South Green
bioinformatics platform

Session de formation 2023

diade

PHIM
Plant
Health
Institute
Montpellier

Alliance
Bioversity & CIAT

agAp

ASTRE
Animal - Santé
Territoires - Risques - Écosystèmes

intertryp
Host-vector-parasite-environment
interactions in neglected tropical diseases
due to trypanosomatids

MIVEGEC
Maladies Infectieuses et Vecteurs
Écologie, Génétique, Évolution et Contrôle

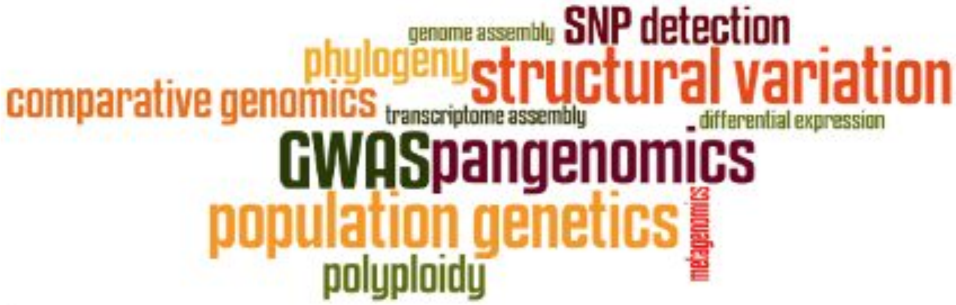
TransVIH
MI

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

South Green
bioinformatics platform

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

ressources
montpellierain
Infrastructure
internationale
formations
génomique
plantes
orienté
développement
support
sud
service
calcul
Réseau
compétences
végétale
plateforme
d'analyses
multi-instituts
Diffusion
communauté
outils
s'appuie
mutualisation
partage



Mutualisation

Cacao

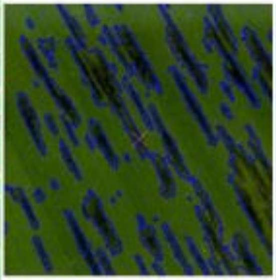
Banana

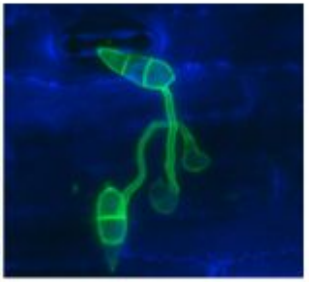
Coffee

Rice

Palm

Cassava


Pseudocercospora


Magnaporthe

South Green

bioinformatics platform

Alliance
Bioversity & CIAT

IRD
Institut de Recherche
pour le Développement
FRANCE

cirad

INRAE

4 institutes

diade

PHIM

agAP

3 research units

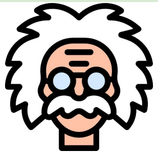
25+

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Storage and computing resources

Tools

Trainings



400+

South Green bioinformatics platform

Storage and computing resources

cirad

RENT

Meso@LR au CINES

1090 threads :

- 35 standard nodes
- 2 bigmem nodes
- 1 GPU node

500 To of replicated storage

IRD

HOSTEL

CINES

1130 threads:

- 30 standard node
- 1 supermem node
- 1 GPU node

150 To on 3 NAS + 210 To scratch

slurm workload manager

400+

600+ tools

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

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Collaborative development of tools

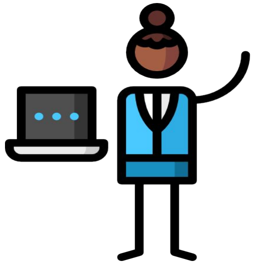
Genomics

- 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/

I-Trop

Plant & Health Bioinformatics Platform

Institut de Recherche pour le Développement
FRANCE

https://bioinfo.ird.fr/

diade

PHIM

TransVIH MI

MiVEGEC
Maladies Infectieuses et Vecteurs
Ecologie, Génétique, Évolution et Contrôle

Support
HPC
Trainings
i-Trop
bioinformatics
Tools
Database

afaq
ISO 9001
Qualité
AFNOR CERTIFICATION

AURORE COMTE	JACQUES DAINAT	ALEXIS DEREEPER	BRUNO GRANOUILLAC	JULIE ORJUELA-	NDOMASSI TANDO	CHRISTINE TRANCHANT

bioinfo@ird.fr

@ItropBioinfo

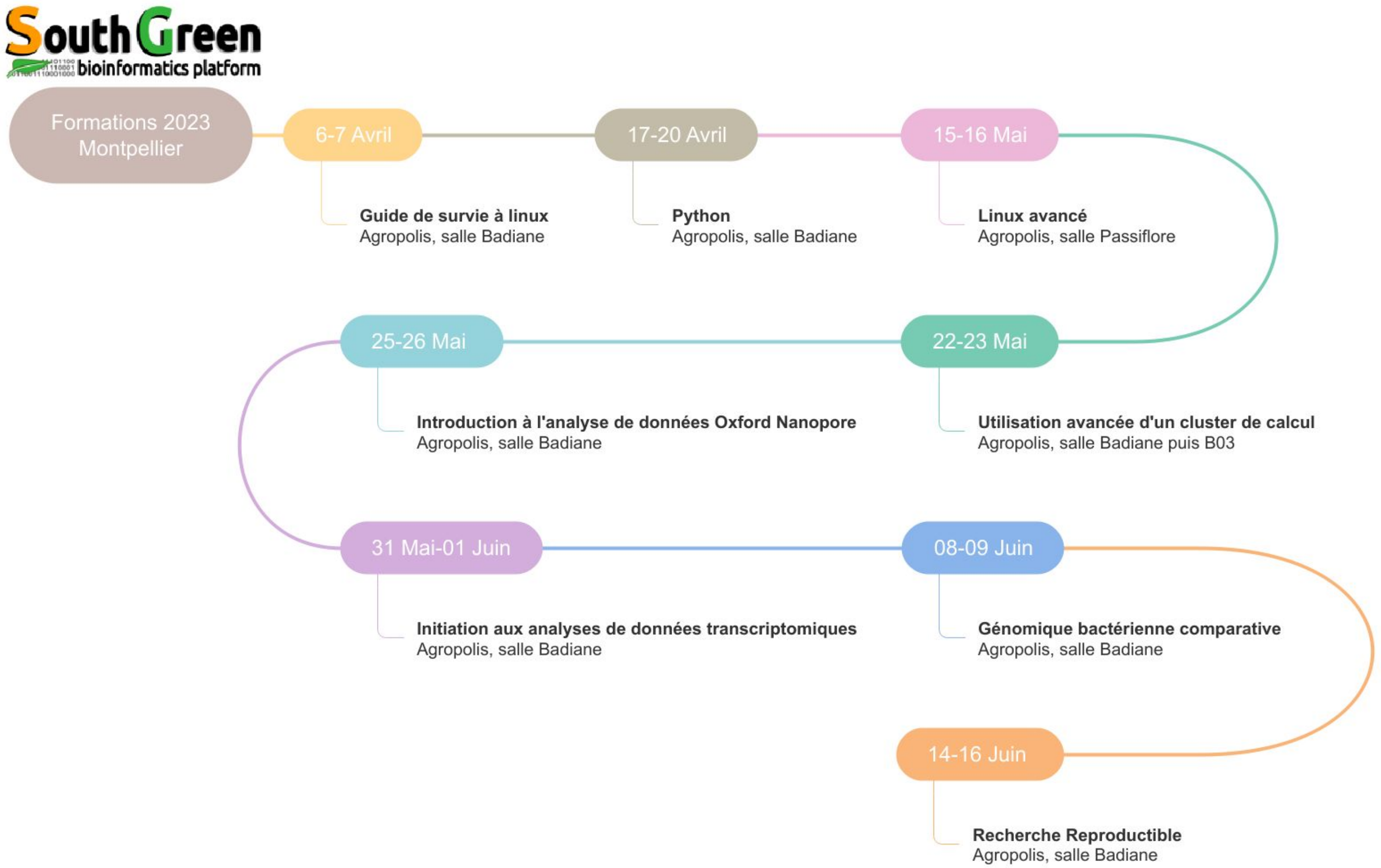
South Green bioinformatics platform

i-Trop bioinformatics

ASTRE Animal · Santé Territoires · Risques · Écosystèmes	Florian Charriat Antoni Exbrayat
intertryp Host-vector-parasite-environment interactions in neglected tropical diseases due to trypanosomatids	Guilhem Sempere
MIVEGEC Maladies Infectieuses et Vecteurs Écologie, Génétique, Évolution et Contrôle	Bruno Granouillac Jacques Dainat
TransVIHMI	Nicolas Fernandez
IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE	Thomas Denecker

And more collaborators !

South Green bioinformatics platform



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

22-23 Mai
Utilisation avancée d'un cluster de calcul
Agropolis, salle Badiane puis B03

25-26 Mai
Introduction à l'analyse de données Oxford Nanopore
Agropolis, salle Badiane

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

Modules de formation 2023

- Toutes nos formations :

  https://southgreenplatform.github.io/trainings/

- Topo & TP :

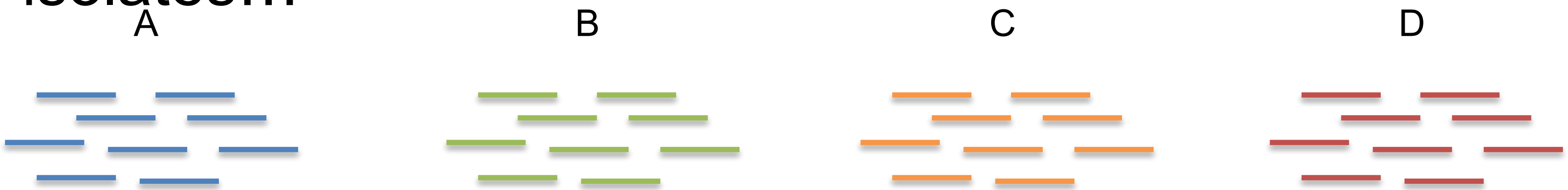
  https://github.com/SouthGreenPlatform/training_ONT_teaching/tree/2023_MTP

- Environnement de travail : Logiciels à installer

Génomique Comparative Bactérienne

Two Approaches to Microbial Genomics

Starting with sets of reads representing your study isolates…



Assembly-based

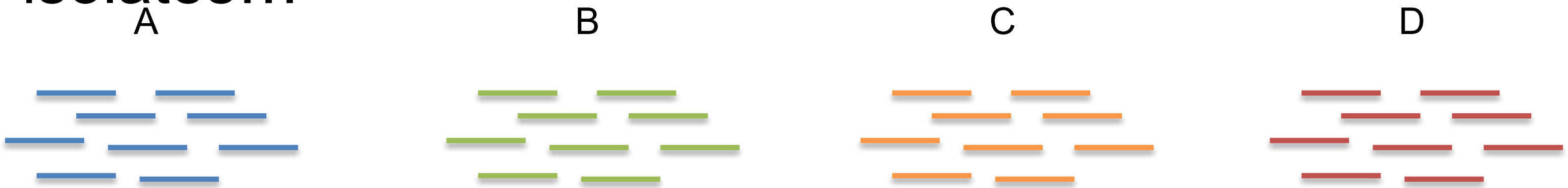
1. Assemble each set of reads into a genome sequence
2. Annotate each genome
3. Cluster genes and compare between each genome

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

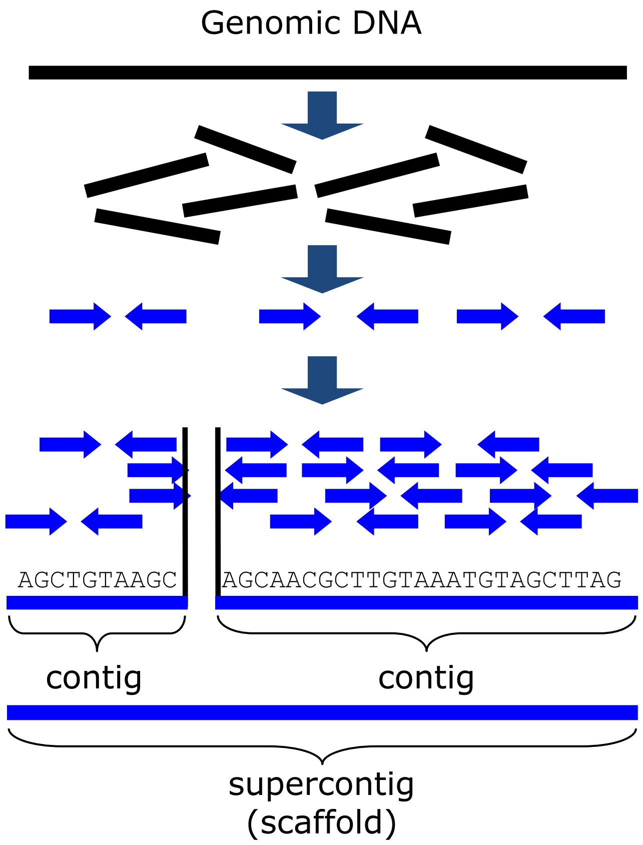
1. Assemble each set of reads into a genome sequence
2. Annotate each genome
3. Cluster genes and compare between each genome

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)



Genomic DNA

AGCTGTAAGC

AGCAACGCTTGTAAATGTAGCTTAG

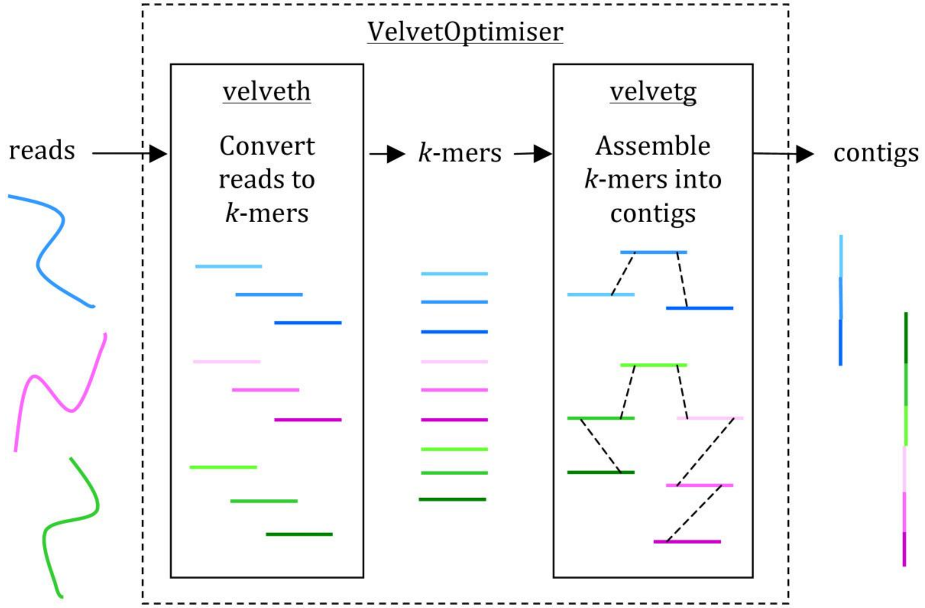
contig

contig

supercontig
(scaffold)

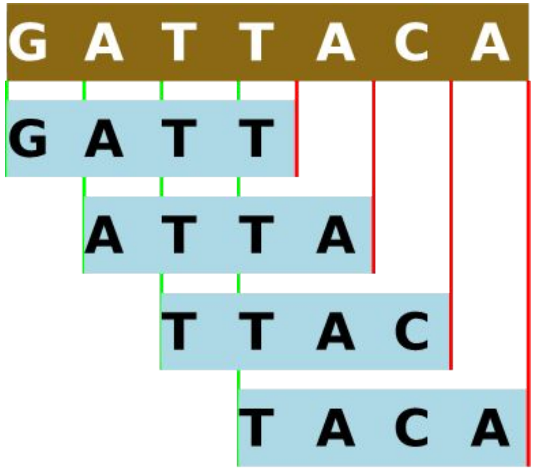
Assembly Methods

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



Edwards and Holt 2013

K-mers



MIE

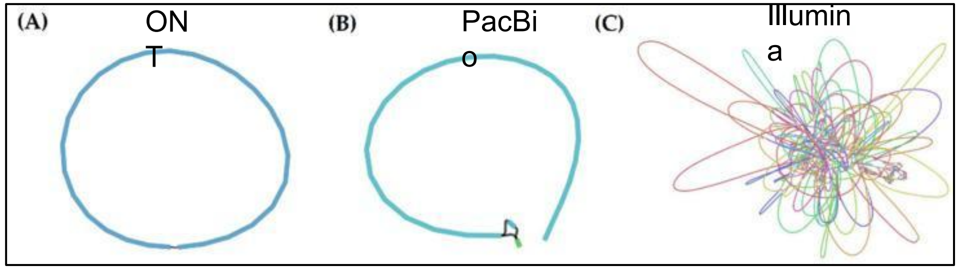
International Journal of *Molecular Sciences*

MDPI

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]



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.

