



# Session de formation 2023



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

réseaux  
génomique  
ressources  
Infrastructure  
montpelliérain  
plantes  
internationale  
orienté  
développement  
service  
calcul  
multi-instituts  
comptences  
végétale  
plateforme  
d'analyses  
communauté  
outils  
s'appuie  
mutualisation  
partage



SNP detection  
phylogeny  
comparative genomics  
GWAS  
population genetics  
structural variation  
transcriptome assembly  
differential expression  
polypliody

### Mutualisation



Cacao

Banana

Coffee

Rice

Palm

Cassava

*Pseudocercospora*

*Magnaporthe*

# South Green

bioinformatics platform



4 institutes



25+



3 research units



Tools

Storage and computing  
resources



400+

Trainings



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

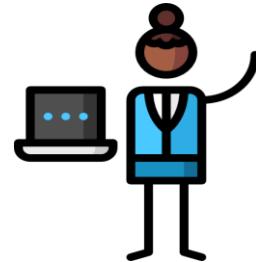
Data mining

## Diversity exploration

genotype manipulation

mosaic manipulation

## Metagenomic



+20  
tools

web applications (16)

visualisation (8)

workflows(5)

packages (4)



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



Plant & Health Bioinformatics Platform



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



AURORE  
COMTE



JACQUES  
DAINAT



ALEXIS  
DEREOPER



BRUNO  
GRANOUILLAC



JULIE  
ORJUELA-



NDOMASSI  
TANDO



CHRISTINE  
TRANCHANT

[bioinfo@ird.fr](mailto:bioinfo@ird.fr)



[@IItropBioinfo](https://twitter.com/IItropBioinfo)





Florian Charriat  
Antoni Exbrayat



Guilhem Sempere



Bruno Granouillac  
Jacques Dainat



Nicolas Fernandez



Thomas Denecker

And more collaborators !

# South Green

bioinformatics platform



Formations 2023  
Montpellier

6-7 Avril

Guide de survie à linux  
Agropolis, salle Badiane

17-20 Avril

Python  
Agropolis, salle Badiane

15-16 Mai

Linux avancé  
Agropolis, salle Passiflore

25-26 Mai

Introduction à l'analyse de données Oxford Nanopore  
Agropolis, salle Badiane

22-23 Mai

Utilisation avancée d'un cluster de calcul  
Agropolis, salle Badiane puis B03

31 Mai-01 Juin

Initiation aux analyses de données transcriptomiques  
Agropolis, salle Badiane

08-09 Juin

Génomique bactérienne comparative  
Agropolis, salle Badiane

14-16 Juin

Recherche Reproductible  
Agropolis, salle Badiane



Plant  
Health  
Institute  
Montpellier



# Modules de formation 2023

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

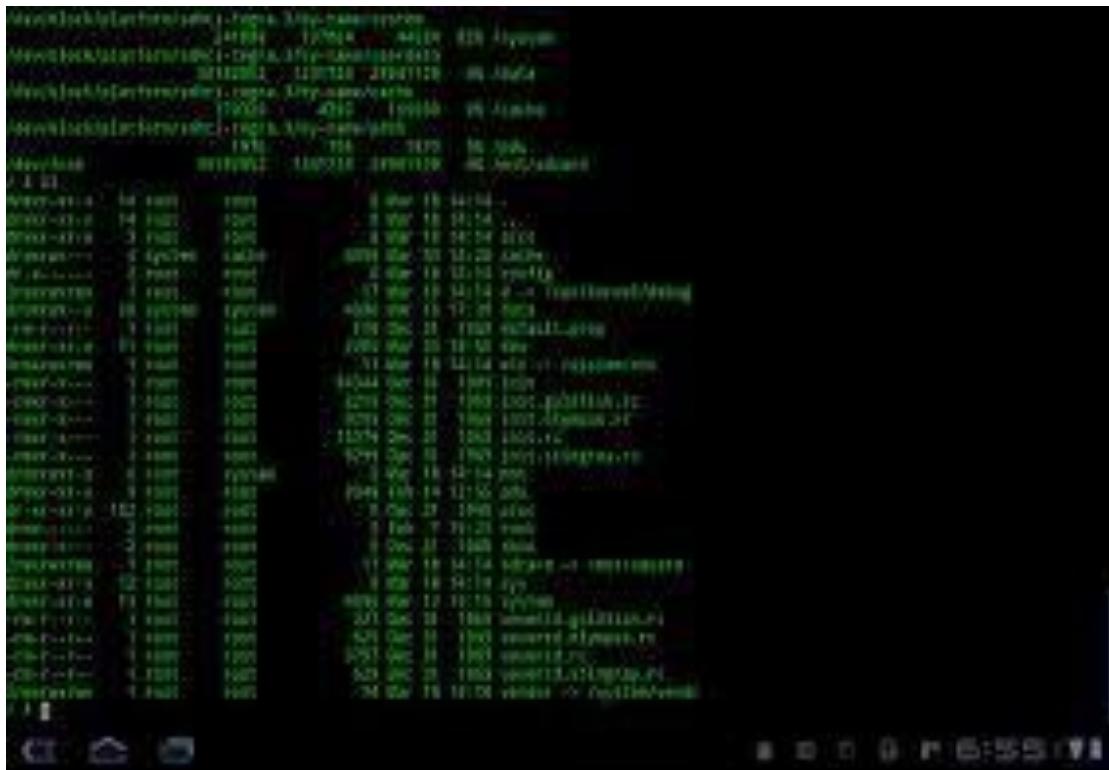


# Génomique Comparative Bactérienne



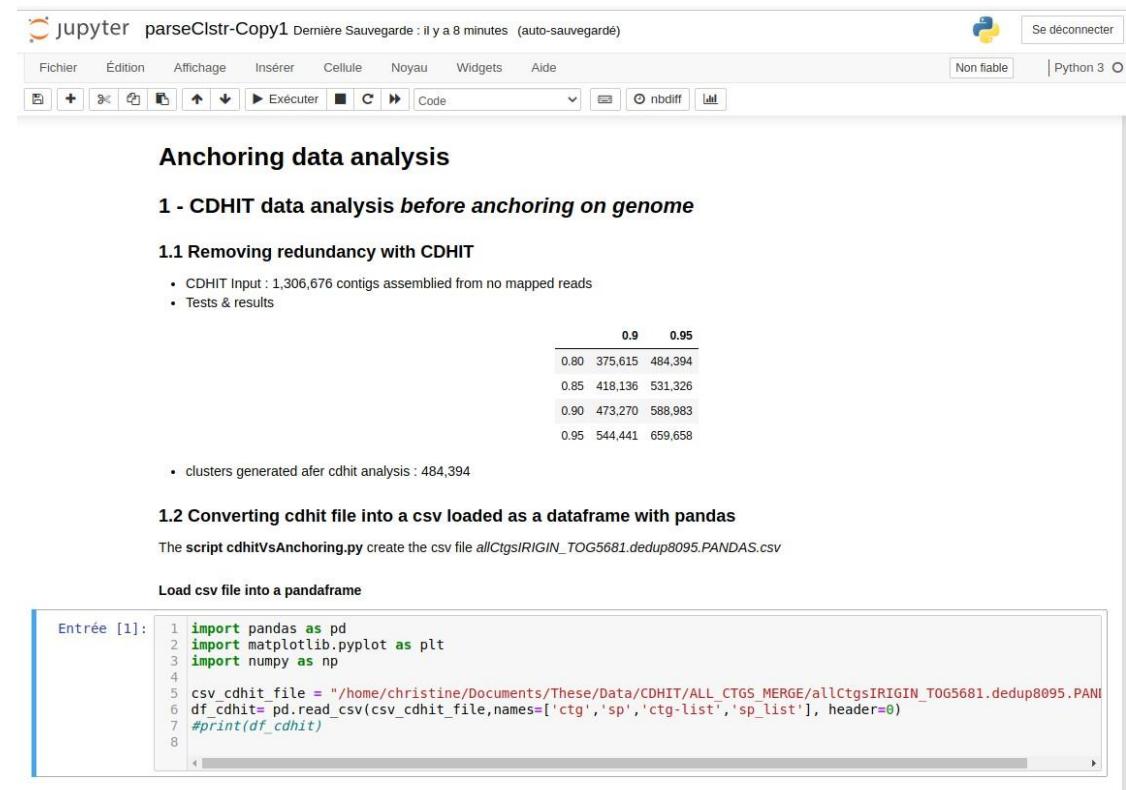
- 2 façons d'utiliser linux :

en *mode console*



A screenshot of a Linux terminal window. The window title is "Terminal". The background is black with white text. The terminal displays a large amount of green and white text, which appears to be command-line output or log data. At the bottom of the terminal window, there is a standard Linux desktop taskbar with icons for file, home, search, and other applications.

en *mode jupyter notebook*



A screenshot of a Jupyter Notebook interface. The title bar says "jupyter parseCstr-Copy1 Dernière Sauvegarde : il y a 8 minutes (auto-sauvegardé)". The menu bar includes Fichier, Édition, Affichage, Insérer, Cellule, Noyau, Widgets, and Aide. The toolbar includes icons for file operations, execution, and code. The main content area is titled "Anchoring data analysis". It contains sections 1.1 and 1.2, and a code cell labeled "Entrée [1]".

