South Green
bioinformatics platform

Modules de formation 2022

agap

Bioversity International

diade

PHIM

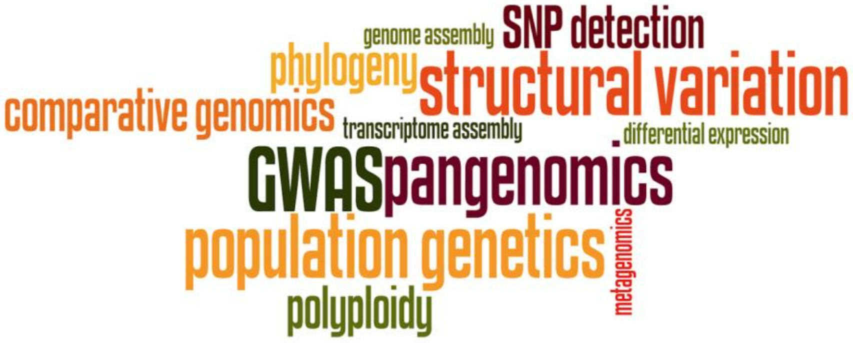
intertryp
Host-vector-parasite-environment interactions in neglected tropical diseases due to trypanosomatids

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

TransVIH MI

South Green bioinformatics platform

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



www.southgreen.fr

Rice

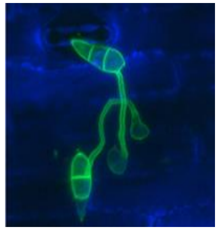
Banana

Palm

Sorghum

Coffee

Cassava


Magnaporthe

South Green bioinformatics platform

Charte

diade

Larmande Pierre
Orjuela-Bouniol Julie
Sabot François
Tando Ndomassi
Tranchant-Dubreuil Christine

PHIM

Comte Aurore
Dereeper Alexis
Ravel Sébastien

agap

Bocs Stephanie
Boizet Alice
De Lamotte Fredéric
Droc Gaetan
Dufayard Jean-François
Hamelin Chantal
Martin Guillaume
Pitollat Bertrand
Ruiz Manuel
Sarah Gautier
Summo Marilyne

Bioversity International

Rouard Mathieu
Guignon Valentin
Catherine Breton

intertryp

Sempere Guilhem

Bioversity International

cirad

IRD Institut de Recherche pour le Développement FRANCE

INRAE

South Green

bioinformatics platform

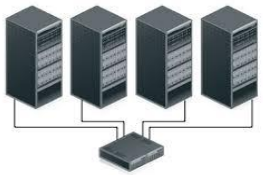
Workflow manager

TOGGLe
Toolbox for generic NGS analyses

SNAKEMAKE

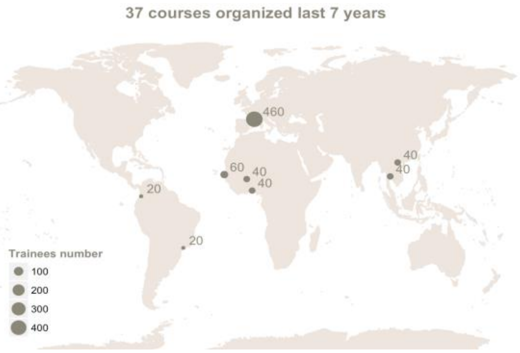
Galaxy

HPC and trainings….



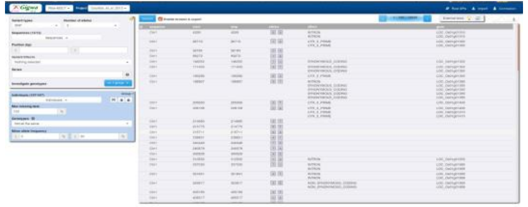
IRD
Institut de Recherche pour le Développement
FRANCE