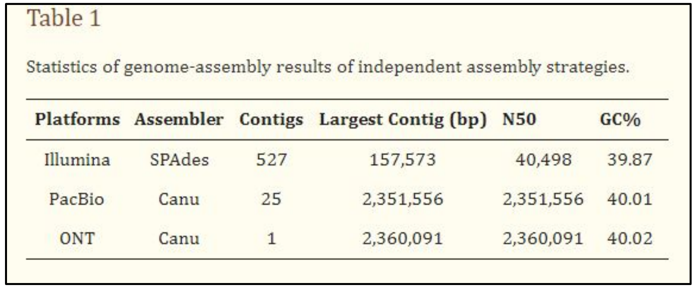


Table 1

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

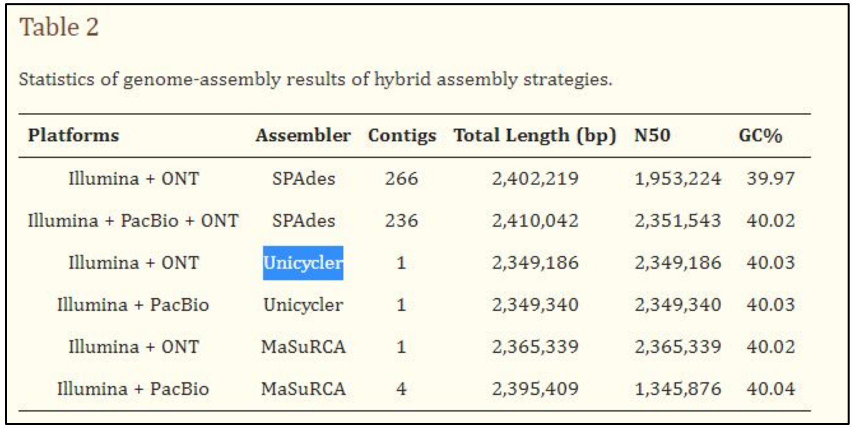
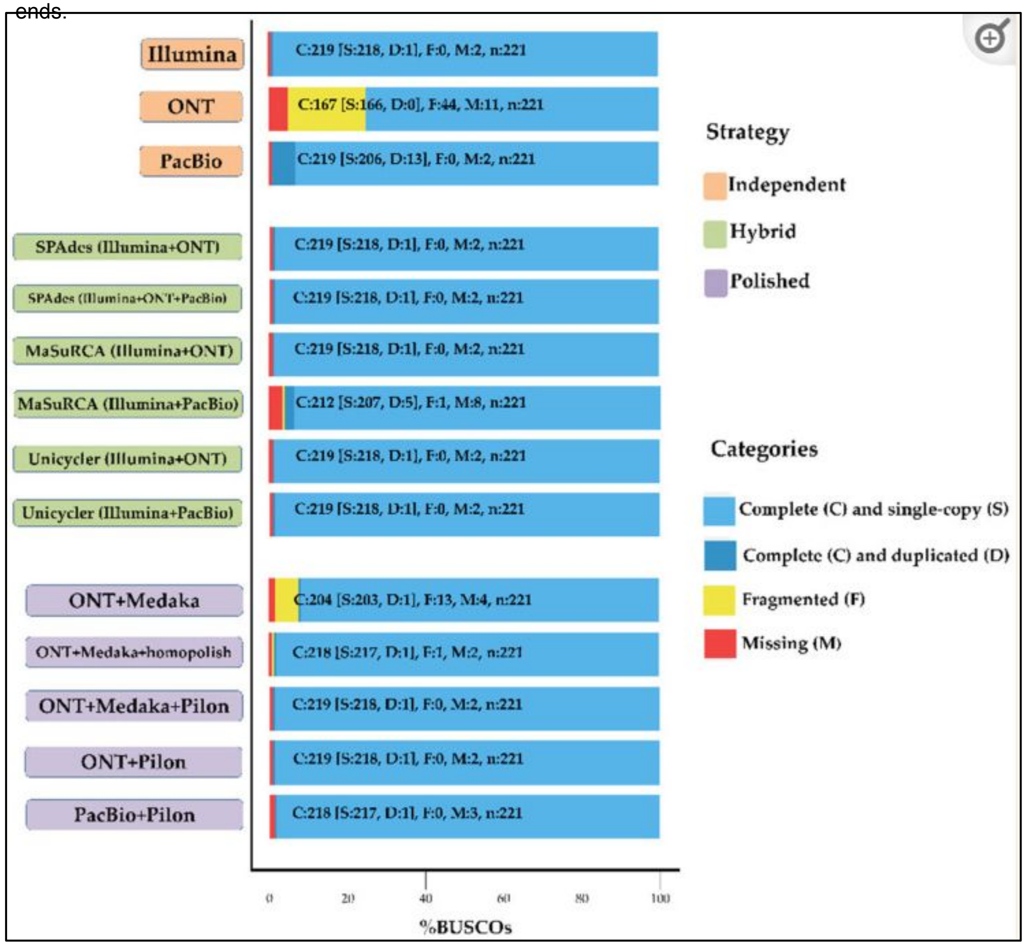


Table 2

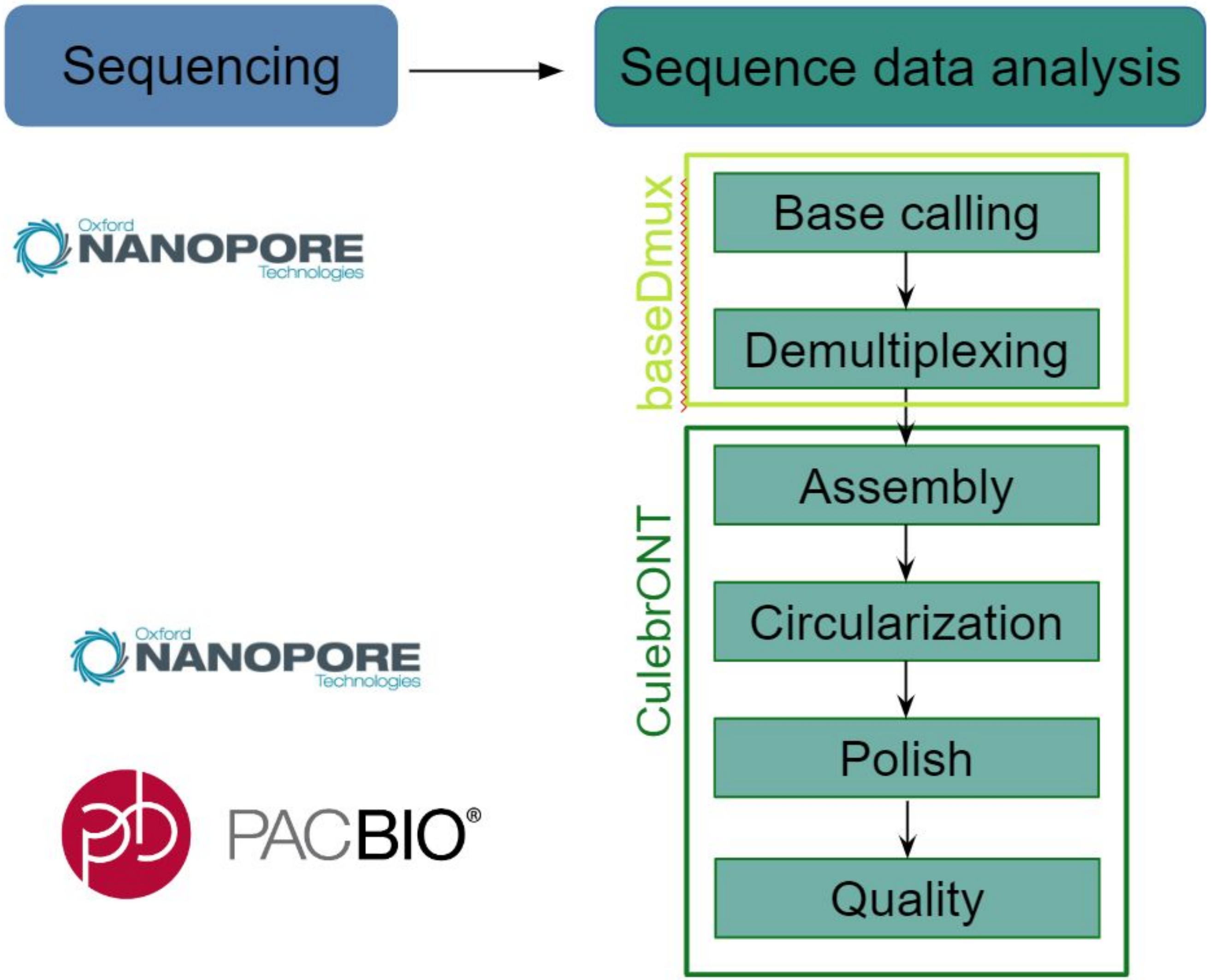
Statistics of genome-assembly results of hybrid assembly strategies.

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



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



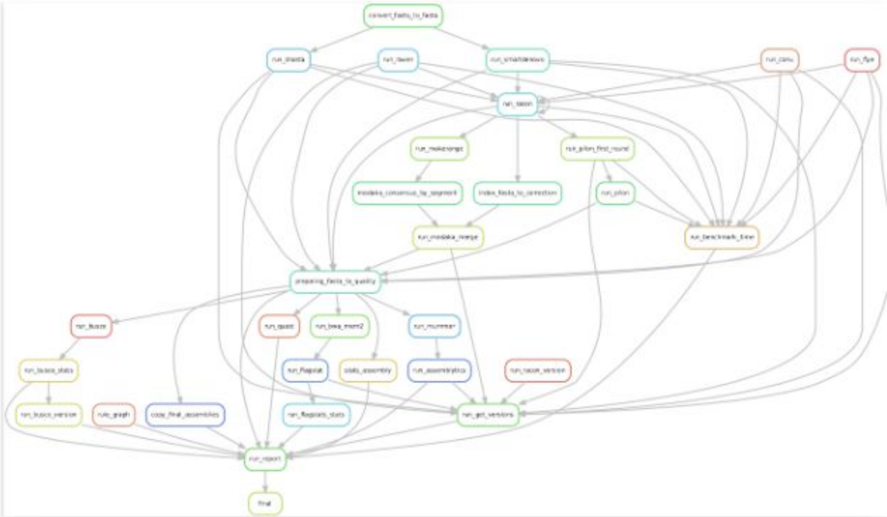
Snakemake

BaseDmux

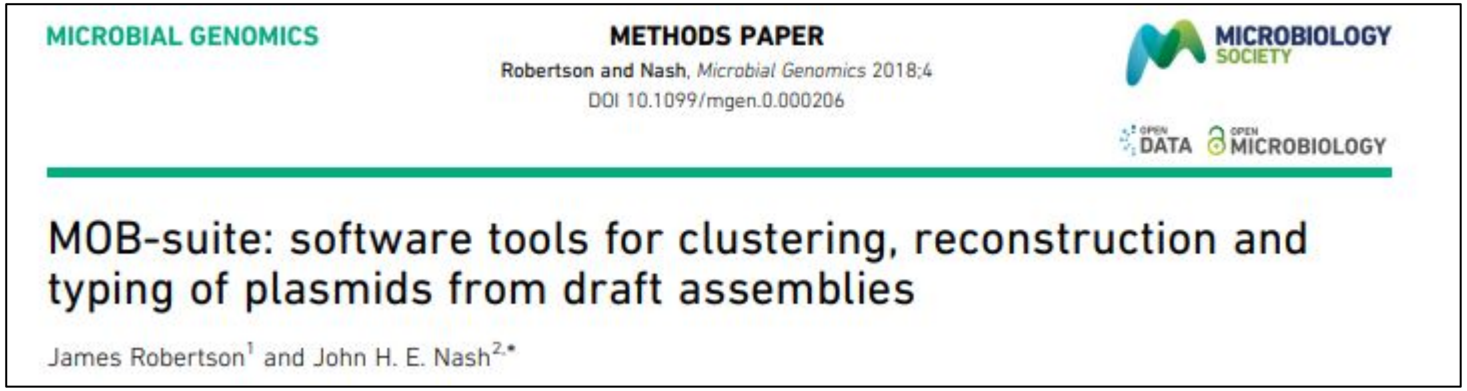
https://github.com/vibaotram/baseDmux

culebrONT

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



2) Separate chromosomal and plasmid scaffolds/contigs



MICROBIAL GENOMICS

METHODS PAPER

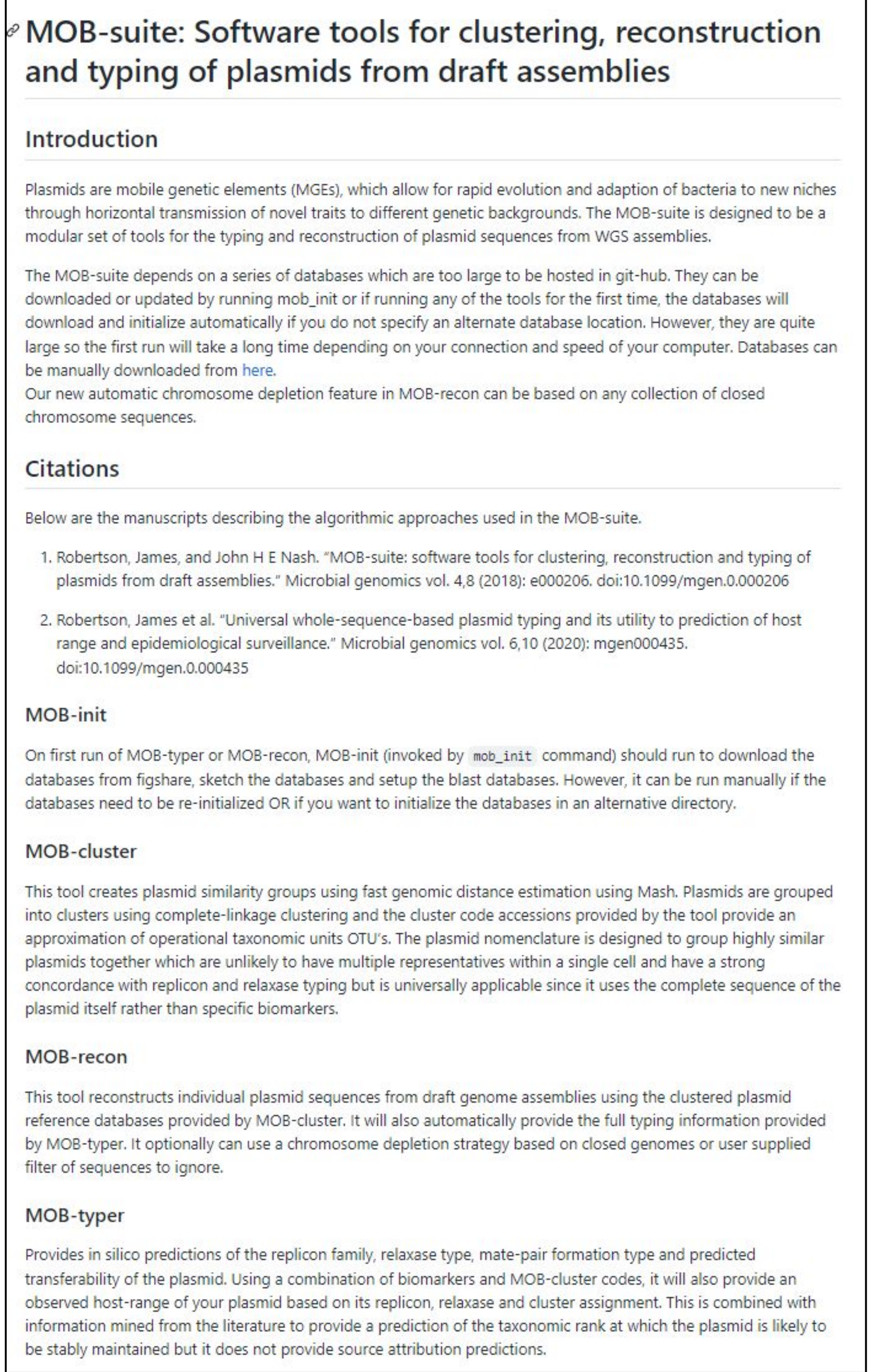
Robertson and Nash, *Microbial Genomics* 2018;4
DOI 10.1099/mgen.0.000206

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

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

James Robertson[1] 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.

