

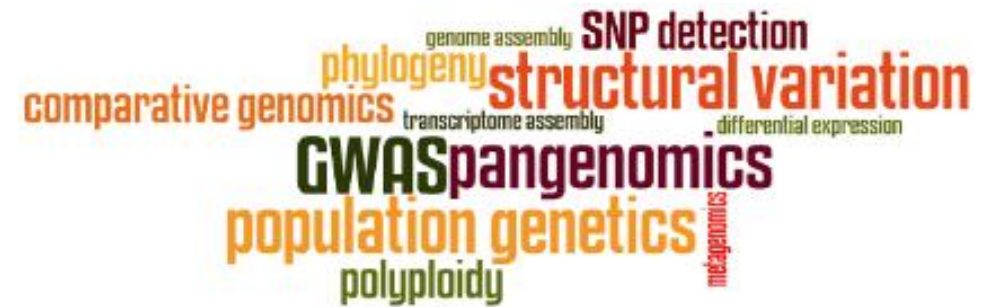
Session de formation 2023



bioinformatics platform dedicated to the genetics and genomics of tropical and Mediterranean plants and their pathogens



Mutualisation



Cacao



Banana



Coffee



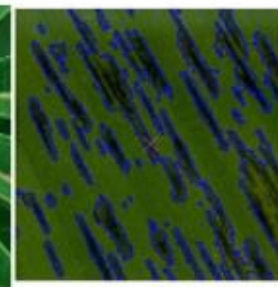
Rice



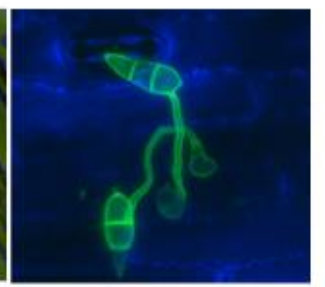
Palm



Cassava



Pseudocercospora



Magnaporthe

South Green

bioinformatics platform



4 institutes



3 research units



25+



Storage and computing resources

Tools

Trainings



400+



Meso@LR au CINES
1090 threads :
 35 standard nodes
 2 bigmem nodes
 1 GPU node
500 To of replicated storage



CINES
1130 threads:
 30 standard node
 1 supermem node
 1 GPU node
150 To on 3 NAS + 210 To scratch



400+



600+ tools

Resources mutualised at Meso@LR through the
Mudis4Ls project (purchase/storage/data)



Collaborative development of tools

Genomics

Pangenomic

Gene families

Comparative

Phylogeny

Assemblies

Annotation

Data mining

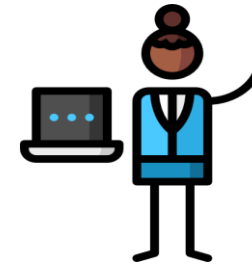
Diversity

genotype manipulation

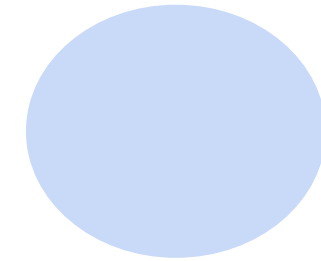
exploration

mosaic manipulation

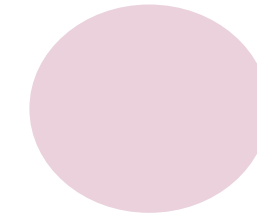
Metagenomic



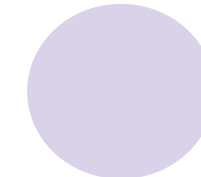
+20
tools



web applications (16)



visualisation (8)



workflows(5)



packages (4)



<https://github.com/SouthGreenPlatform/>

I-Trop

Plant & Health Bioinformatics Platform

<https://bioinfo.ird.fr/>



AUORE
COMTE



JACQUES
DAINAT



ALEXIS
DEREEPER



BRUNO
GRANOULLAC



JULIE
ORJUELA-



NDOMASSI
TANDO



CHRISTINE
TRANCHANT

bioinfo@ird.fr



@ItropBioinfo

South Green

bioinformatics platform



Florian Charriat
Antoni Exbrayat



Guilhem Sempere



Bruno Granouillac
Jacques Dainat



Nicolas Fernandez



Thomas Denecker

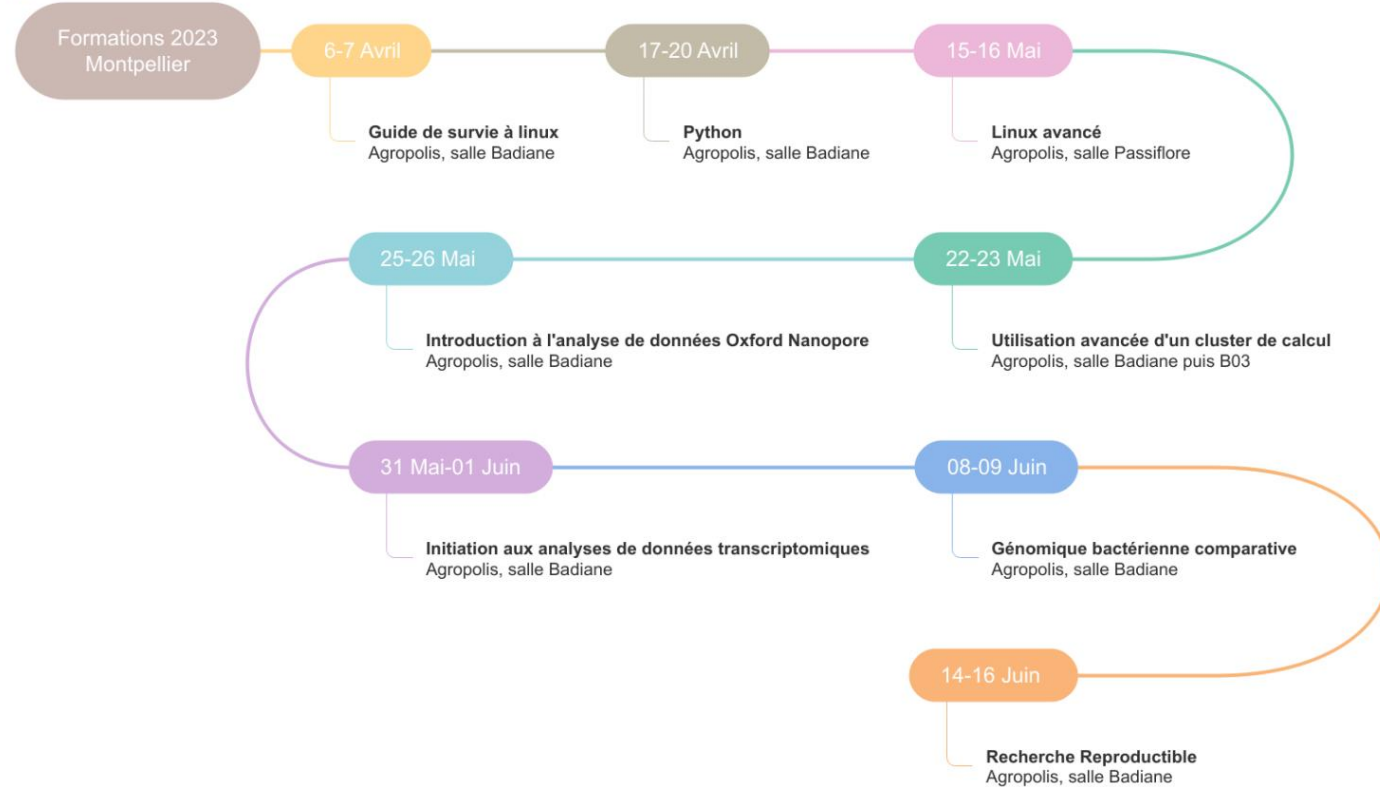
And more collaborators !

South Green

bioinformatics platform

South Green

bioinformatics platform



Plant
Health
Institute
Montpellier



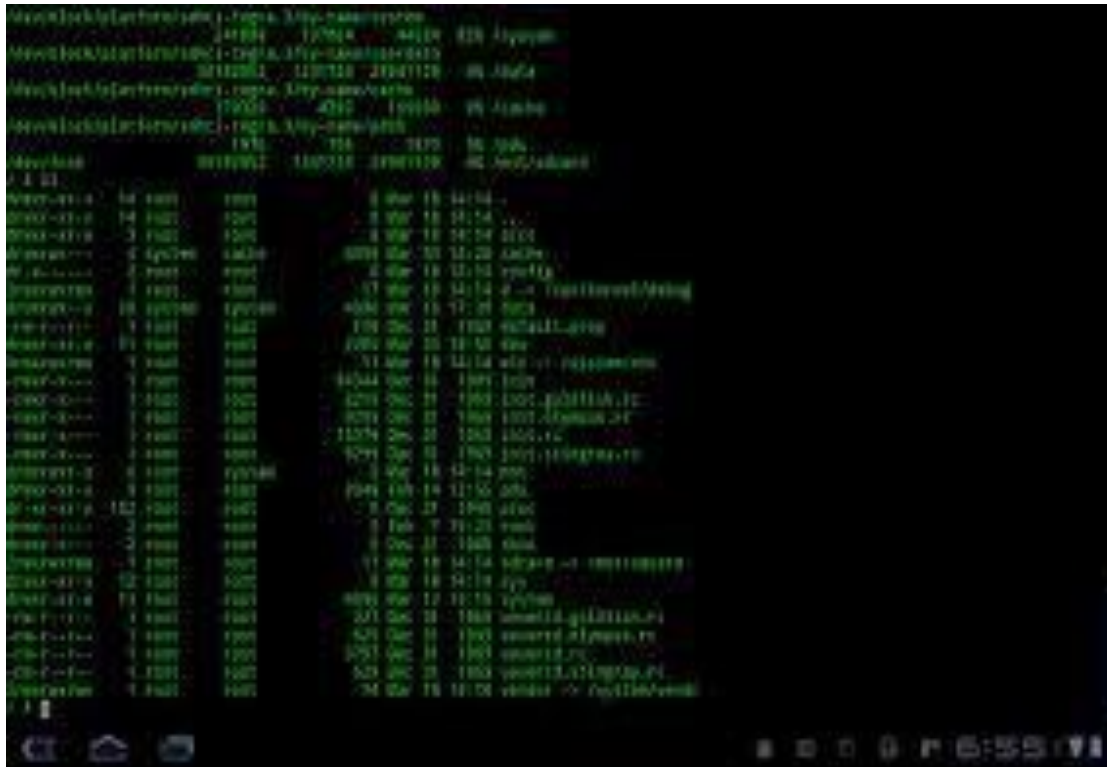
Modules de formation 2023

- Toutes nos formations :
<https://southgreenplatform.github.io/trainings/>
- Topo & TP :
https://github.com/SouthGreenPlatform/training_0NT_teaching/tree/2023_MTP
- Environnement de travail : [Logiciels à installer](#)

Génomique Comparative Bactérienne

- 2 façons d'utiliser linux :

en *mode console*



en *mode jupyter notebook*

jupyter parseClstr-Copy1 Dernière Sauvegarde : il y a 8 minutes (auto-sauvegardé) Se déconnecter

Fichier Édition Affichage Insérer Cellule Noyau Widgets Aide Non fiable | Python 3

Exécuter

Anchoring data analysis

1 - CDHIT data analysis *before anchoring on genome*

1.1 Removing redundancy with CDHIT

- CDHIT Input : 1,306,676 contigs assembled from no mapped reads
- Tests & results

	0.9	0.95
0.80	375,615	484,394
0.85	418,136	531,326
0.90	473,270	588,983
0.95	544,441	659,658

- clusters generated after cdhit analysis : 484,394

1.2 Converting cdhit file into a csv loaded as a dataframe with pandas

The script `cdhitVsAnchoring.py` create the csv file `allCtgsIRIGIN_TOG5681.dedup8095.PANDAS.csv`

Load csv file into a pandaframe

```
Entrée [1]: 1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import numpy as np
4
5 csv_cdhit_file = "/home/christine/Documents/These/Data/CDHIT/ALL_CTGS_MERGE/allCtgsIRIGIN_TOG5681.dedup8095.PAN
6 df_cdhit= pd.read_csv(csv_cdhit_file,names=['ctg','sp','ctg-list','sp-list'], header=0)
7 #print(df_cdhit)
8
```

What is jupyter book ?