cirad



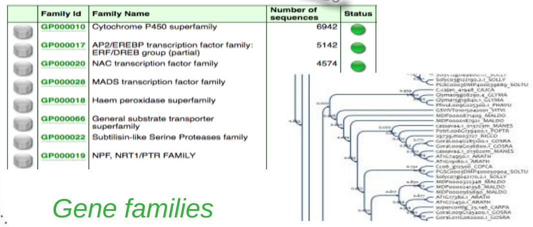
Genome Hubs & Information System

Gigwa



SNPs and Indels

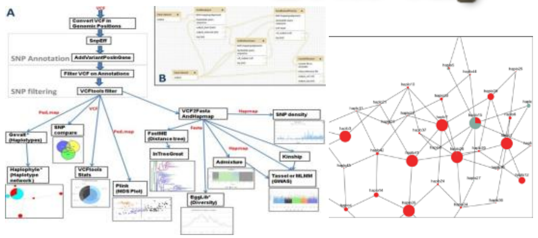
GreenPhyl



	Family Id	Family Name	Number of sequences	Status
	GP000010	Cytochrome P450 superfamily	6942	●
	GP000017	AP2/EREBP transcription factor family: ERF/DREB group (partial)	5142	●
	GP000020	NAC transcription factor family	4574	●
	GP000028	MADS transcription factor family		
	GP000018	Haem peroxidase superfamily		
	GP000066	General substrate transporter superfamily		
	GP000022	Subtilisin-like Serine Proteases family		
	GP000019	NPF, NRT1/PTR FAMILY		

Gene families

SNiPlay



https://github.com/SouthGreenPlatform

@green_bioinfo

The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics, Current Plant Biology, 2016