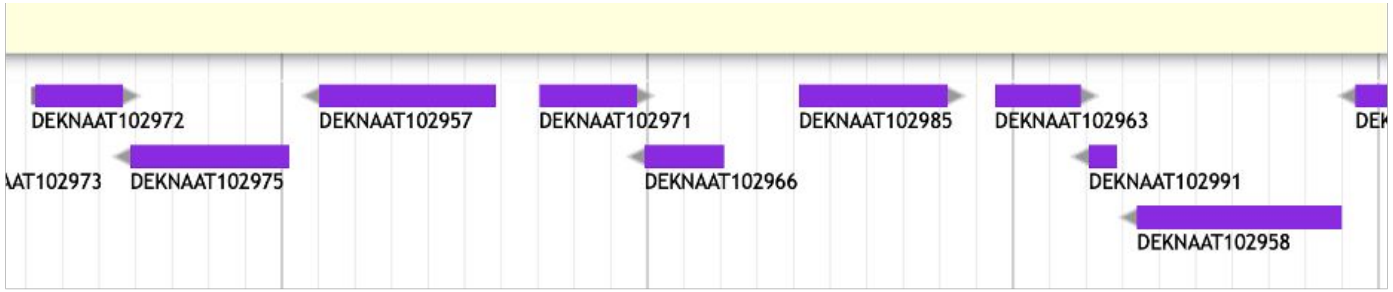
Find out what the regions do. What do they code for?

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

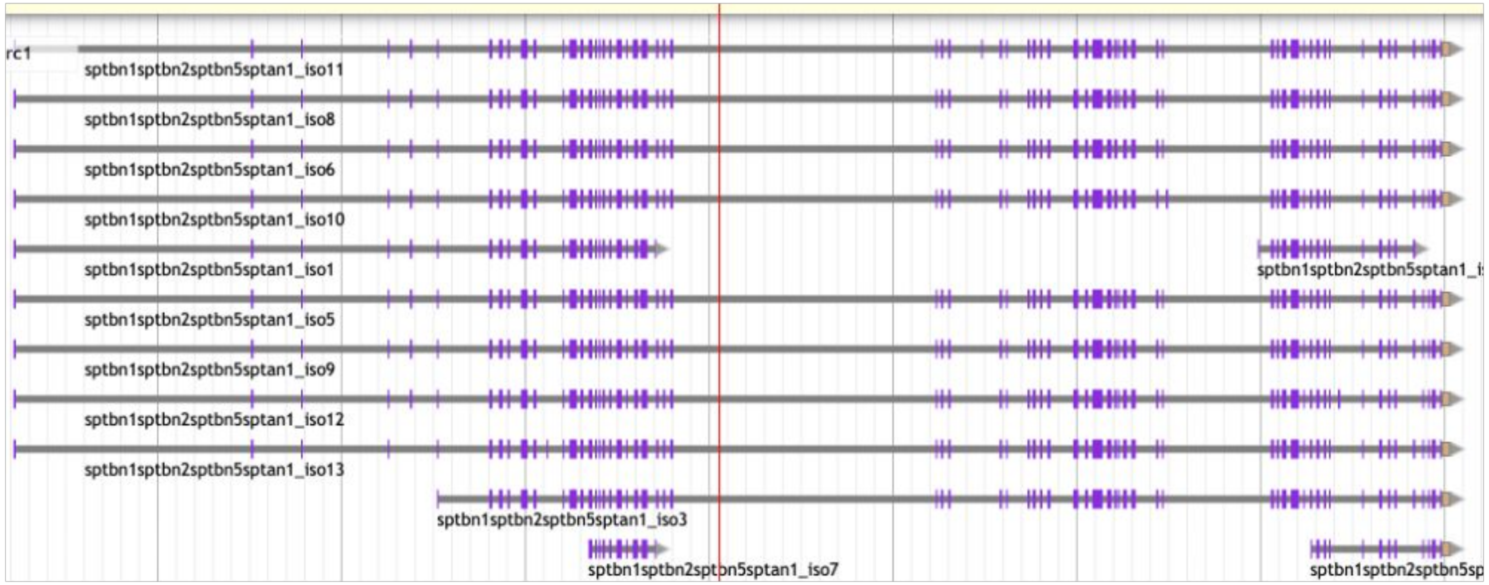
Thx!!
Slide borrowed from NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Organisms differ in genomic complexity

A yeast



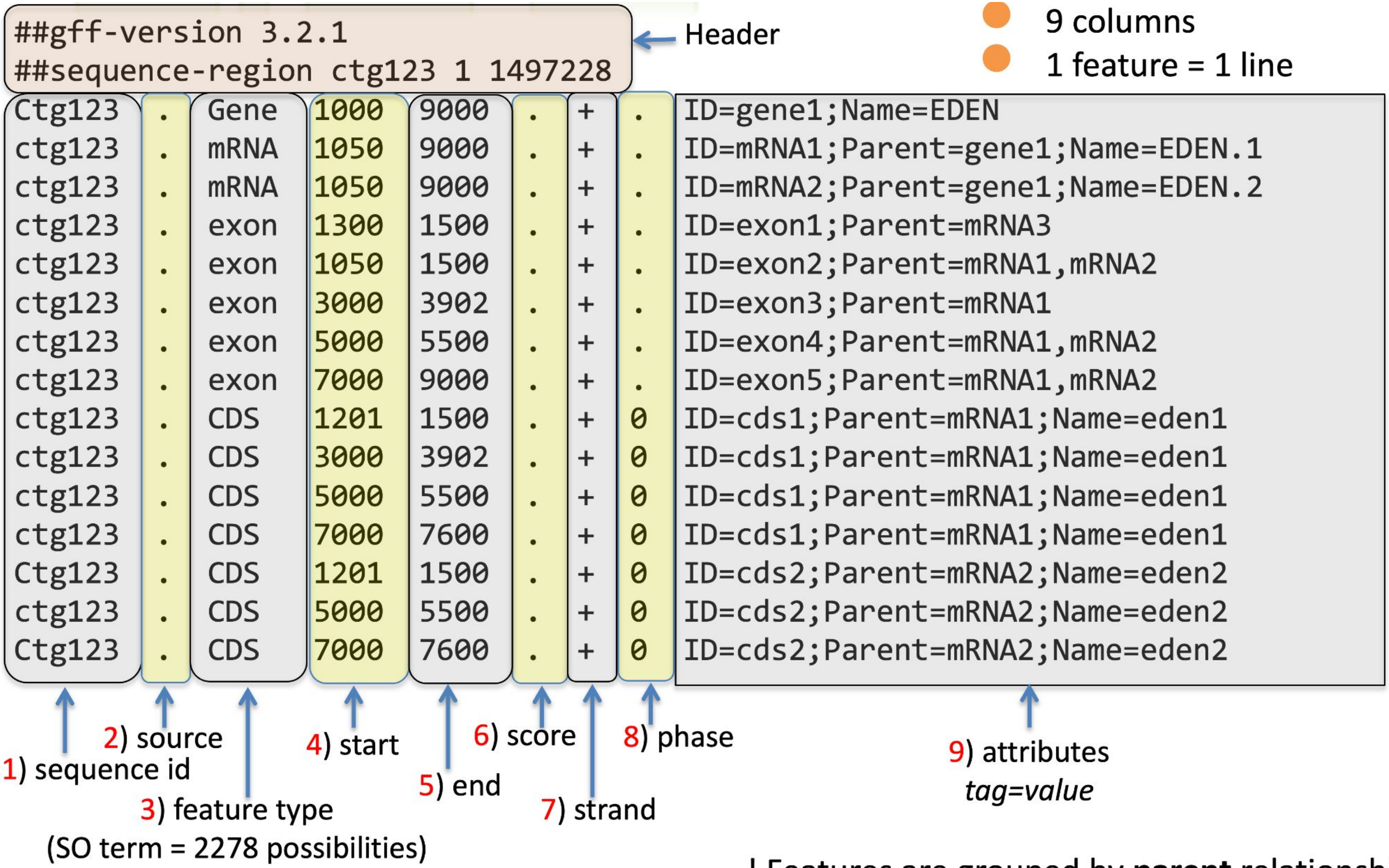
A crustacean


rc1
sptbn1sptbn2sptbn5sptan1_iso11
sptbn1sptbn2sptbn5sptan1_iso8
sptbn1sptbn2sptbn5sptan1_iso6
sptbn1sptbn2sptbn5sptan1_iso10
sptbn1sptbn2sptbn5sptan1_iso1
sptbn1sptbn2sptbn5sptan1_i
sptbn1sptbn2sptbn5sptan1_iso5
sptbn1sptbn2sptbn5sptan1_iso9
sptbn1sptbn2sptbn5sptan1_iso12
sptbn1sptbn2sptbn5sptan1_iso13
sptbn1sptbn2sptbn5sptan1_iso3
sptbn1sptbn2sptbn5sptan1_iso7
sptbn1sptbn2sptbn5sp

7/55

Thx!!
Slide borrowed from NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

- 9 columns
- 1 feature = 1 line



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

Header

1) sequence id

2) source

3) feature type
(SO term = 2278 possibilities)

4) start

5) end

6) score

7) strand

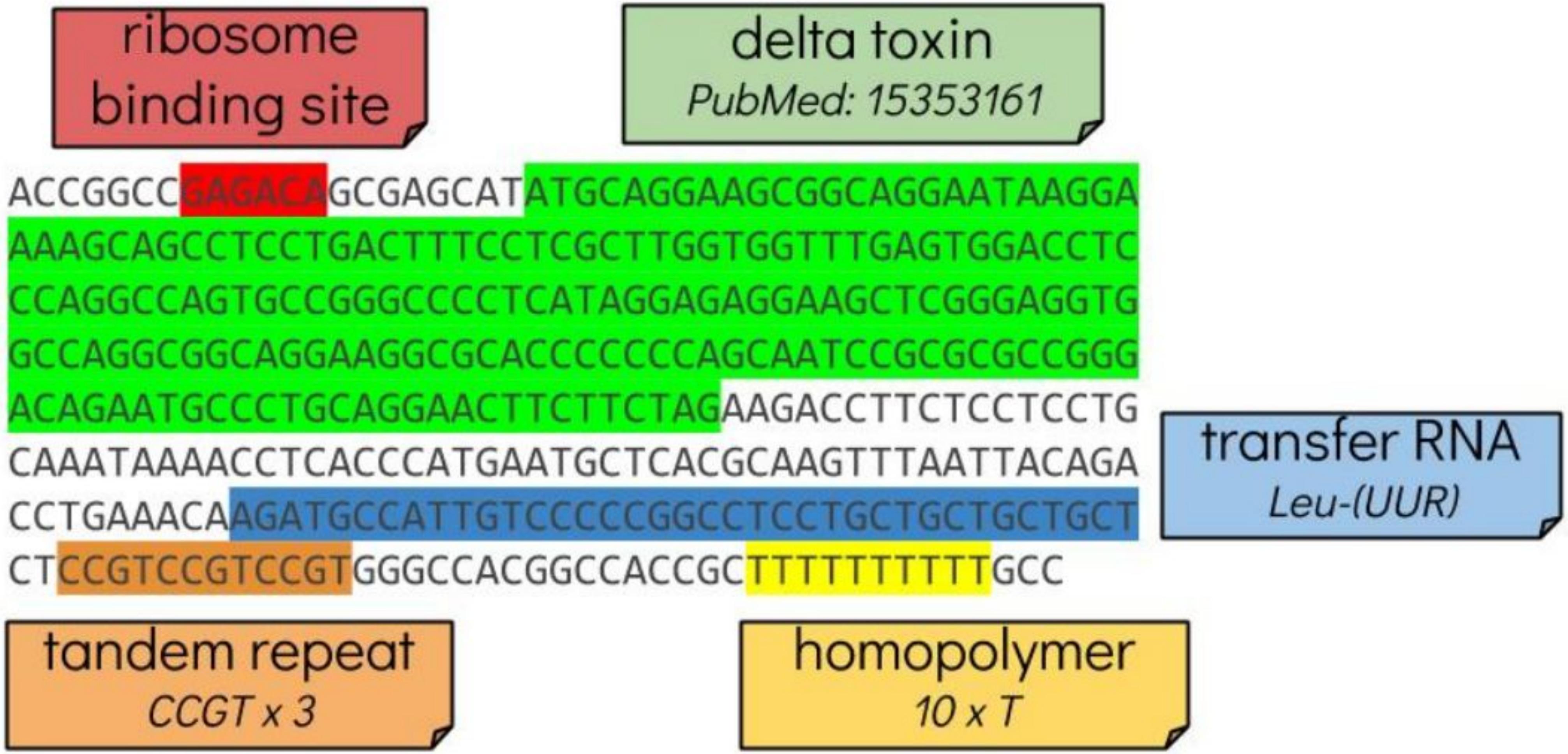
8) phase

9) attributes
tag=value

! Features are grouped by **parent** relationship

Thx!!
Slide borrowed from NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Adding biological info to sequences



ribosome binding site

delta toxin
PubMed: 15353161

ACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGA
AAAGCAGCCTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTC
CCAGGCCAGTGCCGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTG
GCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGG
ACAGAATGCCCTGCAGGAACTTCTTCTAGAAGACCTTCTCCTCCTG
CAAATAAAACCTCACCCATGAATGCTCACGCAAGTTTAATTACAGA
CCTGAAACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCT
CTCCGTCCGTCCGTGGGCCACGGCCACCGCTTTTTTTTTTGCC

transfer RNA
Leu-(UUR)

