instance), genome annotations (GFF3 format) or any metadata (in simple CSV format)
- a valuable resource for results interpretation



Phandango: practice

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

Upload 3 datafiles just by dragging them:

- the Roary gene presence-absence file
- the Roary phylogenetic tree (change the extension file in *.tree*)
- A metadata csv file:
DatasetSalmonella_metadata.csv
Interpret results

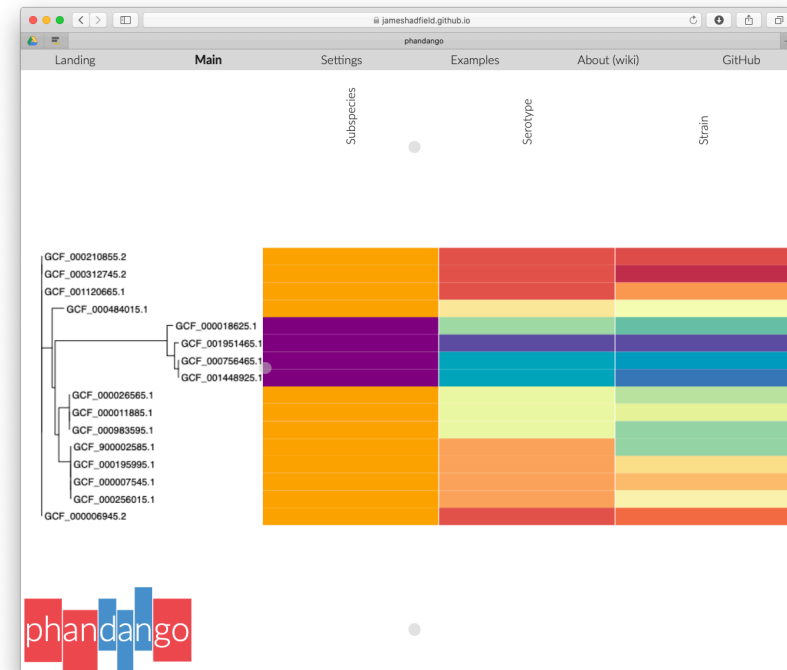


Phandango results on the Salmonella dataset

FastTree results interpretation using Phandango

Upload the following files

- the FastTree phylogenetic tree (change the extension file in *.tree*)
 - the metadata csv file:
DatasetSalmonella_metadata.csv
- Interpret results



FastTree result

Take home message

- Genome comparison is still an ongoing active bioinformatics research field
- Dataset construction, quality and diversity evaluation is a **mandatory** first-step and may be time-consuming
- Dataset de-replication may be helpful for some well-studied organisms
- Comparative strategy depends on the addressed question and on the genome diversity level
- 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](https://doi.org/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>.

Darling, A., B. Mau, F. Blattner, et al. (2004). "Mauve: multiple alignment of conserved genomic sequence with rearrangements". In: *Genome Research* 14(7), pp. 1394-1403. DOI: [10.1101/gr.2289704](https://doi.org/10.1101/gr.2289704).

Delcher, A., S. Kasif, R. Fleischmann, et al. (1999). "Alignment of whole genomes". In: *Nucleic Acids Res* 27(11):, pp. 2369-2376. DOI: [10.1093/nar/27.11.2369](https://doi.org/10.1093/nar/27.11.2369).

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](https://doi.org/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](https://doi.org/10.1093/bioinformatics/btt086). eprint: <https://academic.oup.com/bioinformatics/article-115> / 117

References(2)

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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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 - 594. DOI: <https://doi.org/10.1016/j.gde.2005.09.006>. URL:

References(3)

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](https://doi.org/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](https://doi.org/10.1186/s13059-016-0997-x). URL: <https://doi.org/10.1186/s13059-016-0997-x>.

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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1073/pnas.0506758102).