I-Trop

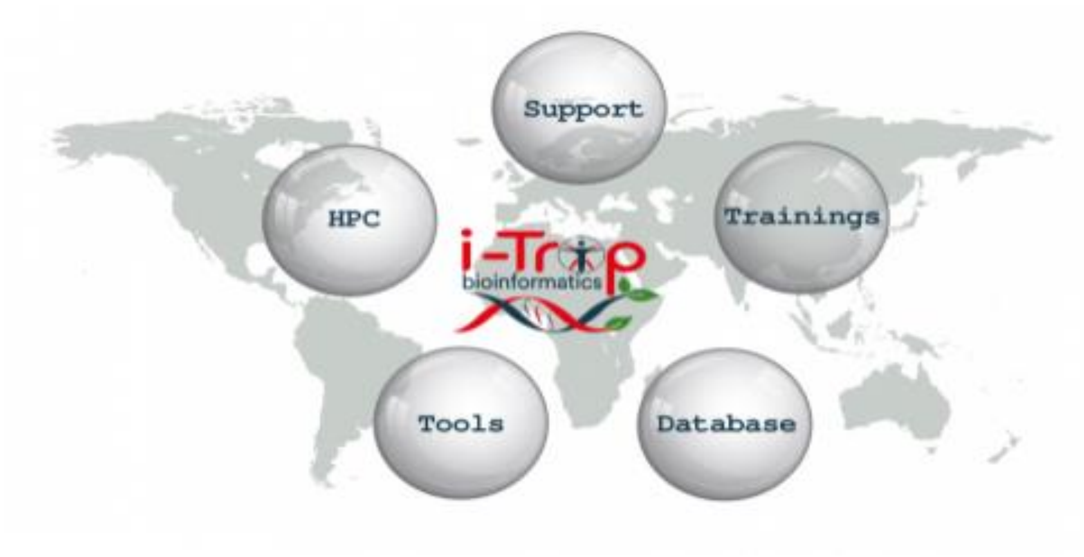
Plant & Health Bioinformatics Platform

diade

PHIM

TransVIH MI

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



Institut de Recherche pour le Développement
FRANCE

EURO-QUALITY SYSTEM
ISO 9001

https://bioinfo.ird.fr/

IE bioinfo

IE bioinfo

IE systèmes d’informatio

IE bioinfo

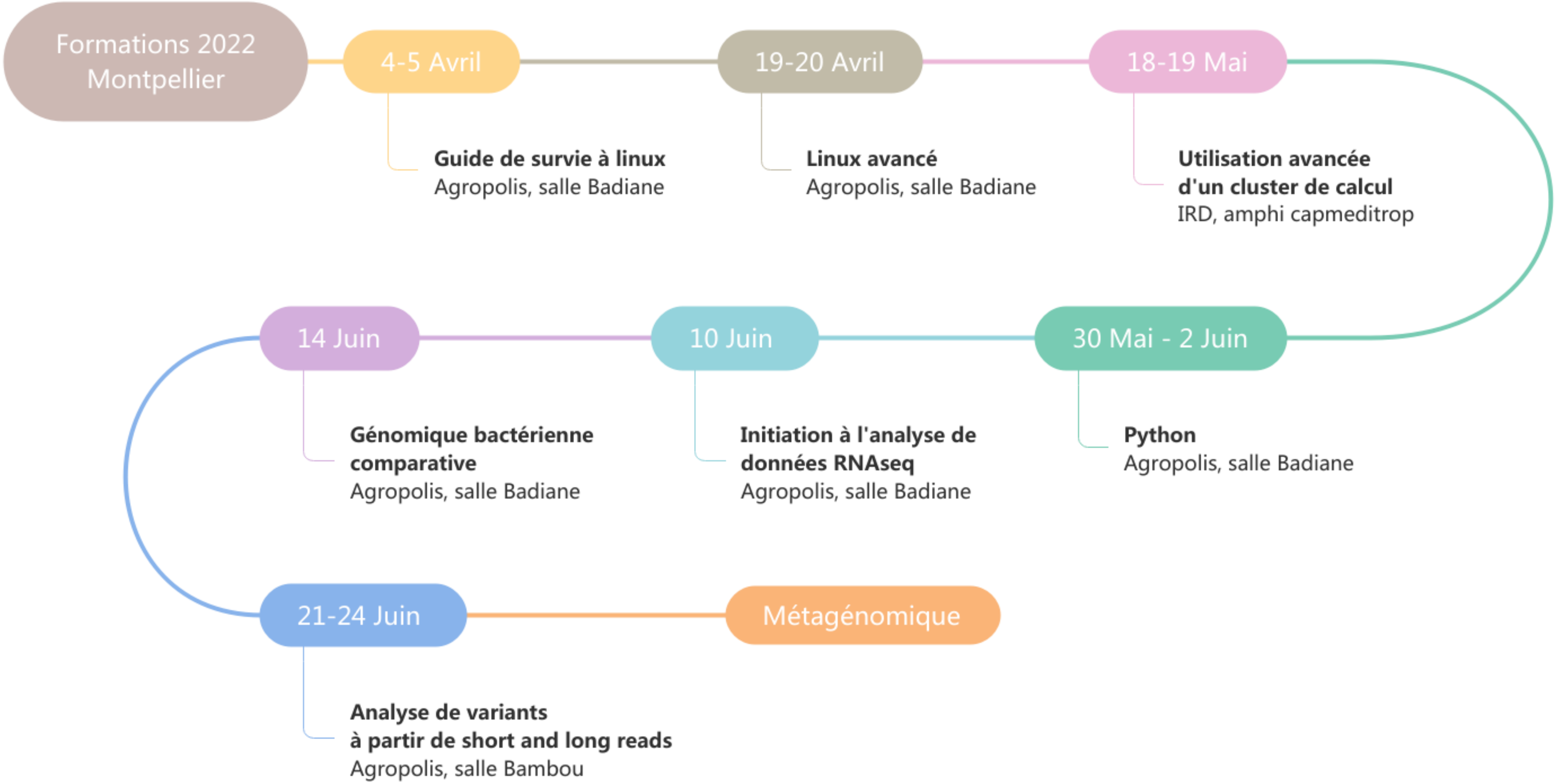
IE systèmes

IR bioinfo

bioinfo@ird.fr

@ItropBioinfo

South Green bioinformatics platform



Formations 2022
Montpellier

4-5 Avril

Guide de survie à linux
Agropolis, salle Badiane

19-20 Avril

Linux avancé
Agropolis, salle Badiane

18-19 Mai

Utilisation avancée d'un cluster de calcul
IRD, amphi capmeditrop

30 Mai - 2 Juin

Python
Agropolis, salle Badiane

10 Juin

Initiation à l'analyse de données RNAseq
Agropolis, salle Badiane

14 Juin

Génomique bactérienne comparative
Agropolis, salle Badiane

21-24 Juin

Analyse de variants à partir de short and long reads
Agropolis, salle Bambou

Métagénomique

agap Bioversity International intertryp diade MiVEGEC PHIM TransVIH MI

South Green bioinformatics platform

Modules de formation 2022

- Toutes nos formations :