tandem repeat
CCGT x 3

homopolymer
10 x T

Annotation 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

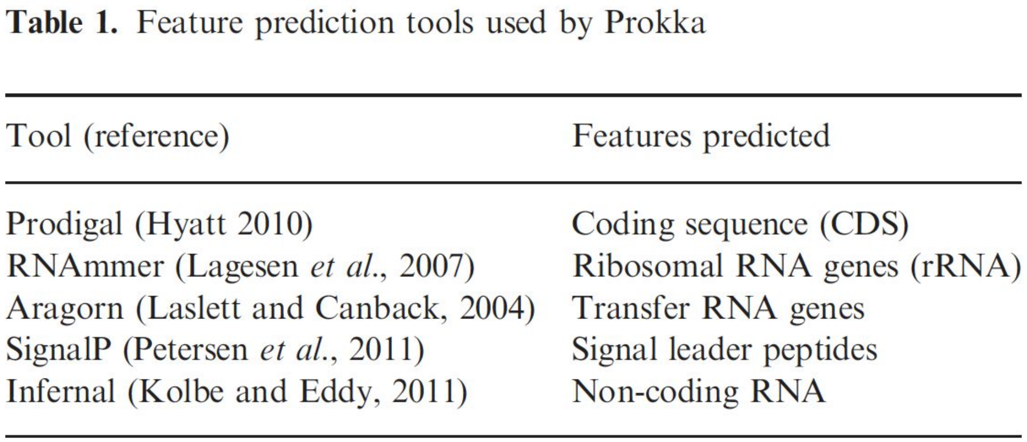
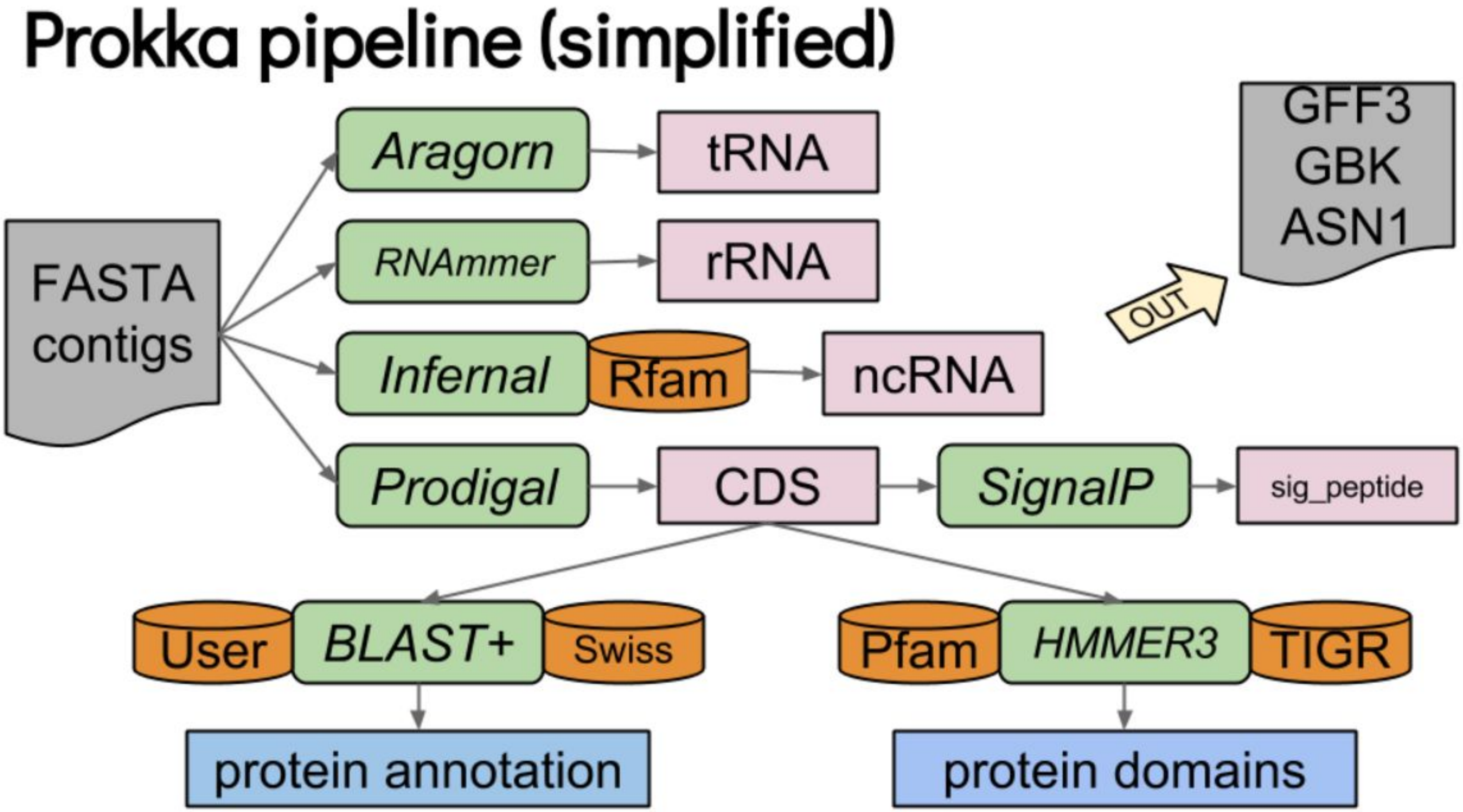


Table 1. Feature prediction tools used by Prokka

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

Prokka



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

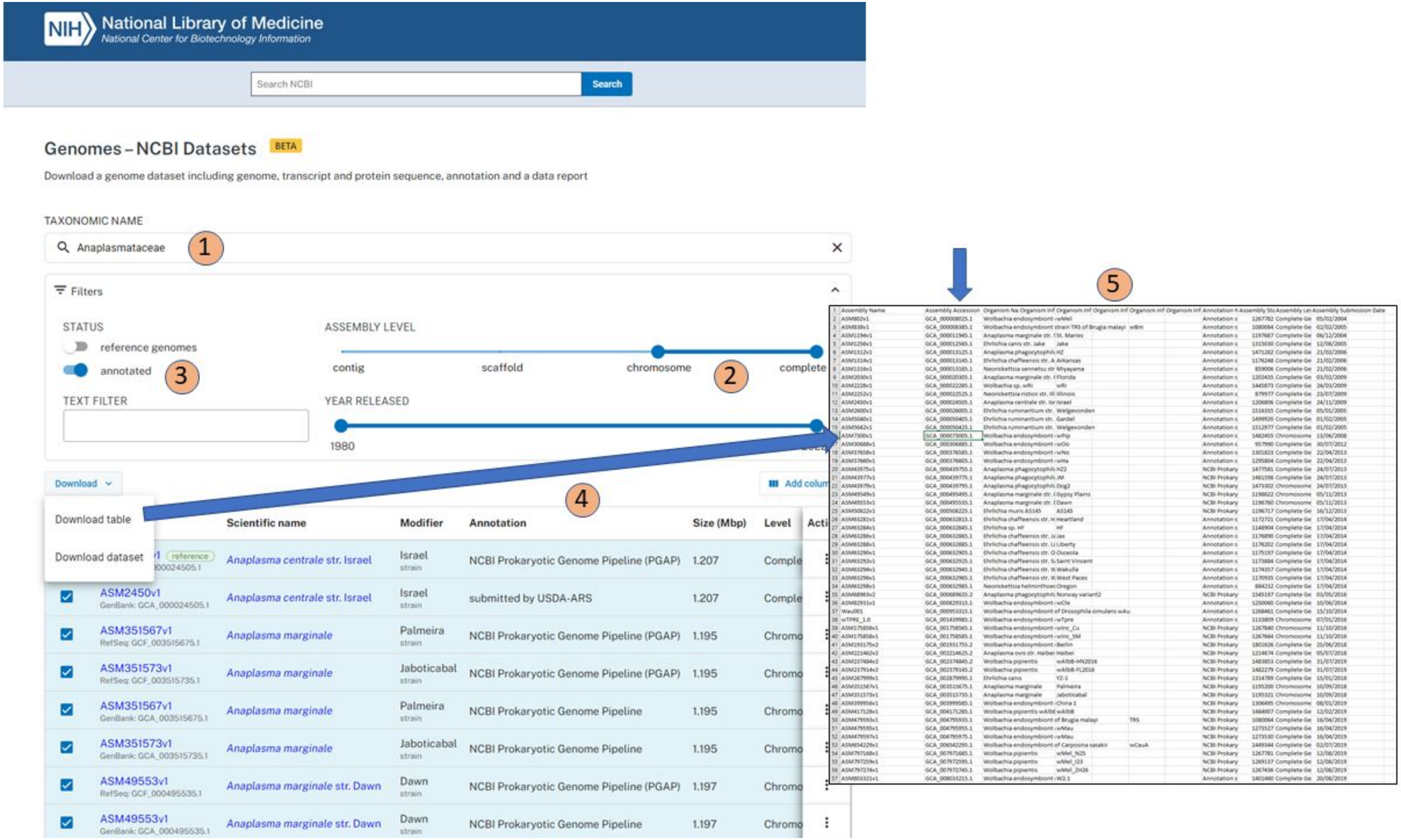
- tRNAscan-SE
- Aragorn
- INFERNAL
- PILER-CR
- Prodigal
- Hmmer
- Diamond
- Blast+
- AMRFinderPlus
- DeepSig

Databases

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

4) Public genomes retrieval

https://www.ncbi.nlm.nih.gov/datasets/genomes



NIH 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: Anaplasmataceae

STATUS
- reference genomes
- annotated

ASSEMBLY LEVEL: contig, scaffold, chromosome, complete

TEXT FILTER

YEAR RELEASED: 1980

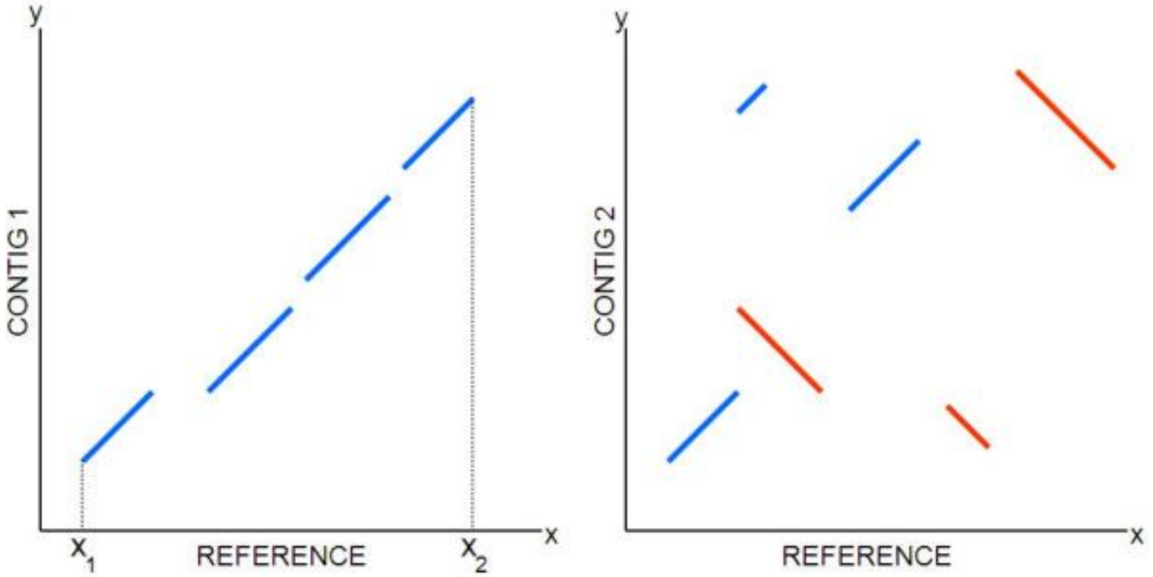
Download
Download table
Download dataset
Add colum

	Scientific name	Modifier	Annotation	Size (Mbp)	Level
1 reference 000024505.1	*Anaplasma centrale* str. Israel	Israel strain	NCBI Prokaryotic Genome Pipeline (PGAP)	1.207	Comple
☑ ASM2450v1 GenBank: GCA_000024505.1	*Anaplasma centrale* str. Israel	Israel strain	submitted by USDA-ARS	1.207	Comple
☑ ASM351567v1 RefSeq: GCF_003515675.1	*Anaplasma marginale*	Palmeira strain	NCBI Prokaryotic Genome Pipeline (PGAP)	1.195	Chromo
☑ ASM351573v1 RefSeq: GCF_003515735.1	*Anaplasma marginale*	Jaboticabal strain	NCBI Prokaryotic Genome Pipeline (PGAP)	1.195	Chromo
☑ ASM351567v1 GenBank: GCA_003515675.1	*Anaplasma marginale*	Palmeira strain	NCBI Prokaryotic Genome Pipeline	1.195	Chromo
☑ ASM351573v1 GenBank: GCA_003515735.1	*Anaplasma marginale*	Jaboticabal strain	NCBI Prokaryotic Genome Pipeline	1.195	Chromo
☑ ASM49553v1 RefSeq: GCF_004495535.1	*Anaplasma marginale* str. Dawn	Dawn strain	NCBI Prokaryotic Genome Pipeline (PGAP)	1.197	Chromo
☑ ASM49553v1 GenBank: GCA_004495535.1	*Anaplasma marginale* str. Dawn	Dawn strain	NCBI Prokaryotic Genome Pipeline	1.197	Chromo