**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 creates the csv file allCtgtsIRIGIN\_TOG5681.dedup8095.PANDAS.csv

Load csv file into a pandasframe

```
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_Ctgts MERGE/allCtgtsIRIGIN_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 used through a simple web browser

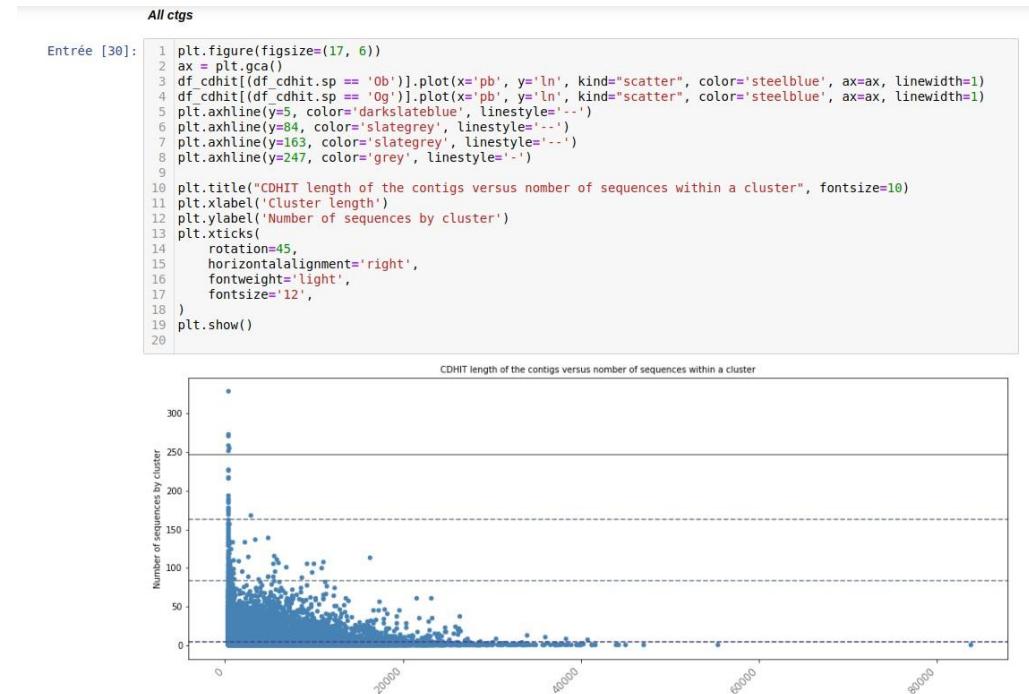
ex : Bash (Linux), Python, Java, R, Julia, Matlab, Octave, Scheme, Processing, Scala



# What is jupyter book ?

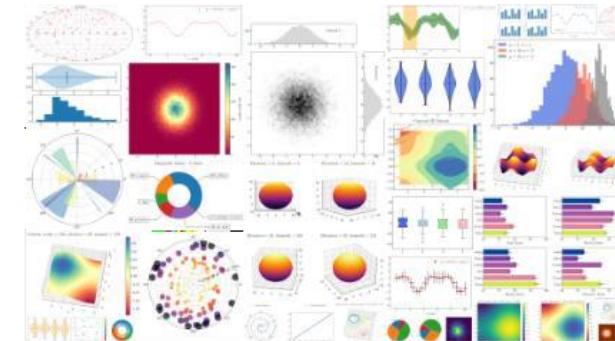
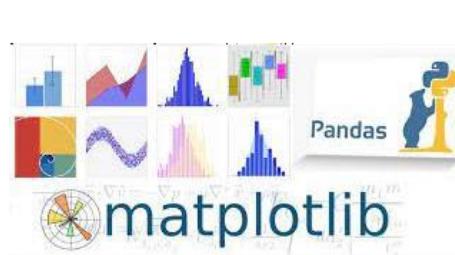
- An unique interface/file where text, code and output codes can be mixed :

- 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

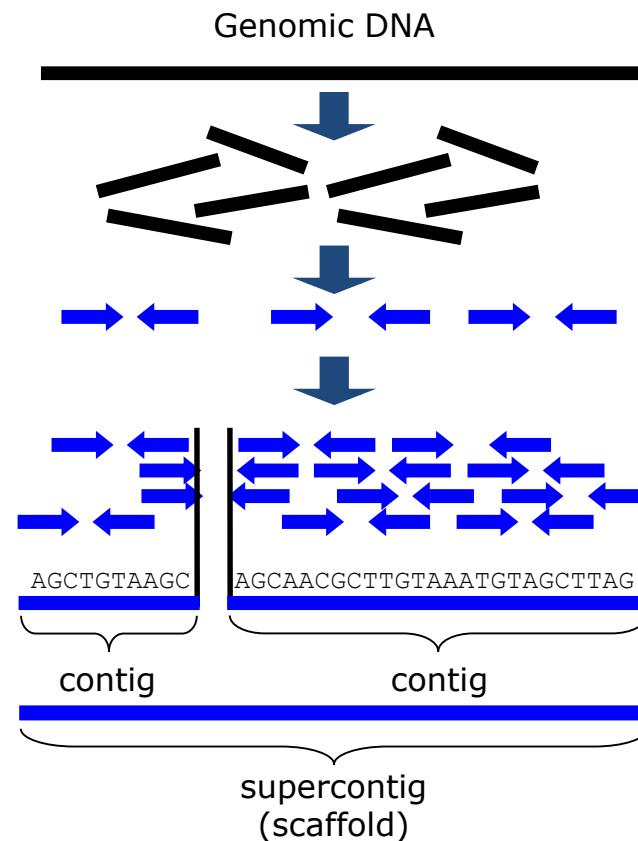
1. Assemble each set of reads into a genome sequence
2. Annotate each genome
3. Cluster genes and compare between each genome

## Variant-based

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

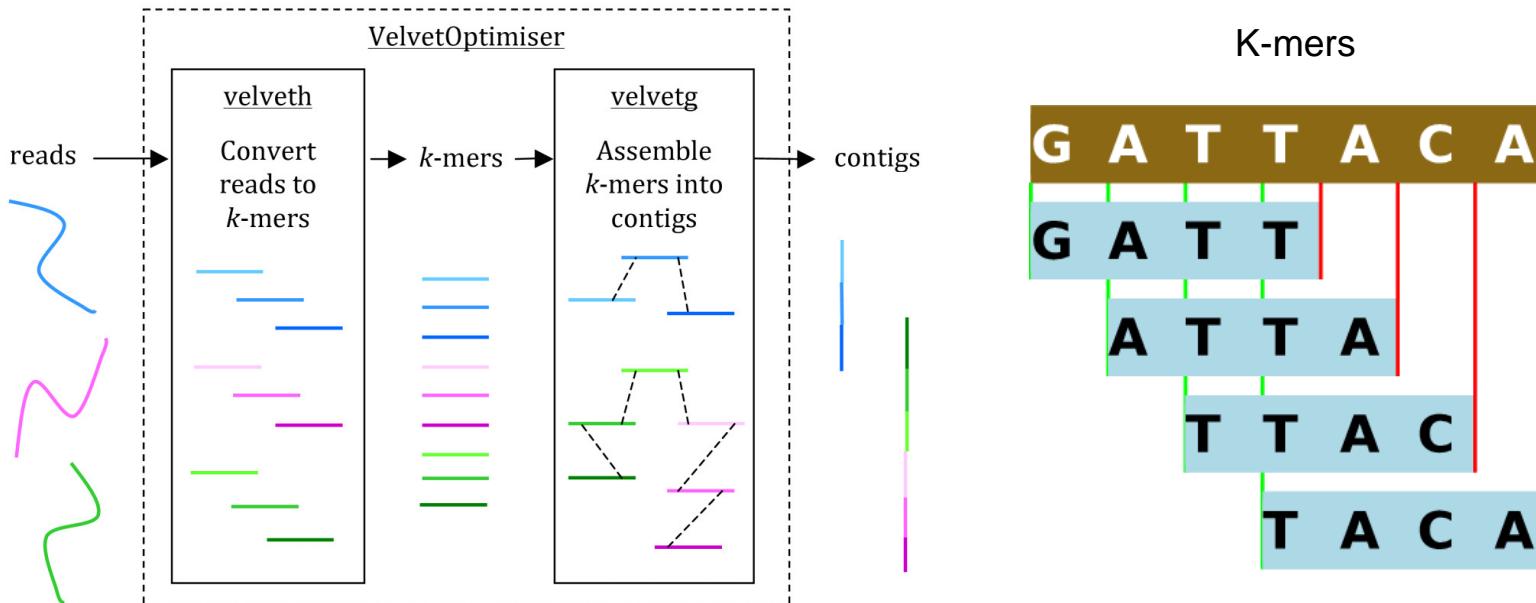
# 1) Assembly

# 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<sup>1,2,†</sup>, Dike Jiang<sup>1,2,†</sup>, Yin Wang<sup>1,2,\*</sup>, Xueping Yao<sup>1,2</sup>, Yan Luo<sup>1,2</sup> and Zexiao Yang<sup>1,2</sup>

Table 1

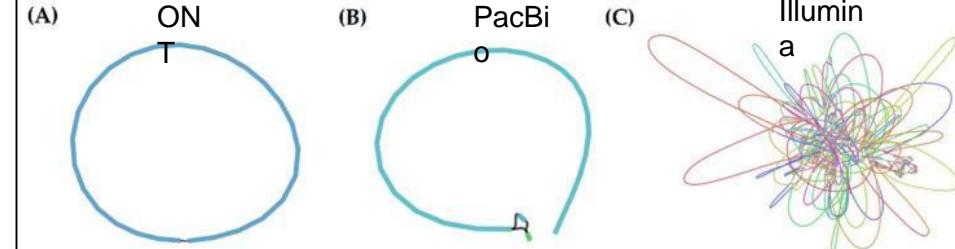
Statistics of genome-assembly results of independent assembly strategies.

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

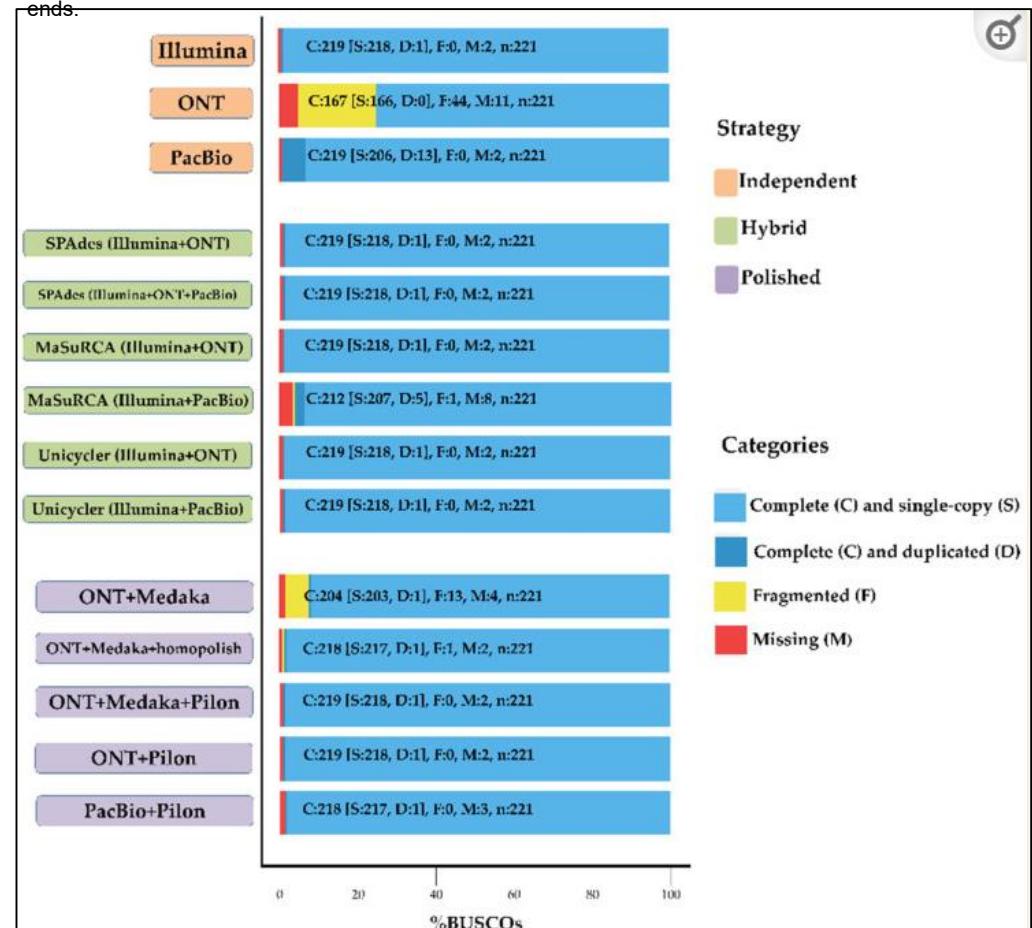
Table 2

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

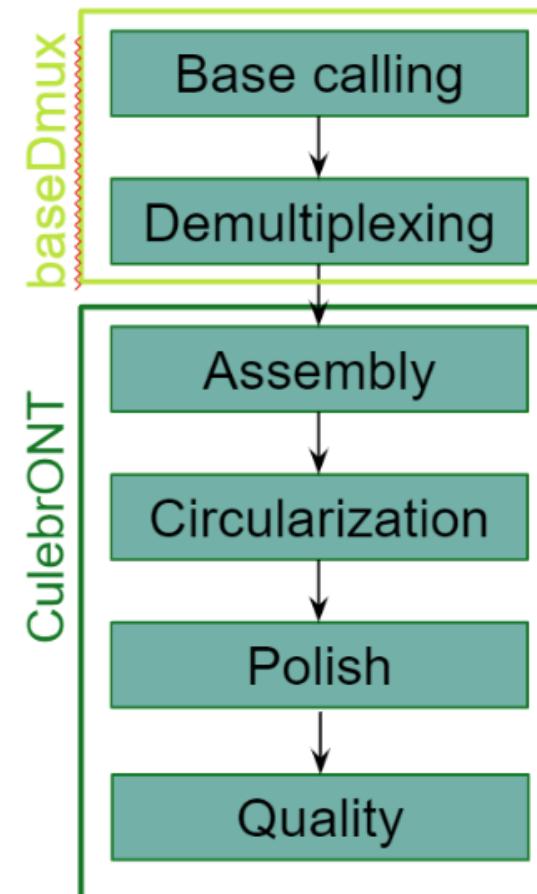
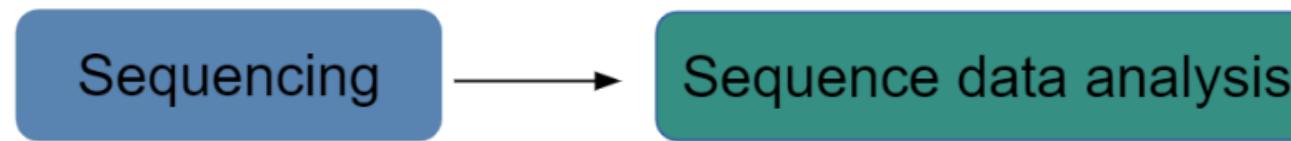


Comparison of results of independent assembly strategies. (A) Genome assembled with nanopore reads; (B) longest contig assembled with PacBio reads; (C) genome assembled with Illumina reads. Plots were obtained by using Bandage on the “assembly\_graph.gfa” output file from SPAdes or the “contig.gfa” output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