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

Version 2020.1

SouthGreen bioinformatics platform

Génomique Comparative Bactérienne

Bioversity International

cirad

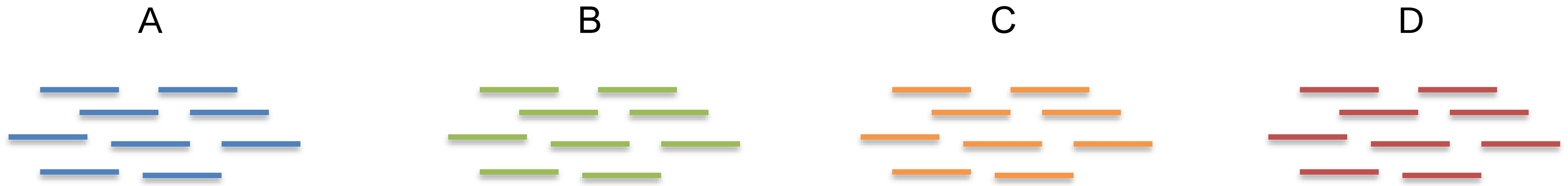
INRAE

IRD Institut de Recherche pour le Développement FRANCE

Montpellier SupAgro

Two Approaches to Microbial Genomics

Starting with sets of reads representing your study isolates…



Assembly-based

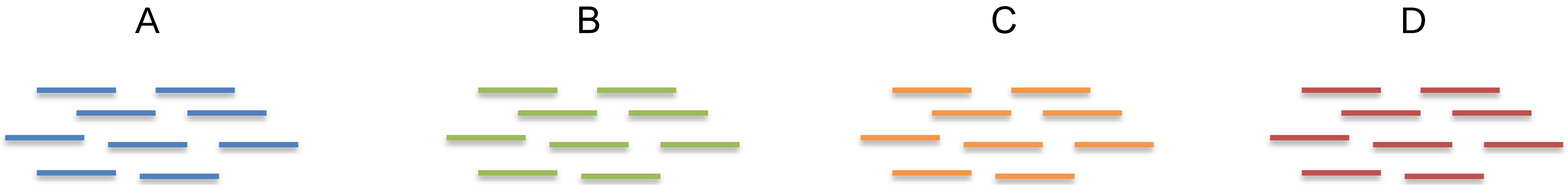
1. Assemble each set of reads into a genome sequence
2. Annotate each genome
3. Cluster genes and compare between each genome

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…



<u>Assembly-based</u>

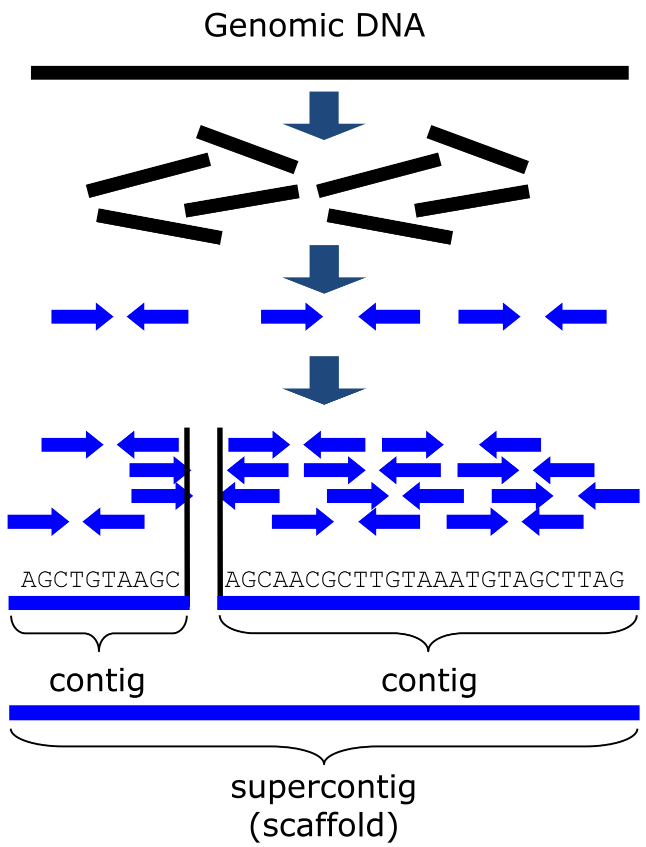
1. Assemble each set of reads into a genome sequence
2. Annotate each genome
3. Cluster genes and compare between each genome