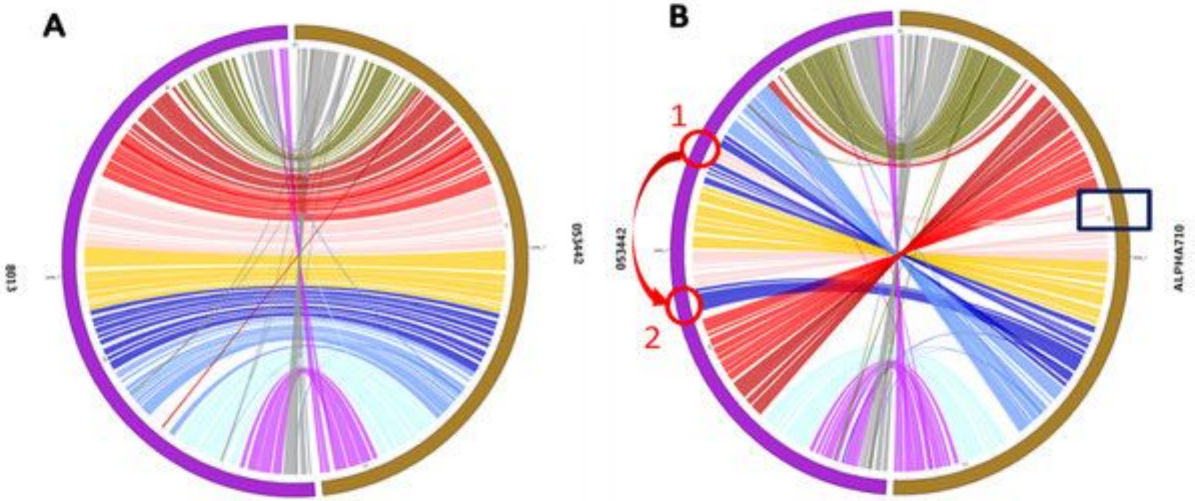
1	Assembly Name	Assembly Accession	Organism Na	Organism Inf	Organism Inf	Organism Inf	Organism Inf	Organism Inf	Annotation R	Assembly Sta	Assembly Lev	Assembly Submission Date
2	ASM802v1	GCA_000008025.1	Wolbachia endosymbiont -		wMel				Annotation s	1267782	Complete Ge	05/02/2004
3	ASM838v1	GCA_000008385.1	Wolbachia endosymbiont strain TRS of Brugia malayi				wBm		Annotation s	1080084	Complete Ge	02/02/2005
4	ASM1194v1	GCA_000011945.1	Anaplasma marginale str. f		St. Maries				Annotation s	1197687	Complete Ge	06/12/2004
5	ASM1256v1	GCA_000012565.1	Ehrlichia canis str. Jake		Jake				Annotation s	1315030	Complete Ge	12/08/2005
6	ASM1312v1	GCA_000013125.1	Anaplasma phagocytophil		HZ				Annotation s	1471282	Complete Ge	21/02/2006
7	ASM1314v1	GCA_000013145.1	Ehrlichia chaffeensis str. A		Arkansas				Annotation s	1176248	Complete Ge	21/02/2006
8	ASM1316v1	GCA_000013165.1	Neorickettsia sennetsu str		Miyayama				Annotation s	859006	Complete Ge	21/02/2006
9	ASM2030v1	GCA_000020305.1	Anaplasma marginale str. f		Florida				Annotation s	1202435	Complete Ge	03/02/2009
10	ASM2228v1	GCA_000022285.1	Wolbachia sp. wRi		wRi				Annotation s	1445873	Complete Ge	24/03/2009
11	ASM2252v1	GCA_000022525.1	Neorickettsia risticii str. Ill		Illinois				Annotation s	879977	Complete Ge	23/07/2009
12	ASM2450v1	GCA_000024505.1	Anaplasma centrale str. Isr		Israel				Annotation s	1206806	Complete Ge	24/11/2009
13	ASM2600v1	GCA_000026005.1	Ehrlichia ruminantium str.		Welgevonden				Annotation s	1516355	Complete Ge	05/01/2005
14	ASM5040v1	GCA_000050405.1	Ehrlichia ruminantium str.		Gardel				Annotation s	1499920	Complete Ge	01/02/2005
15	ASM5042v1	GCA_000050425.1	Ehrlichia ruminantium str.		Welgevonden				Annotation s	1512977	Complete Ge	01/02/2005
	ASM7300v1	GCA_000073005.1	Wolbachia endosymbiont -		wPip				Annotation s	1482455	Chromosome	13/06/2008
17	ASM30688v1	GCA_000306885.1	Wolbachia endosymbiont -		wOo				Annotation s	957990	Complete Ge	30/07/2012
18	ASM37658v1	GCA_000376585.1	Wolbachia endosymbiont -		wNo				Annotation s	1301823	Complete Ge	22/04/2013
19	ASM37660v1	GCA_000376605.1	Wolbachia endosymbiont -		wHa				Annotation s	1295804	Complete Ge	22/04/2013
20	ASM43975v1	GCA_000439755.1	Anaplasma phagocytophil		HZ2				NCBI Prokary	1477581	Complete Ge	24/07/2013
21	ASM43977v1	GCA_000439775.1	Anaplasma phagocytophil		JM				NCBI Prokary	1481598	Complete Ge	24/07/2013
22	ASM43979v1	GCA_000439795.1	Anaplasma phagocytophil		Dog2				NCBI Prokary	1471302	Chromosome	24/07/2013
23	ASM49549v1	GCA_000495495.1	Anaplasma marginale str. (		Gypsy Plains				NCBI Prokary	1198622	Chromosome	05/11/2013
24	ASM49553v1	GCA_000495535.1	Anaplasma marginale str. I		Dawn				NCBI Prokary	1196760	Chromosome	05/11/2013
25	ASM50822v1	GCA_000508225.1	Ehrlichia muris AS145		AS145				NCBI Prokary	1196717	Complete Ge	16/12/2013
26	ASM63281v1	GCA_000632815.1	Ehrlichia chaffeensis str. H		Heartland				Annotation s	1172721	Complete Ge	17/04/2014
27	ASM63284v1	GCA_000632845.1	Ehrlichia sp. HF		HF				Annotation s	1148904	Complete Ge	17/04/2014
28	ASM63286v1	GCA_000632865.1	Ehrlichia chaffeensis str. Ja		Jax				Annotation s	1176890	Complete Ge	17/04/2014
29	ASM63288v1	GCA_000632885.1	Ehrlichia chaffeensis str. Li		Liberty				Annotation s	1176202	Complete Ge	17/04/2014
30	ASM63290v1	GCA_000632905.1	Ehrlichia chaffeensis str. O		Osceola				Annotation s	1175197	Complete Ge	17/04/2014
31	ASM63292v1	GCA_000632925.1	Ehrlichia chaffeensis str. S		Saint Vincent				Annotation s	1173884	Complete Ge	17/04/2014
32	ASM63294v1	GCA_000632945.1	Ehrlichia chaffeensis str. W		Wakulla				Annotation s	1174357	Complete Ge	17/04/2014
33	ASM63296v1	GCA_000632965.1	Ehrlichia chaffeensis str. W		West Paces				Annotation s	1170935	Complete Ge	17/04/2014
34	ASM63298v1	GCA_000632985.1	Neorickettsia helminthoec		Oregon				Annotation s	884232	Complete Ge	17/04/2014
35	ASM68963v2	GCA_000689635.2	Anaplasma phagocytophil		Norway variant2				NCBI Prokary	1545197	Complete Ge	03/05/2016
36	ASM82931v1	GCA_000829315.1	Wolbachia endosymbiont -		wCle				Annotation s	1250060	Complete Ge	10/06/2014
37	Wau001	GCA_000953315.1	Wolbachia endosymbiont of Drosophila simulans wAu						Annotation s	1268461	Complete Ge	15/10/2014
38	wTPRE_1.0	GCA_001439985.1	Wolbachia endosymbiont -		wTpre				Annotation s	1133809	Chromosome	07/01/2016
39	ASM175856v1	GCA_001758565.1	Wolbachia endosymbiont -		winc_Cu				NCBI Prokary	1267840	Chromosome	11/10/2016
40	ASM175858v1	GCA_001758585.1	Wolbachia endosymbiont -		winc_SM				NCBI Prokary	1267664	Chromosome	11/10/2016
41	ASM193175v2	GCA_001931755.2	Wolbachia endosymbiont -		Berlin				NCBI Prokary	1801626	Complete Ge	25/06/2018
42	ASM221462v2	GCA_002214625.2	Anaplasma ovis str. Haibei		Haibei				NCBI Prokary	1214674	Complete Ge	05/07/2018
43	ASM237484v2	GCA_002374845.2	Wolbachia pipientis		wAlbB-HN2016				NCBI Prokary	1483853	Complete Ge	31/07/2019
44	ASM237914v2	GCA_002379145.2	Wolbachia pipientis		wAlbB-FL2016				NCBI Prokary	1482279	Complete Ge	31/07/2019
45	ASM287999v1	GCA_002879995.1	Ehrlichia canis		YZ-1				NCBI Prokary	1314789	Complete Ge	15/01/2018
46	ASM351567v1	GCA_003515675.1	Anaplasma marginale		Palmeira				NCBI Prokary	1195200	Chromosome	10/09/2018
47	ASM351573v1	GCA_003515735.1	Anaplasma marginale		Jaboticabal				NCBI Prokary	1195321	Chromosome	10/09/2018
48	ASM399958v1	GCA_003999585.1	Wolbachia endosymbiont -		China 1				NCBI Prokary	1306495	Chromosome	08/01/2019
49	ASM417128v1	GCA_004171285.1	Wolbachia pipientis wAlbI		wAlbB				NCBI Prokary	1484007	Complete Ge	12/02/2019
50	ASM479593v1	GCA_004795935.1	Wolbachia endosymbiont of Brugia malayi				TRS		NCBI Prokary	1080064	Complete Ge	16/04/2019
51	ASM479595v1	GCA_004795955.1	Wolbachia endosymbiont -		wMau				NCBI Prokary	1273527	Complete Ge	16/04/2019
52	ASM479597v1	GCA_004795975.1	Wolbachia endosymbiont -		wMau				NCBI Prokary	1273530	Complete Ge	16/04/2019
53	ASM654229v1	GCA_006542295.1	Wolbachia endosymbiont of Carposina sasakii				wCauA		NCBI Prokary	1449344	Complete Ge	02/07/2019
54	ASM797168v1	GCA_007971685.1	Wolbachia pipientis		wMel_N25				NCBI Prokary	1267781	Complete Ge	12/08/2019
55	ASM797259v1	GCA_007972595.1	Wolbachia pipientis		wMel_I23				NCBI Prokary	1269137	Complete Ge	12/08/2019
56	ASM797274v1	GCA_007972745.1	Wolbachia pipientis		wMel_ZH26				NCBI Prokary	1267436	Complete Ge	12/08/2019
57	ASM803321v1	GCA_008033215.1	Wolbachia endosymbiont -		W2.1				Annotation s	1401460	Complete Ge	20/08/2019

5) Pairwise genome alignment

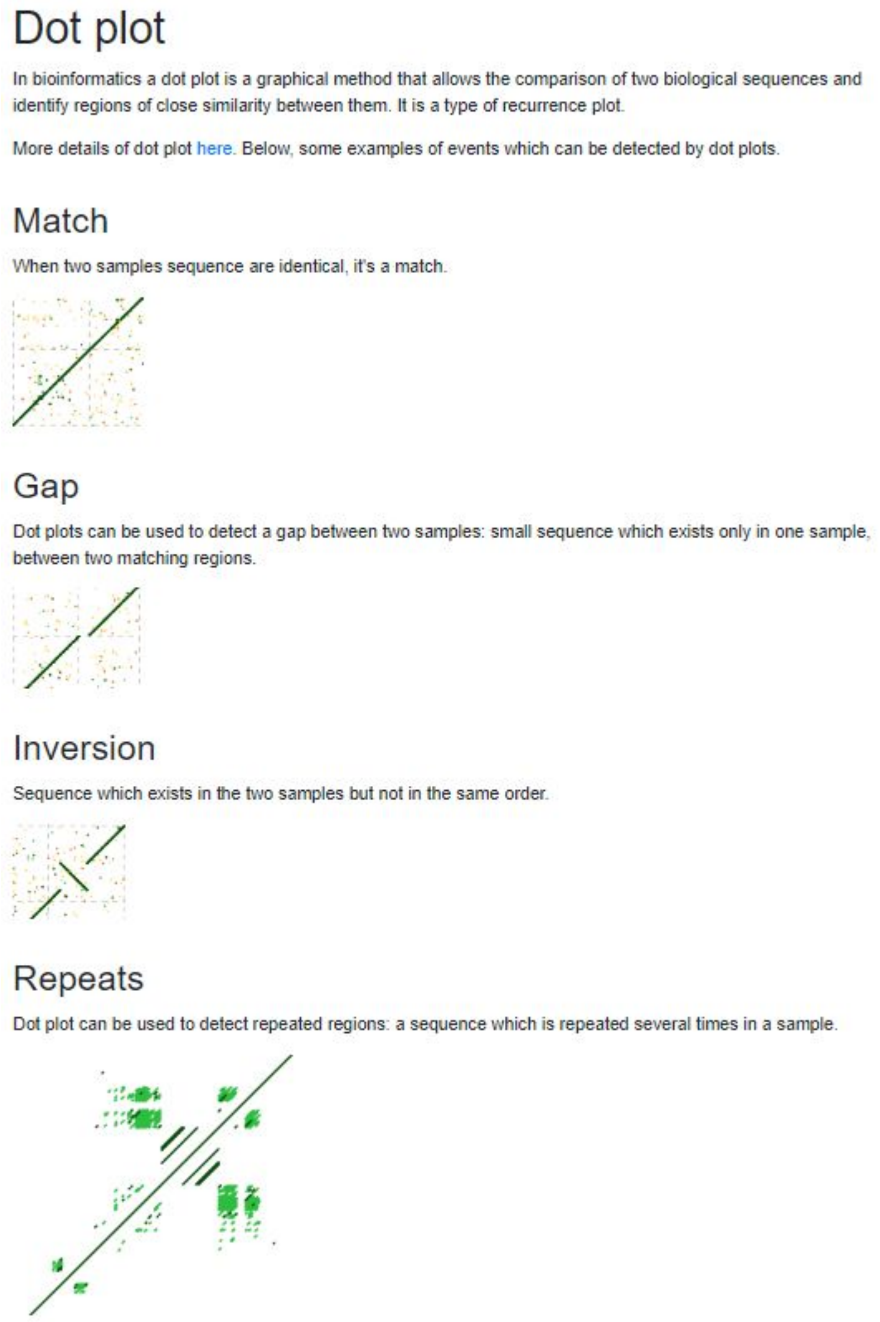
Dot plot



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

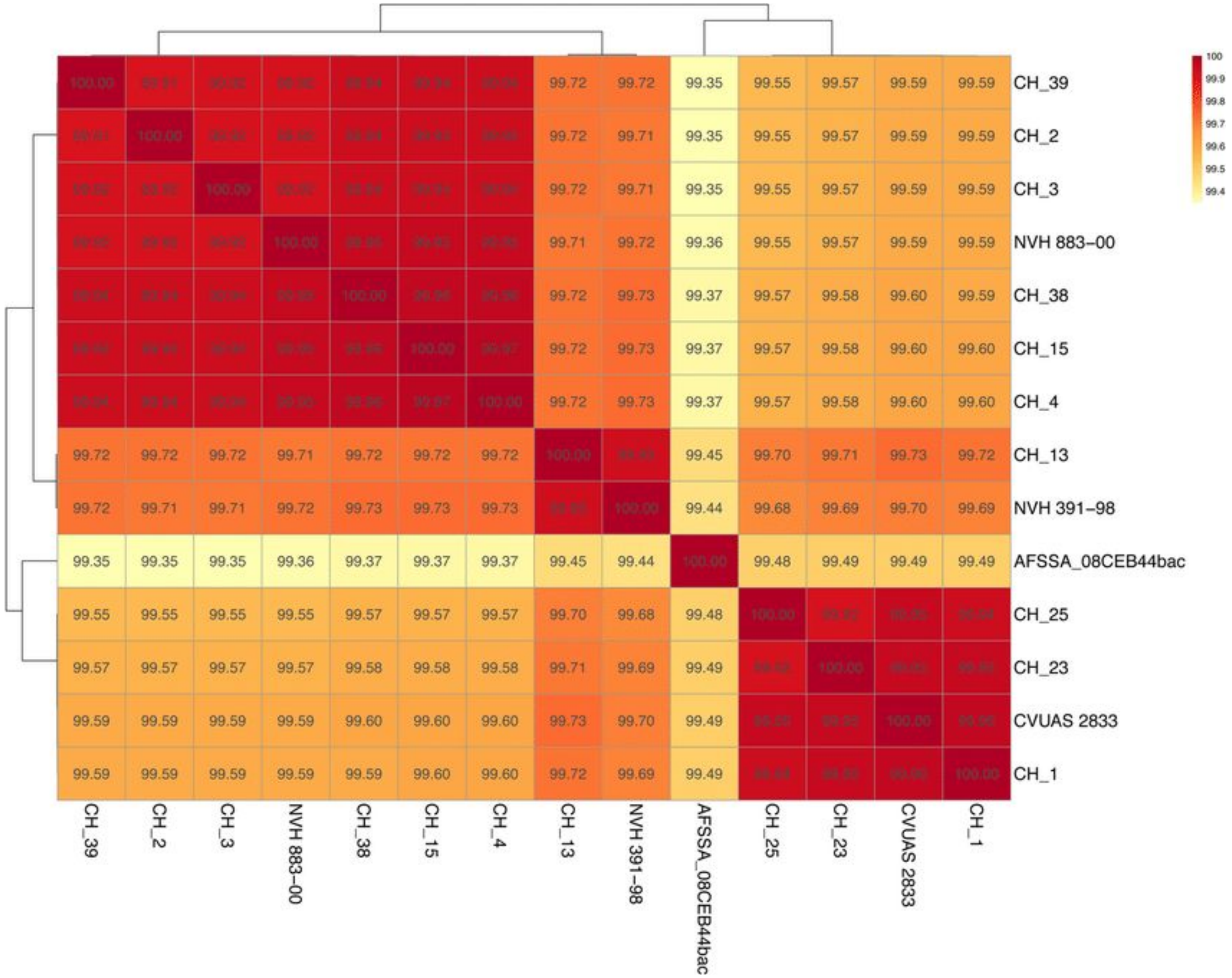
Dot plot can be used to detect repeated regions: a sequence which is repeated several times in a sample.