### 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.

## MICROBIAL GENOMICS

### METHODS PAPER

Robertson and Nash, *Microbial Genomics* 2018;4  
DOI 10.1099/mgen.0.000206



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

James Robertson<sup>1</sup> and John H. E. Nash<sup>2,\*</sup>

### 3) Genome Annotation

# What is annotation ?

Structural annotation:

**VS**

Functional annotation:

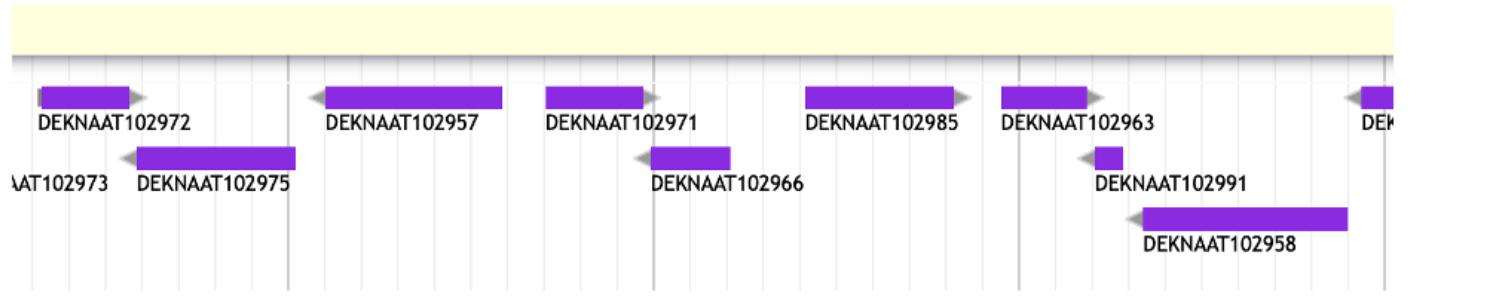
Find out where the regions of interest (usually genes) are in the sequence data and what they look like.

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

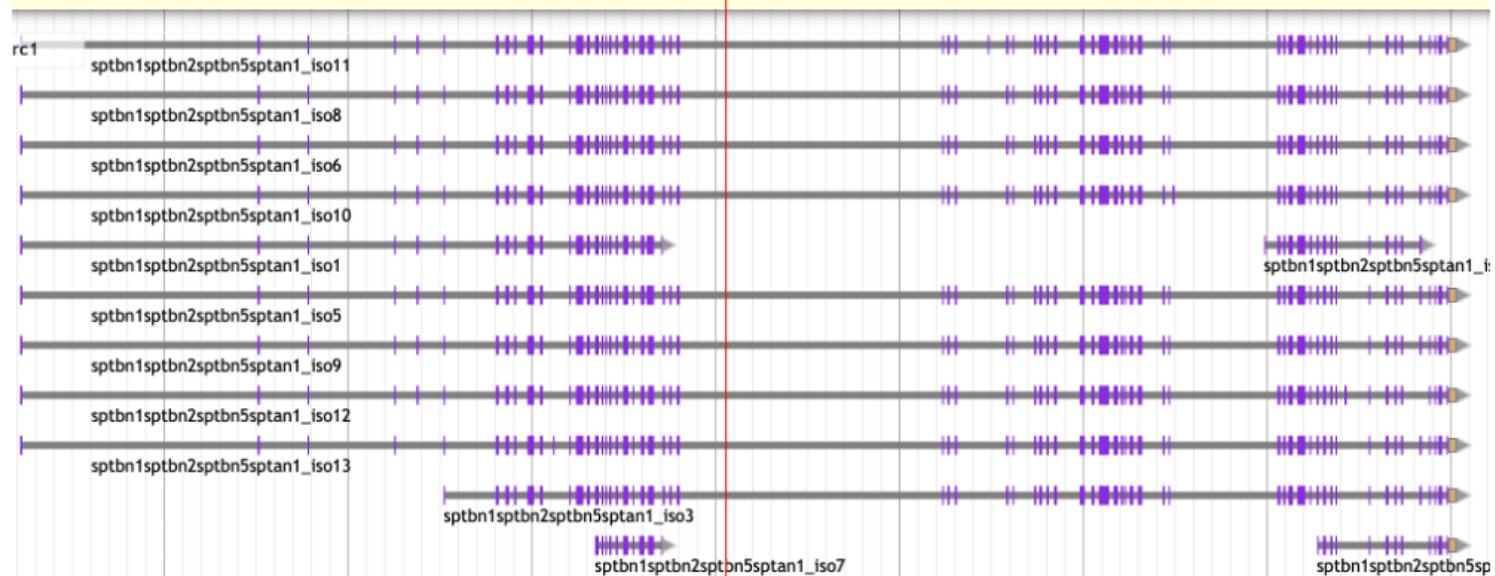
*It is the **annotation** that bridges the gap from the sequence to the biology of the organism*

# 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

(SO term = 2278 possibilities)

9) attributes  
*tag=value*

! Features are grouped by **parent** relationship

# Adding biological info to sequences

ribosome  
binding site

delta toxin  
*PubMed: 15353161*

ACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGA  
AAAGCAGCCTCCTGACTTCCCTCGCTTGGTGGTTGAGTGGACCTC  
CCAGGCCAGTGCCGGGCCCCCTCATAGGAGAGGAAGCTGGGAGGTG  
GCCAGGGCGCAGGAAGGCGCACCCCCCCCAGCAATCCGCGCGCCGGG  
ACAGAATGCCCTGCAGGAATTCTTCTAGAACGACCTTCCTCCTG  
CAAATAAAACCTCACCCATGAATGCTCACGCAAGTTAATTACAGA  
CCTGAAACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCT  
CTCCGTCCGTCCGTGGGCCACGGCCACCGCTTTTTTTGCC

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

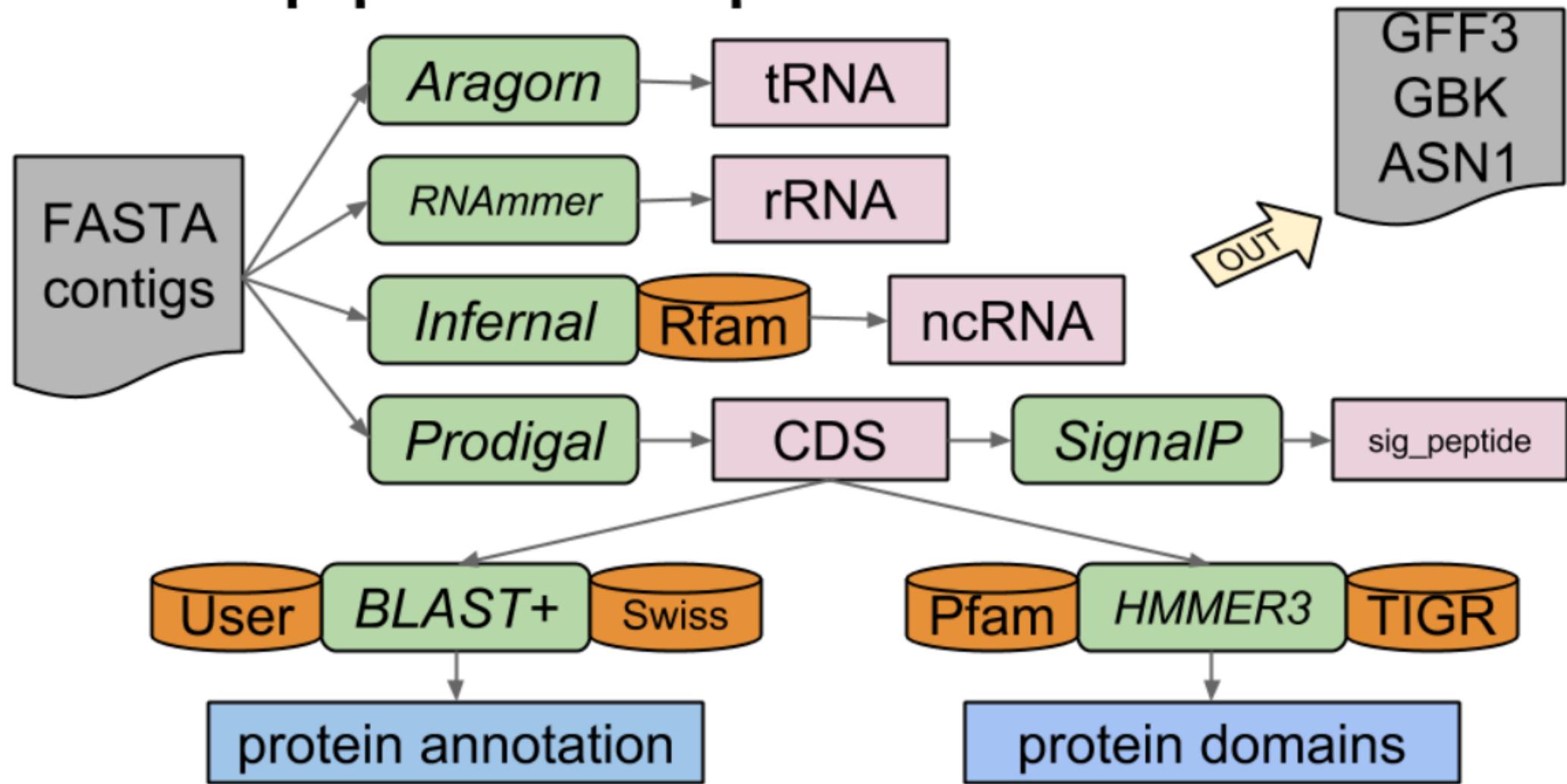
## Prokaryote annotation:

- 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

National Library of Medicine  
National Center for Biotechnology Information

Search NCBI Search

## Genomes – NCBI Datasets BETA

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

TAXONOMIC NAME:  1

STATUS:  reference genomes  annotated 3

ASSEMBLY LEVEL: contig scaffold chromosome 2

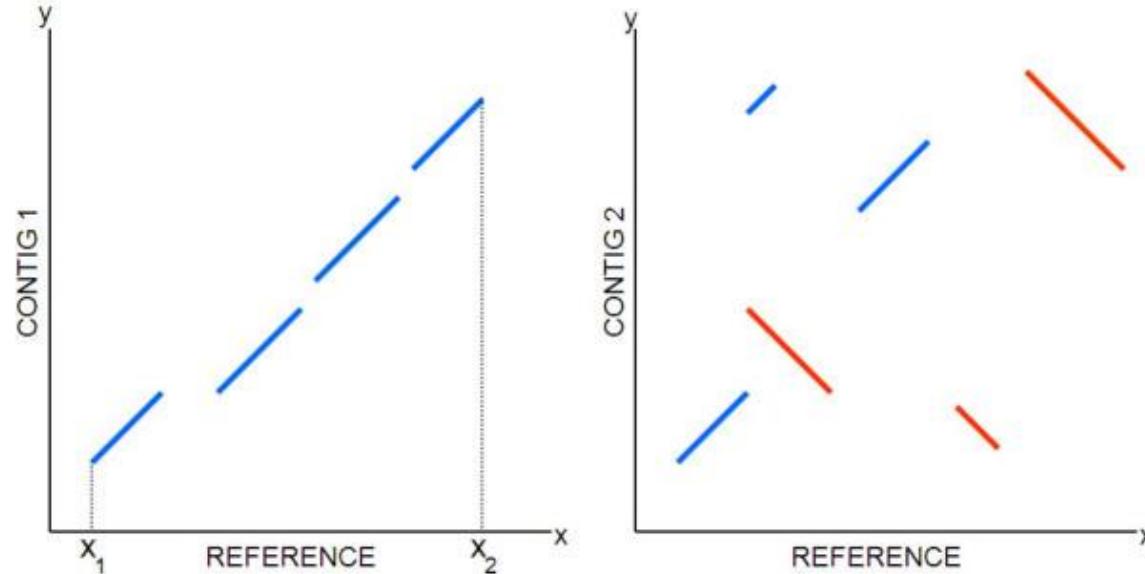
YEAR RELEASED:  4