- One of the most popular tool among data scientists to perform data analysis
- Provides a complete environment in which numerous programming languages can be use through a simple web browser

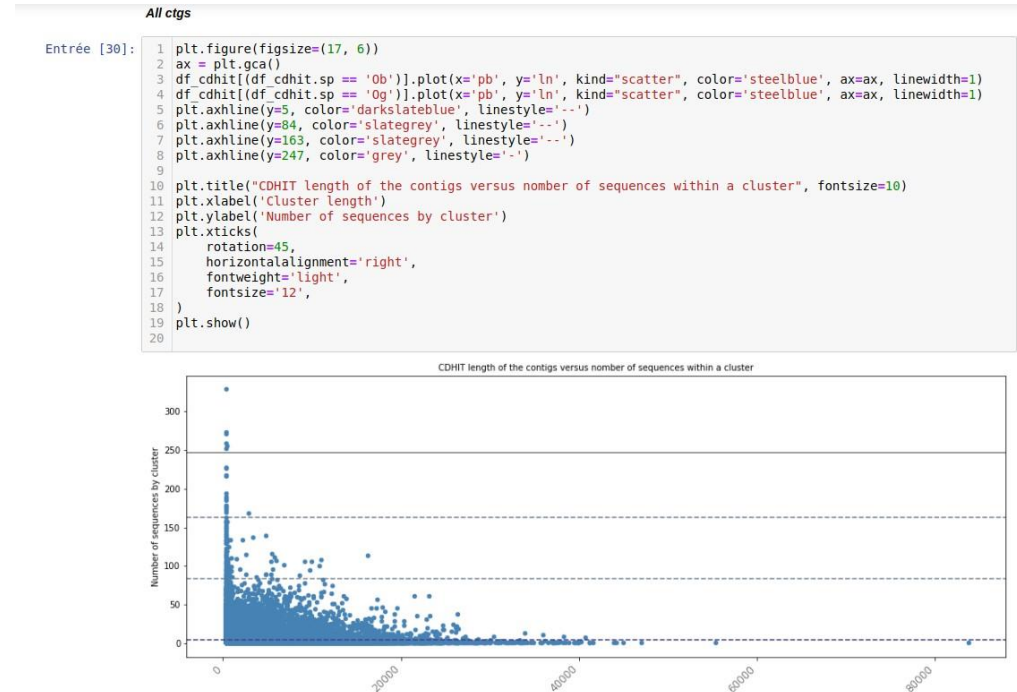
ex : Bash (Linux), Python, Java, R,
Julia, Matlab, Octave, Scheme,
Processing, Scala



What is jupyter book ?

- An unique interface/file where text, code and output codes can be mixed :

- code can be executed inside each cell of the notebook
- code output is directly displayed in the notebook



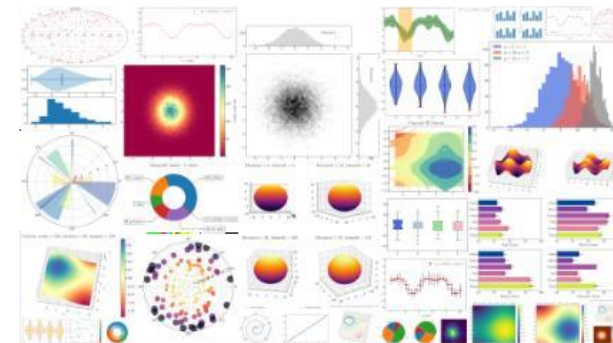
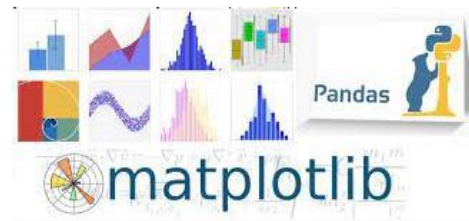
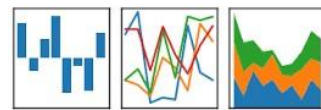
- facilement importer des fichiers tabulés dans des dataframes, similaires aux dataframes sous R.

(et exporter)

- manipuler ces tableaux de données / DataFrames
- facilement tracer des graphes à partir de ces DataFrames grâce à matplotlib

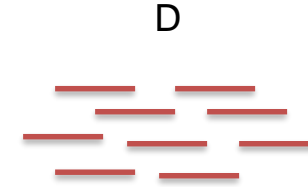
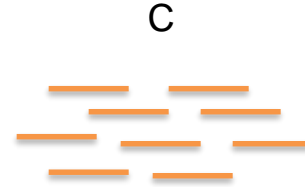
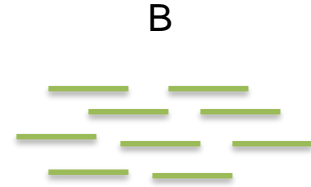
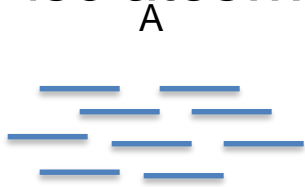
pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Two Approaches to Microbial Genomics

Starting with sets of reads representing your study isolates...



Assembly-based

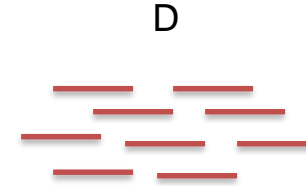
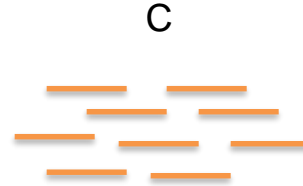
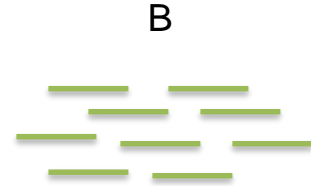
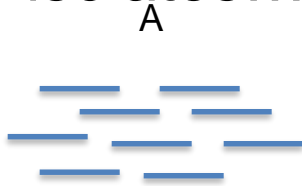
1. Assemble each set of reads into a genome sequence
2. Annotate each genome
3. Cluster genes and compare between each genome

Variant-based

1. Compare each read set to a reference genome assembly
2. Directly compare variants between each genome

Two Approaches to Microbial Genomics

Starting with sets of reads representing your study isolates...



Assembly-based

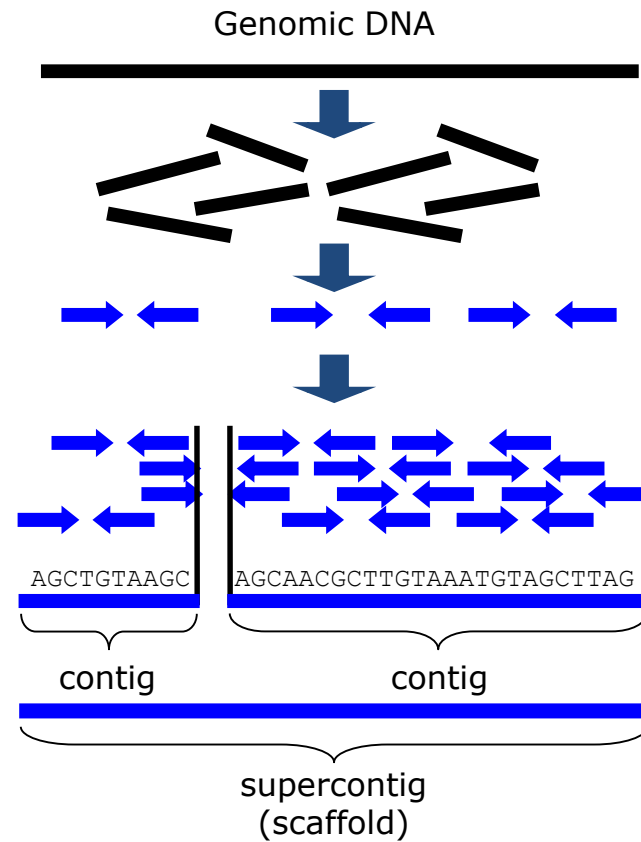
1. Assemble each set of reads into a genome sequence
2. Annotate each genome
3. Cluster genes and compare between each genome

Variant-based

1. Compare each read set to a reference genome assembly
2. Directly compare variants between each genome

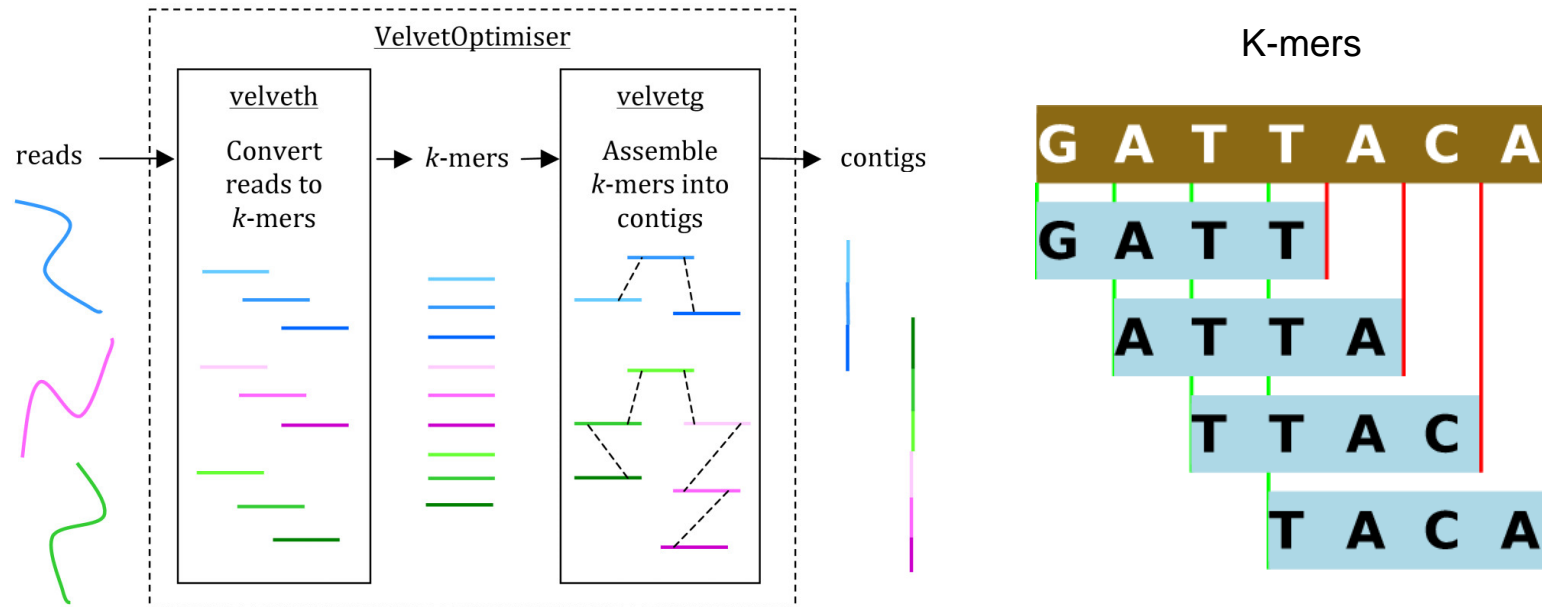
1) Assembly

Assembly Basics (de-novo assembly)



Assembly Methods

- SPAdes (<http://cab.spbu.ru/software/spades/>)
- Velvet (<https://www.ebi.ac.uk/~zerbino/velvet/>)
- Both are De Bruijn graph assemblers



Brief Report

Comparison of De Novo Assembly Strategies for Bacterial Genomes

Pengfei Zhang ^{1,2,†}, Dike Jiang ^{1,2,†}, Yin Wang ^{1,2,*}, Xueping Yao ^{1,2}, Yan Luo ^{1,2} and Zexiao Yang ^{1,2}

Table 1

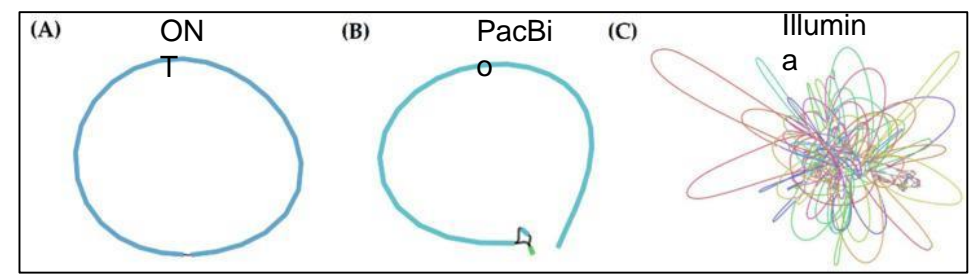
Statistics of genome-assembly results of independent assembly strategies.