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

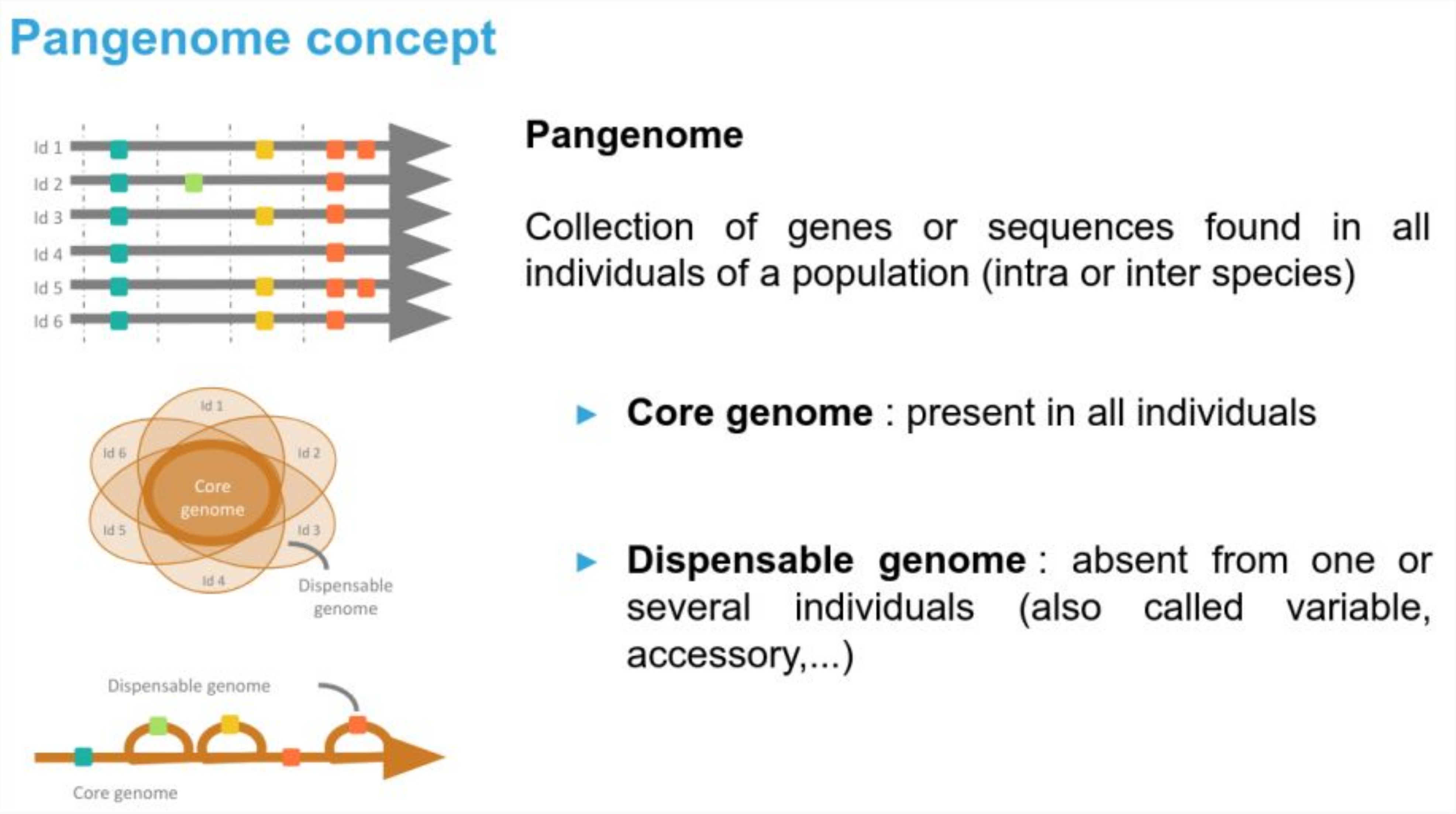


	CH_39	CH_2	CH_3	NVH 883–00	CH_38	CH_15	CH_4	CH_13	NVH 391–98	AFSSA_08CEB44bac	CH_25	CH_23	CVUAS 2833	CH_1
CH_39	100.00	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	99.72	99.72	99.35	99.55	99.57	99.59	99.59
CH_2	[illegible]	100.00	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	99.72	99.71	99.35	99.55	99.57	99.59	99.59
CH_3	[illegible]	[illegible]	100.00	[illegible]	[illegible]	[illegible]	[illegible]	99.72	99.71	99.35	99.55	99.57	99.59	99.59
NVH 883–00	[illegible]	[illegible]	[illegible]	100.00	[illegible]	[illegible]	[illegible]	99.71	99.72	99.36	99.55	99.57	99.59	99.59
CH_38	[illegible]	[illegible]	[illegible]	[illegible]	100.00	[illegible]	[illegible]	99.72	99.73	99.37	99.57	99.58	99.60	99.59
CH_15	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	100.00	99.97	99.72	99.73	99.37	99.57	99.58	99.60	99.60
CH_4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	99.97	100.00	99.72	99.73	99.37	99.57	99.58	99.60	99.60
CH_13	99.72	99.72	99.72	99.71	99.72	99.72	99.72	100.00	[illegible]	99.45	99.70	99.71	99.73	99.72
NVH 391–98	99.72	99.71	99.71	99.72	99.73	99.73	99.73	[illegible]	100.00	99.44	99.68	99.69	99.70	99.69
AFSSA_08CEB44bac	99.35	99.35	99.35	99.36	99.37	99.37	99.37	99.45	99.44	100.00	99.48	99.49	99.49	99.49
CH_25	99.55	99.55	99.55	99.55	99.57	99.57	99.57	99.70	99.68	99.48	100.00	[illegible]	[illegible]	[illegible]
CH_23	99.57	99.57	99.57	99.57	99.58	99.58	99.58	99.71	99.69	99.49	[illegible]	100.00	[illegible]	[illegible]
CVUAS 2833	99.59	99.59	99.59	99.59	99.60	99.60	99.60	99.73	99.70	99.49	[illegible]	[illegible]	100.00	[illegible]
CH_1	99.59	99.59	99.59	99.59	99.59	99.60	99.60	99.72	99.69	99.49	[illegible]	[illegible]	[illegible]	100.00

Heat map of the average nucleotide identity (ANI) for strains of the species B. cytotoxicus *(Stevens et al., 20.19)*

7) Pan-genome and Gene clustering

Le Pangénome, un concept simple et complexe



Pangenome concept

Pangenome

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

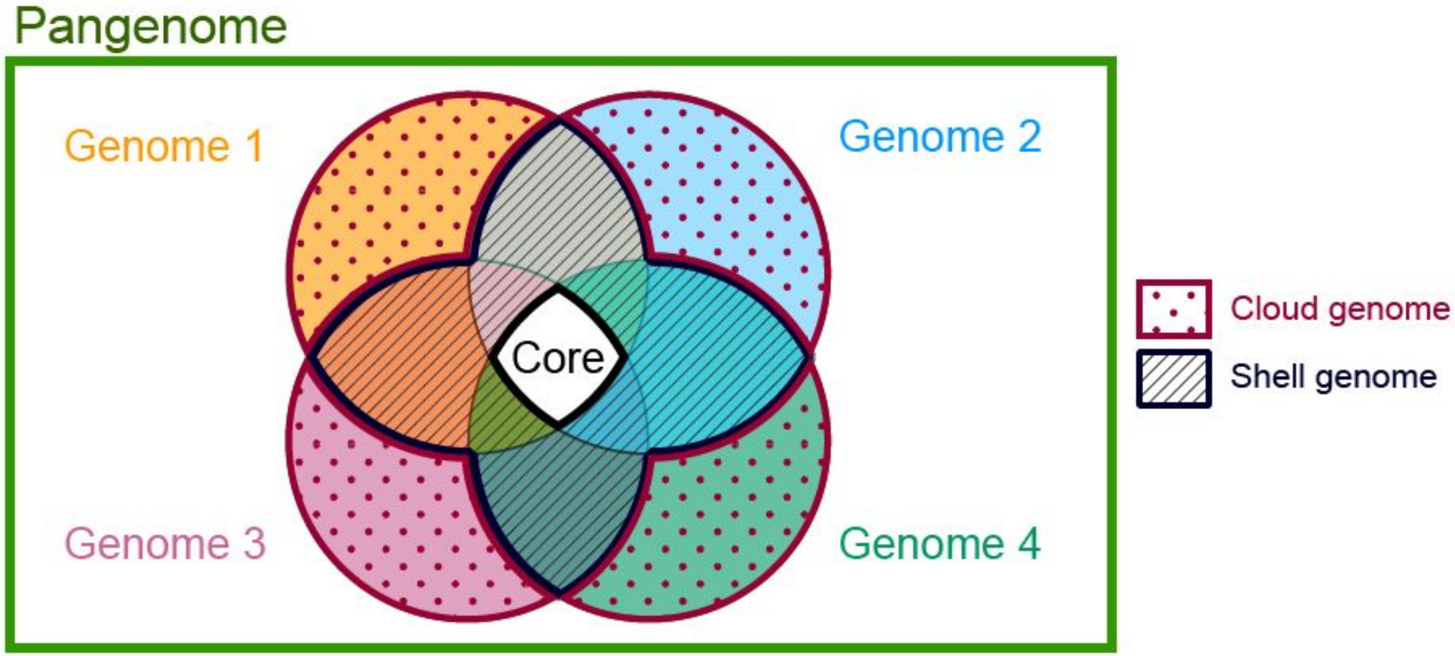
- **Core genome** : present in all individuals
- **Dispensable genome** : absent from one or several individuals (also called variable, accessory,...)

Tranchant-Dubreuil, Rouard, Sabot

Gene Clustering - how it works

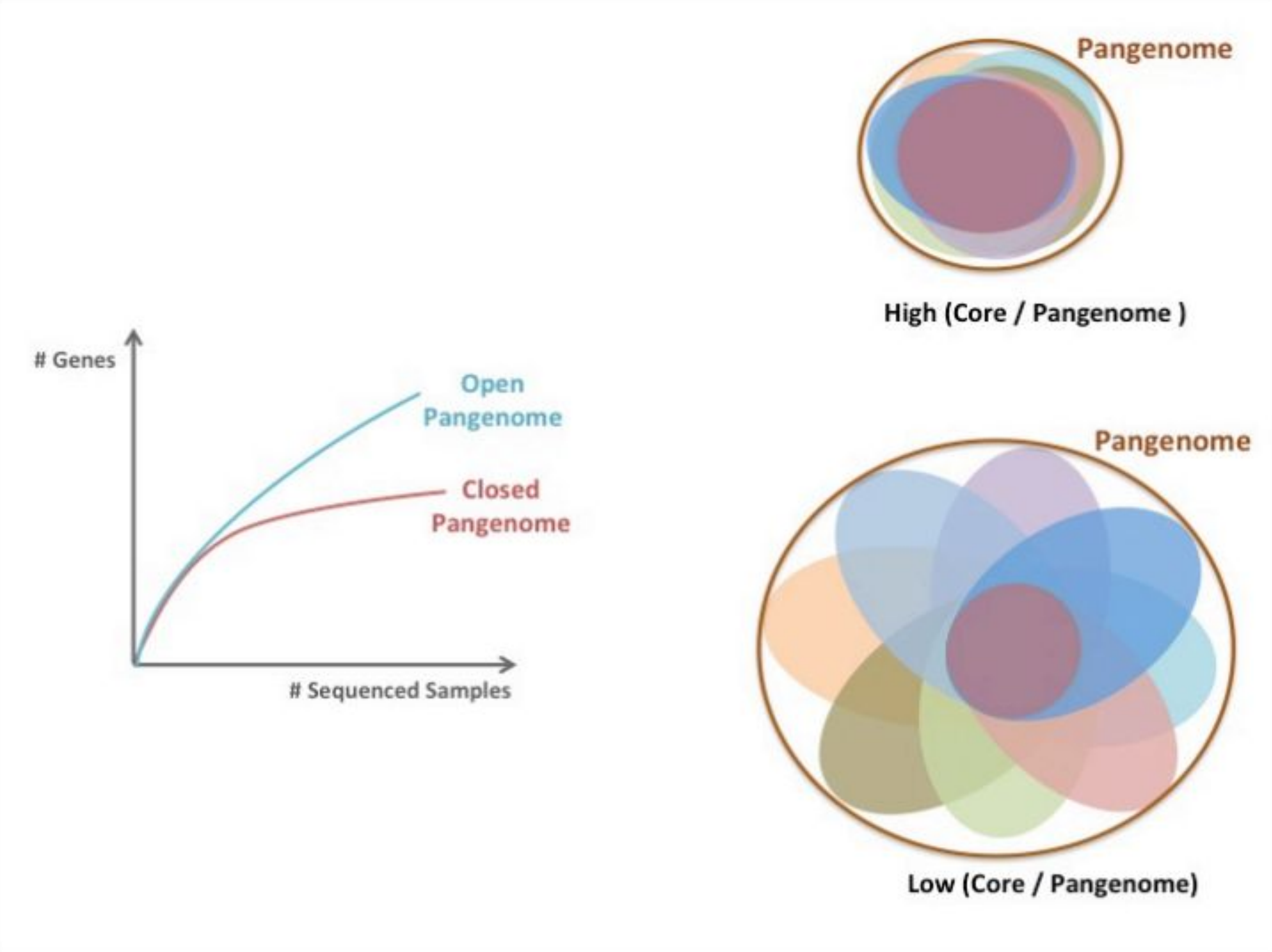
- Assess the similarity of every gene to every other gene
    - e.g., using BLAST
- Use that similarity to join pairs of genes
    - e.g., using Reciprocal Best Hits
- Connect the gene pairs into larger clusters
    - e.g., using Reciprocal Best Hits or Markov clustering

