SouthGreen bioinformatics platform

Session de formation 2023

diade

PHIM Plant Health Institute Montpellier

Alliance Bioversity & CIAT

agAP

ASTRE Animal - Santé Territoires - Risques - Ecosystèmes

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

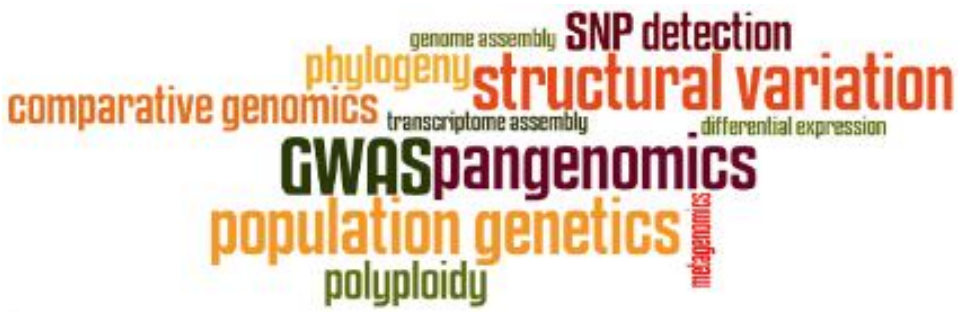
MiVEGEC Maladies Infectieuses et Vecteurs Écologie, Génétique, Évolution et Contrôle

TransVIHMI

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

SouthGreen bioinformatics platform

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



Mutualisation

SouthGreen 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+

Storage and computing resources

cirad

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

IRD

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)

Collaborative development of tools

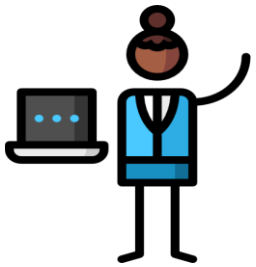
Genomics

- 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/

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

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

i-Trop bioinformatics

TransVIH MI

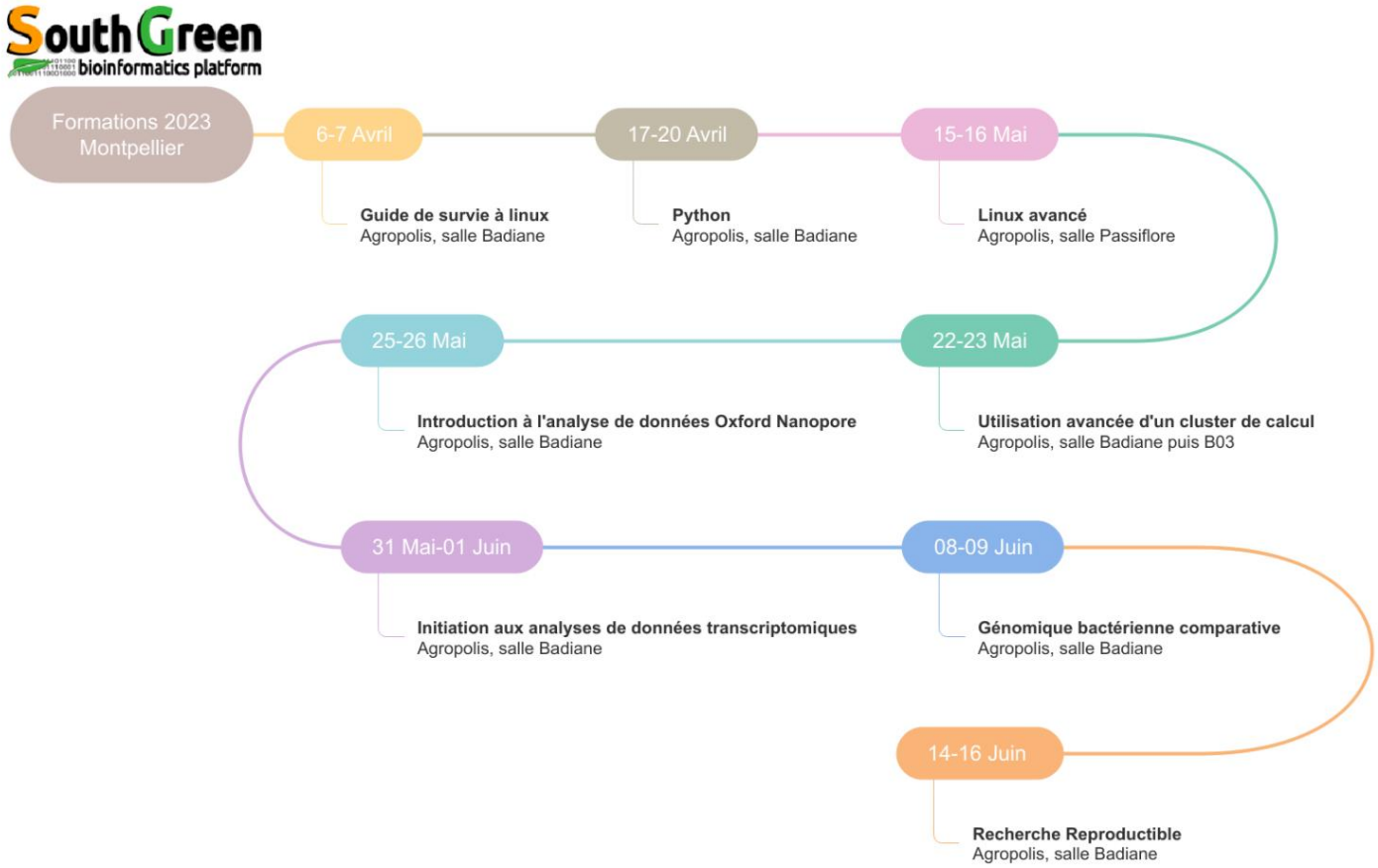
Nicolas Fernandez

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Thomas Denecker

And more collaborators !

South Green bioinformatics platform



Modules de formation 2023

- Toutes nos formations :
  https://southgreenplatform.github.io/trainings/

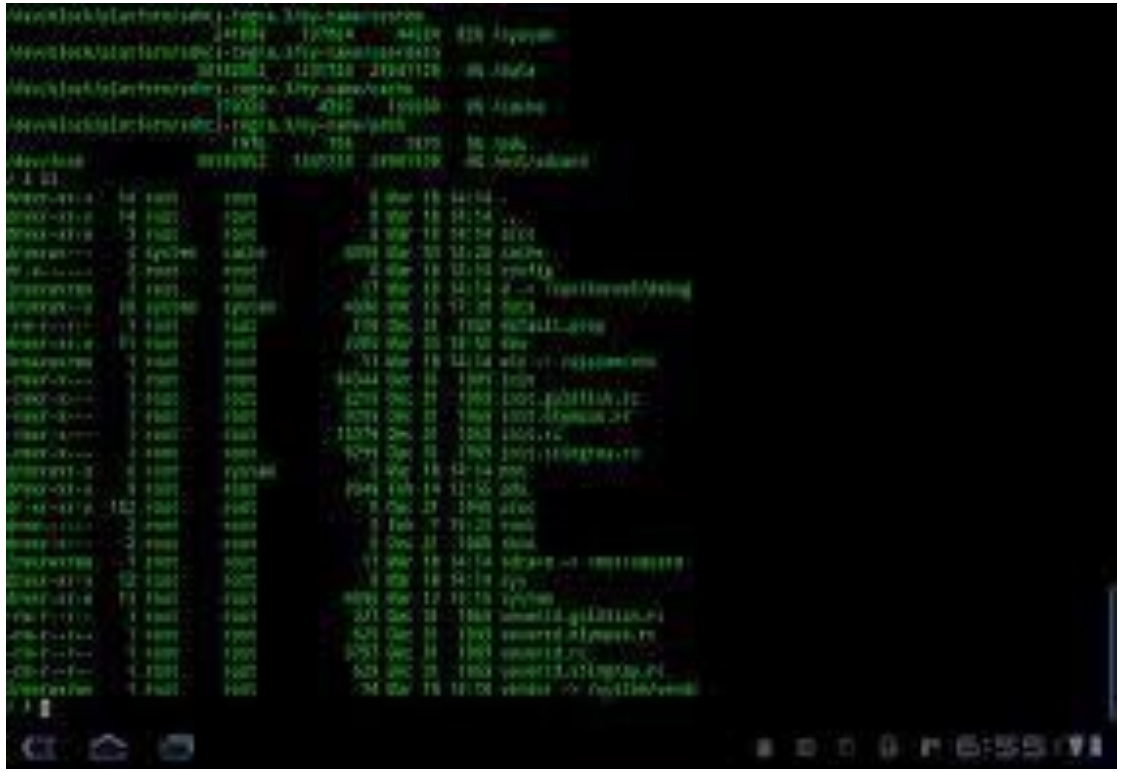
- Topo & TP :
  https://github.com/SouthGreenPlatform/training_ONT_teaching/tree/2023_MTP

- Environnement de travail : Logiciels à installer

Génomique Comparative Bactérienne

- 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 Code nbdiff

Anchoring data analysis

1 - CDHIT data analysis *before anchoring on genome*

1.1 Removing redundancy with CDHIT

- CDHIT Input : 1,306,676 contigs assemblied 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 afer 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]:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

csv_cdhit_file = "/home/christine/Documents/These/Data/CDHIT/ALL_CTGS_MERGE/allCtgsIRIGIN_TOG5681.dedup8095.PANI
df_cdhit= pd.read_csv(csv_cdhit_file,names=['ctg','sp','ctg-list','sp_list'], header=0)
#print(df_cdhit)

```

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

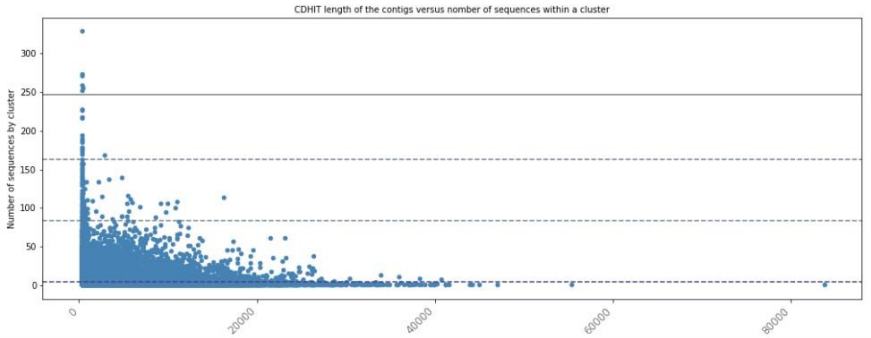
  - code output is directly displayed in the notebook

All ctgs

Entrée [30]:

```
plt.figure(figsize=(17, 6))
ax = plt.gca()
df_cdhit[(df_cdhit.sp == 'Ob')].plot(x='pb', y='ln', kind="scatter", color='steelblue', ax=ax, linewidth=1)
df_cdhit[(df_cdhit.sp == 'Og')].plot(x='pb', y='ln', kind="scatter", color='steelblue', ax=ax, linewidth=1)
plt.axhline(y=5, color='darkslateblue', linestyle='--')
plt.axhline(y=84, color='slategrey', linestyle='--')
plt.axhline(y=163, color='slategrey', linestyle='--')
plt.axhline(y=247, color='grey', linestyle='-')

plt.title("CDHIT length of the contigs versus nomber of sequences within a cluster", fontsize=10)
plt.xlabel('Cluster length')
plt.ylabel('Number of sequences by cluster')
plt.xticks(
    rotation=45,
    horizontalalignment='right',
    fontweight='light',
    fontsize='12',
)
plt.show()
```



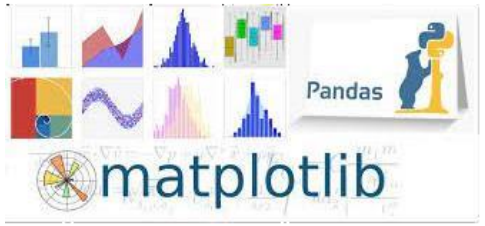
How to become a super datascientist ?

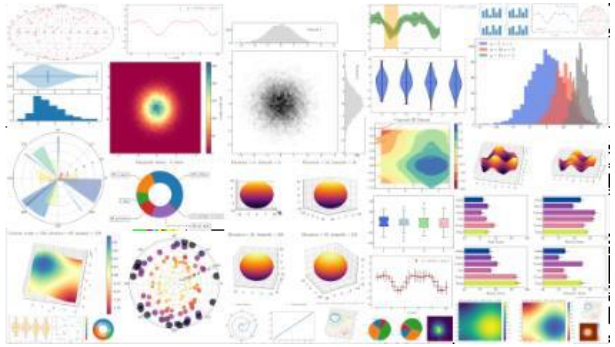
- facilement importer des fichiers tabulés dans des dataframes, similaires aux dataframes sous R.

  (et exporter)

- manipuler ces tableaux de données / DataFrames

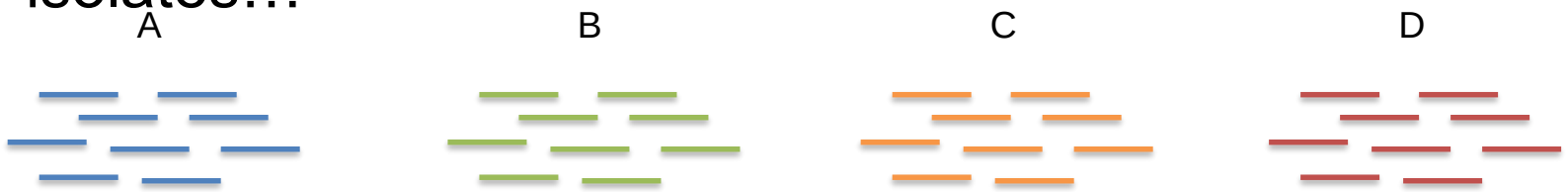