Download table 5

Assembly Name	Assembly Accession	Organism	Assembly Inf	Annotation	Assembly Sub	Assembly Len	Assembly Submission Date
1. ASM802v1	GCA_000008025.1	Wolbachia endosymbiont wMel		Annotation s	326792	Complete Ge	05/07/2006
2. ASM1194v1	GCA_000011945.1	Anaplasma marginale str. TSE, Mares		Annotation s	1291787	Complete Ge	02/07/2007
3. ASM1202v1	GCA_000012025.1	Anaplasma phagocytophila		Annotation s	1291787	Complete Ge	06/12/2006
4. ASM1318v1	GCA_000013185.1	Ehrlichia canis str. Jake		Annotation s	1315030	Complete Ge	11/08/2005
5. ASM1319v1	GCA_000013195.1	Anaplasma phagocytophila		Annotation s	3471282	Complete Ge	21/03/2006
6. ASM1318v1	GCA_000013185.1	Ehrlichia chaffeensis str. Arkansas		Annotation s	1176248	Complete Ge	21/03/2006
7. ASM1318v1	GCA_000013185.1	Neorickettsia helvetica str. Miyazaki		Annotation s	839906	Complete Ge	21/03/2006
8. ASM2036v1	GCA_000020365.1	Anaplasma marginale str. Florida		Annotation s	1302483	Complete Ge	03/07/2009
9. ASM2228v1	GCA_000022285.1	Wolbachia sp. wili		Annotation s	1845373	Complete Ge	24/03/2009
10. ASM2228v1	GCA_000022285.1	Neorickettsia ritaense str. illinois		Annotation s	879977	Complete Ge	23/07/2009
11. ASM2232v1	GCA_000022325.1	Anaplasma centrale str. Isr Israel		Annotation s	1206806	Complete Ge	24/11/2009
12. ASM2405v1	GCA_000024055.1	Ehrlichia ruminantium str. Welgevonden		Annotation s	1318355	Complete Ge	05/11/2005
13. ASM2504v1	GCA_000025045.1	Ehrlichia ruminantium str. Gassel		Annotation s	1206806	Complete Ge	05/11/2005
14. ASM2504v1	GCA_000025045.1	Wolbachia ruminantium str. Welgevonden		Annotation s	1512997	Complete Ge	31/12/2005
15. ASM2504v1	GCA_000025045.1	Wolbachia endosymbiont wRi		Annotation s	1482454	Chromosome	23/04/2006
16. ASM2668v1	GCA_000026685.1	Wolbachia endosymbiont wOr		Annotation s	957990	Complete Ge	30/11/2012
17. ASM3765v1	GCA_000376545.1	Wolbachia endosymbiont wNo		Annotation s	1301823	Complete Ge	22/04/2013