=> Programs: OrthoMCL, Roary, PGAP…



Accessory genome
=
Dispensable genome

Le pangénome ouvert, fermé, le ratio C/P



Tranchant-Dubreuil, Rouard & Sabot, 2018

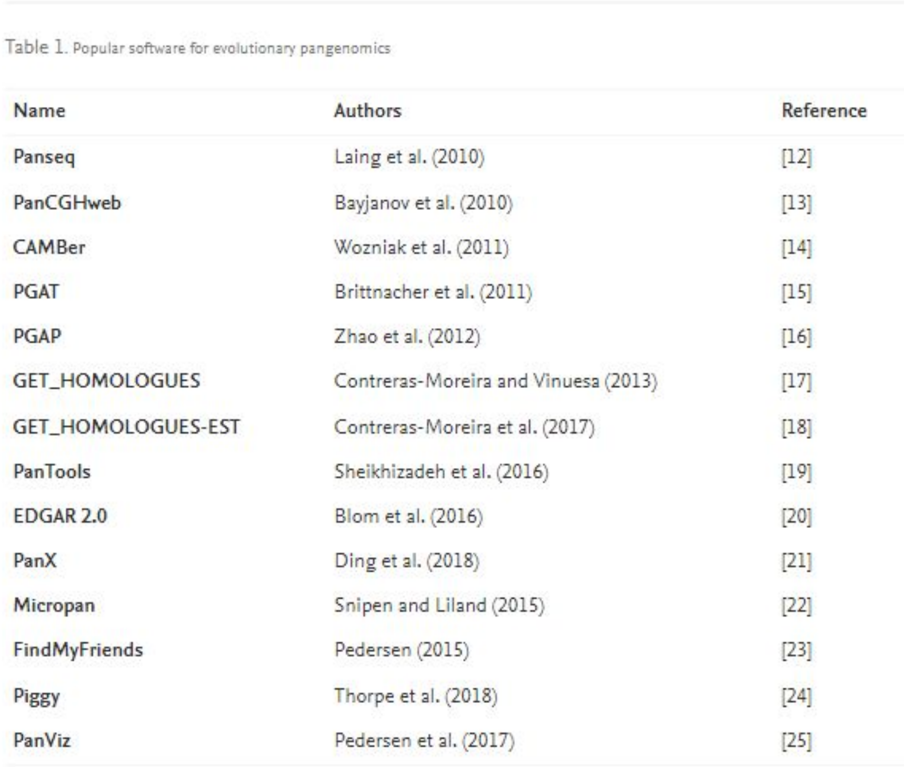
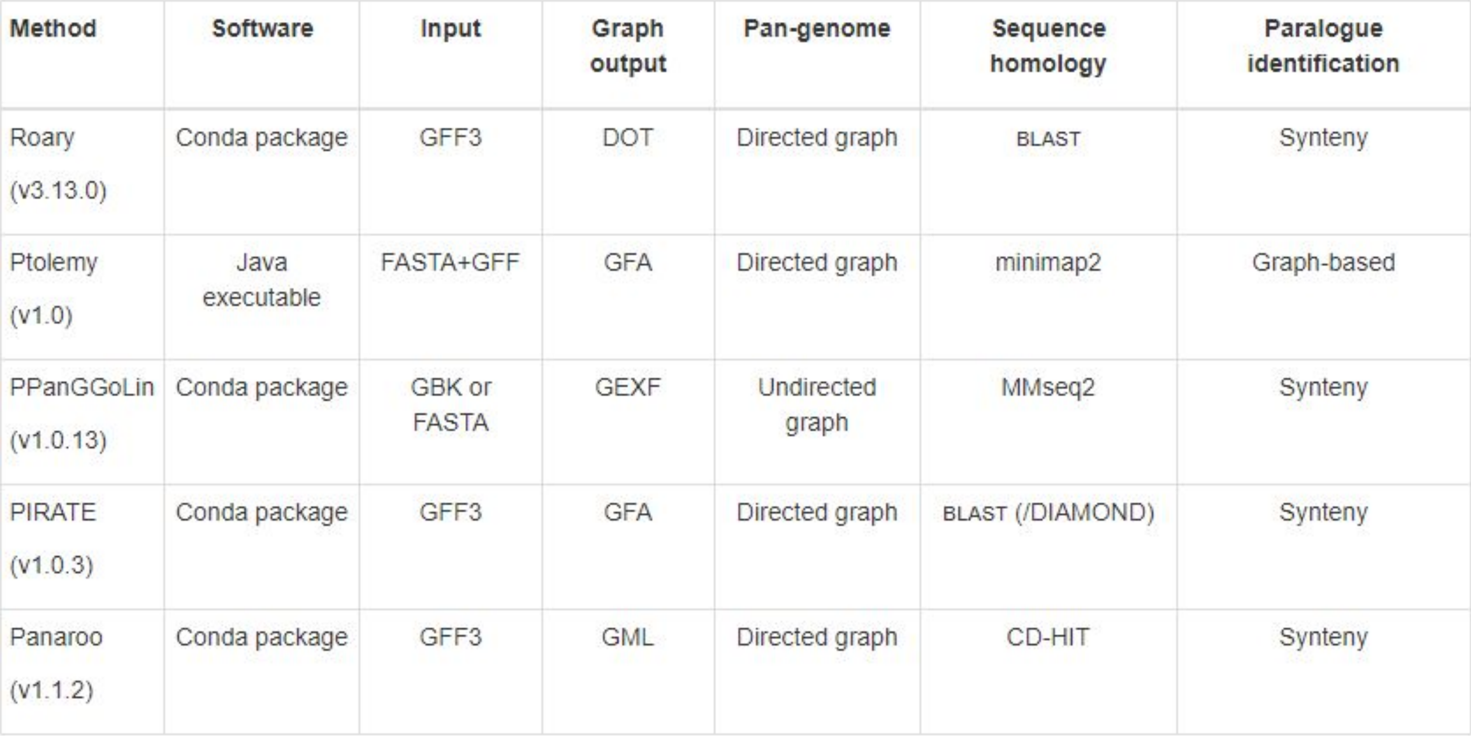
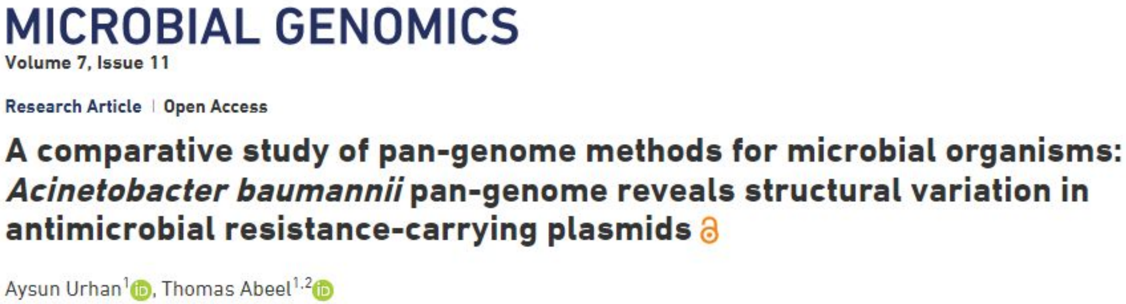


Table 1. Popular software for evolutionary pangenomics

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



MICROBIAL GENOMICS
Volume 7, Issue 11

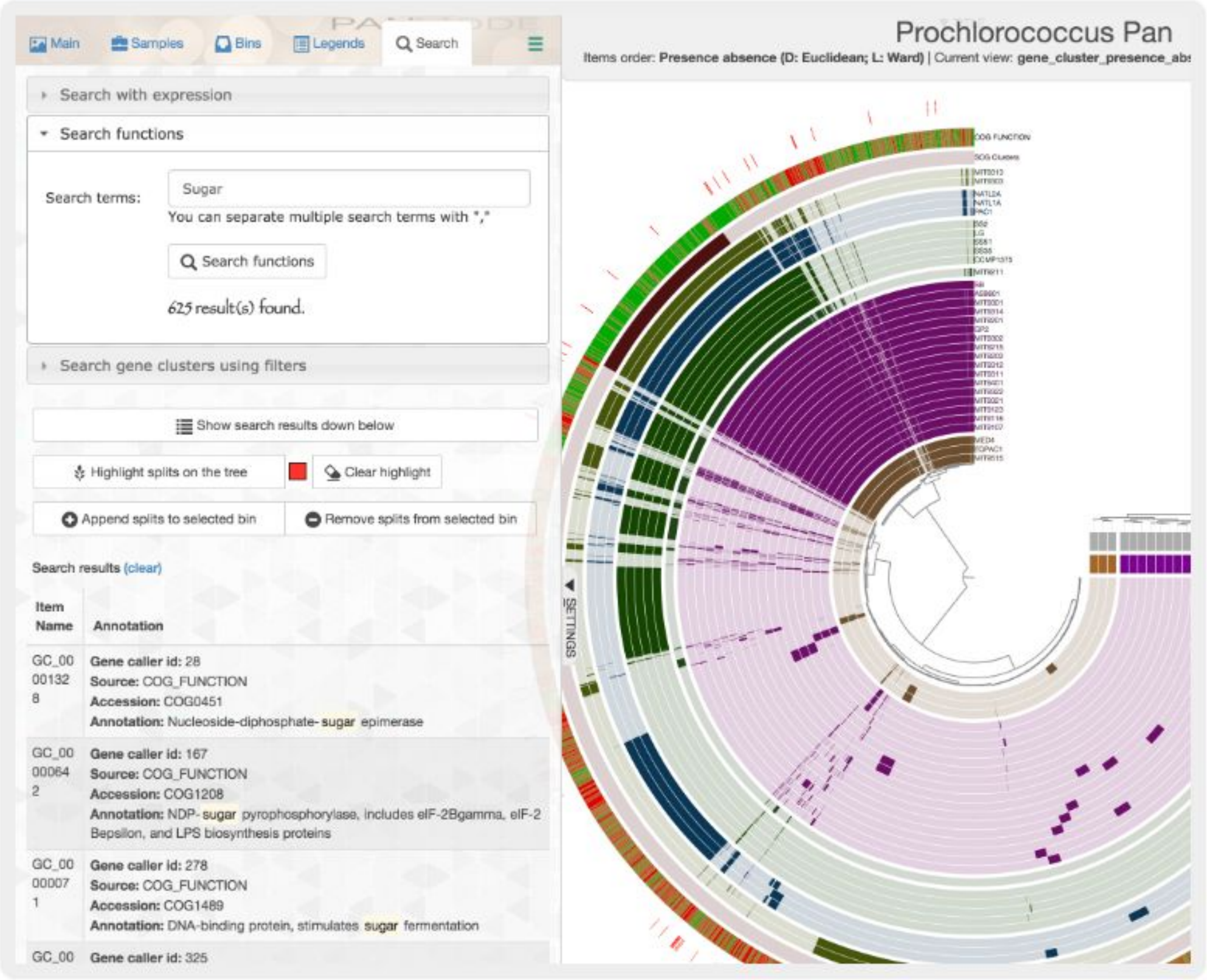
Research Article | Open Access

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

Aysun Urhan[1], Thomas Abeel[1,2]

An anvi'o workflow for microbial pangenomics

https://merenlab.org/2016/11/08/pangenomics-v2/



Prochlorococcus Pan

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

Search terms: Sugar

You can separate multiple search terms with ","

625 result(s) found.

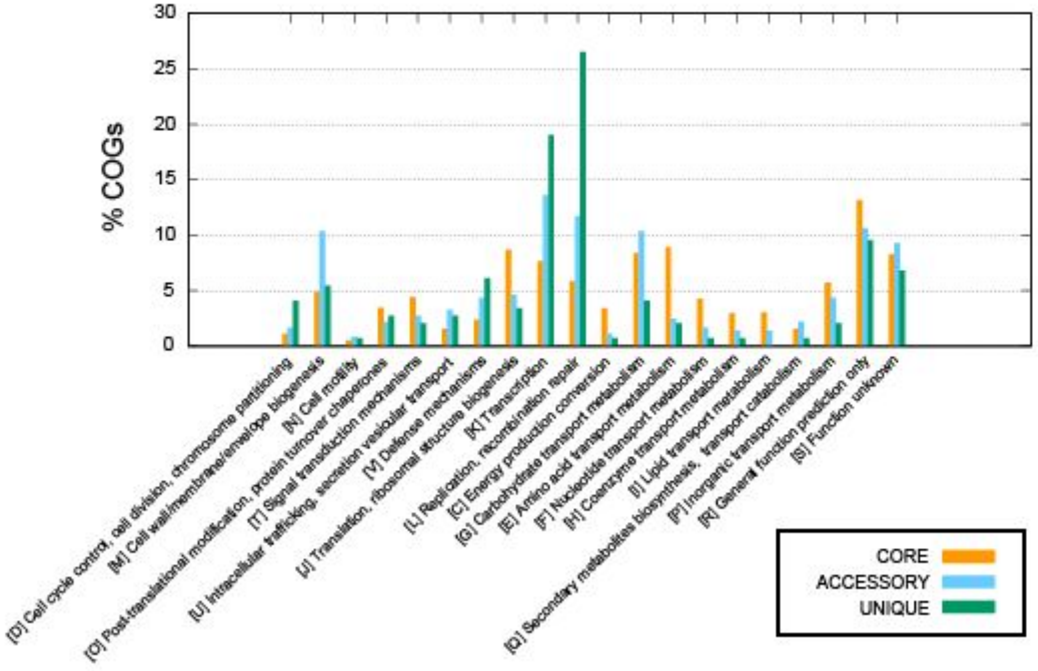
Search results (clear)

Item Name	Annotation
GC_00001328	**Gene caller id:** 28 **Source:** COG_FUNCTION **Accession:** COG0451 **Annotation:** Nucleoside-diphosphate- sugar epimerase
GC_00000642	**Gene caller id:** 167 **Source:** COG_FUNCTION **Accession:** COG1208 **Annotation:** NDP- sugar pyrophosphorylase, includes eIF-2Bgamma, eIF-2Bepsilon, and LPS biosynthesis proteins
GC_00000071	**Gene caller id:** 278 **Source:** COG_FUNCTION **Accession:** COG1489 **Annotation:** DNA-binding protein, stimulates sugar fermentation
GC_00	**Gene caller id:** 325

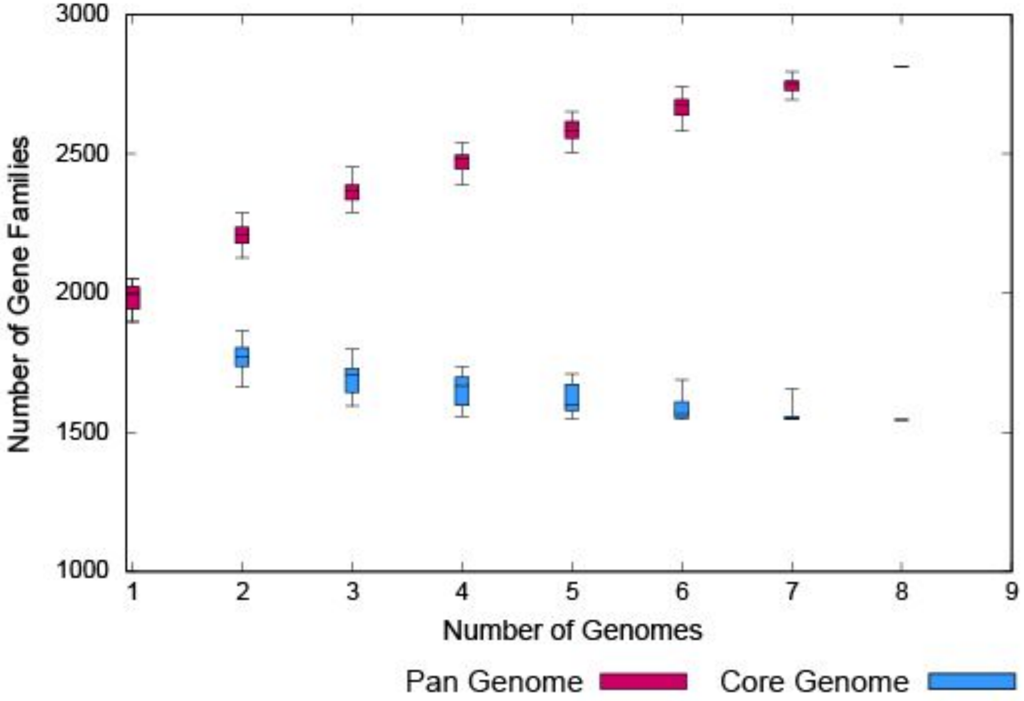
BPGA (Bacterial Pan Genome Analysis tool)