Platforms	Assembler	Contigs	Largest Contig (bp)	N50	GC%
Illumina	SPAdes	527	157,573	40,498	39.87
PacBio	Canu	25	2,351,556	2,351,556	40.01
ONT	Canu	1	2,360,091	2,360,091	40.02

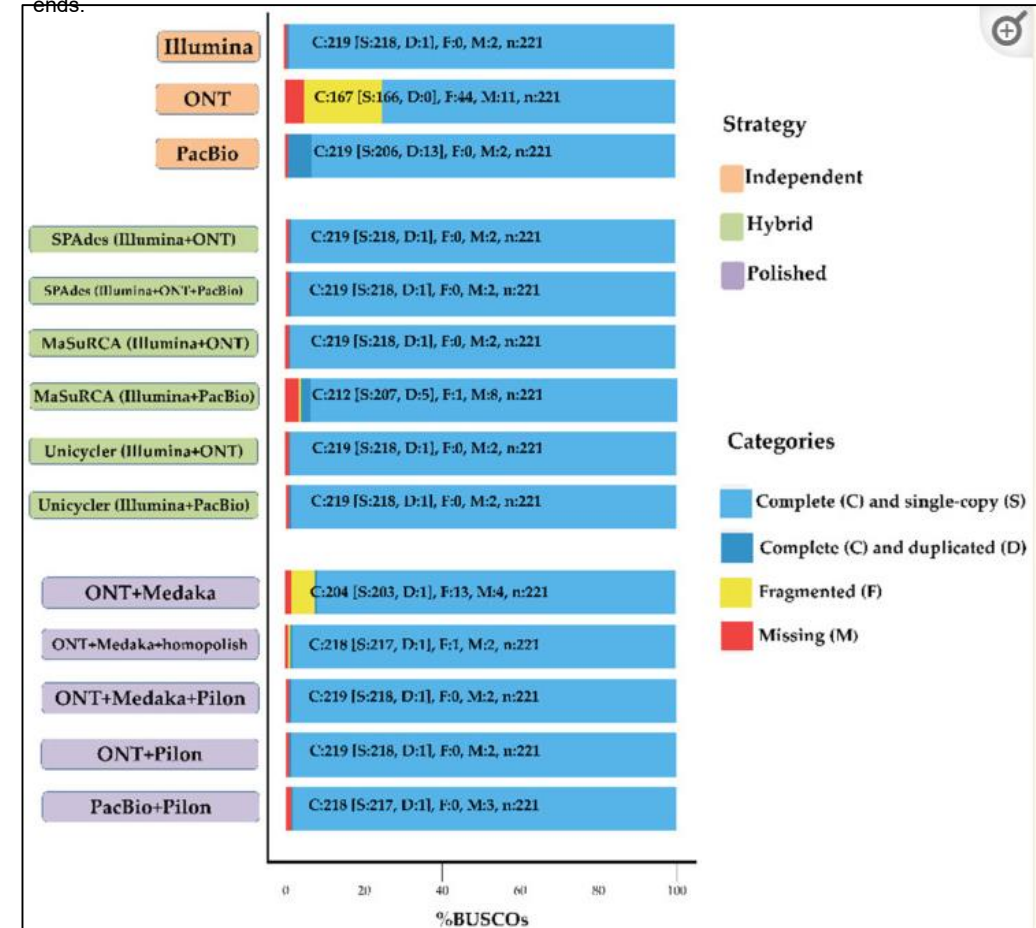
Table 2

Statistics of genome-assembly results of hybrid assembly strategies.

Platforms	Assembler	Contigs	Total Length (bp)	N50	GC%
Illumina + ONT	SPAdes	266	2,402,219	1,953,224	39.97
Illumina + PacBio + ONT	SPAdes	236	2,410,042	2,351,543	40.02
Illumina + ONT	Unicycler	1	2,349,186	2,349,186	40.03
Illumina + PacBio	Unicycler	1	2,349,340	2,349,340	40.03
Illumina + ONT	MaSuRCA	1	2,365,339	2,365,339	40.02
Illumina + PacBio	MaSuRCA	4	2,395,409	1,345,876	40.04

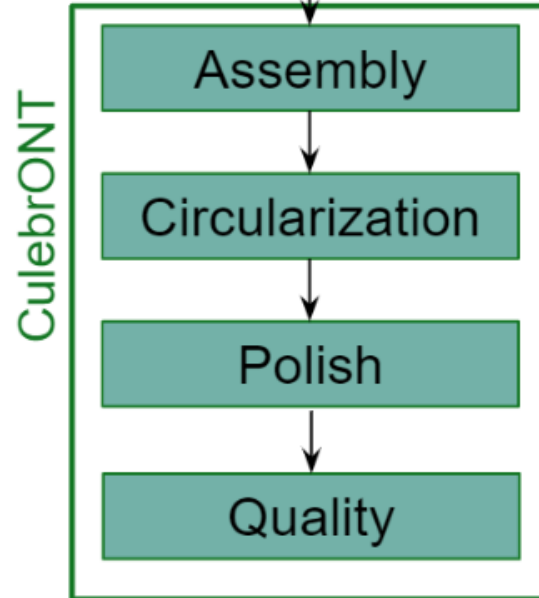
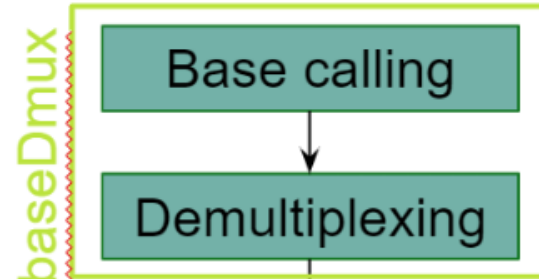


Comparison of results of independent assembly strategies. (A) Genome assembled with nanopore reads; (B) longest contig assembled with PacBio reads; (C) genome assembled with Illumina reads. Plots were obtained by using Bandage on the “assembly_graph.gfa” output file from SPAdes or the “contig.gfa” output file from Canu. Connections between contigs represent overlaps between contig ends.



Evaluation of completeness of assembly results of different strategies. Assessments of the completeness of the assembly genomes with the datasets of proteobacteria_odb9 lineage. Bar charts produced with BUSCO plotting tool to show proportions that were classified as complete (C, blue), complete single copy (S, light blue), complete duplicated (D, dark blue), fragmented (F, yellow), and missing (M, red).

Bioinformatic Workflows: assembly



Snakemake



<https://github.com/vibaotram/baseDmux>



<https://culebront-pipeline.readthedocs.io/en/latest/>



2) Separate chromosomal and plasmid scaffolds/contigs

MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies

James Robertson¹ and John H. E. Nash^{2,*}

MOB-suite: Software tools for clustering, reconstruction and typing of plasmids from draft assemblies

Introduction

Plasmids are mobile genetic elements (MGEs), which allow for rapid evolution and adaption of bacteria to new niches through horizontal transmission of novel traits to different genetic backgrounds. The MOB-suite is designed to be a modular set of tools for the typing and reconstruction of plasmid sequences from WGS assemblies.

The MOB-suite depends on a series of databases which are too large to be hosted in git-hub. They can be downloaded or updated by running `mob_init` or if running any of the tools for the first time, the databases will download and initialize automatically if you do not specify an alternate database location. However, they are quite large so the first run will take a long time depending on your connection and speed of your computer. Databases can be manually downloaded from [here](#).

Our new automatic chromosome depletion feature in MOB-recon can be based on any collection of closed chromosome sequences.

Citations

Below are the manuscripts describing the algorithmic approaches used in the MOB-suite.

1. Robertson, James, and John H E Nash. "MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies." *Microbial genomics* vol. 4,8 (2018): e000206. doi:10.1099/mgen.0.000206
2. Robertson, James et al. "Universal whole-sequence-based plasmid typing and its utility to prediction of host range and epidemiological surveillance." *Microbial genomics* vol. 6,10 (2020): mgen000435. doi:10.1099/mgen.0.000435

MOB-init

On first run of MOB-typer or MOB-recon, MOB-init (invoked by `mob_init` command) should run to download the databases from figshare, sketch the databases and setup the blast databases. However, it can be run manually if the databases need to be re-initialized OR if you want to initialize the databases in an alternative directory.

MOB-cluster

This tool creates plasmid similarity groups using fast genomic distance estimation using Mash. Plasmids are grouped into clusters using complete-linkage clustering and the cluster code accessions provided by the tool provide an approximation of operational taxonomic units OTU's. The plasmid nomenclature is designed to group highly similar plasmids together which are unlikely to have multiple representatives within a single cell and have a strong concordance with replicon and relaxase typing but is universally applicable since it uses the complete sequence of the plasmid itself rather than specific biomarkers.

MOB-recon

This tool reconstructs individual plasmid sequences from draft genome assemblies using the clustered plasmid reference databases provided by MOB-cluster. It will also automatically provide the full typing information provided by MOB-typer. It optionally can use a chromosome depletion strategy based on closed genomes or user supplied filter of sequences to ignore.

MOB-typer

Provides in silico predictions of the replicon family, relaxase type, mate-pair formation type and predicted transferability of the plasmid. Using a combination of biomarkers and MOB-cluster codes, it will also provide an observed host-range of your plasmid based on its replicon, relaxase and cluster assignment. This is combined with information mined from the literature to provide a prediction of the taxonomic rank at which the plasmid is likely to be stably maintained but it does not provide source attribution predictions.

3) Genome Annotation

What is annotation ?

Structural annotation:

VS

Functional annotation:

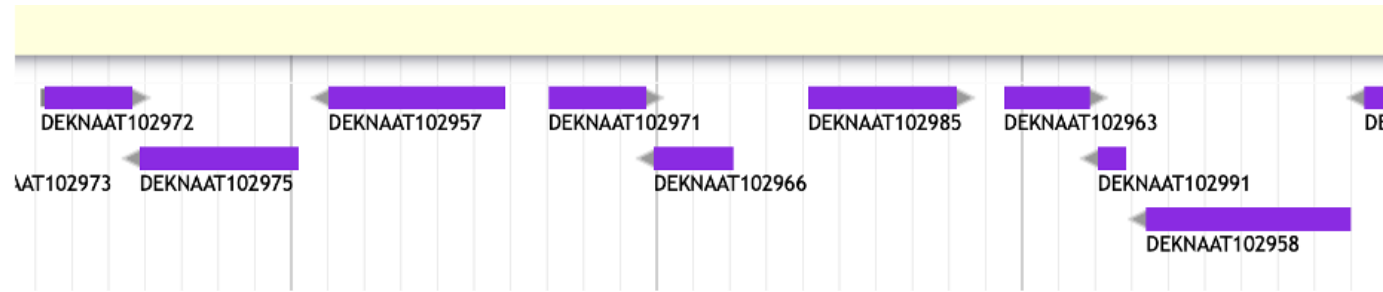
Find out where the regions of interest (usually genes) are in the sequence data and what they look like.

Find out what the regions do. What do they code for?