18. ASM3766v1	GCA_000376625.1	Wolbachia endosymbiont wHa		Annotation s	1293804	Complete Ge	22/04/2013
19. ASM3975v1	GCA_000439755.1	Anaplasma phagocytophila		NCBI Prokary	5477545	Complete Ge	24/07/2013
20. ASM3975v1	GCA_000439775.1	Anaplasma phagocytophila		NCBI Prokary	3481398	Complete Ge	24/07/2013
21. ASM3975v1	GCA_000439795.1	Anaplasma phagocytophila		NCBI Prokary	1471302	Chromosome	24/07/2013
22. ASM4954v1	GCA_000495495.1	Anaplasma marginale str. Gossy Plants		NCBI Prokary	3198622	Chromosome	05/11/2013
23. ASM4954v1	GCA_000495495.1	Anaplasma marginale str. Dawn		NCBI Prokary	1196760	Chromosome	05/11/2013
24. ASM4955v1	GCA_000495535.1	Anaplasma marginale str. AS45		NCBI Prokary	1196760	Chromosome	05/11/2013
25. ASM4955v1	GCA_000495535.1	Anaplasma marginale str. AS45		NCBI Prokary	1177221	Complete Ge	23/04/2014
26. ASM4955v1	GCA_000495535.1	Anaplasma marginale str. AS45		NCBI Prokary	1148904	Complete Ge	23/04/2014
27. ASM4956v1	GCA_000495625.1	Ehrlichia sp. str. wI		Annotation s	1176890	Complete Ge	23/04/2014
28. ASM4956v1	GCA_000495625.1	Ehrlichia chaffeensis str. U.I.Liberty		Annotation s	1176202	Complete Ge	23/04/2014
29. ASM4956v1	GCA_000495625.1	Ehrlichia chaffeensis str. O.Osceola		Annotation s	1175197	Complete Ge	23/04/2014
30. ASM4956v1	GCA_000495625.1	Ehrlichia chaffeensis str. S.Saint Vincent		Annotation s	1173884	Complete Ge	23/04/2014
31. ASM4956v1	GCA_000495625.1	Ehrlichia chaffeensis str. W.Wakula		Annotation s	1174357	Complete Ge	23/04/2014
32. ASM4956v1	GCA_000495625.1	Ehrlichia chaffeensis str. W.West Paes		Annotation s	1170935	Complete Ge	23/04/2014
33. ASM4956v1	GCA_000495625.1	Neorickettsia helvetica Oregon		Annotation s	884232	Complete Ge	17/04/2014
34. ASM4956v1	GCA_000495625.1	Anaplasma phagocytophila, Asym variant2		NCBI Prokary	1345197	Complete Ge	03/05/2018
35. ASM4956v1	GCA_000495625.1	Wolbachia endosymbiont of Brucella suis var.		Annotation s	1296461	Complete Ge	13/10/2018