<u>Variant-based</u>

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

1) Assembly

Assembly Basics (de-novo assembly)



Genomic DNA

AGCTGTAAGC

AGCAACGCTTGTAAATGTAGCTTAG

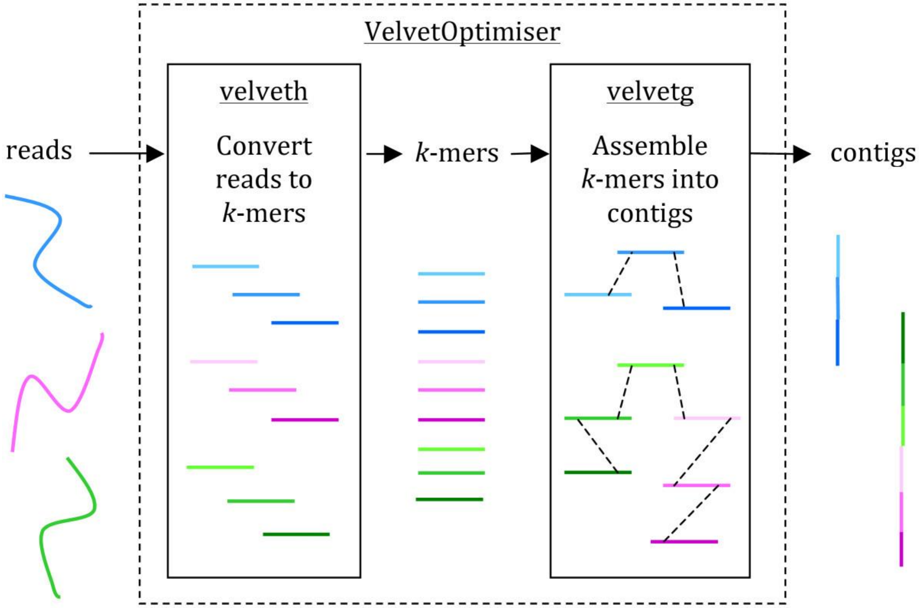
contig

contig

supercontig
(scaffold)

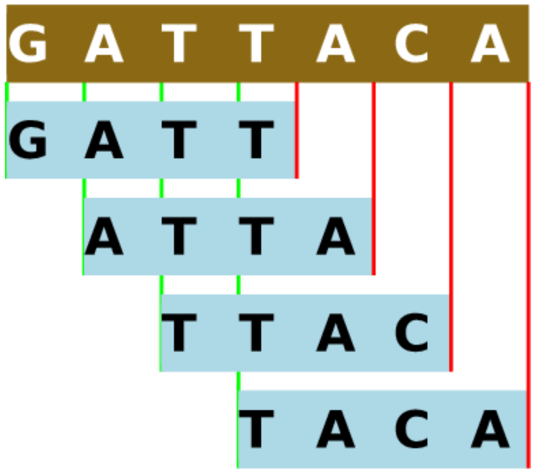
Assembly Methods

- 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 *MIE*

K-mers



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]