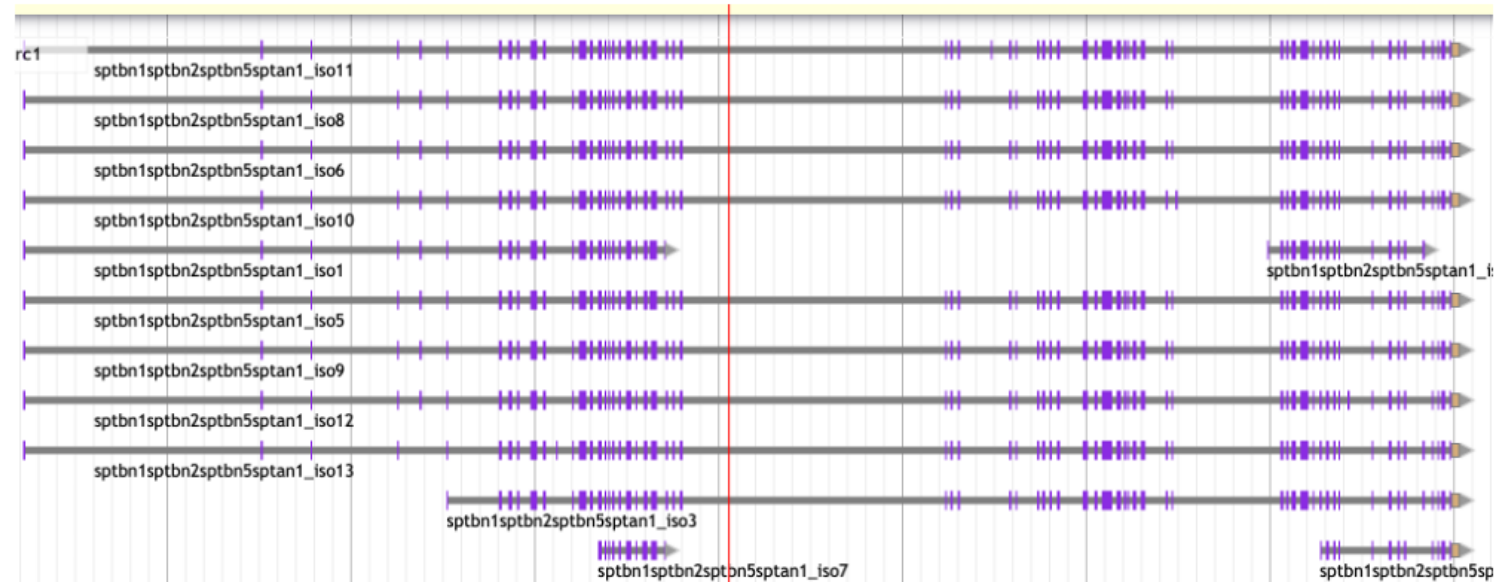
*It is the **annotation** that bridges the gap from the sequence to the biology of the organism*

Organisms differ in genomic complexity

A yeast



A crustacean




```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
```

← Header

- 9 columns
- 1 feature = 1 line

Ctg123	.	Gene	1000	9000	.	+	.	ID=gene1;Name=EDEN
ctg123	.	mRNA	1050	9000	.	+	.	ID=mRNA1;Parent=gene1;Name=EDEN.1
ctg123	.	mRNA	1050	9000	.	+	.	ID=mRNA2;Parent=gene1;Name=EDEN.2
ctg123	.	exon	1300	1500	.	+	.	ID=exon1;Parent=mRNA3
ctg123	.	exon	1050	1500	.	+	.	ID=exon2;Parent=mRNA1,mRNA2
ctg123	.	exon	3000	3902	.	+	.	ID=exon3;Parent=mRNA1
ctg123	.	exon	5000	5500	.	+	.	ID=exon4;Parent=mRNA1,mRNA2
ctg123	.	exon	7000	9000	.	+	.	ID=exon5;Parent=mRNA1,mRNA2
ctg123	.	CDS	1201	1500	.	+	0	ID=cds1;Parent=mRNA1;Name=eden1
ctg123	.	CDS	3000	3902	.	+	0	ID=cds1;Parent=mRNA1;Name=eden1
ctg123	.	CDS	5000	5500	.	+	0	ID=cds1;Parent=mRNA1;Name=eden1
ctg123	.	CDS	7000	7600	.	+	0	ID=cds1;Parent=mRNA1;Name=eden1
Ctg123	.	CDS	1201	1500	.	+	0	ID=cds2;Parent=mRNA2;Name=eden2
ctg123	.	CDS	5000	5500	.	+	0	ID=cds2;Parent=mRNA2;Name=eden2
Ctg123	.	CDS	7000	7600	.	+	0	ID=cds2;Parent=mRNA2;Name=eden2

- 1) sequence id 2) source 3) feature type 4) start 5) end 6) score 7) strand 8) phase 9) attributes
tag=value

(SO term = 2278 possibilities)

! Features are grouped by **parent** relationship

Adding biological info to sequences

ribosome
binding site

delta toxin
PubMed: 15353161

ACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGA
AAAGCAGCCTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTC
CCAGGCCAGTGCCGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTG
GCCAGGCGGCAGGAAGGCGCACCCCCCAGCAATCCGCGCGCCGGG
ACAGAATGCCCTGCAGGAACTTCTTCTAGAAGACCTTCTCCTCCTG
CAAATAAAACCTCACCCATGAATGCTCACGCAAGTTTAATTACAGA
CCTGAAACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTGCT
CTCCGTCCGTCCGTGGGCCACGGCCACCGCTTTTTTTTTTTGCC

transfer RNA
Leu-(UUR)

tandem repeat
CCGT x 3

homopolymer
10 x T

Annotation Methods

- There are different annotation algorithms for protein-coding genes, tRNAs, rRNAs, other non-coding RNAs
- Pipelines exist for performing several in one go

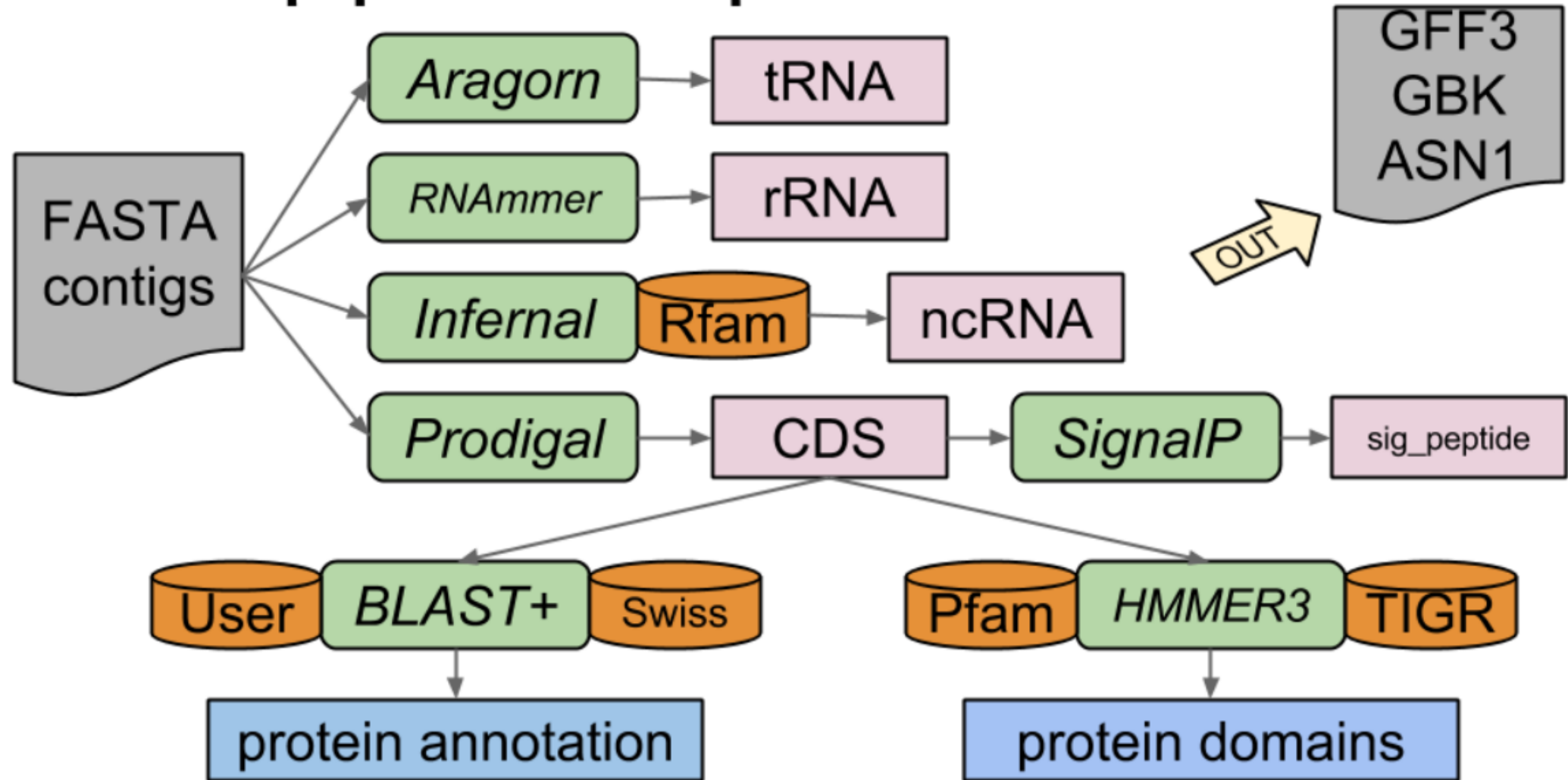
Prokaryote annotation:

- Prokka (<http://www.vicbioinformatics.com/software/prokka.shtml>) is an all-in-one wrapper for these tools

Table 1. Feature prediction tools used by Prokka

Tool (reference)	Features predicted
Prodigal (Hyatt 2010)	Coding sequence (CDS)
RNAmmer (Lagesen <i>et al.</i> , 2007)	Ribosomal RNA genes (rRNA)
Aragorn (Laslett and Canback, 2004)	Transfer RNA genes
SignalP (Petersen <i>et al.</i> , 2011)	Signal leader peptides
Infernal (Kolbe and Eddy, 2011)	Non-coding RNA

Prokka pipeline (simplified)



Prokaryote annotation:

- Bakta: rapid & standardized annotation of bacterial genomes, MAGs & plasmids (<https://github.com/oschwengers/bakta>)

Schwengers O., Jelonek L., Dieckmann M. A., Beyvers S., Blom J., Goesmann A. (2021). Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics*, 7(11). <https://doi.org/10.1099/mgen.0.000685>

Tools

- tRNAscan-SE
- Aragorn
- INFERNAL
- PILER-CR
- Prodigal
- Hmmer
- Diamond
- Blast+
- AMRFinderPlus
- DeepSig

Databases

- Rfam
- DoriC: AntiFam
- UniProt
- RefSeq
- COG
- KEGG
- PHROG
- AMRFinder
- ISFinder
- Pfam
- VFDB

4) Public genomes retrieval

Search NCBI

Search

Genomes - NCBI Datasets BETA

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

TAXONOMIC NAME

🔍 Anaplasmataceae 1

Filters

STATUS

reference genomes

annotated 3

ASSEMBLY LEVEL

contig scaffold chromosome 2 complete

TEXT FILTER

YEAR RELEASED

1980

Download

Download table

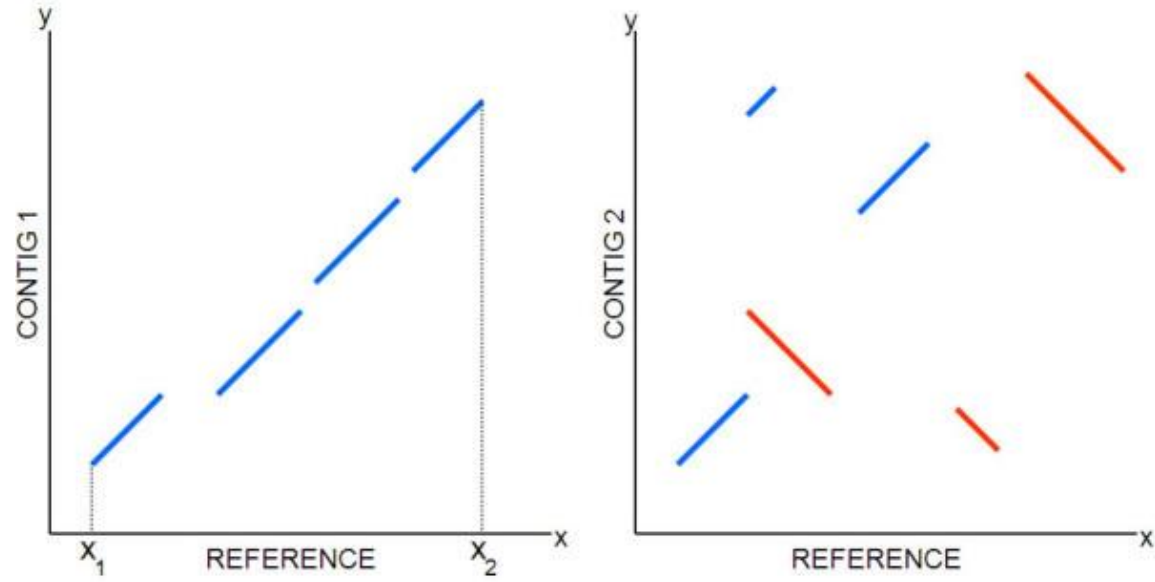
Download dataset

Scientific name	Modifier	Annotation	Size (Mbp)	Level	Acti
<i>Anaplasma centrale</i> str. Israel	Israel strain	NCBI Prokaryotic Genome Pipeline (PGAP)	1.207	Comple	
<i>Anaplasma centrale</i> str. Israel	Israel strain	submitted by USDA-ARS	1.207	Comple	
<i>Anaplasma marginale</i>	Palmeira strain	NCBI Prokaryotic Genome Pipeline (PGAP)	1.195	Chromo	
<i>Anaplasma marginale</i>	Jaboticabal strain	NCBI Prokaryotic Genome Pipeline (PGAP)	1.195	Chromo	
<i>Anaplasma marginale</i>	Palmeira strain	NCBI Prokaryotic Genome Pipeline	1.195	Chromo	
<i>Anaplasma marginale</i>	Jaboticabal strain	NCBI Prokaryotic Genome Pipeline	1.195	Chromo	
<i>Anaplasma marginale</i> str. Dawn	Dawn strain	NCBI Prokaryotic Genome Pipeline (PGAP)	1.197	Chromo	
<i>Anaplasma marginale</i> str. Dawn	Dawn strain	NCBI Prokaryotic Genome Pipeline	1.197	Chromo	