- facilement tracer des graphes à partir de ces DataFrames grâce à matplotlib





Two Approaches to Microbial Genomics

Starting with sets of reads representing your study isolates…



<u>Assembly-based</u>

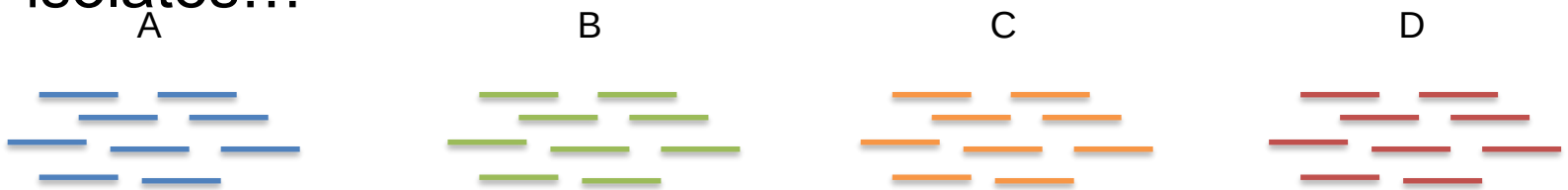
1. Assemble each set of reads into a genome sequence
2. Annotate each genome
3. Cluster genes and compare between each genome

<u>Variant-based</u>

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

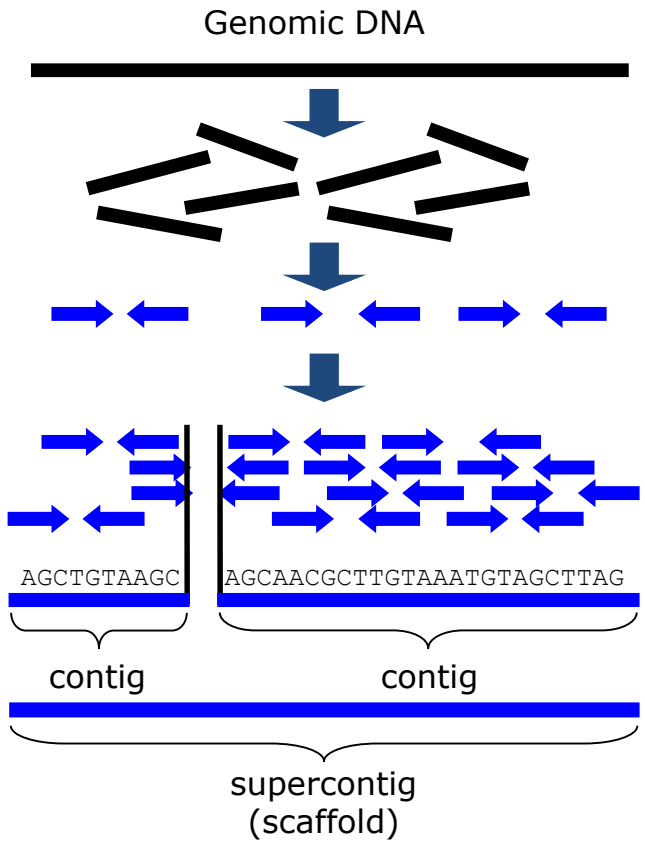
1. Assemble each set of reads into a genome sequence
2. Annotate each genome
3. Cluster genes and compare between each genome

Variant-based

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

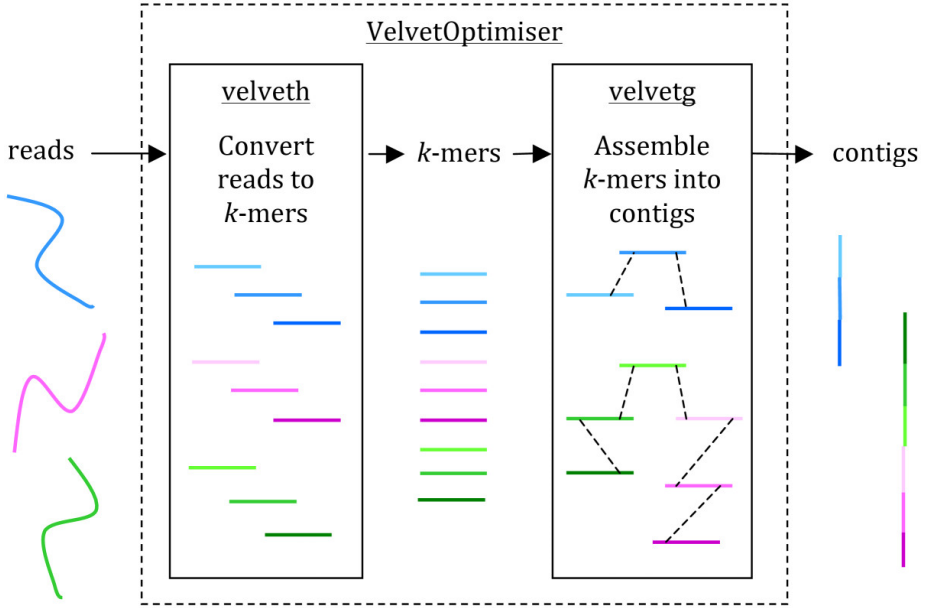
1) Assembly

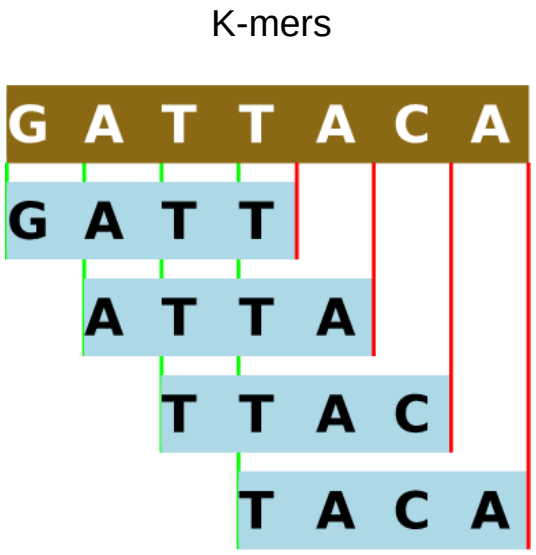
Assembly Basics (de-novo assembly)



Assembly Methods

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





Edwards and Holt 2013

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]

Table 1

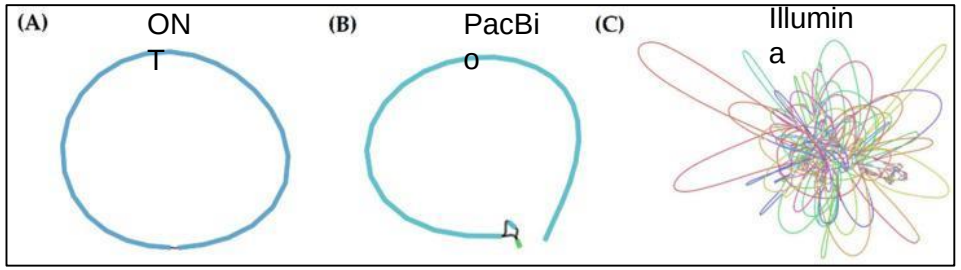
Statistics of genome-assembly results of independent assembly strategies.