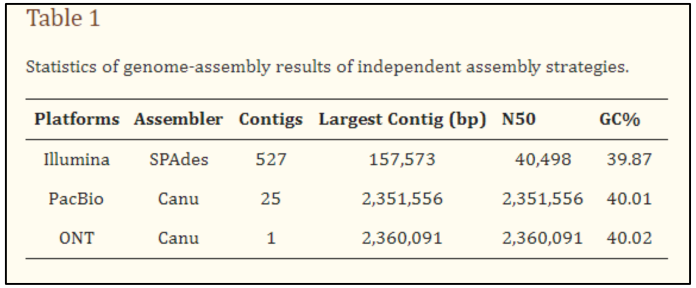


Table 1

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

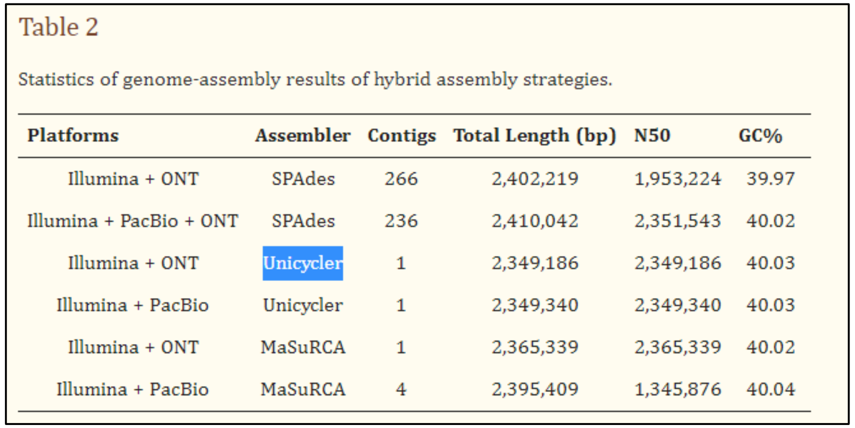
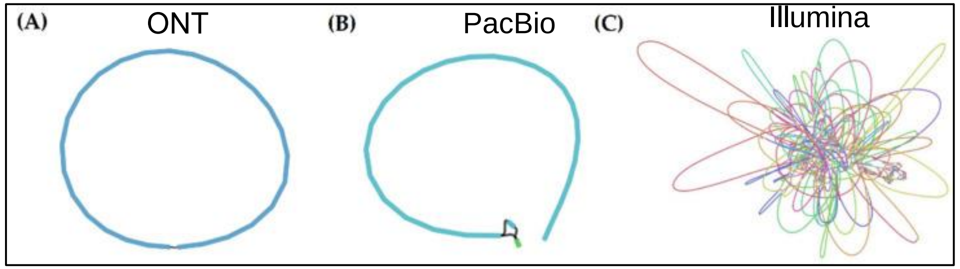