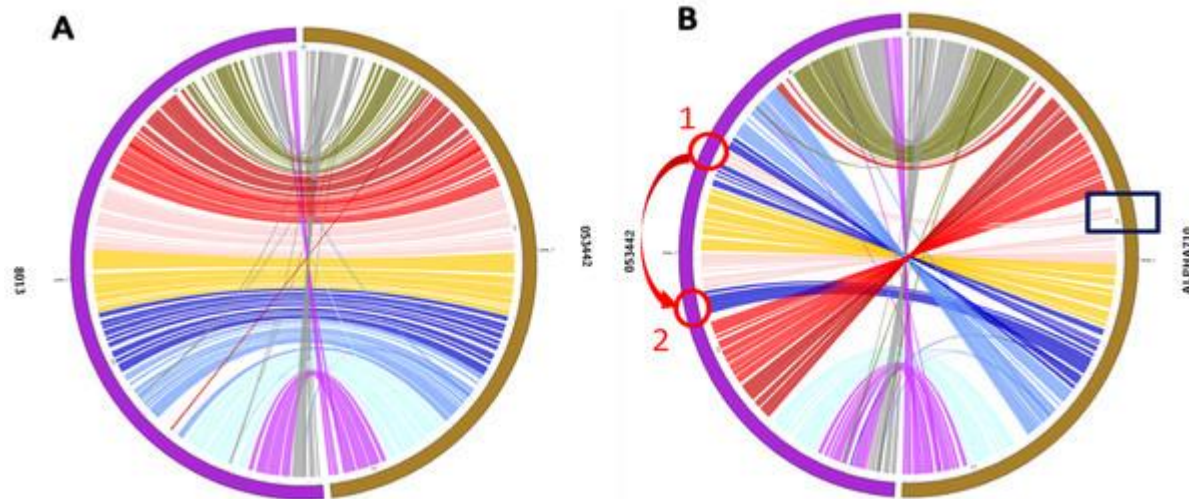
Assembly name	Assembly Accession	Organism tax	Organism inf	Organism int	Organism inf	Organism int	Annotation	Assembly Sta	Assembly Len	Assembly Submission Date
1	ASM802v1	GCA_000008025.1	Wolbachia endosymbiont -vMell				Annotation	5	326782	Complete Ge 05/02/2004
2	ASM128v1	GCA_000008283.1	Wolbachia endosymbiont strain TR5 of <i>Brugia malayi</i> wMell				Annotation	5	326782	Complete Ge 02/02/2005
3	ASM1154v1	GCA_000011543.1	<i>Anaplasma marginale</i> str. 1St. Marie				Annotation	5	1519687	Complete Ge 06/12/2004
4	ASM1256v1	GCA_000012563.1	<i>Ehrlichia canis</i> str. Jake Jake				Annotation	5	531030	Complete Ge 13/08/2005
5	ASM1312v1	GCA_000013123.1	<i>Anaplasma phagocytophilum</i> H2				Annotation	5	347282	Complete Ge 21/02/2006
6	ASM1339v1	GCA_000013393.1	<i>Ehrlichia chaffeensis</i> str. Arkansas				Annotation	5	317048	Complete Ge 21/02/2006
7	ASM1339v1	GCA_000013393.1	<i>Neorickettsia sennetsu</i> str. Miyayama				Annotation	5	83906	Complete Ge 21/02/2006
8	ASM2030v1	GCA_000020303.1	<i>Anaplasma marginale</i> str. Florida				Annotation	5	120483	Complete Ge 03/02/2009
9	ASM2228v1	GCA_000022283.1	<i>Wolbachia</i> sp. wMell				Annotation	5	348571	Complete Ge 24/03/2009
10	ASM2228v1	GCA_000022283.1	<i>Neorickettsia risticii</i> str. Illinois				Annotation	5	87977	Complete Ge 25/07/2009
11	ASM2450v1	GCA_000024503.1	<i>Anaplasma centrale</i> str. Israel				Annotation	5	326806	Complete Ge 24/11/2009
12	ASM2450v1	GCA_000024503.1	<i>Ehrlichia sumanum</i> str. Waikgonden				Annotation	5	531833	Complete Ge 05/02/2009
13	ASM5040v1	GCA_000050403.1	<i>Ehrlichia sumanum</i> str. Gardel				Annotation	5	149930	Complete Ge 01/02/2005
14	ASM5042v1	GCA_000050423.1	<i>Ehrlichia sumanum</i> str. Waikgonden				Annotation	5	531297	Complete Ge 01/02/2005
15	ASM7300v1	GCA_000073003.1	<i>Wolbachia endosymbiont -vPip</i>				Annotation	5	348255	Chromosome 13/06/2008
16	ASM7300v1	GCA_000073003.1	<i>Wolbachia endosymbiont -vD0</i>				Annotation	5	93799	Complete Ge 30/07/2012
17	ASM7300v1	GCA_000073003.1	<i>Wolbachia endosymbiont -vHq2</i>				Annotation	5	130323	Complete Ge 22/04/2013
18	ASM7300v1	GCA_000073003.1	<i>Wolbachia endosymbiont -vHq4</i>				Annotation	5	129504	Complete Ge 22/04/2013
19	ASM7300v1	GCA_000073003.1	<i>Anaplasma phagocytophilum</i> H2				NCBI Prokary	5	347754	Complete Ge 24/07/2013
20	ASM7300v1	GCA_000073003.1	<i>Anaplasma phagocytophilum</i> H2				NCBI Prokary	5	348358	Complete Ge 24/07/2013
21	ASM7300v1	GCA_000073003.1	<i>Anaplasma phagocytophilum</i> H2				NCBI Prokary	5	347302	Chromosome 24/07/2013
22	ASM7300v1	GCA_000073003.1	<i>Anaplasma marginale</i> str. Ogeya Plains				NCBI Prokary	5	119822	Chromosome 05/11/2013
23	ASM7300v1	GCA_000073003.1	<i>Anaplasma marginale</i> str. Dawn				NCBI Prokary	5	119670	Chromosome 05/11/2013
24	ASM7300v1	GCA_000073003.1	<i>Ehrlichia muris</i> AS145 AS145				NCBI Prokary	5	119671	Complete Ge 16/12/2013
25	ASM7300v1	GCA_000073003.1	<i>Ehrlichia chaffeensis</i> str. Heartland				Annotation	5	117271	Complete Ge 17/04/2014
26	ASM7300v1	GCA_000073003.1	<i>Ehrlichia sp. wMell</i>				Annotation	5	314894	Complete Ge 17/04/2014
27	ASM7300v1	GCA_000073003.1	<i>Ehrlichia chaffeensis</i> str. JJ Lee				Annotation	5	117899	Complete Ge 17/04/2014
28	ASM7300v1	GCA_000073003.1	<i>Ehrlichia chaffeensis</i> str. U.Liberty				Annotation	5	117802	Complete Ge 17/04/2014
29	ASM7300v1	GCA_000073003.1	<i>Ehrlichia chaffeensis</i> str. O.Oceola				Annotation	5	117197	Complete Ge 17/04/2014
30	ASM7300v1	GCA_000073003.1	<i>Ehrlichia chaffeensis</i> str. St. Vincent				Annotation	5	117384	Complete Ge 17/04/2014
31	ASM7300v1	GCA_000073003.1	<i>Ehrlichia chaffeensis</i> str. W.Wakulla				Annotation	5	117817	Complete Ge 17/04/2014
32	ASM7300v1	GCA_000073003.1	<i>Ehrlichia chaffeensis</i> str. W.West Plains				Annotation	5	117953	Complete Ge 17/04/2014
33	ASM7300v1	GCA_000073003.1	<i>Neorickettsia helminthos</i> Oregon				Annotation	5	84232	Complete Ge 17/04/2014
34	ASM7300v1	GCA_000073003.1	<i>Anaplasma phagocytophilum</i> Norway variant2				NCBI Prokary	5	348187	Complete Ge 03/05/2015
35	ASM7300v1	GCA_000073003.1	<i>Wolbachia endosymbiont -vC6</i>				Annotation	5	325060	Complete Ge 10/06/2016
36	WTTRE_1.0	GCA_000499953.1	<i>Wolbachia endosymbiont of Drosophila simulans</i> wku				Annotation	5	326841	Complete Ge 13/10/2016
37	WTTRE_1.0	GCA_000499953.1	<i>Wolbachia endosymbiont -vTyr</i>				Annotation	5	313802	Chromosome 07/01/2016
38	ASM17056v1	GCA_001705651.1	<i>Wolbachia endosymbiont -vTm_Cu</i>				NCBI Prokary	5	126786	Chromosome 11/10/2016
39	ASM17056v1	GCA_001705651.1	<i>Wolbachia endosymbiont -vTm_SM</i>				NCBI Prokary	5	126764	Chromosome 11/10/2016
40	ASM19117v1	GCA_001911752.1	<i>Wolbachia endosymbiont -vBurlin</i>				NCBI Prokary	5	180528	Chromosome 25/06/2018
41	ASM2228v2	GCA_002228623.1	<i>Anaplasma ovis</i> str. Heiler Heiler				NCBI Prokary	5	321404	Complete Ge 03/07/2018
42	ASM2228v2	GCA_002228623.1	<i>Wolbachia pipiensis</i> wMell-wk2018				NCBI Prokary	5	348353	Complete Ge 11/07/2019
43	ASM2228v2	GCA_002228623.1	<i>Wolbachia pipiensis</i> wMell-wk2018				NCBI Prokary	5	348279	Complete Ge 11/07/2019
44	ASM2228v2	GCA_002228623.1	<i>Ehrlichia canis</i> Y2-1				NCBI Prokary	5	113478	Complete Ge 15/01/2018
45	ASM2228v2	GCA_002228623.1	<i>Anaplasma marginale</i> Palmeira				NCBI Prokary	5	119520	Chromosome 10/09/2018
46	ASM2228v2	GCA_002228623.1	<i>Anaplasma marginale</i> Jaboticabal				NCBI Prokary	5	119533	Chromosome 10/09/2018
47	ASM2228v2	GCA_002228623.1	<i>Wolbachia pipiensis</i> wMell-wk2018				NCBI Prokary	5	350495	Chromosome 08/01/2019
48	ASM41122v1	GCA_004112281.1	<i>Wolbachia endosymbiont -vMau</i>				NCBI Prokary	5	348407	Complete Ge 12/02/2019
49	ASM41122v1	GCA_004112281.1	<i>Wolbachia endosymbiont -vMau</i>				NCBI Prokary	5	348004	Complete Ge 16/04/2019
50	ASM41122v1	GCA_004112281.1	<i>Wolbachia endosymbiont -vMau</i>				NCBI Prokary	5	348004	Complete Ge 16/04/2019
51	ASM41122v1	GCA_004112281.1	<i>Wolbachia endosymbiont -vMau</i>				NCBI Prokary	5	348004	Complete Ge 16/04/2019
52	ASM41122v1	GCA_004112281.1	<i>Wolbachia endosymbiont -vMau</i>				NCBI Prokary	5	348004	Complete Ge 16/04/2019
53	ASM41122v1	GCA_004112281.1	<i>Wolbachia endosymbiont -vMau</i>				NCBI Prokary	5	348004	Complete Ge 16/04/2019
54	ASM41122v1	GCA_004112281.1	<i>Wolbachia pipiensis</i> wMell_NZ3				NCBI Prokary	5	126781	Complete Ge 12/08/2019
55	ASM41122v1	GCA_004112281.1	<i>Wolbachia pipiensis</i> wMell_Q28				NCBI Prokary	5	126817	Complete Ge 12/08/2019
56	ASM41122v1	GCA_004112281.1	<i>Wolbachia pipiensis</i> wMell_Q28				NCBI Prokary	5	126783	Complete Ge 12/08/2019
57	ASM41122v1	GCA_004112281.1	<i>Wolbachia endosymbiont -vWJ.1</i>				Annotation	5	348340	Complete Ge 20/09/2019

5) Pairwise genome alignment

Dot plot



Circos link



Dgenies: <https://dgenies.toulouse.inra.fr>

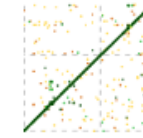
Dot plot

In bioinformatics a dot plot is a graphical method that allows the comparison of two biological sequences and identify regions of close similarity between them. It is a type of recurrence plot.