36. Wv-003	GCA_000499915.1	Wolbachia endosymbiont of Brucella suis var.		Annotation s	1133809	Chromosome	01/07/2018
37. WvF-003	GCA_001499855.1	Wolbachia endosymbiont -wType		NCBI Prokary	1267840	Chromosome	11/10/2016
38. WvF-1_0	GCA_001499855.1	Wolbachia endosymbiont -wType_Cu		NCBI Prokary	1267664	Chromosome	11/10/2016
39. ASM172506v1	GCA_001725065.1	Wolbachia endosymbiont wMef_SM		NCBI Prokary	1805436	Complete Ge	25/06/2018
40. ASM172506v1	GCA_001725065.1	Wolbachia endosymbiont wMef_Berlin		NCBI Prokary	1214874	Complete Ge	05/07/2018
41. ASM171175v2	GCA_001711755.2	Anaplasma ovis str. Haller Haben		NCBI Prokary	1485853	Complete Ge	31/07/2019
42. ASM221462v2	GCA_002214625.2	Anaplasma ovis str. Haller Haben		NCBI Prokary	3482279	Complete Ge	31/07/2019
43. ASM221462v2	GCA_002214625.2	Wolbachia pipiens wAIB-wN2018		NCBI Prokary	1314799	Complete Ge	15/03/2018
44. ASM221462v2	GCA_002214625.2	Wolbachia pipiens wAIB-wL2018		NCBI Prokary	1195156	Chromosome	10/09/2018
45. ASM287995v1	GCA_002879955.1	Ehrlichia canis		NCBI Prokary	1195156	Chromosome	10/09/2018
46. ASM287995v1	GCA_002879955.1	Anaplasma marginale Palmeira		NCBI Prokary	1300495	Chromosome	09/10/2018
47. ASM303156v1	GCA_003031565.1	Anaplasma marginale Palmeira		NCBI Prokary	1300495	Chromosome	09/10/2018
48. ASM399950v1	GCA_003999505.1	Wolbachia endosymbiont Chiona 3		NCBI Prokary	1300495	Chromosome	08/10/2018
49. ASM441723v1	GCA_004417235.1	Wolbachia pipiens wMef_W488		NCBI Prokary	1484007	Complete Ge	13/03/2019
50. ASM49553v1	GCA_004955355.1	Wolbachia endosymbiont of Brugia malayi		NCBI Prokary	1080064	Complete Ge	16/04/2019
51. ASM49553v1	GCA_004955355.1	Wolbachia endosymbiont wMau		NCBI Prokary	1275327	Complete Ge	16/04/2019
52. ASM49553v1	GCA_004955355.1	Wolbachia endosymbiont wMau		NCBI Prokary	1275320	Complete Ge	16/04/2019
53. ASM584229v1	GCA_005842295.1	Wolbachia endosymbiont of Carpocoris sativus	wCauA	NCBI Prokary	1449344	Complete Ge	02/07/2019
54. ASM797258v1	GCA_007972585.1	Wolbachia pipiens wMet_N2S		NCBI Prokary	1267781	Complete Ge	12/08/2019