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

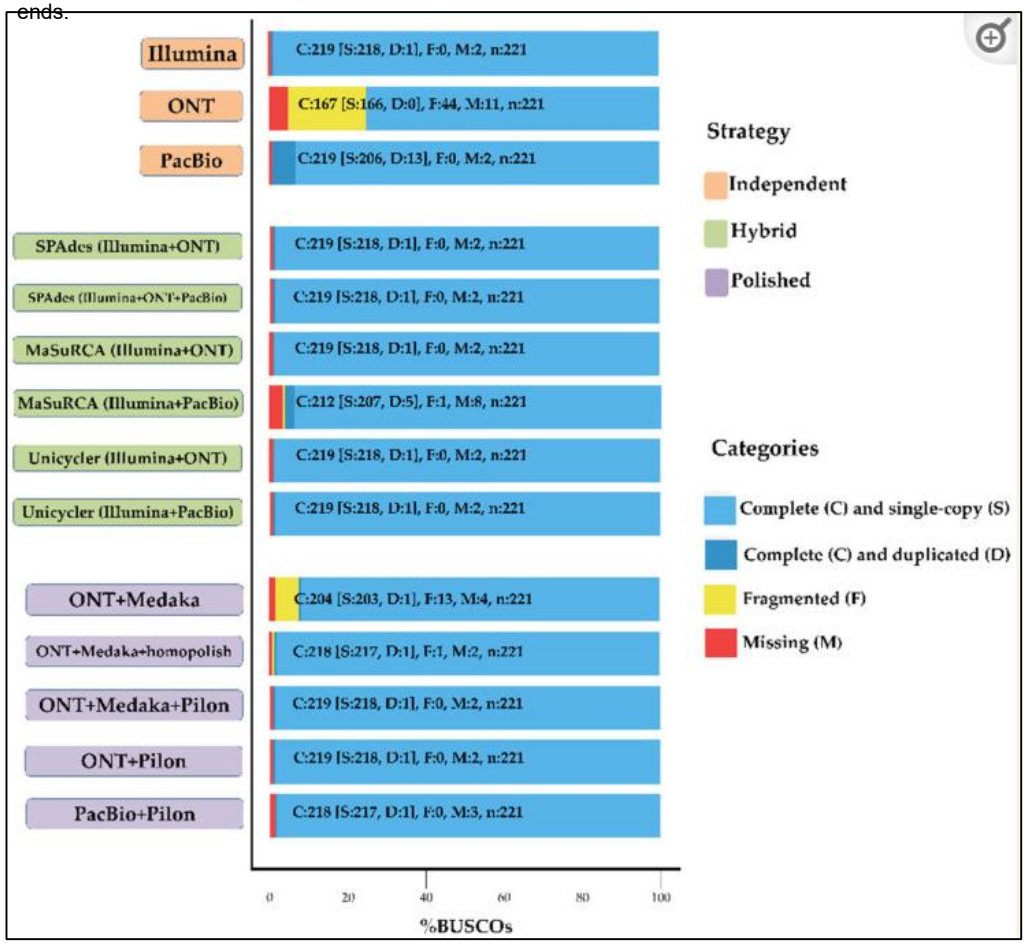
Table 2

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

Sequencing → Sequence data analysis

Oxford NANOPORE Technologies

Oxford NANOPORE Technologies

PACBIO®

baseDmux
- Base calling
- Demultiplexing

CulebrONT
- Assembly
- Circularization
- Polish
- Quality

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

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 ?

<u>Structural annotation:</u>

VS

<u>Functional annotation:</u>

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

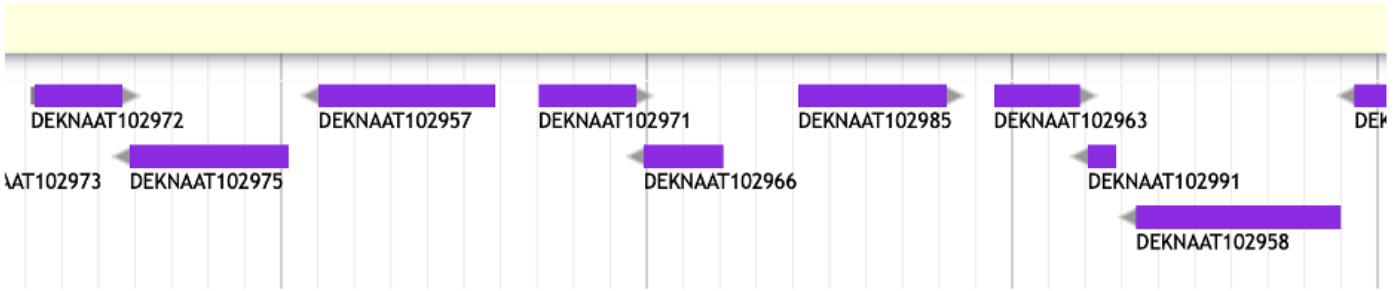
Find out what the regions do. What do they code for?

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

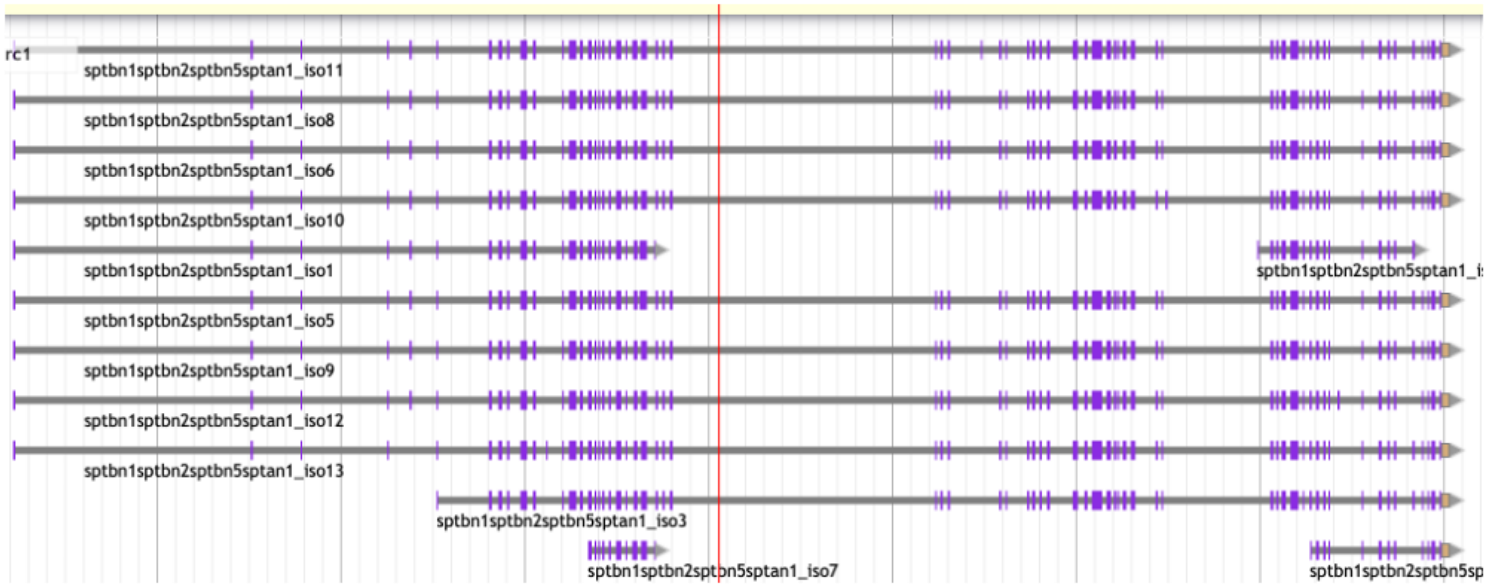
Thx!!
Slide borrowed to NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Organisms differ in genomic complexity

A yeast

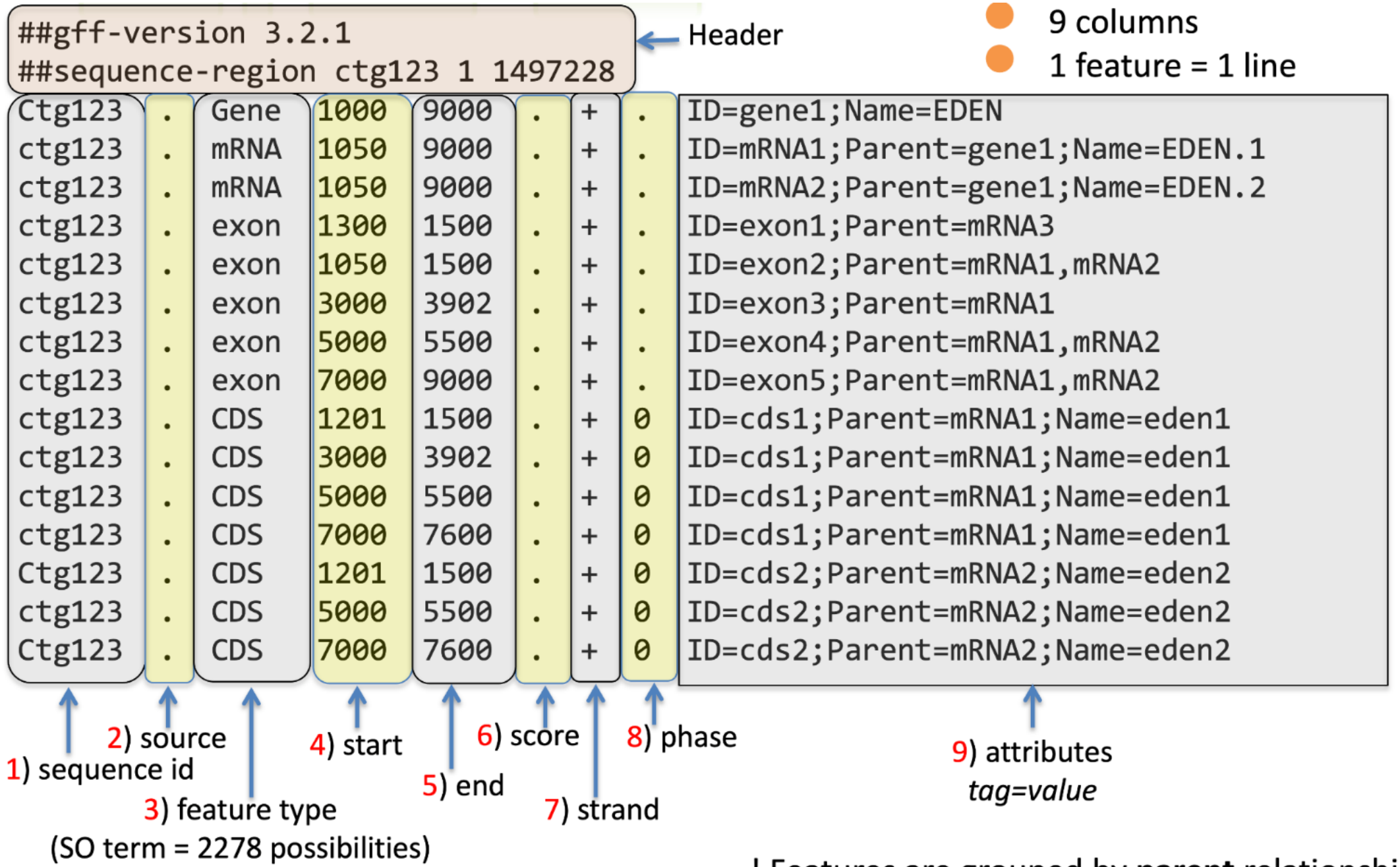


A crustacean



7/55

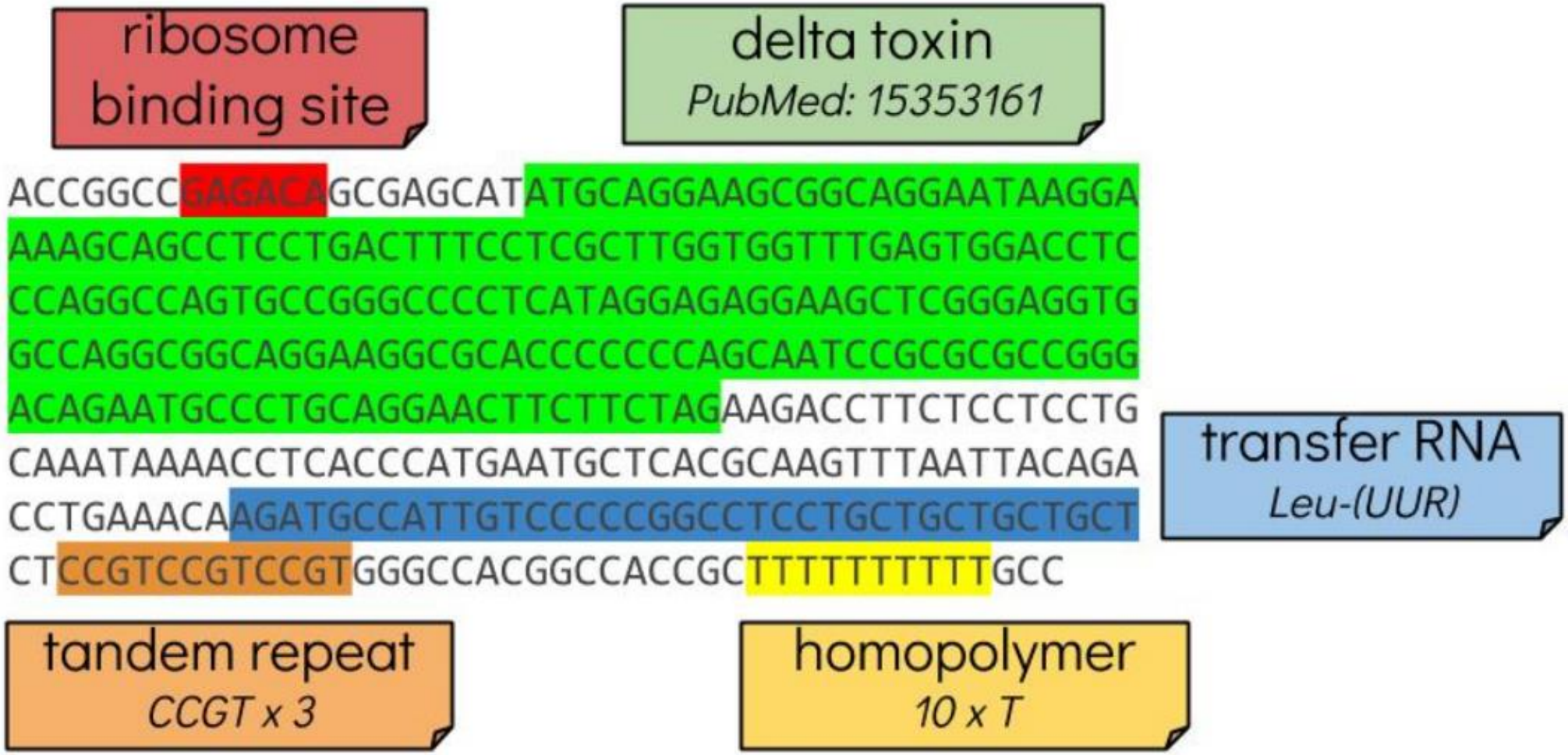
Thx!!