More details of dot plot [here](#). Below, some examples of events which can be detected by dot plots.

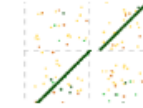
Match

When two samples sequence are identical, it's a match.



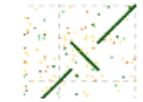
Gap

Dot plots can be used to detect a gap between two samples: small sequence which exists only in one sample, between two matching regions.



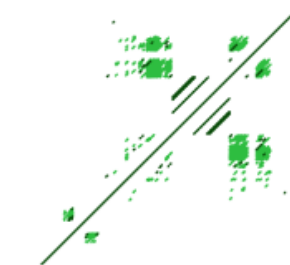
Inversion

Sequence which exists in the two samples but not in the same order.



Repeats

Dot plot can be used to detect repeated regions: a sequence which is repeated several times in a sample.

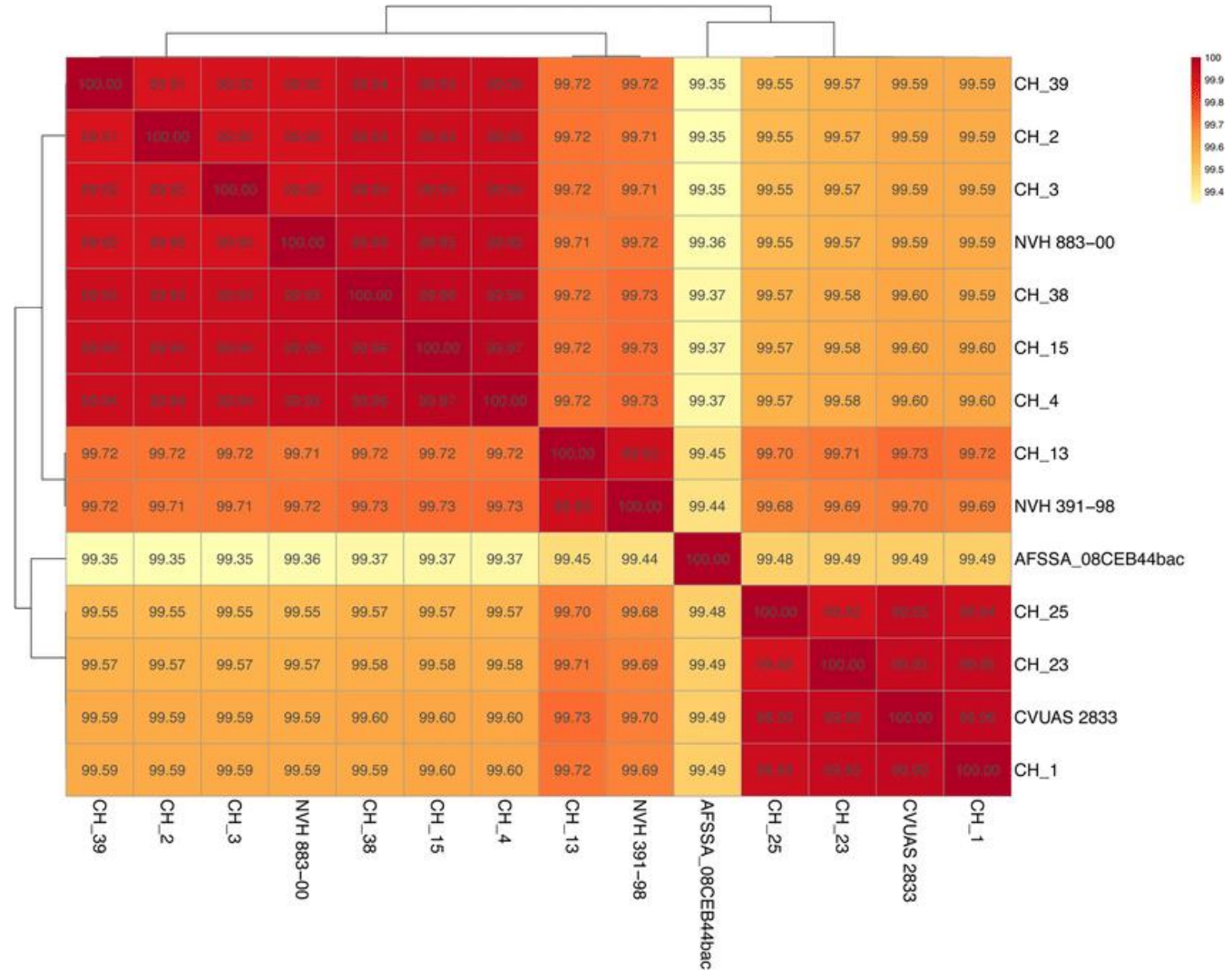


6) Pairwise Average Nucleotide Identity (ANI)

ANI: Average Nucleotide Identity

The average nucleotide identity (ANI) is a similarity index between a given pair of genomes that can be applicable to prokaryotic organisms independently of their G+C content, and a cutoff score of >95% indicates that they belong to the same species

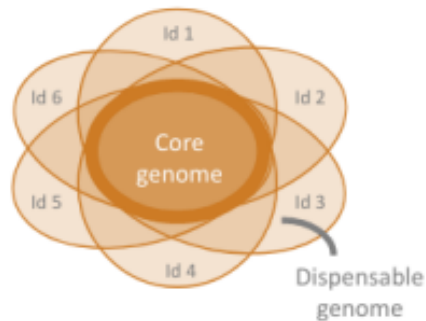
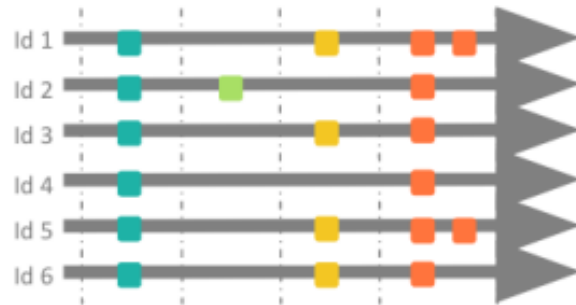
Program: FastANI



Heat map of the average nucleotide identity (ANI) for strains of the species *B. cytotoxicus* (Stevens et al., 2019)

7) Pan-genome and Gene clustering

Pangenome concept



Pangenome

Collection of genes or sequences found in all individuals of a population (intra or inter species)

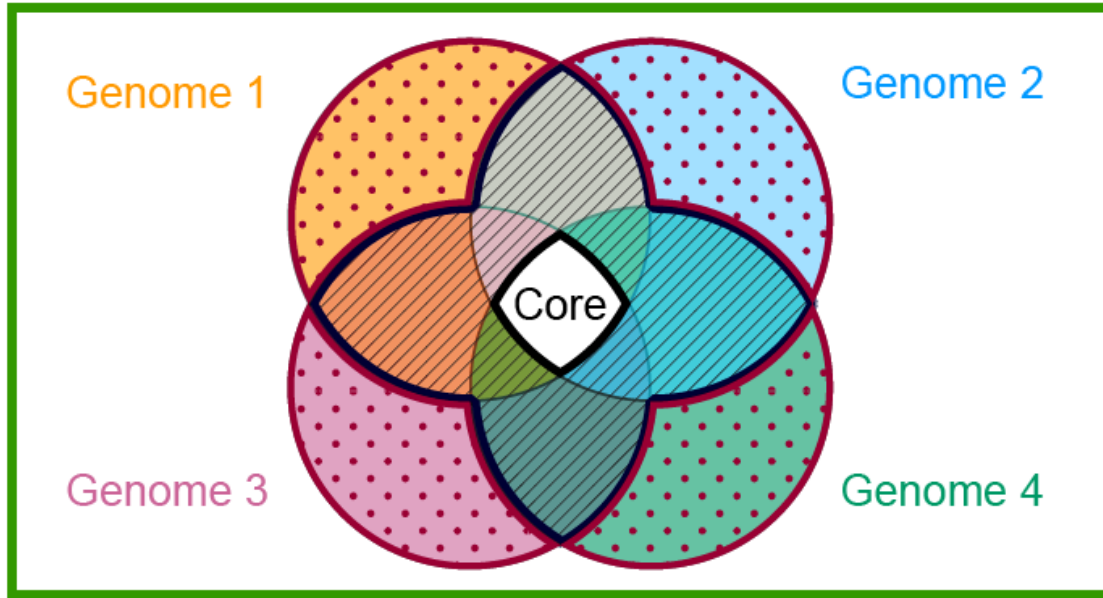
- ▶ **Core genome** : present in all individuals
- ▶ **Dispensable genome** : absent from one or several individuals (also called variable, accessory,...)

Gene Clustering - how it works

- Assess the similarity of every gene to every other gene
 - e.g., using BLAST
- Use that similarity to join pairs of genes
 - e.g., using Reciprocal Best Hits
- Connect the gene pairs into larger clusters
 - e.g., using Reciprocal Best Hits or Markov clustering

=> Programs: [OrthoMCL](#), [Roary](#), PGAP...

Pangenome



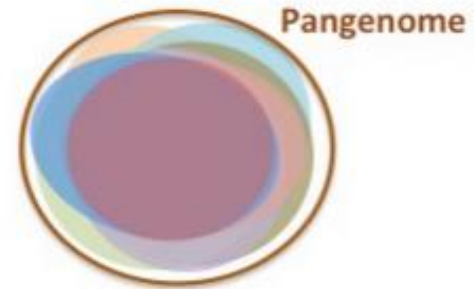
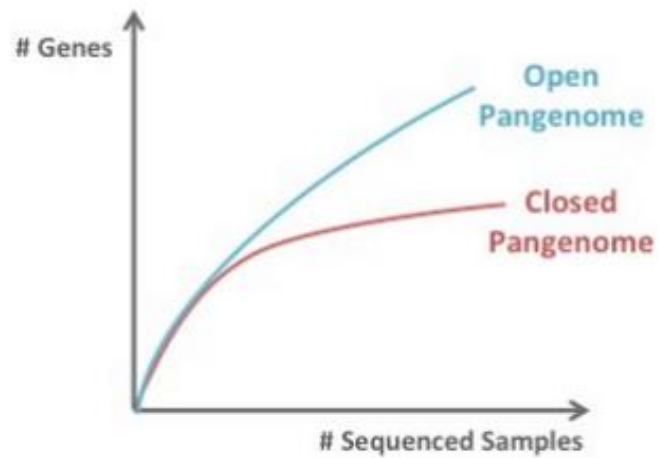
Cloud genome



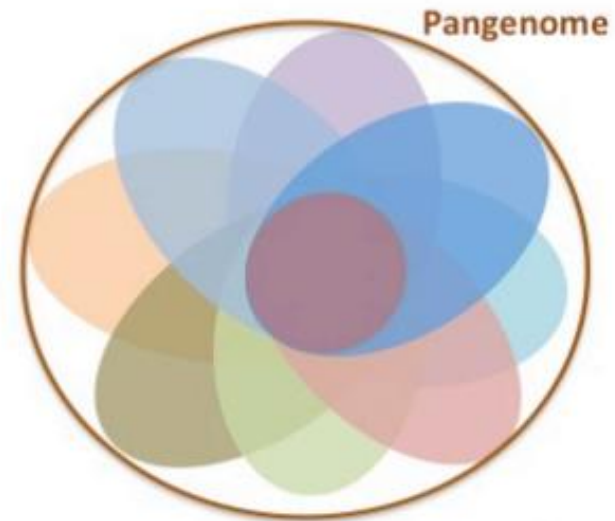
Shell genome



Accessory genome
=
Dispensable genome



High (Core / Pangenome)



Low (Core / Pangenome)

Table 1. Popular software for evolutionary pangenomics

Name	Authors	Reference
Panseq	Laing et al. (2010)	[12]
PanCGHweb	Bayjanov et al. (2010)	[13]
CAMBer	Wozniak et al. (2011)	[14]
PGAT	Brittnacher et al. (2011)	[15]
PGAP	Zhao et al. (2012)	[16]
GET_HOMOLOGUES	Contreras-Moreira and Vinuesa (2013)	[17]
GET_HOMOLOGUES-EST	Contreras-Moreira et al. (2017)	[18]
PanTools	Sheikhzadeh et al. (2016)	[19]
EDGAR 2.0	Blom et al. (2016)	[20]
PanX	Ding et al. (2018)	[21]
Micropan	Snipen and Liland (2015)	[22]
FindMyFriends	Pedersen (2015)	[23]
Piggy	Thorpe et al. (2018)	[24]
PanViz	Pedersen et al. (2017)	[25]

Method	Software	Input	Graph output	Pan-genome	Sequence homology	Paralogue identification
Roary	Conda package	GFF3	DOT	Directed graph	BLAST	Synteny
(v3.13.0)						
Ptolemy	Java executable	FASTA+GFF	GFA	Directed graph	minimap2	Graph-based
(v1.0)						
PPanGGoLin	Conda package	GBK or FASTA	GEXF	Undirected graph	MMseq2	Synteny
(v1.0.13)						
PIRATE	Conda package	GFF3	GFA	Directed graph	BLAST (/DIAMOND)	Synteny
(v1.0.3)						
Panaroo	Conda package	GFF3	GML	Directed graph	CD-HIT	Synteny
(v1.1.2)						

MICROBIAL GENOMICS

Volume 7, Issue 11

Research Article | Open Access

A comparative study of pan-genome methods for microbial organisms: *Acinetobacter baumannii* pan-genome reveals structural variation in antimicrobial resistance-carrying plasmids 

Aysun Urhan¹ , Thomas Abeel^{1,2} 

[Main](#)
[Samples](#)
[Bins](#)
[Legends](#)

Search with expression

Search functions

Search terms:

You can separate multiple search terms with "*,*"

625 result(s) found.