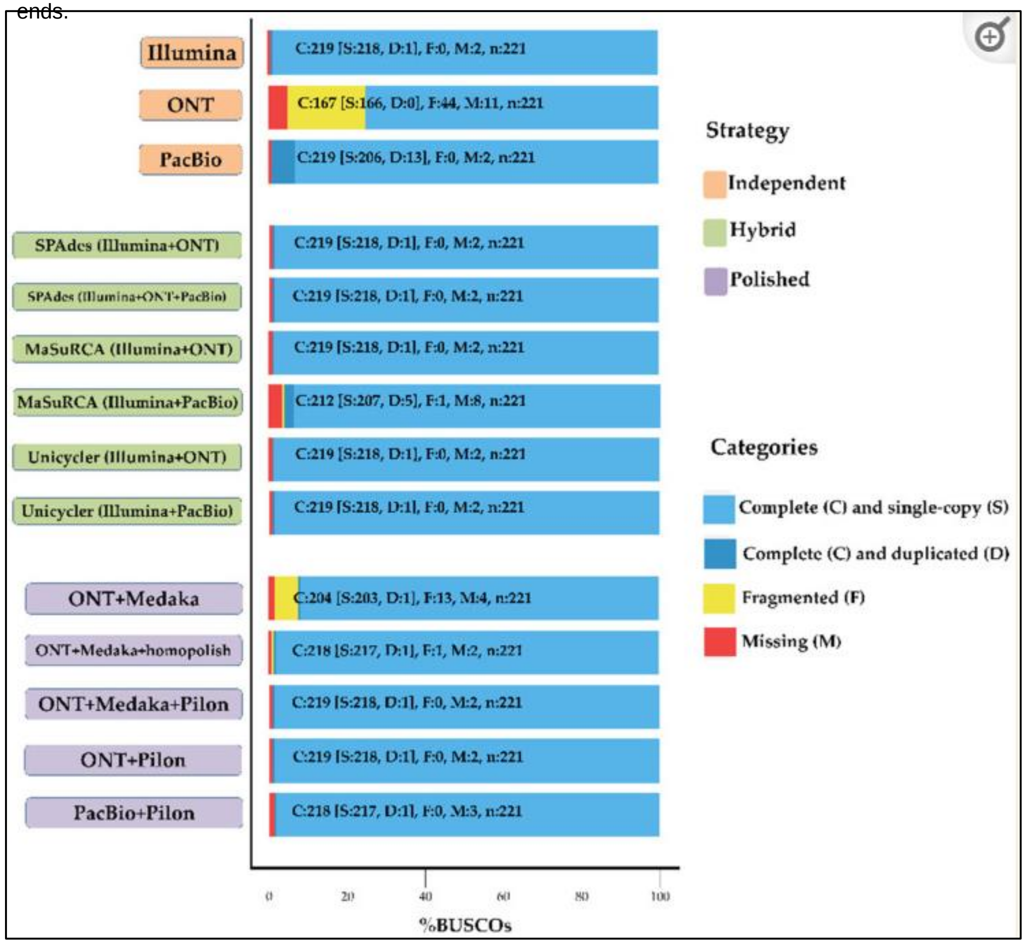
Table 2

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

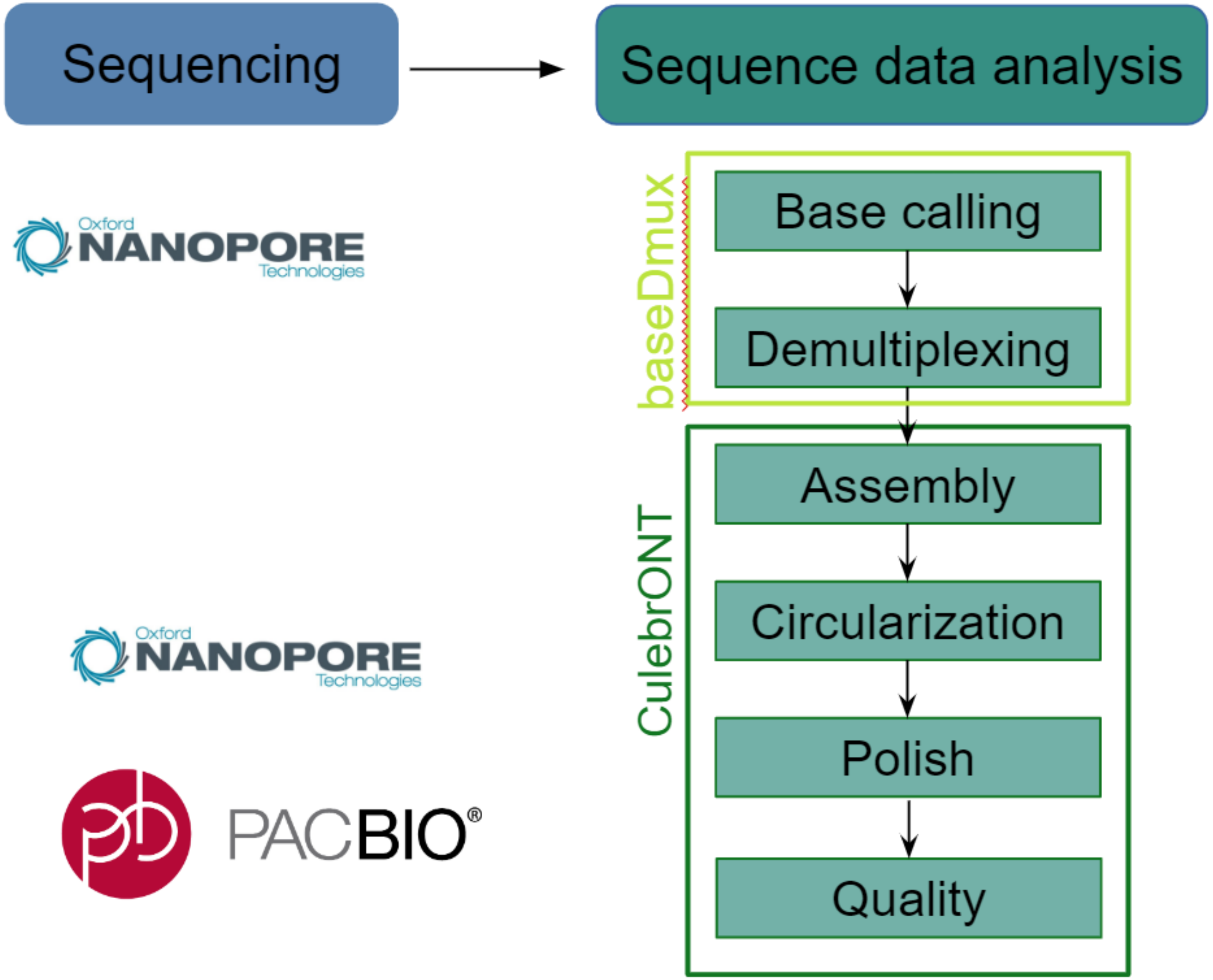


Comparison of results of independent assembly strategies. (**A**) Genome assembled with nanopore reads; (**B**) longest contig assembled with PacBio reads; (**C**) genome assembled with Illumina reads. Plots were obtained by using Bandage on the “assembly_graph.gfa” output