Slide borrowed to NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN



! Features are grouped by **parent** relationship

Thx!!
Slide borrowed to NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Adding biological info to sequences



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

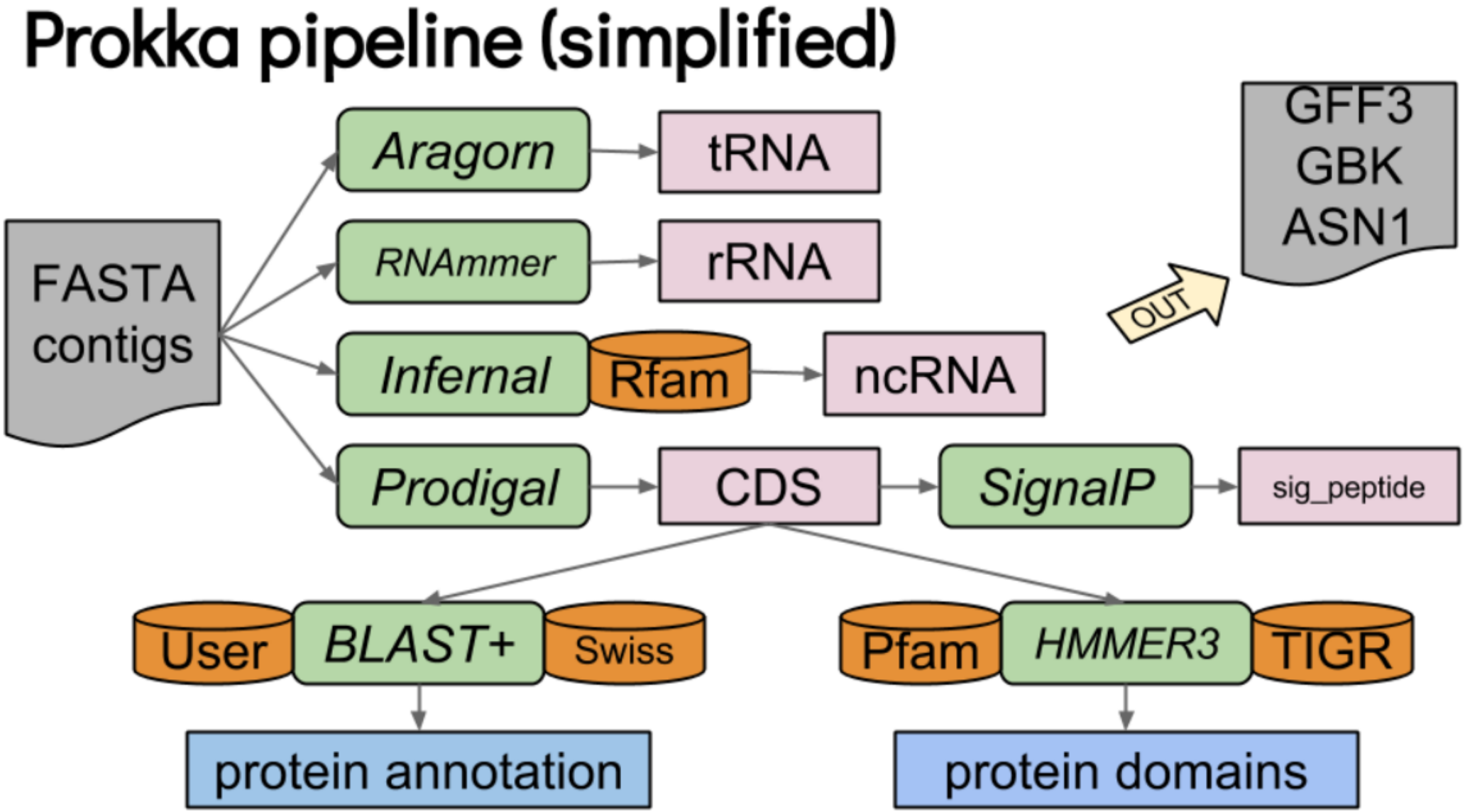
Prokaryote annotation:

- 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)
FASTA contigs
Aragorn
tRNA
RNAmmer
rRNA
Infernal
Rfam
ncRNA
Prodigal
CDS
SignalP
sig_peptide
OUT
GFF3
GBK
ASN1
User
BLAST+
Swiss
protein annotation
Pfam
HMMER3
TIGR
protein domains

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

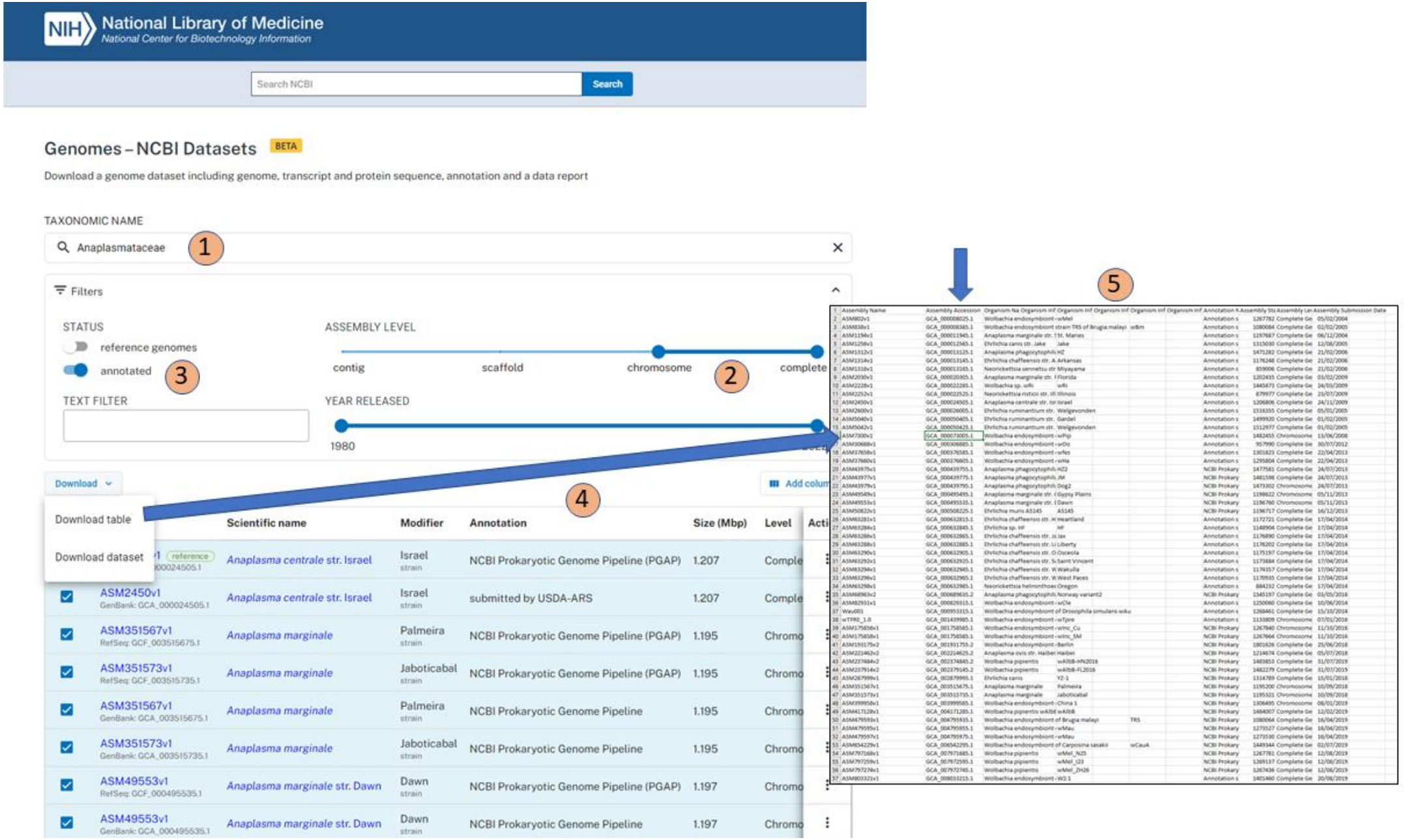
- tRNAscan-SE
- Aragorn
- INFERNAL
- PILER-CR
- Prodigal
- Hmmer
- Diamond
- Blast+
- AMRFinderPlus
- DeepSig

Databases

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

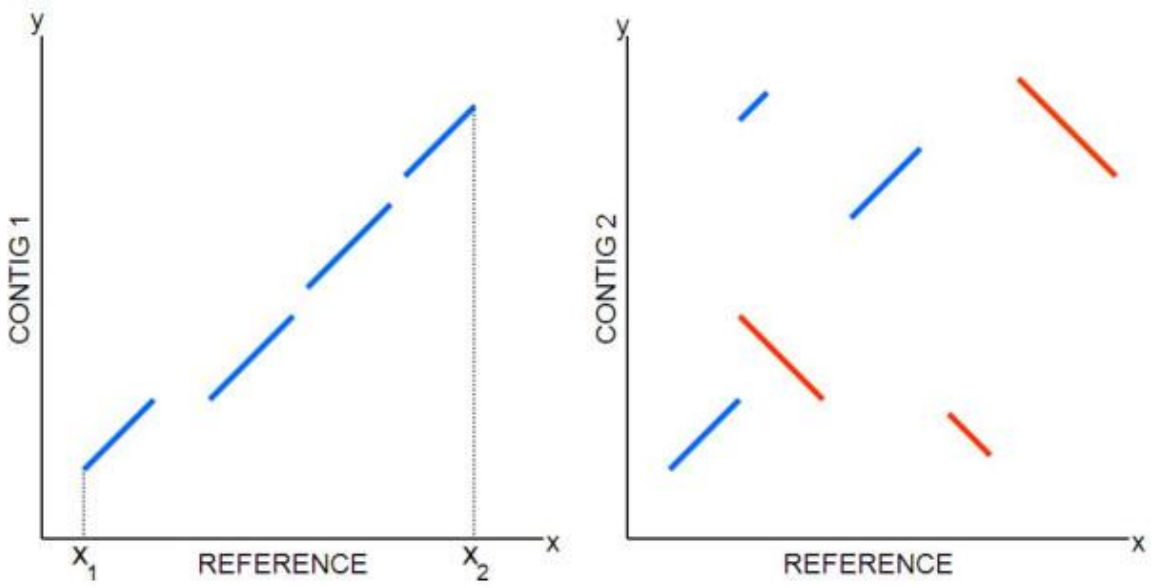
4) Public genomes retrieval

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

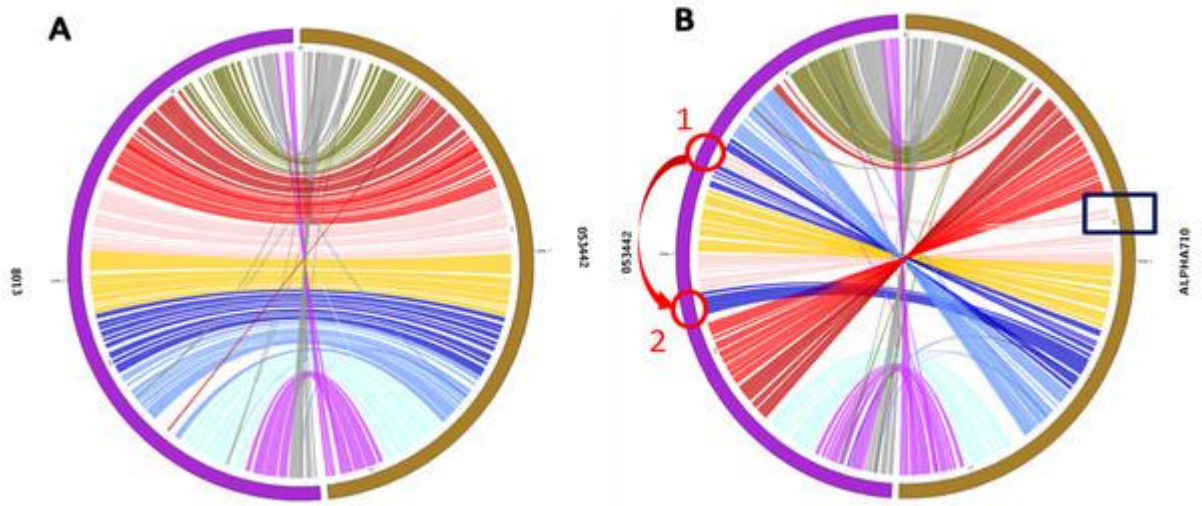


5) Pairwise genome alignment

Dot plot



Circos link