55. ASM797259v1	GCA_007972595.1	Wolbachia pipiens wMet_O3		NCBI Prokary	1269137	Complete Ge	12/08/2019
56. ASM797259v1	GCA_007972595.1	Wolbachia pipiens wMet_D28		NCBI Prokary	1267346	Complete Ge	12/08/2019
57. ASM800321v1	GCA_008003215.1	Wolbachia endosymbiont wO21		NCBI Prokary	1401540	Complete Ge	30/10/2019

## 5) Pairwise genome alignment

## Dot plot



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

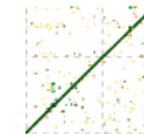
## Dot plot

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.

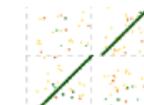
### Match

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.

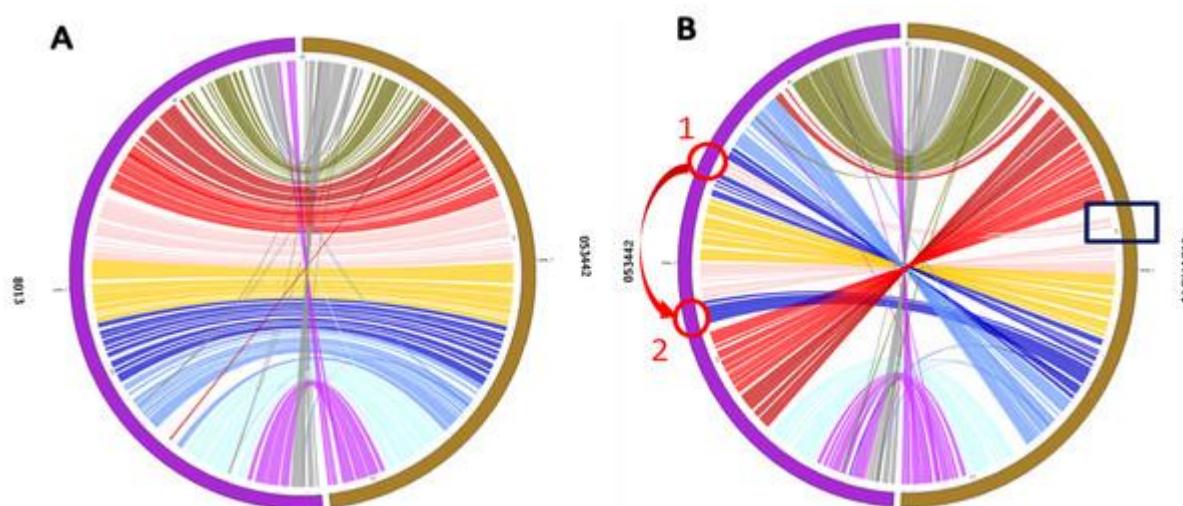


### Repeats

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



## Circos link

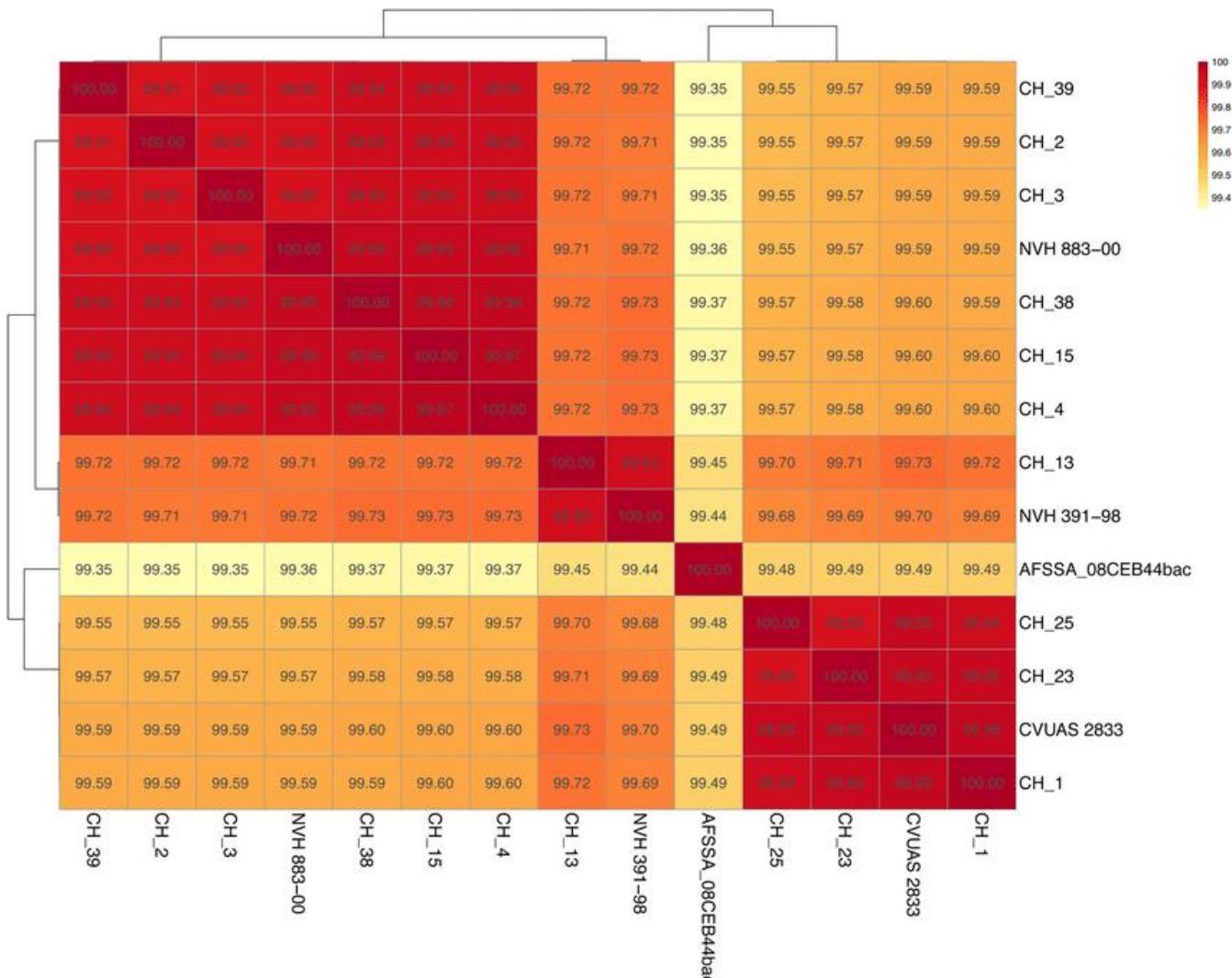


## 6) Pairwise Average Nucleotide Identity (ANI)

## 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

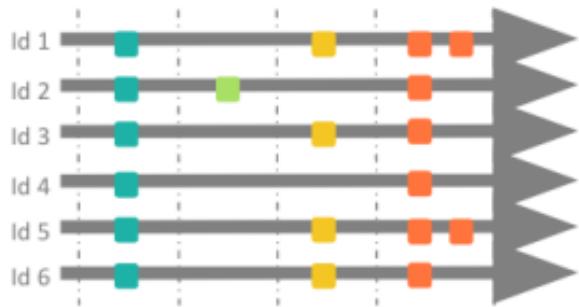
Program: FastANI



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

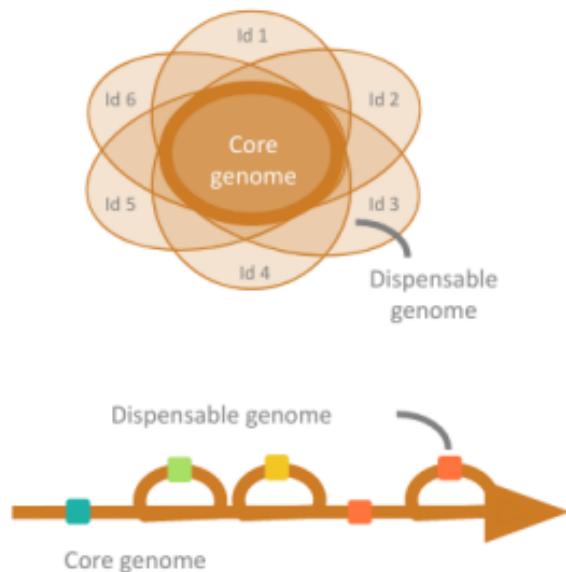
## 7) Pan-genome and Gene clustering

## Pangenome concept



### Pangenome

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



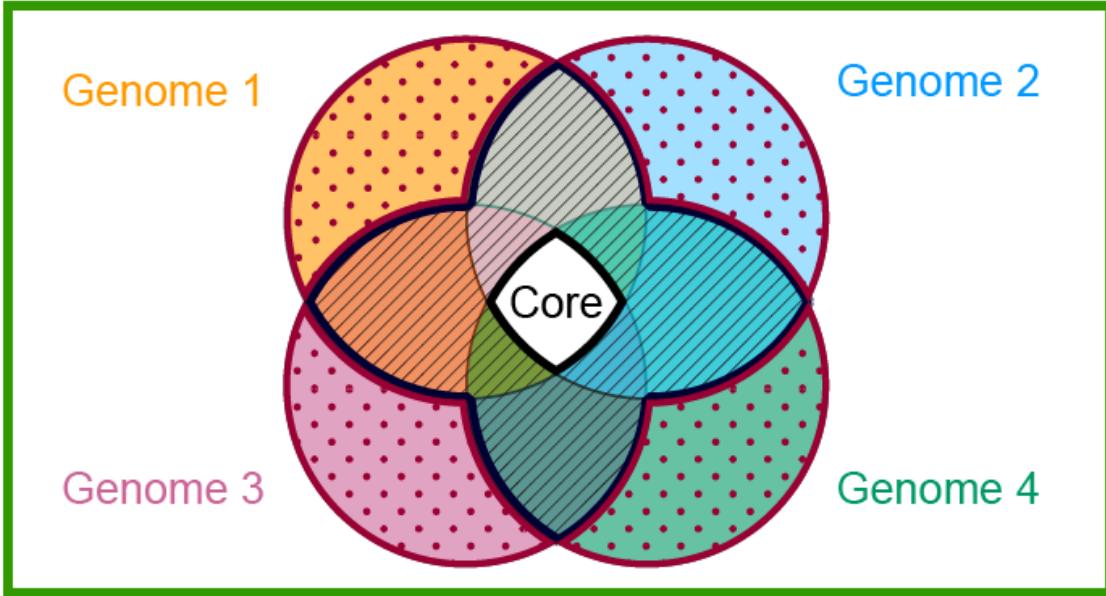
- ▶ **Core genome** : present in all individuals
- ▶ **Disposable genome** : absent from one or several individuals (also called variable, accessory,...)

# 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



[Legend:  
Cloud genome (dotted pattern)  
Shell genome (diagonal lines)]

Accessory genome  
=  
Disposable genome

# Le pangénome ouvert, fermé, le ratio C/P

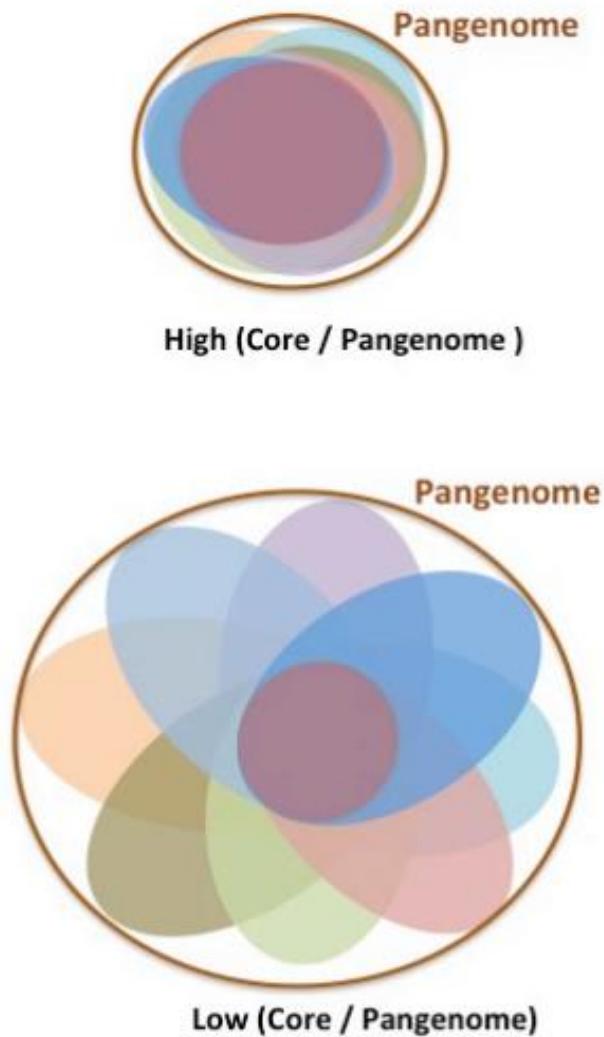
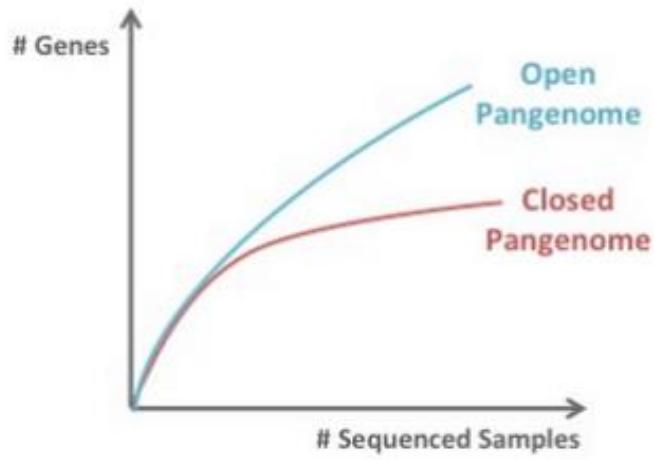


Table 1. Popular software for evolutionary pangenomics

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 (v3.13.0)	GFF3	DOT	Directed graph	BLAST	Synteny