Search gene clusters using filters

Highlight splits on the tree

Append splits to selected bin
 Remove splits from selected bin

Search results (clear)

Item Name	Annotation
GC_0000132	Gene caller id: 28 Source: COG_FUNCTION 8 Accession: COG0451 Annotation: Nucleoside-diphosphate-sugar epimerase
GC_0000064	Gene caller id: 167 Source: COG_FUNCTION 2 Accession: COG1208 Annotation: NDP-sugar pyrophosphorylase, includes eIF-2Bgamma, eIF-2Bepsilon, and LPS biosynthesis proteins
GC_0000007	Gene caller id: 278 Source: COG_FUNCTION 1 Accession: COG1489 Annotation: DNA-binding protein, stimulates sugar fermentation
GC_0000325	Gene caller id: 325

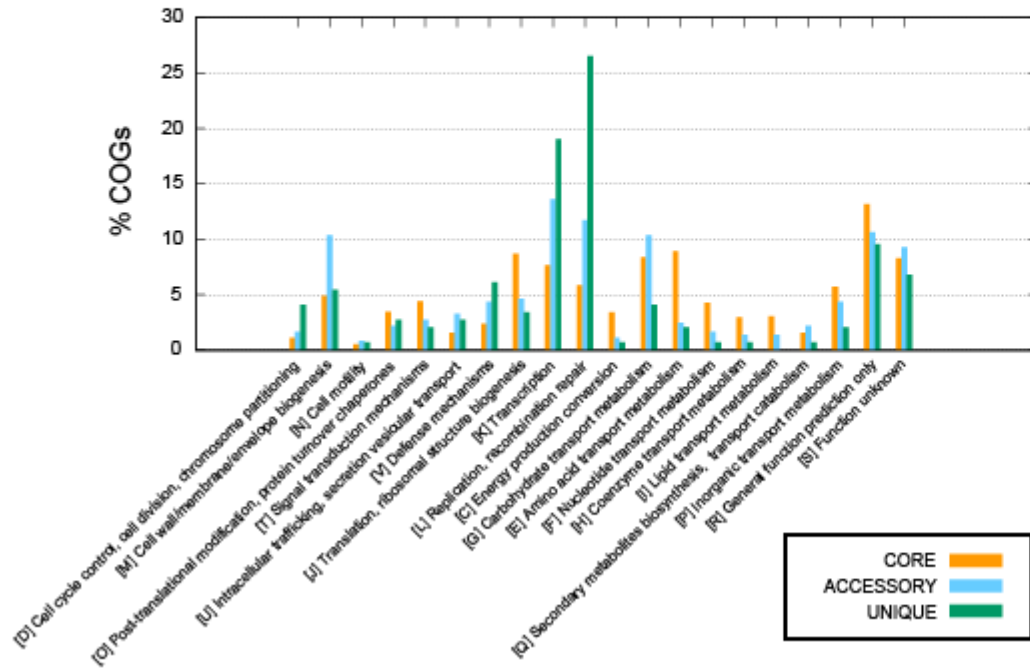
Prochlorococcus Pan

Items order: Presence absence (D: Euclidean; L: Ward) | Current view: gene_cluster_presence_abi

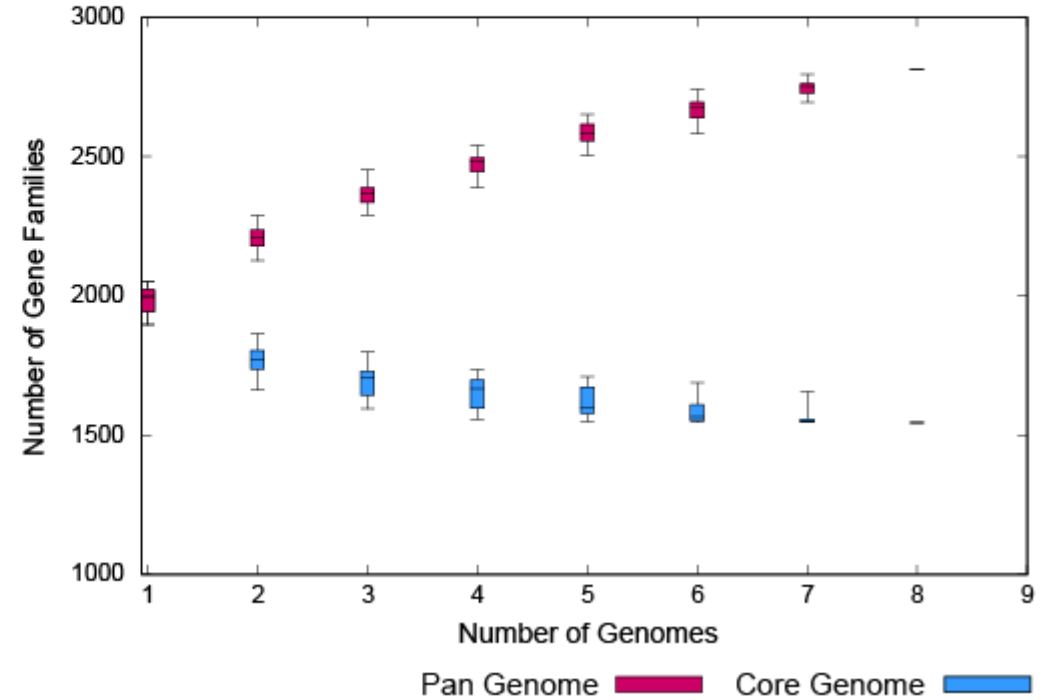
BPGA (Bacterial Pan Genome Analysis tool)

Streptococcus agalactiae

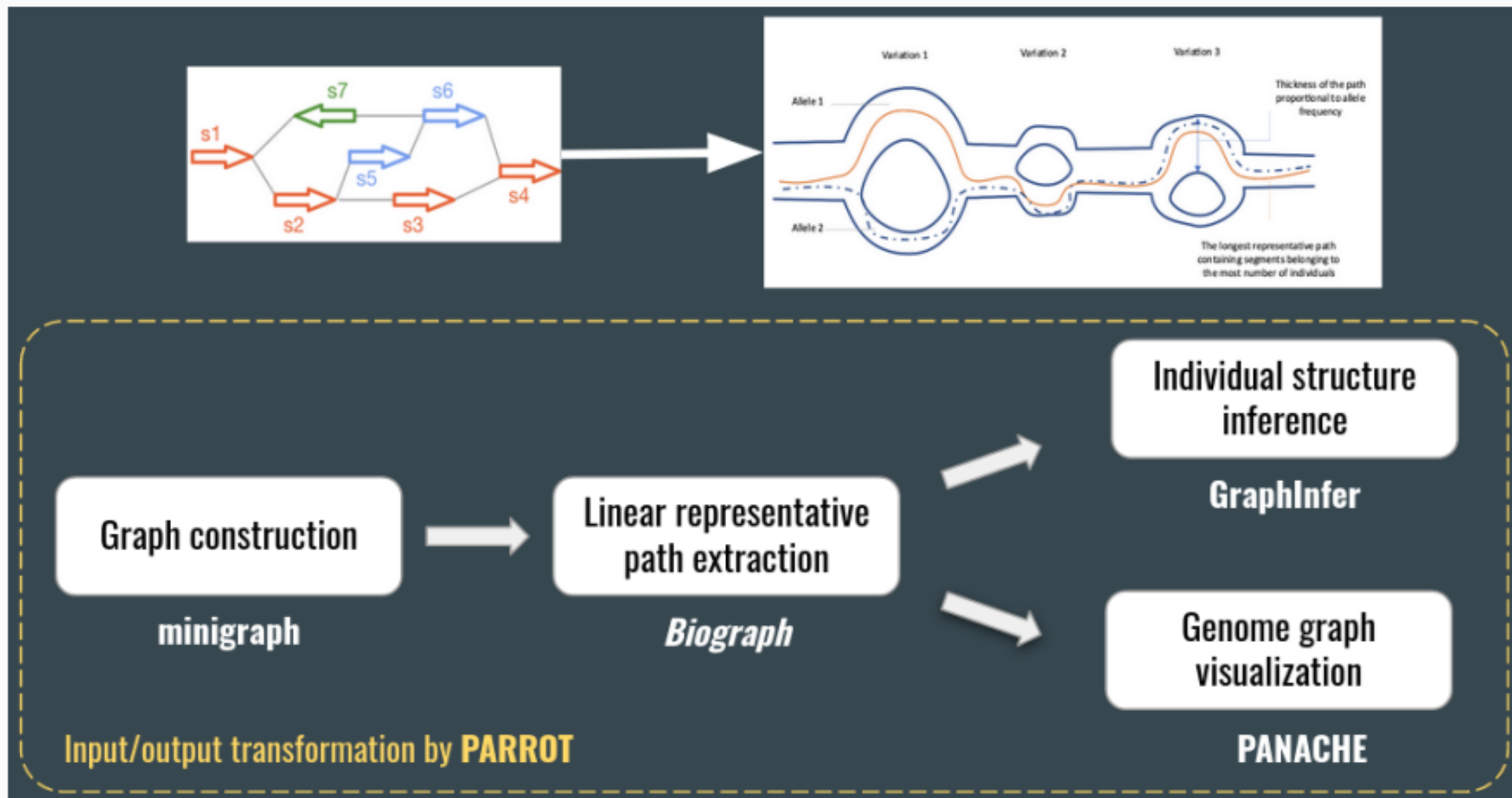
COG Distribution



Pan and Core Genome Plot



Comment manipuler le graphe pour les biologistes ?