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

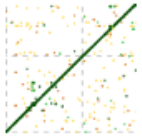
Dot plot

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.

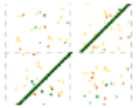
Match

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



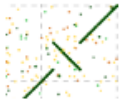
Gap

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

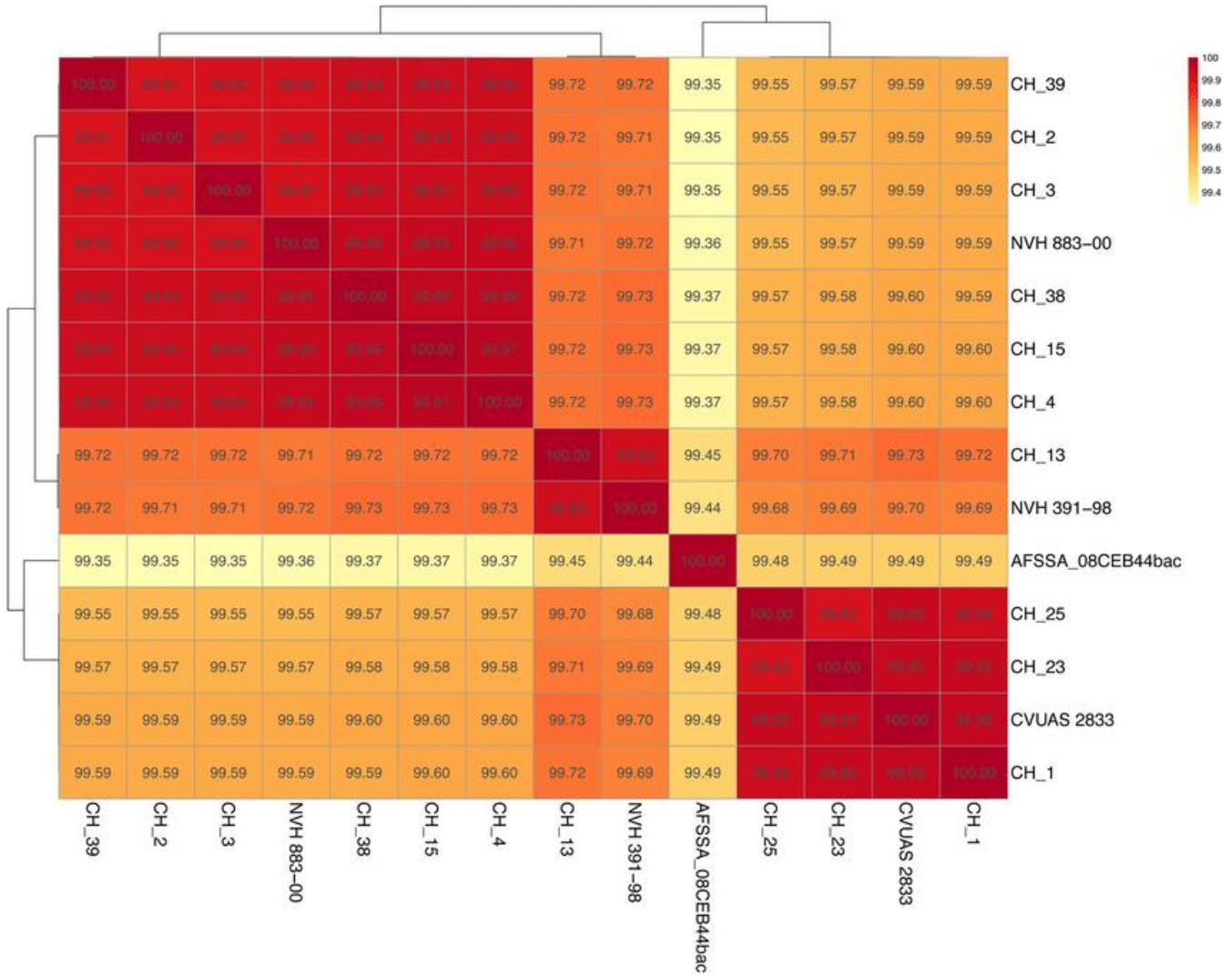
Dot plot can be used to detect repeated regions: a sequence which is repeated several times in a sample.

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

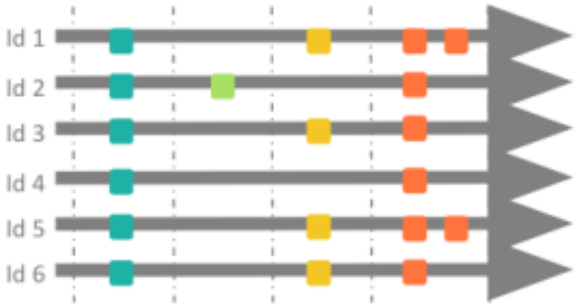
Program: FastANI

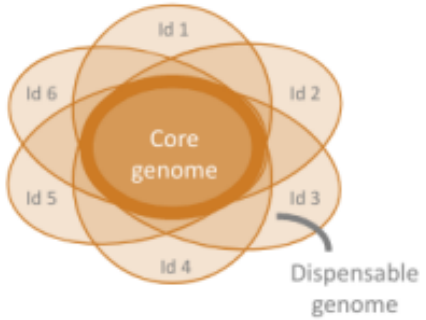


Heat map of the average nucleotide identity (ANI) for strains of the species B. cytotoxicus *(Stevens et al., 20.19)*

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)

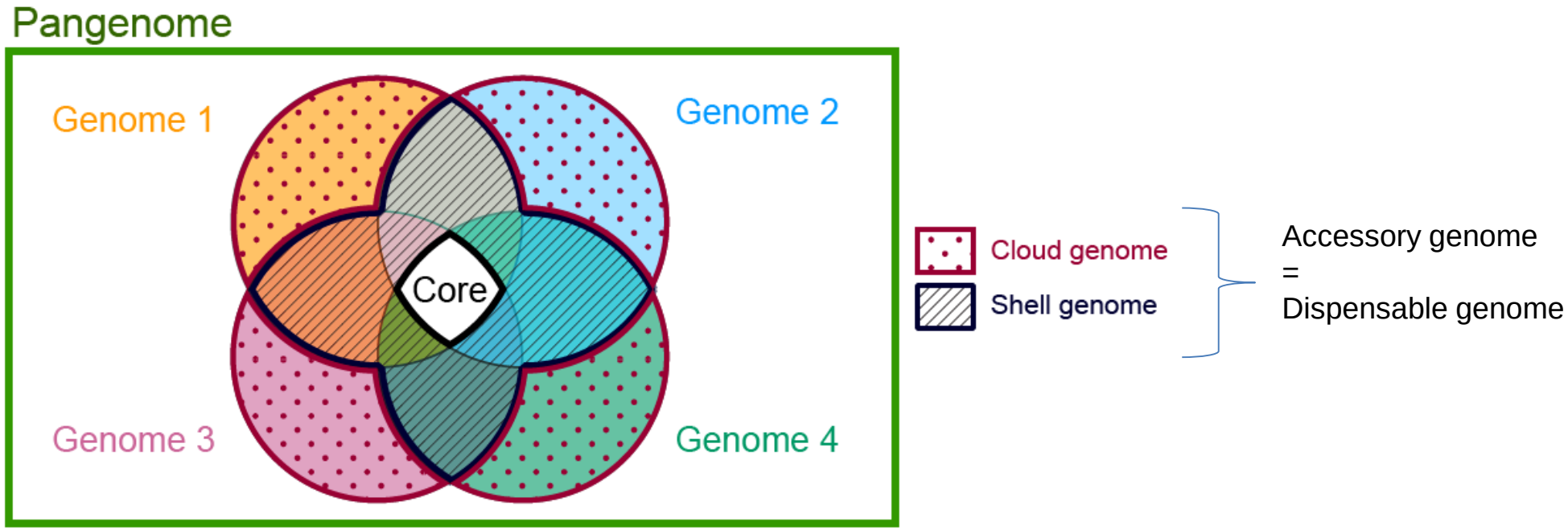
- **Core genome** : present in all individuals
- **Dispensable genome** : absent from one or several individuals (also called variable, accessory,...)

Tranchant-Dubreuil, Rouard, Sabot