Streptococcus agalactiae

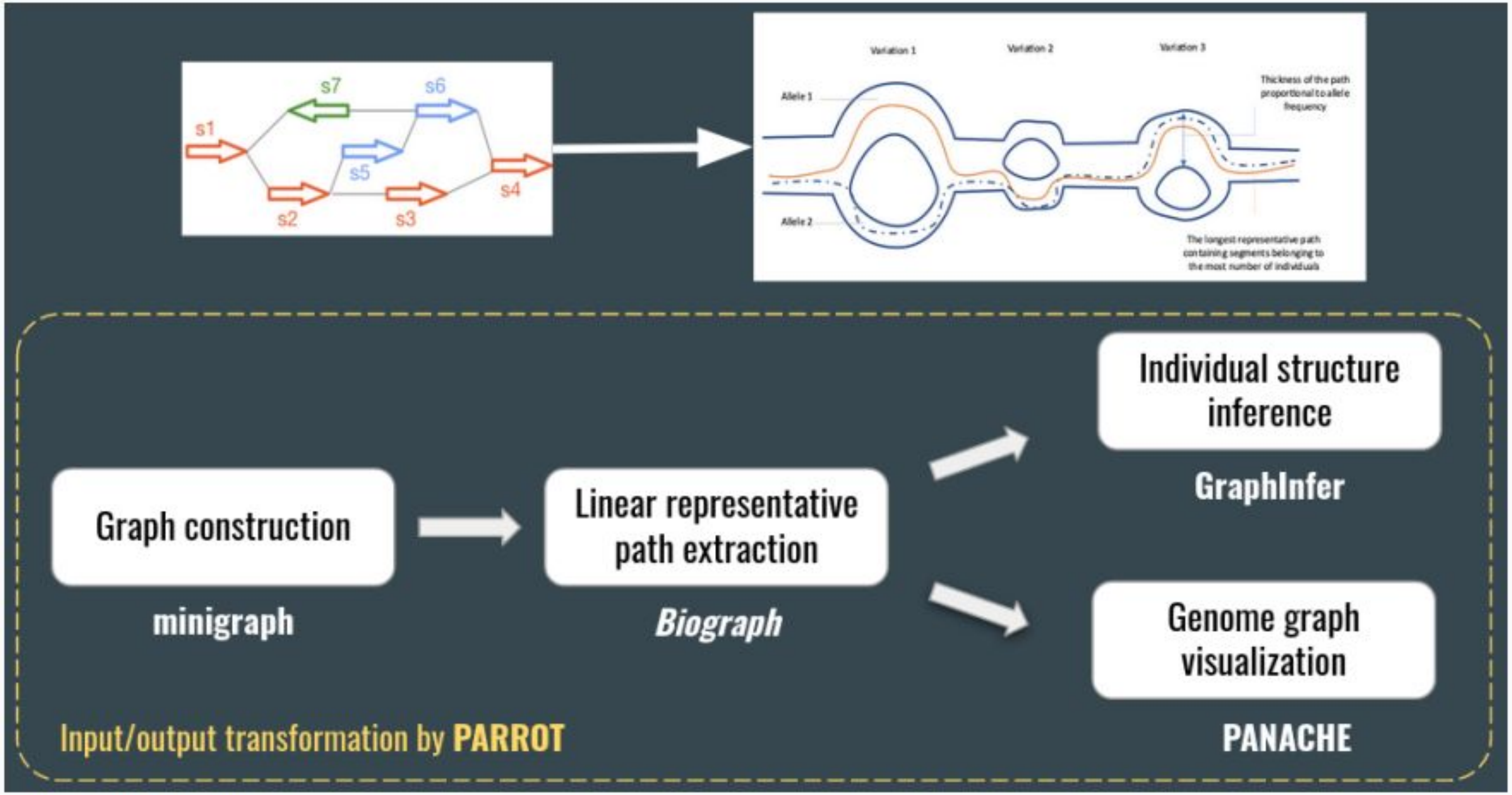
COG Distribution



Pan and Core Genome Plot



Comment manipuler le graphe pour les biologistes ?


s1
s7
s6
s5
s2
s3
s4
Variation 1
Variation 2
Variation 3
Allele 1
Allele 2
Thickness of the path proportional to allele frequency
The longest representative path containing segments belonging to the most number of individuals

Graph construction

minigraph

Linear representative path extraction

Biograph

Individual structure inference

GraphInfer

Genome graph visualization

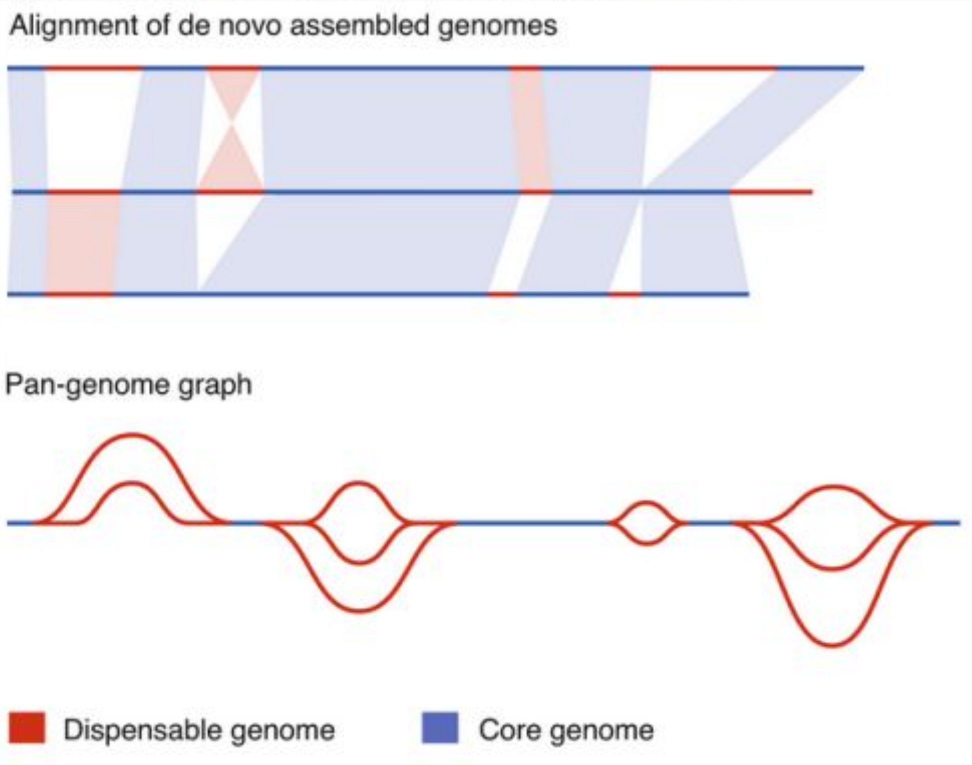
PANACHE

Input/output transformation by **PARROT**

Dang, Do and Sabot, 2022

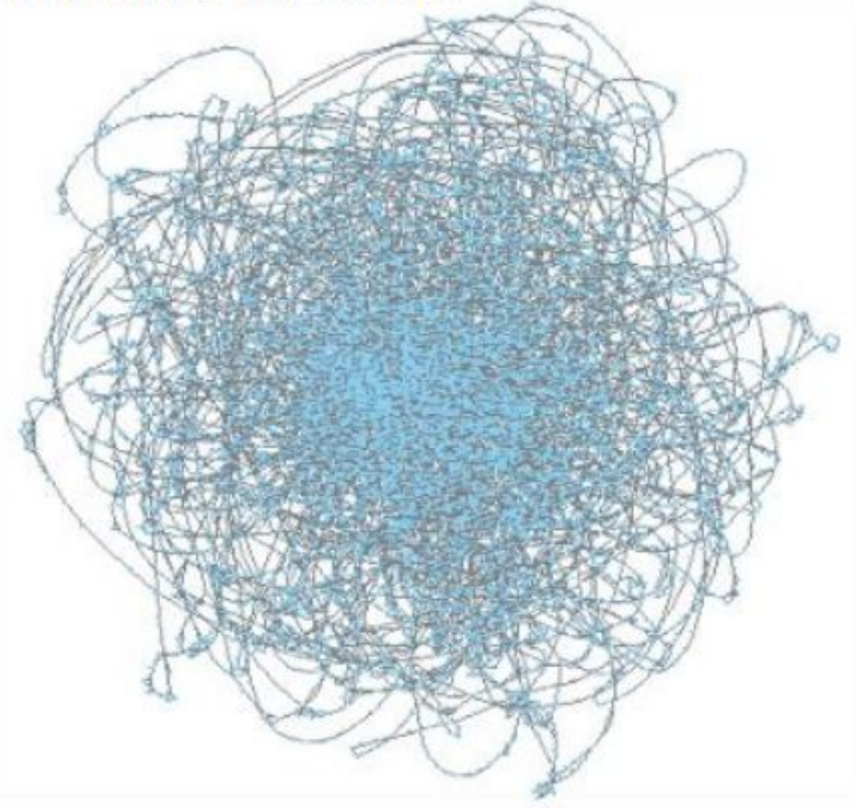
La représentation

Concept du graphe de génome



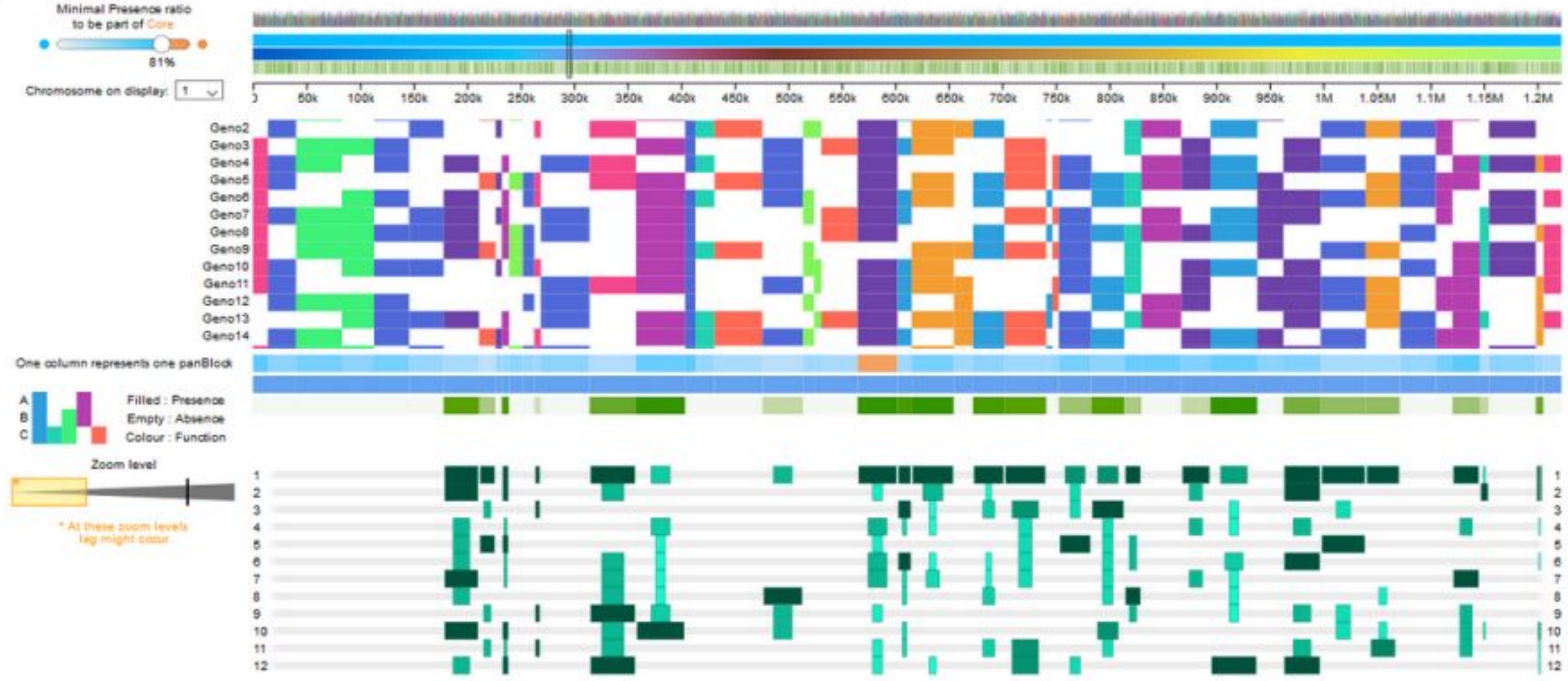
Bayer et al., 2020

The HairBall effect



GigaSciences blog

Un exemple linéaire, Panache



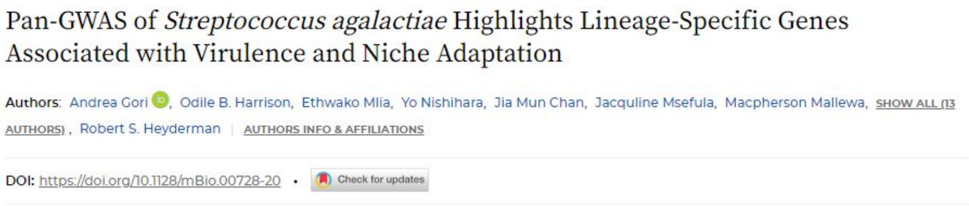
Minimal Presence ratio to be part of Core: 81%

Chromosome on display: 1

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, Jacquline 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

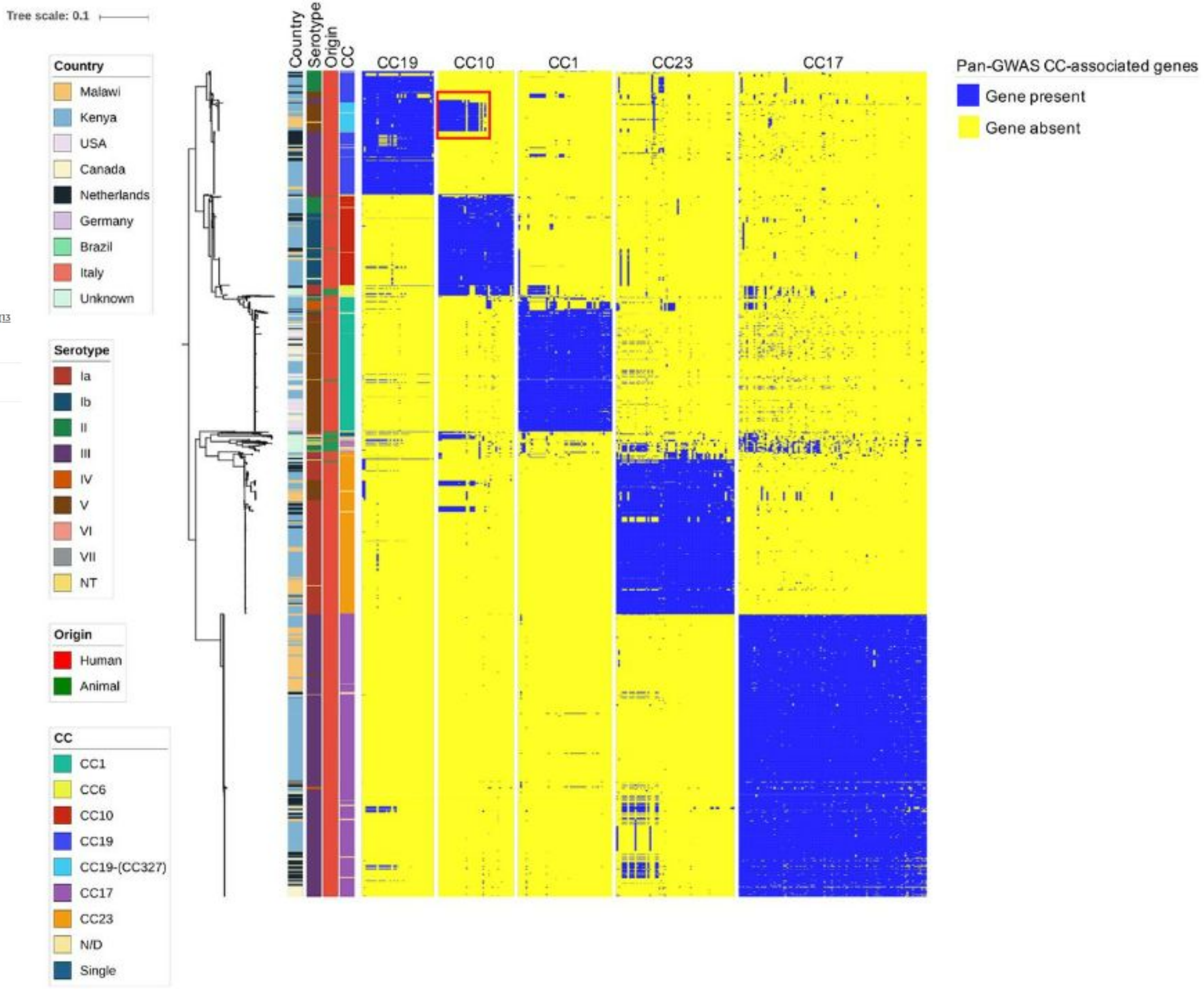


FIG 2 Core genome-based population structure of GBS. The phylogenetic tree is annotated with 4 colored strips representing the clonal complex, the country of isolation, the origin, and the serotype of each strain. The three binary heatmaps represent the presence (blue) or absence (yellow) of the genes 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.

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.

South Green bioinformatics platform

Merci pour votre attention !

SUIVEZ NOUS SUR TWITTER !

South Green : @green_bioinfo

I-Trop : @ItropBioinfo

Merci de prendre 5 min pour remplir l’enquête

https://itrop-survey.ird.fr/index.php/515725?lang=fr

South Green bioinformatics platform

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”