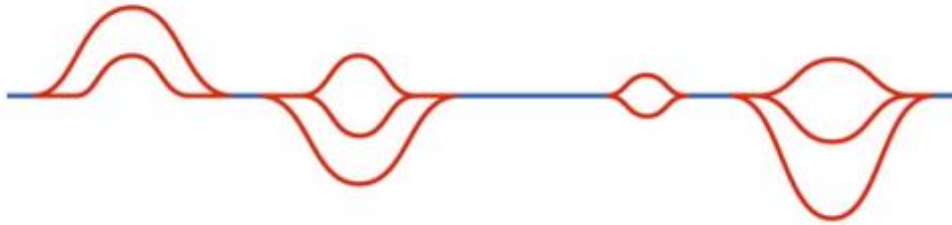


Concept du graphe de génome

Alignment of de novo assembled genomes



Pan-genome graph

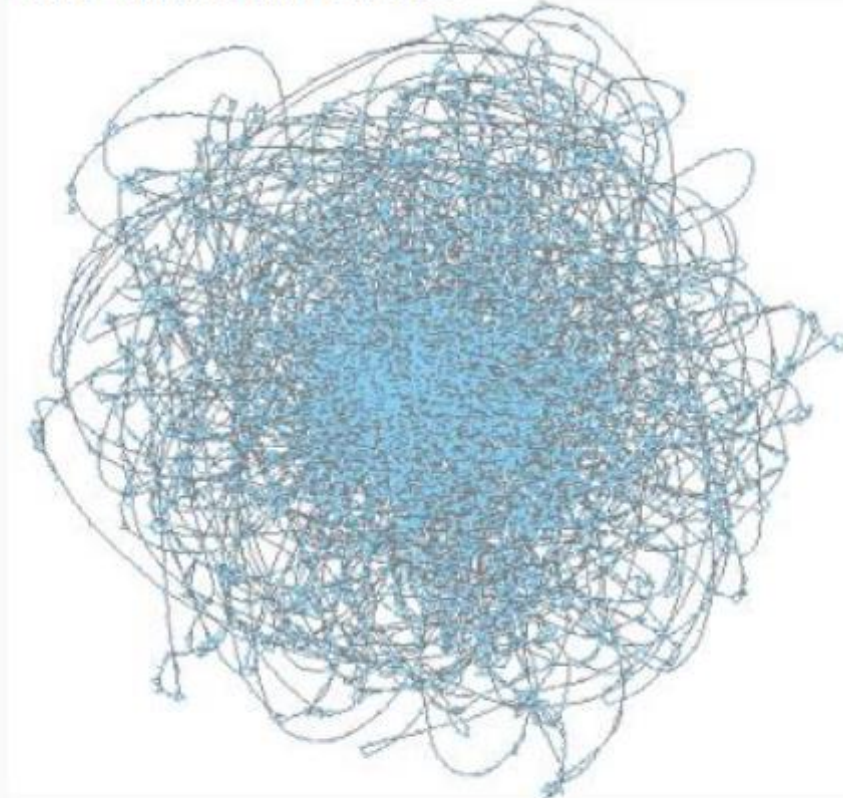


■ Dispensable genome

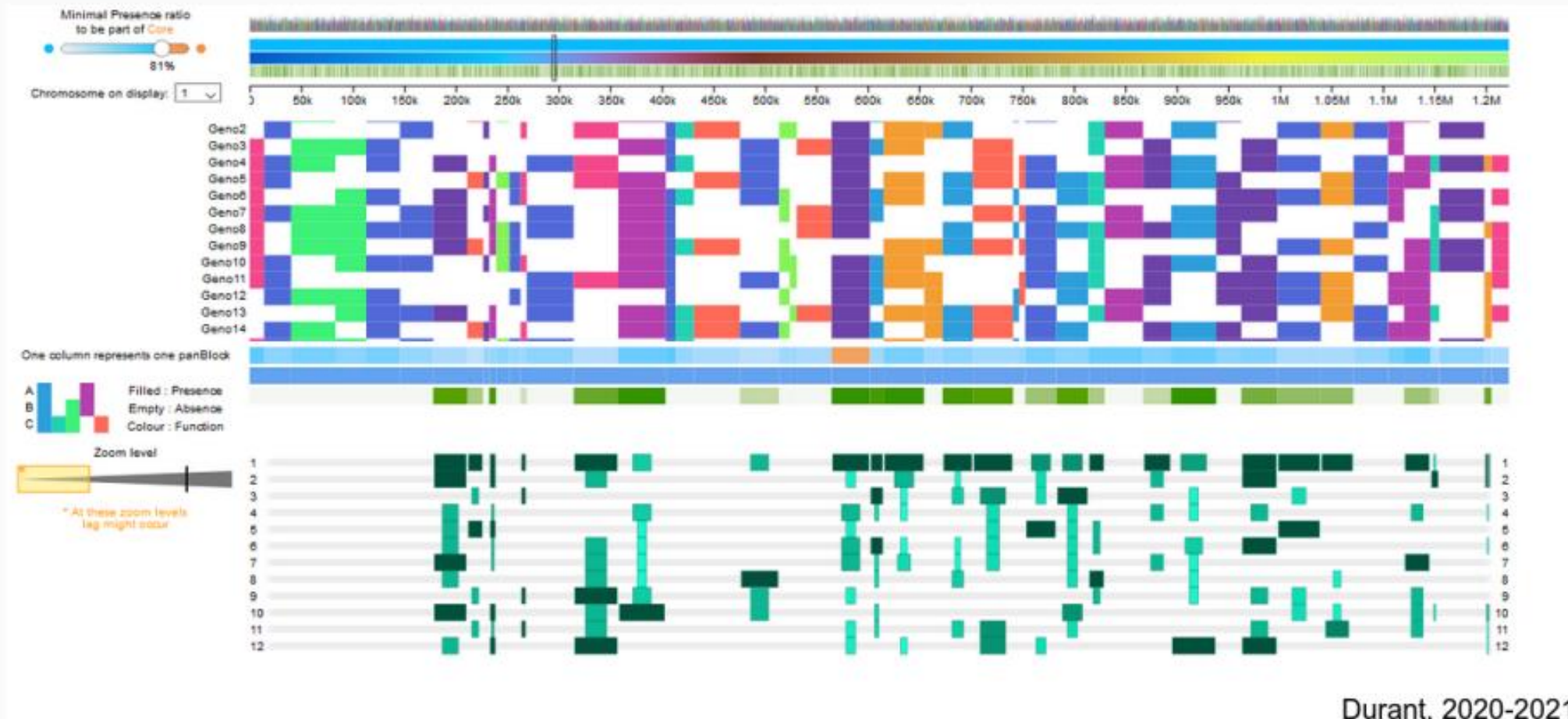
■ Core genome

Bayer et al., 2020

The HairBall effect



Un exemple linéaire, Panache



8) Pan-GWAS

Pan-GWAS

Pan-GWAS of *Streptococcus agalactiae* Highlights Lineage-Specific Genes Associated with Virulence and Niche Adaptation

Authors: Andrea Gori, Odile B. Harrison, Ethwako Mlia, Yo Nishihara, Jia Mun Chan, Jacqueline Msefula, Macpherson Mallewa, SHOW ALL (3 AUTHORS), Robert S. Heyderman | AUTHORS INFO & AFFILIATIONS

DOI: <https://doi.org/10.1128/mBio.00728-20> • [Check for updates](#)

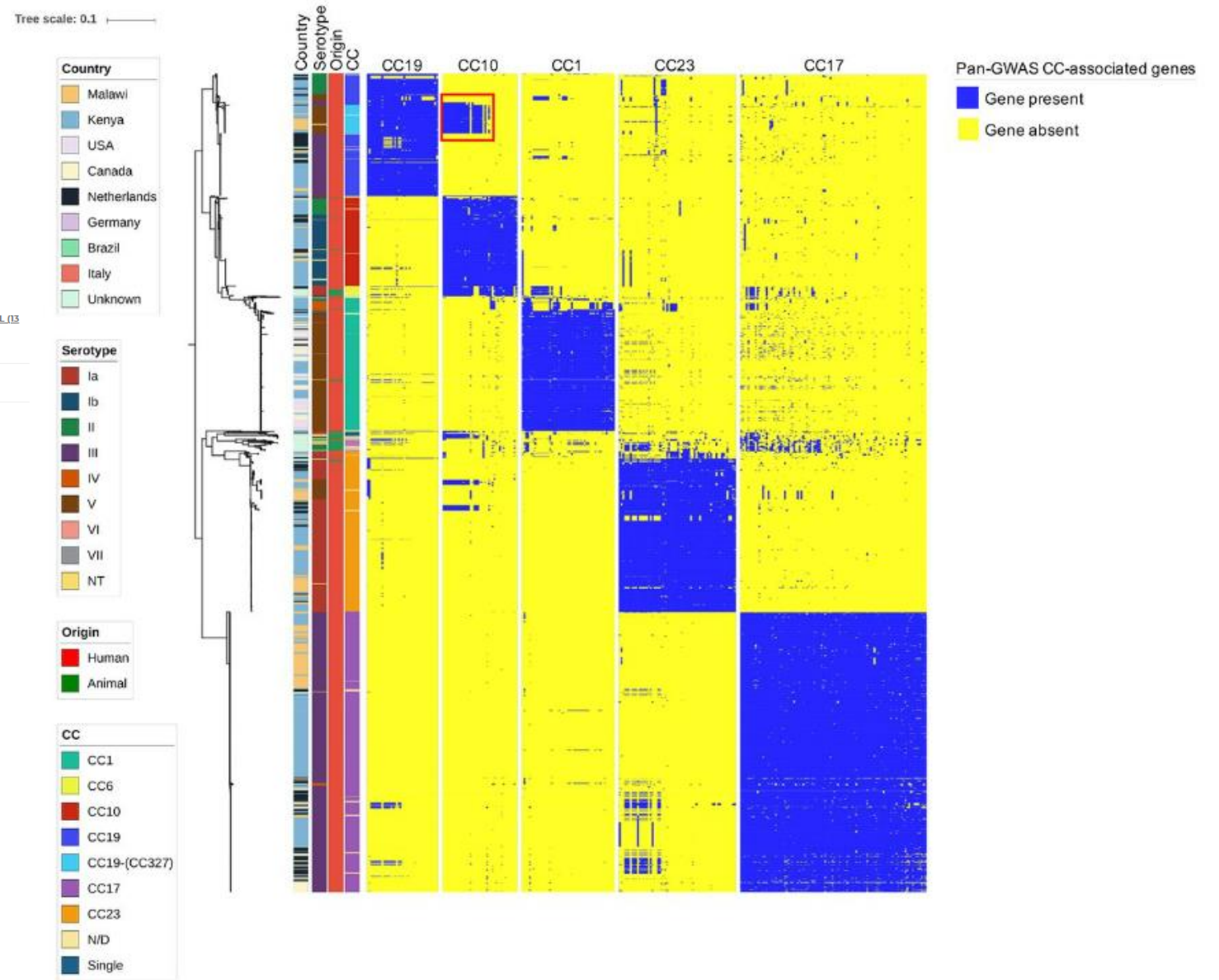


FIG 2 Core genome-based population structure of GBS. The phylogenetic tree is annotated with 4 colored strips representing the clonal complex, the country of isolation, the origin, and the serotype of each strain. The three binary heatmaps represent the presence (blue) or absence (yellow) of the genes identified by the pan-GWAS pipeline. The tree is rooted at midpoint. The reference strain used in this analysis was COH1, reference HG939456. The red square in the CC10 heatmap highlights the cluster of CC10-associated genes found in CC19 clones. Trees built with different reference strains are shown in Fig. S1 in the supplemental material and show analogous topology.



“*Scoary is designed to take the gene_presence_absence.csv file from [Roary](#) as well as a traits file created by the user and calculate the associations between all genes in the accessory genome and the traits. It reports a list of genes sorted by strength of association per trait.*”

=> Provides:

odds ratios

Un *odds ratio* :

< 1 signifie que l'événement est moins fréquent dans le groupe A que dans le groupe B ;

= 1 signifie que l'événement est aussi fréquent dans les deux groupes ;

> 1 signifie que l'événement est plus fréquent dans le groupe A que dans le groupe B.

p-value and p-value adjusted with Bonferroni's method

The `traits.csv` file needs to be formatted in a specific way.

- It must use the same delimiter as the `gene_presence_absence.csv` file
- The names of your isolates need to be identical in the two files
- The rows should correspond to your isolates, the columns to the different traits
- Traits needs to be dichotomized. Use "0" to indicate absence and "1" to indicate presence of the trait
- All isolates and traits should be uniquely named and not contain any weird characters (e.g. %; /&[]@? etc)
- The top left cell should be left blank

It should look something like this:

	Trait1	Trait2	...	TraitM
Strain1	1	0	...	1
Strain2	1	1	...	0
Strain3	0	0	...	1
...
StrainN	1	0	...	0

Merci pour votre attention !

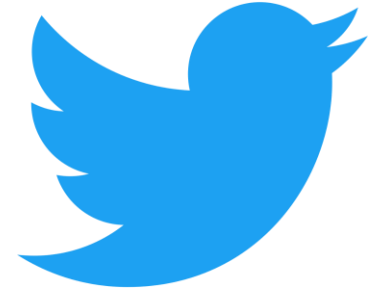


Le matériel pédagogique utilisé pour ces enseignements est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions (BY-NC-SA) 4.0 International:

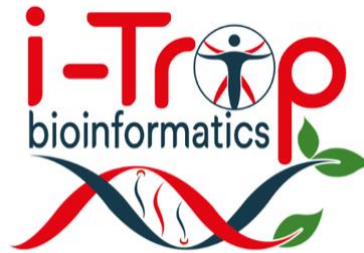
<http://creativecommons.org/licenses/by-nc-sa/4.0/>



SUIVEZ NOUS SUR TWITTER !



South Green : [@green_bioinfo](#)



I-Trop : [@ltropBioinfo](#)

N'oubliez pas de nous citer !

Comment citer les clusters?

"The authors acknowledge the IRD i-Trop HPC at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: <http://bioinfo.ird.fr/> "

"The authors acknowledge the CIRAD UMR-AGAP HPC (South Green Platform) at CIRAD montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: <http://www.southgreen.fr>"