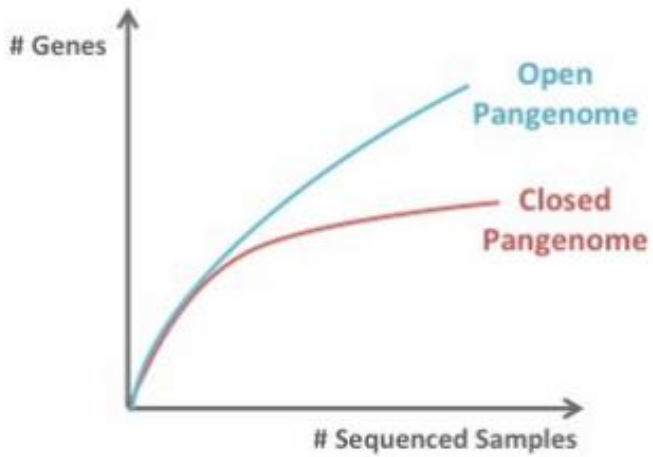
Gene Clustering - how it works

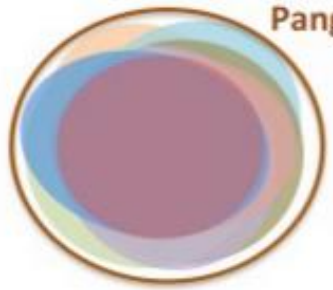
- Assess the similarity of every gene to every other gene
  - e.g., using BLAST
- Use that similarity to join pairs of genes
  - e.g., using Reciprocal Best Hits
- Connect the gene pairs into larger clusters
  - e.g., using Reciprocal Best Hits or Markov clustering

=> Programs: OrthoMCL, Roary, PGAP…

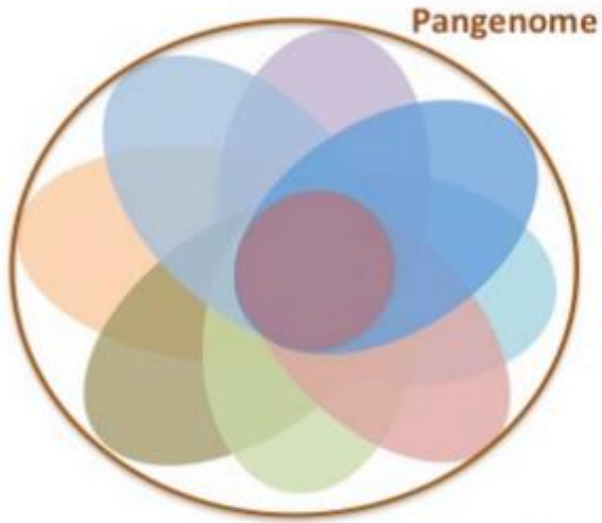

Pangenome
Genome 1
Genome 2
Core
Genome 3
Genome 4
Cloud genome
Shell genome
Accessory genome
=
Dispensable genome

Le pangénome ouvert, fermé, le ratio C/P





High (Core / Pangenome)



Low (Core / Pangenome)

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

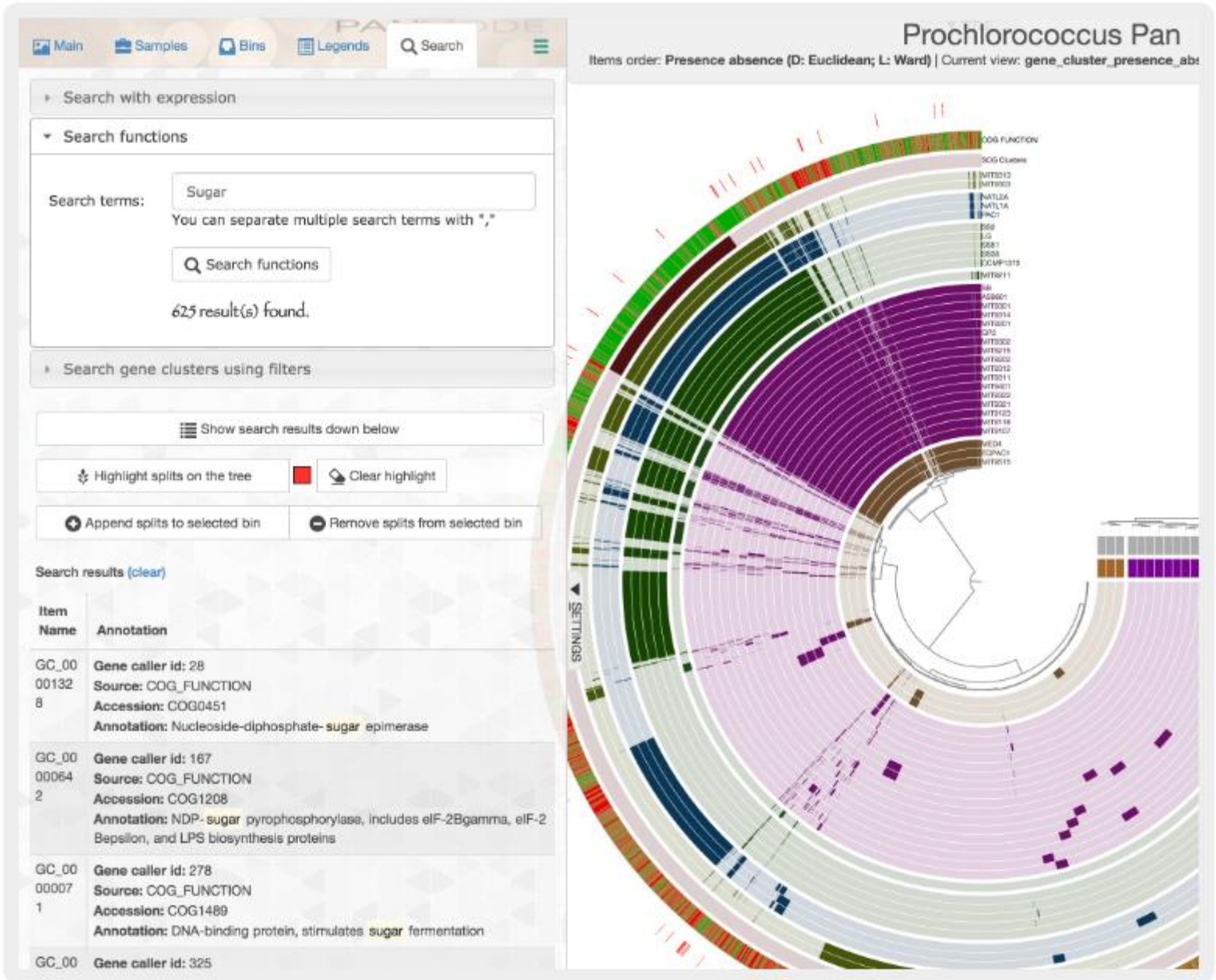
MICROBIAL GENOMICS
Volume 7, Issue 11

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]

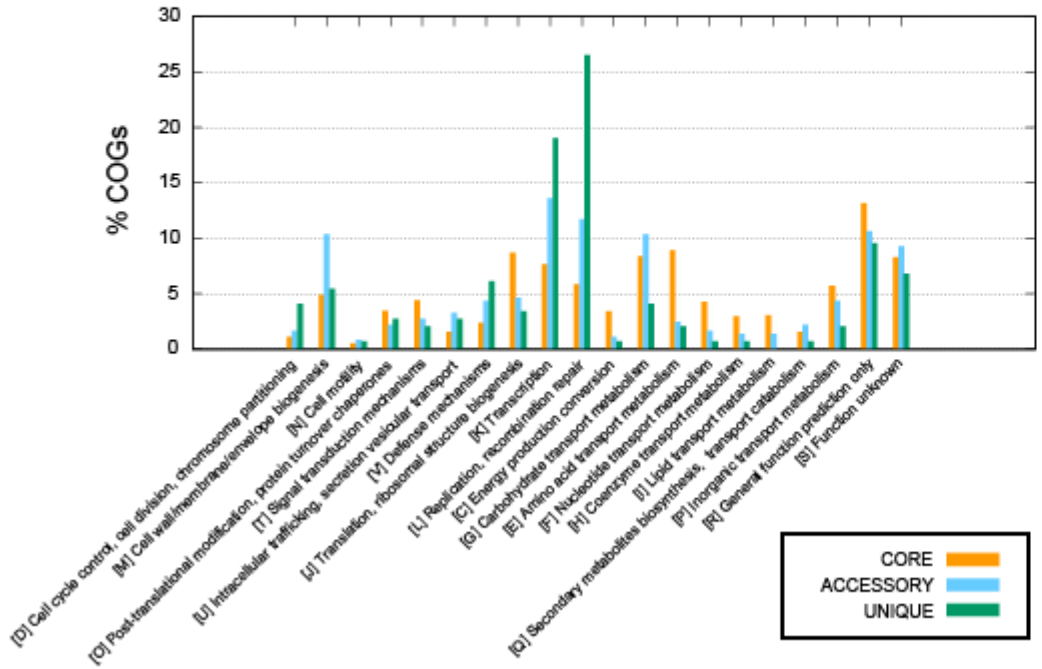
An **anvi'o** workflow for microbial pangenomics https://merenlab.org/2016/11/08/pangenomics-v2/



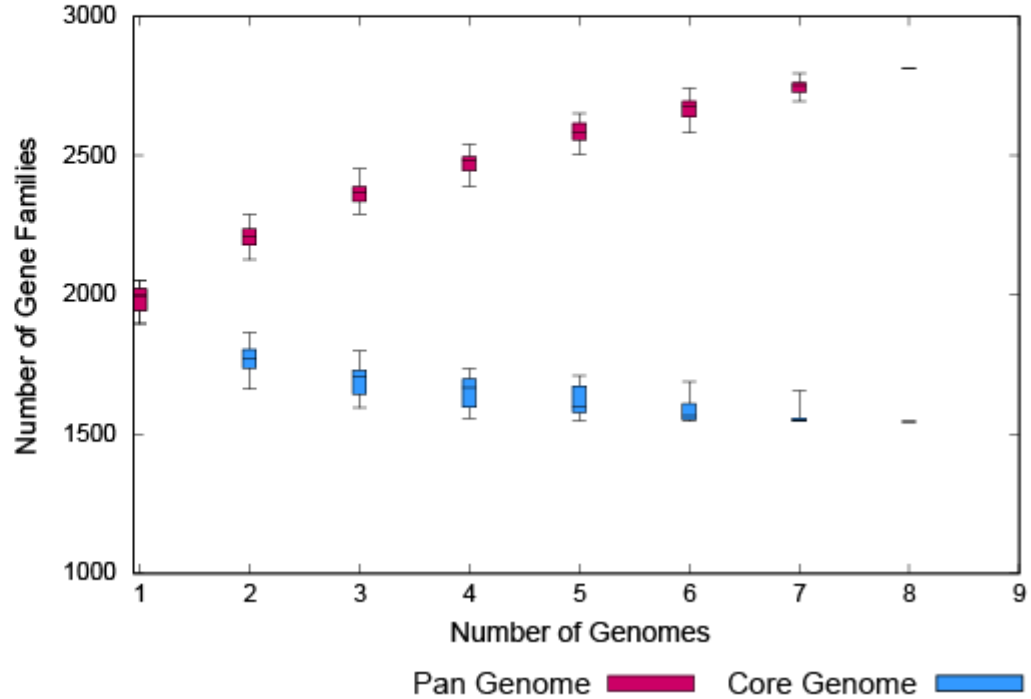
BPGA (Bacterial Pan Genome Analysis tool)