Ptolemy	Java executable (v1.0)	FASTA+GFF	GFA	Directed graph	minimap2	Graph-based
PPanGGoLin	Conda package (v1.0.13)	GBK or FASTA	GEXF	Undirected graph	MMseq2	Synteny
PIRATE	Conda package (v1.0.3)	GFF3	GFA	Directed graph	BLAST (/DIAMOND)	Synteny
Panaroo	Conda package (v1.1.2)	GFF3	GML	Directed graph	CD-HIT	Synteny

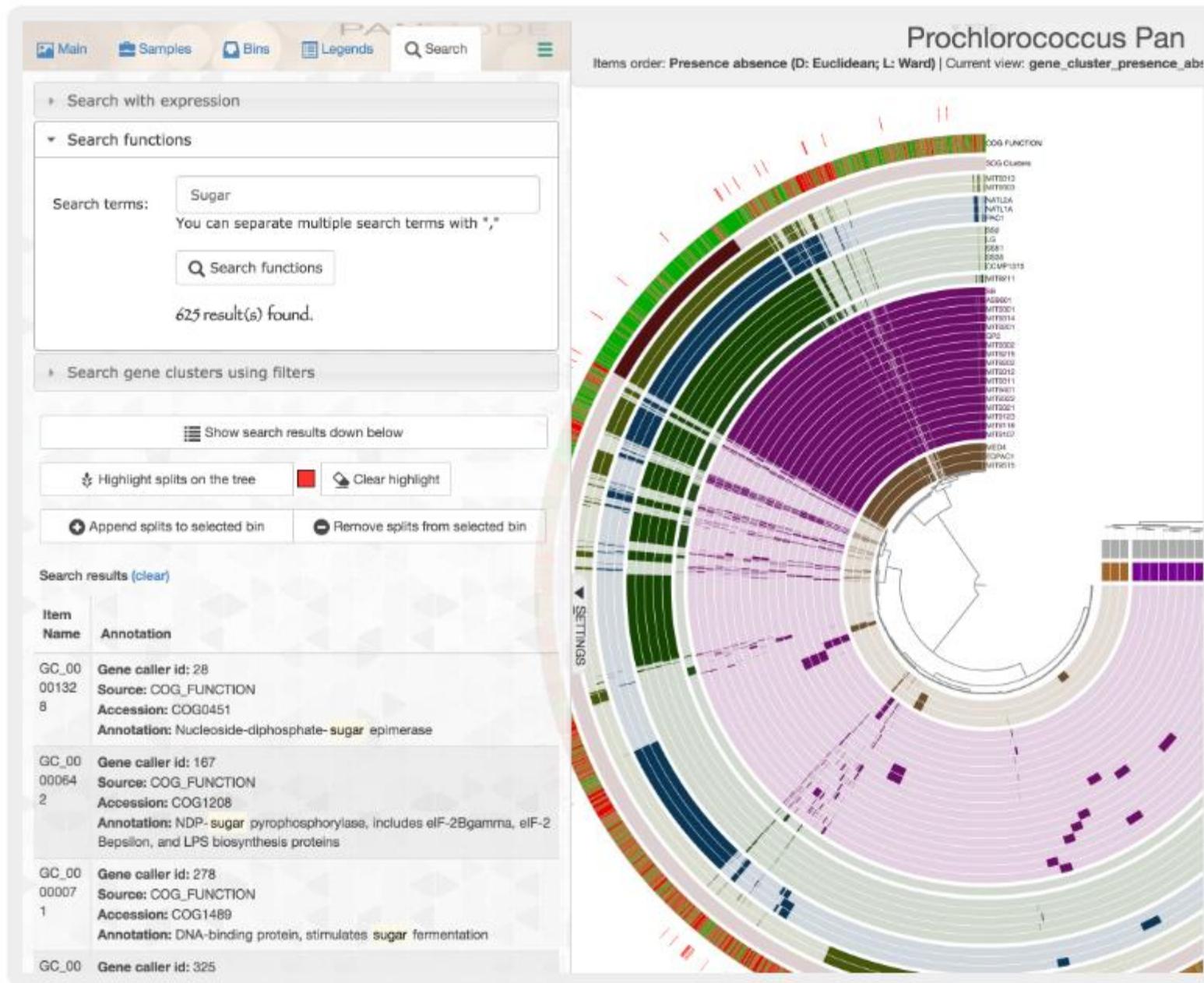
## MICROBIAL GENOMICS

Volume 7, Issue 11

Research Article | Open Access

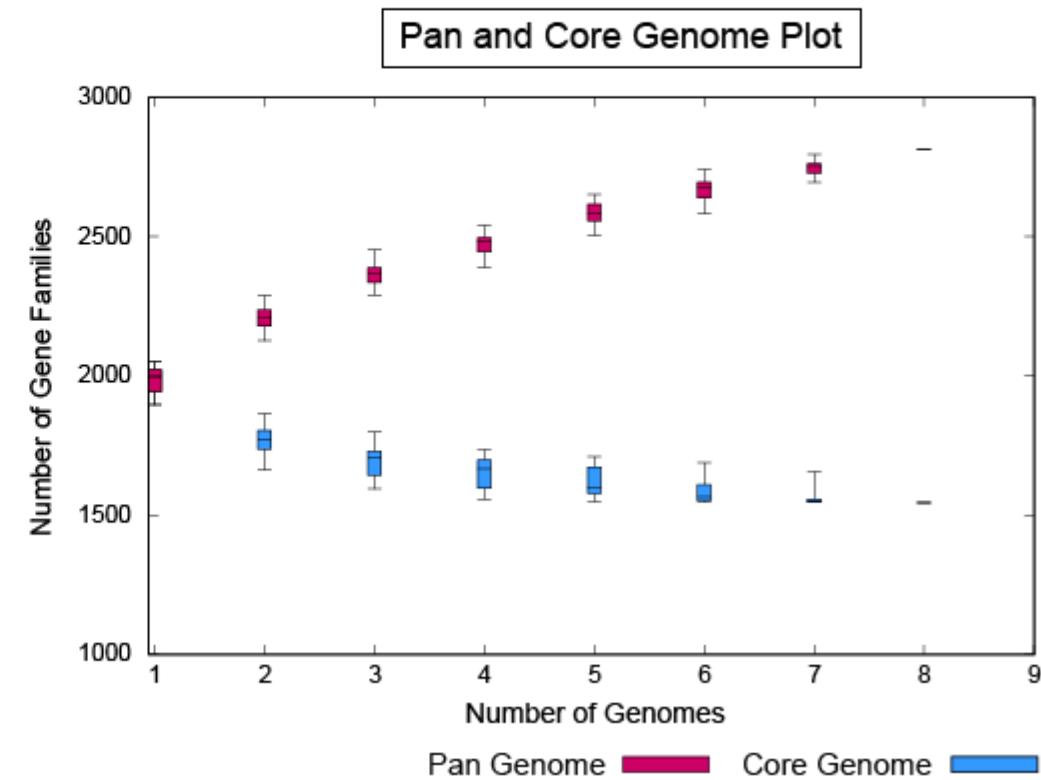
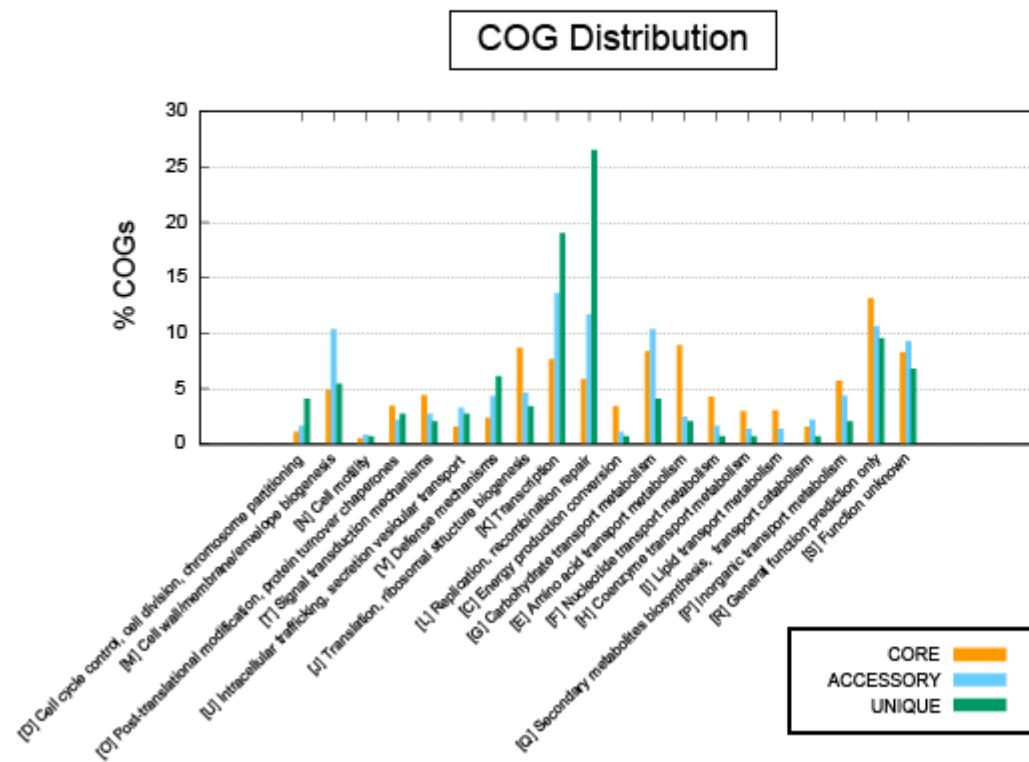
**A comparative study of pan-genome methods for microbial organisms: *Acinetobacter baumannii* pan-genome reveals structural variation in antimicrobial resistance-carrying plasmids** 

Aysun Urhan<sup>1</sup> , Thomas Aebel<sup>1,2</sup> 

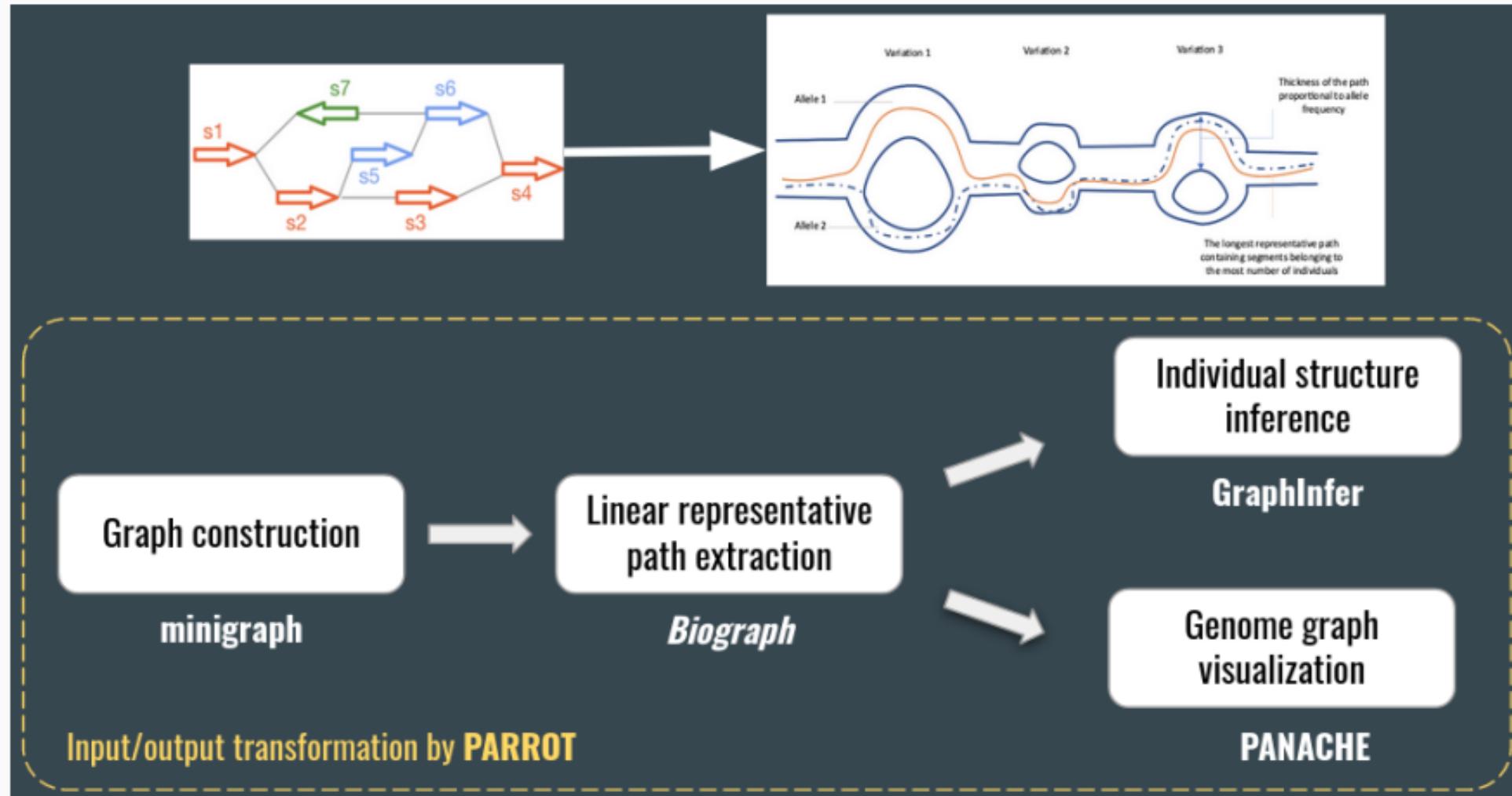


# BPGA (Bacterial Pan Genome Analysis tool)

*Streptococcus agalactiae*

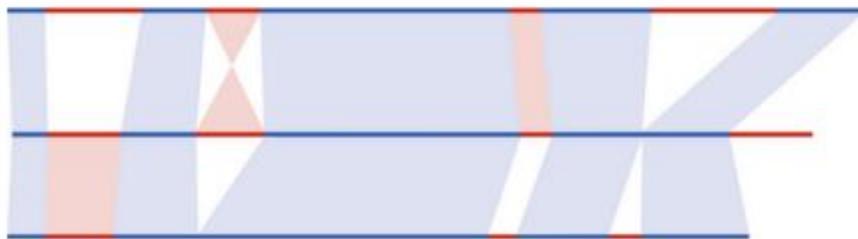


# Comment manipuler le graphe pour les biologistes ?

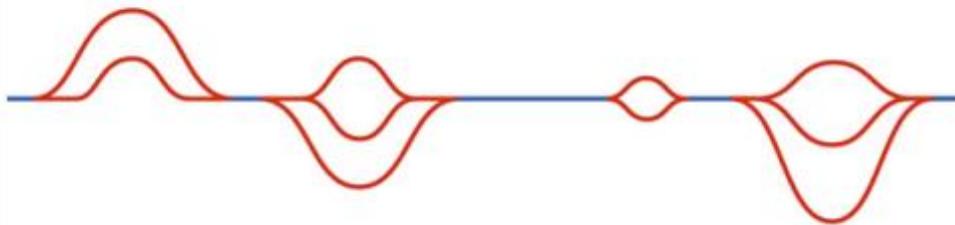


## Concept du graphe de génome

Alignment of de novo assembled genomes



Pan-genome graph

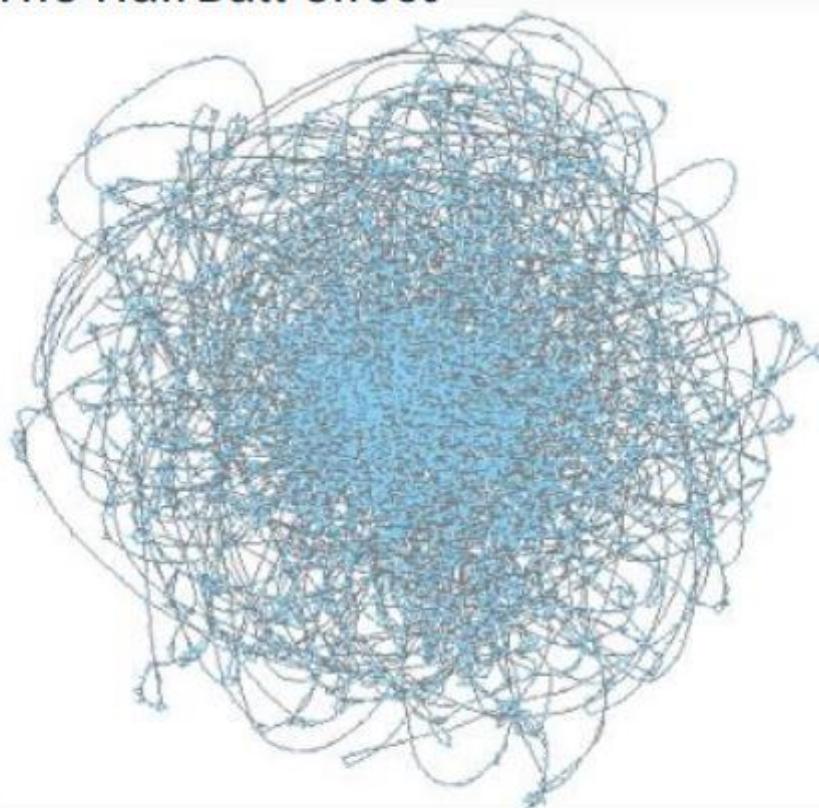


■ Dispensable genome

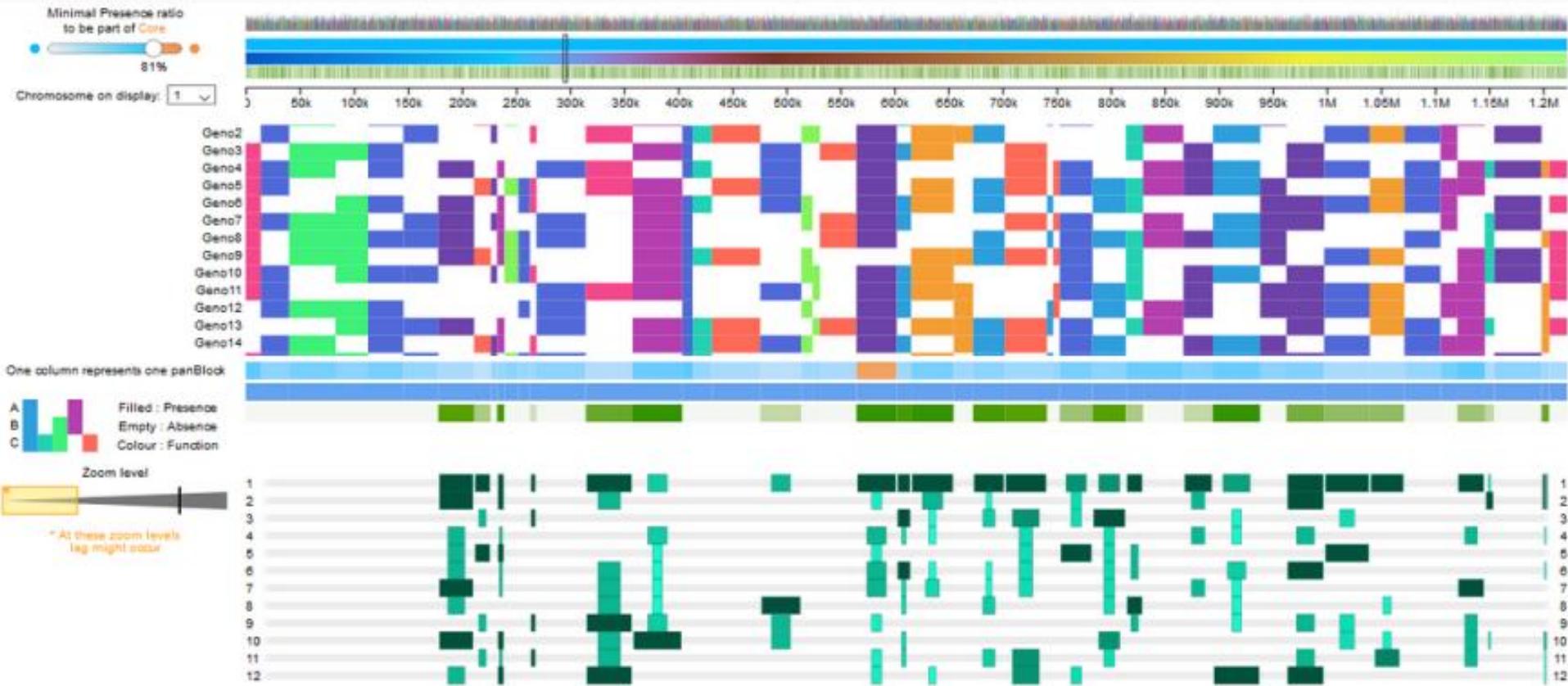
■ Core genome

Bayer et al., 2020

The HairBall effect



# Un exemple linéaire, Panache



Durant, 2020-2021

## 8) Pan-GWAS

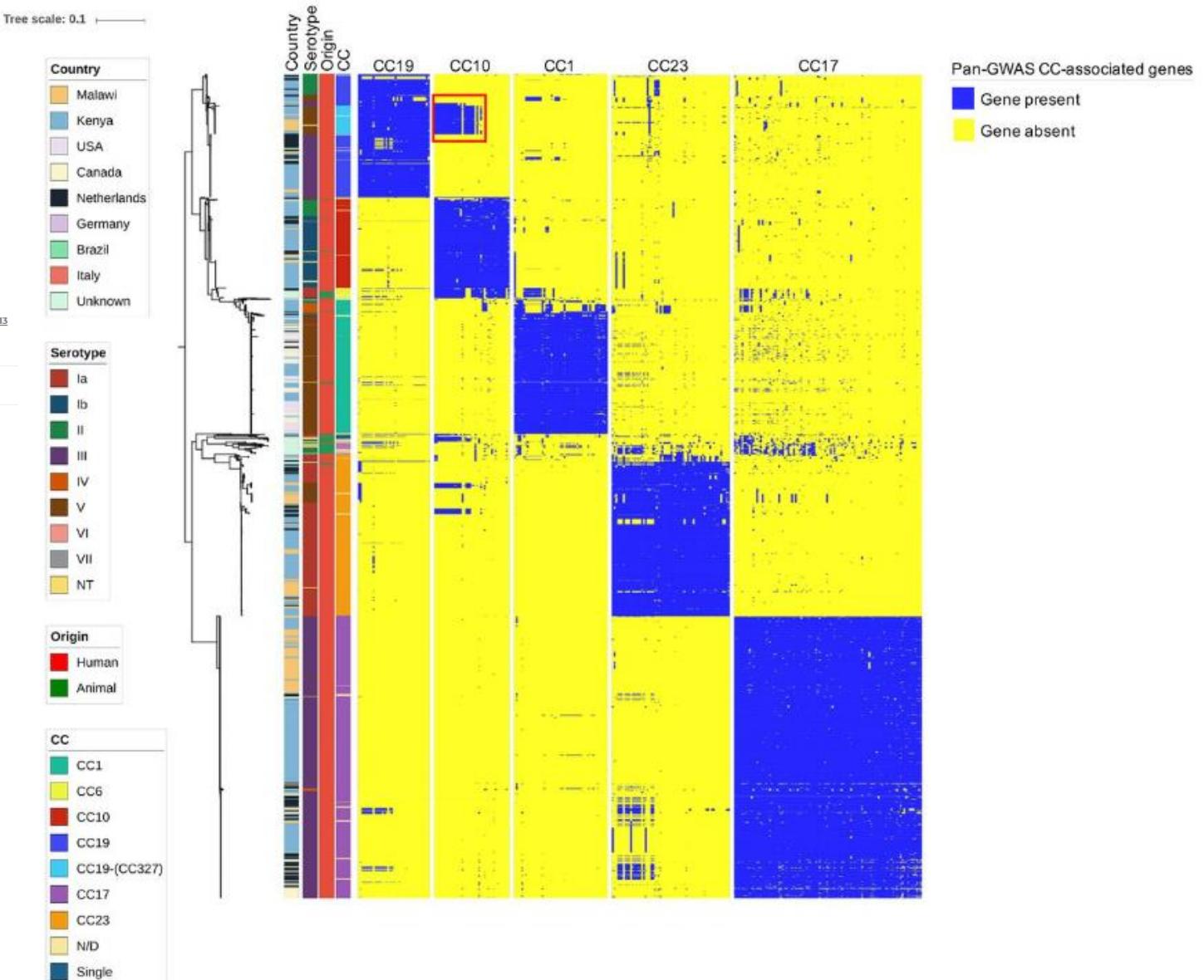
# Pan-GWAS

## 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 \(13\)](#)

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



## microbial pan-GWAS

*“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.”*

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

=> Provides:

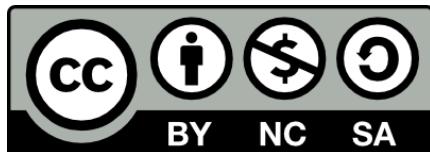
### odds ratios

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

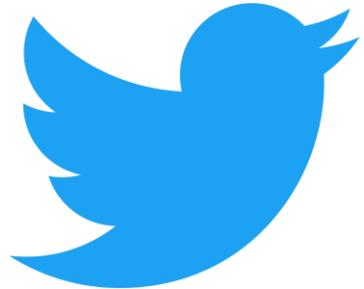
# Merci pour votre attention !



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 : [@ItropBioinfo](#)



## 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>”