Streptococcus agalactiae

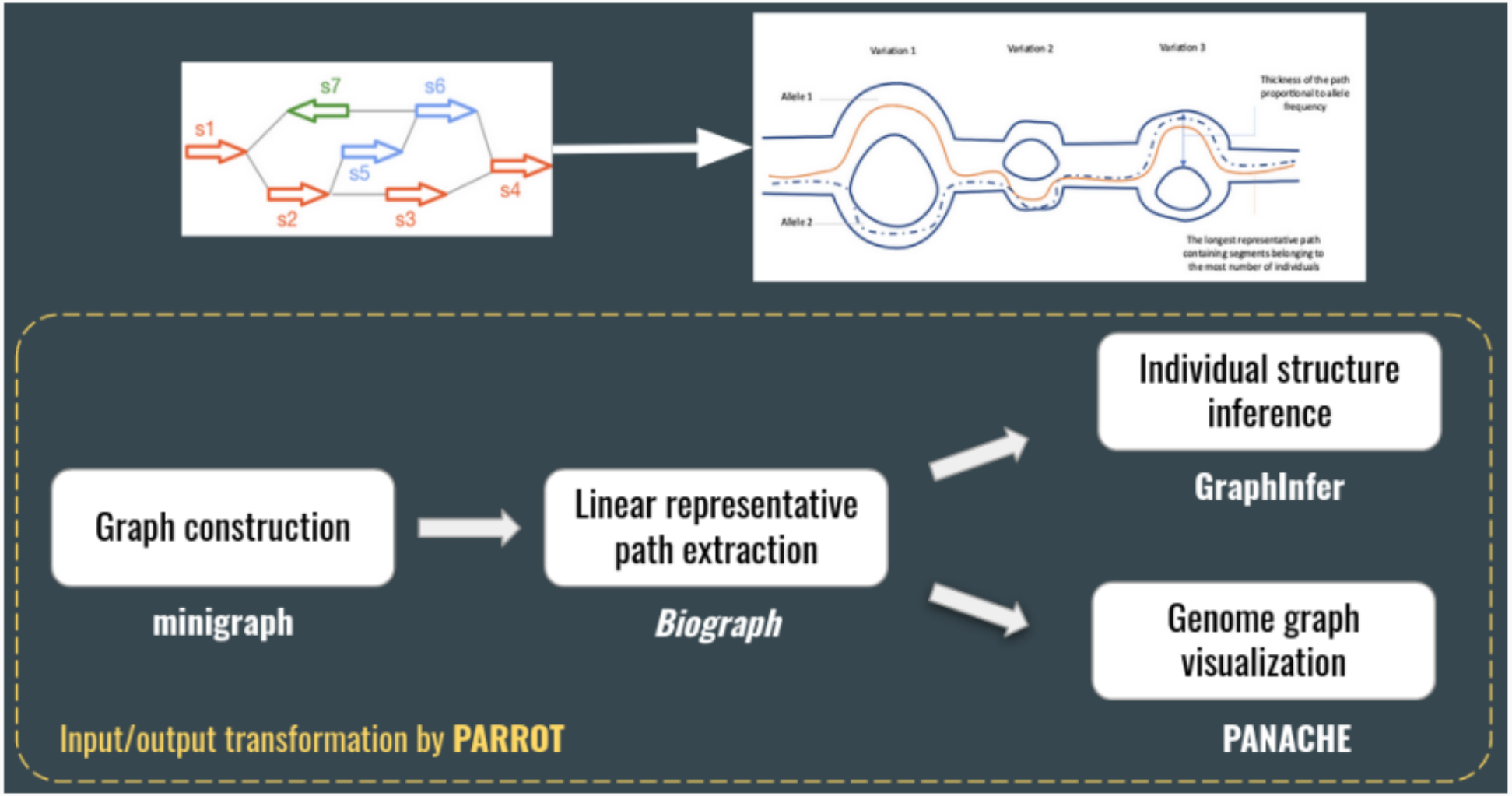
COG Distribution



Pan and Core Genome Plot



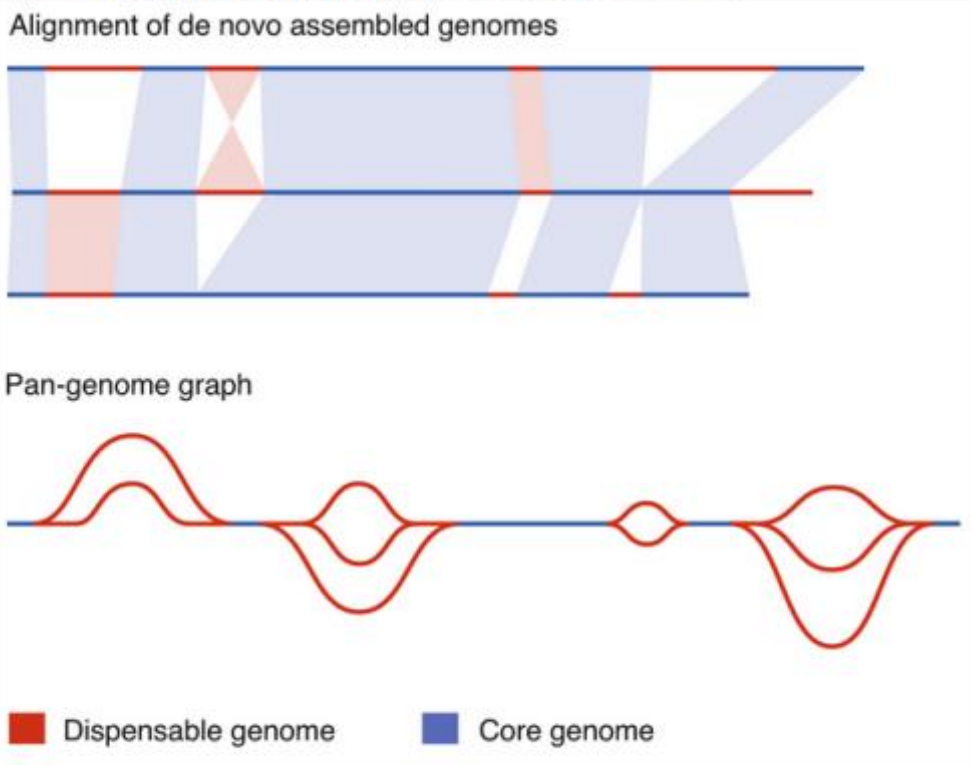
Comment manipuler le graphe pour les biologistes ?



Dang, Do and Sabot, 2022

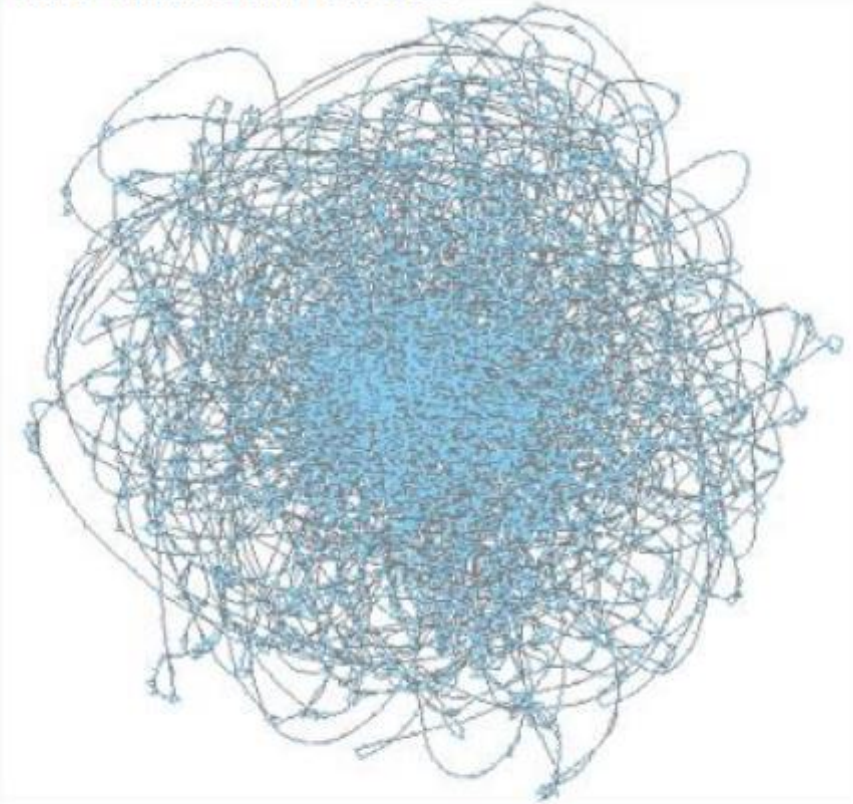
La représentation

Concept du graphe de génome



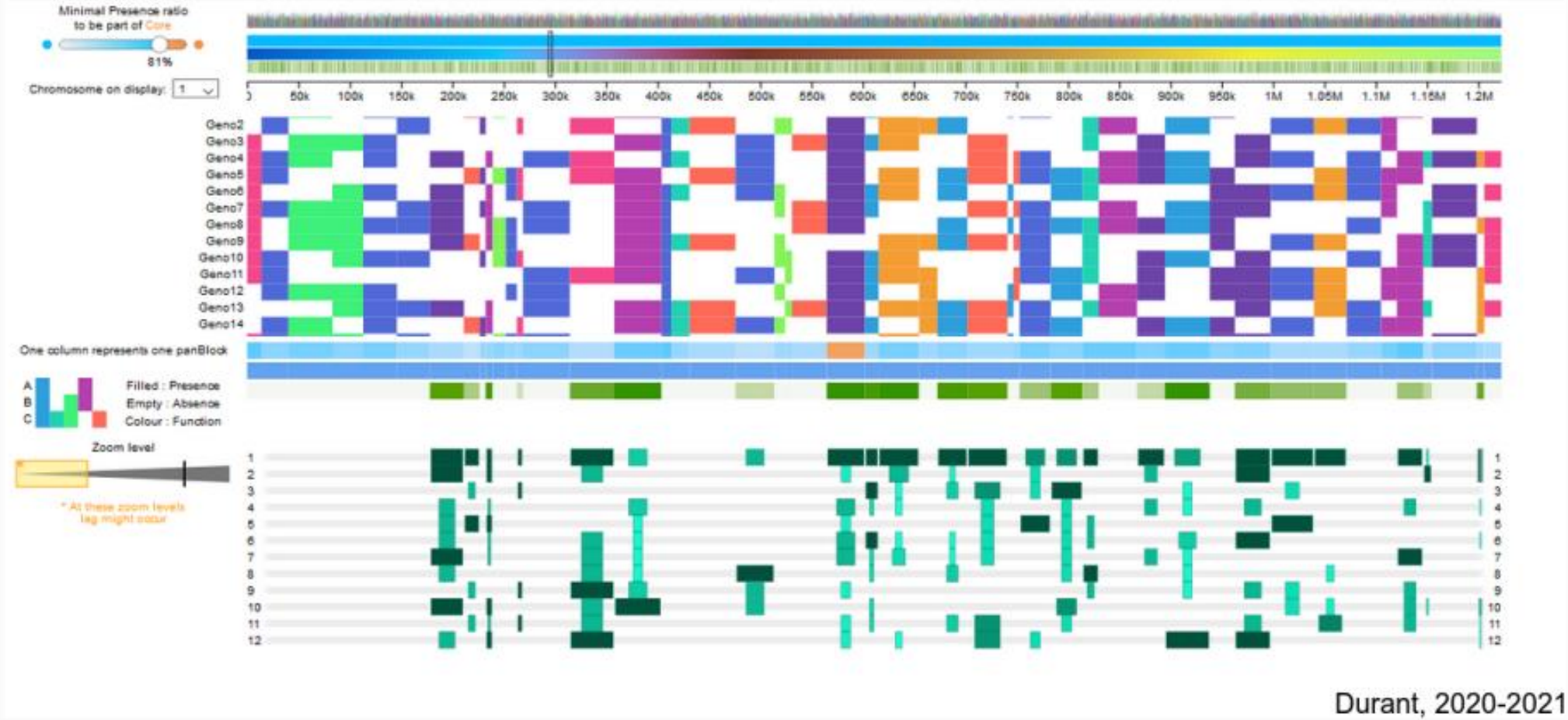
Bayer et al., 2020

The HairBall effect



GigaSciences blog

Un exemple linéaire, Panache



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

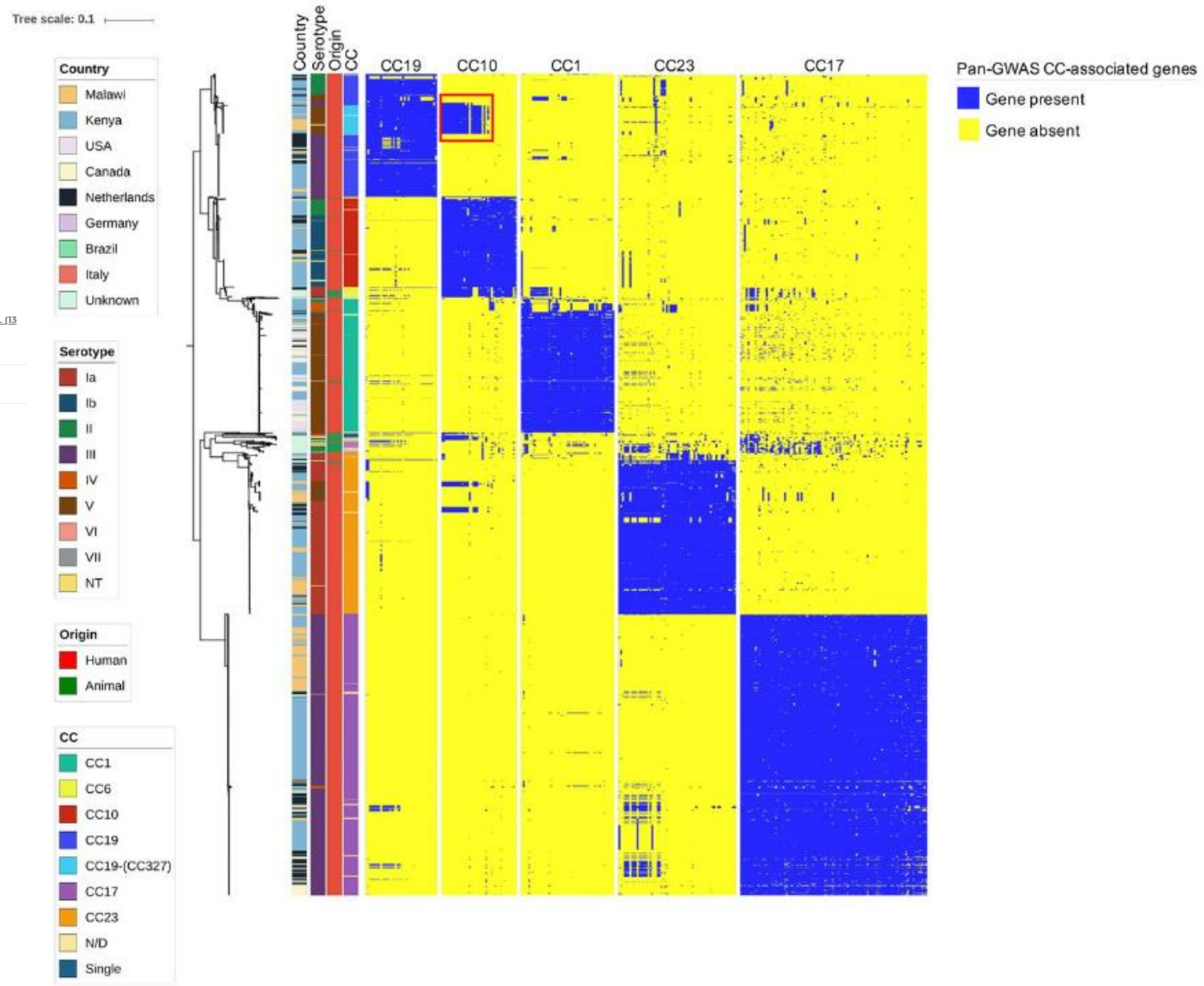


FIG 2 Core genome-based population structure of GBS. The phylogenetic tree is annotated with 4 colored strips representing the clonal complex, the country of isolation, the origin, and the serotype of each strain. The three binary heatmaps represent the presence (blue) or absence (yellow) of the genes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

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 assocations 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 !

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”