file from SPAdes or the “contig.gfa”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

Bioinformatic Workflows: assembly



Oxford NANOPORE Technologies

Oxford NANOPORE Technologies

PACBIO®

Snakemake

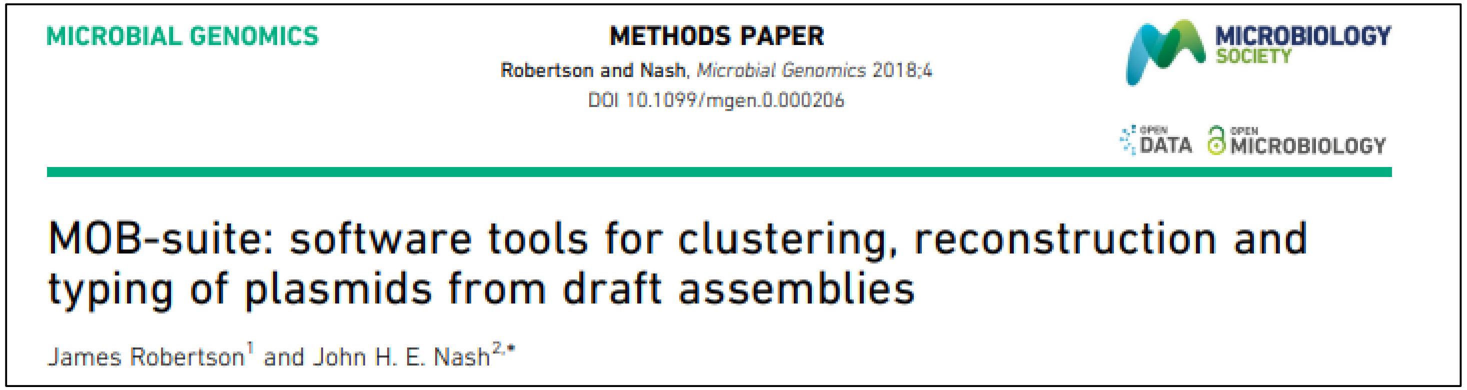
BaseDmux

https://github.com/vibaotram/baseDmux

culebrONT

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

2) Separate chromosomal and plasmid scaffolds/contigs



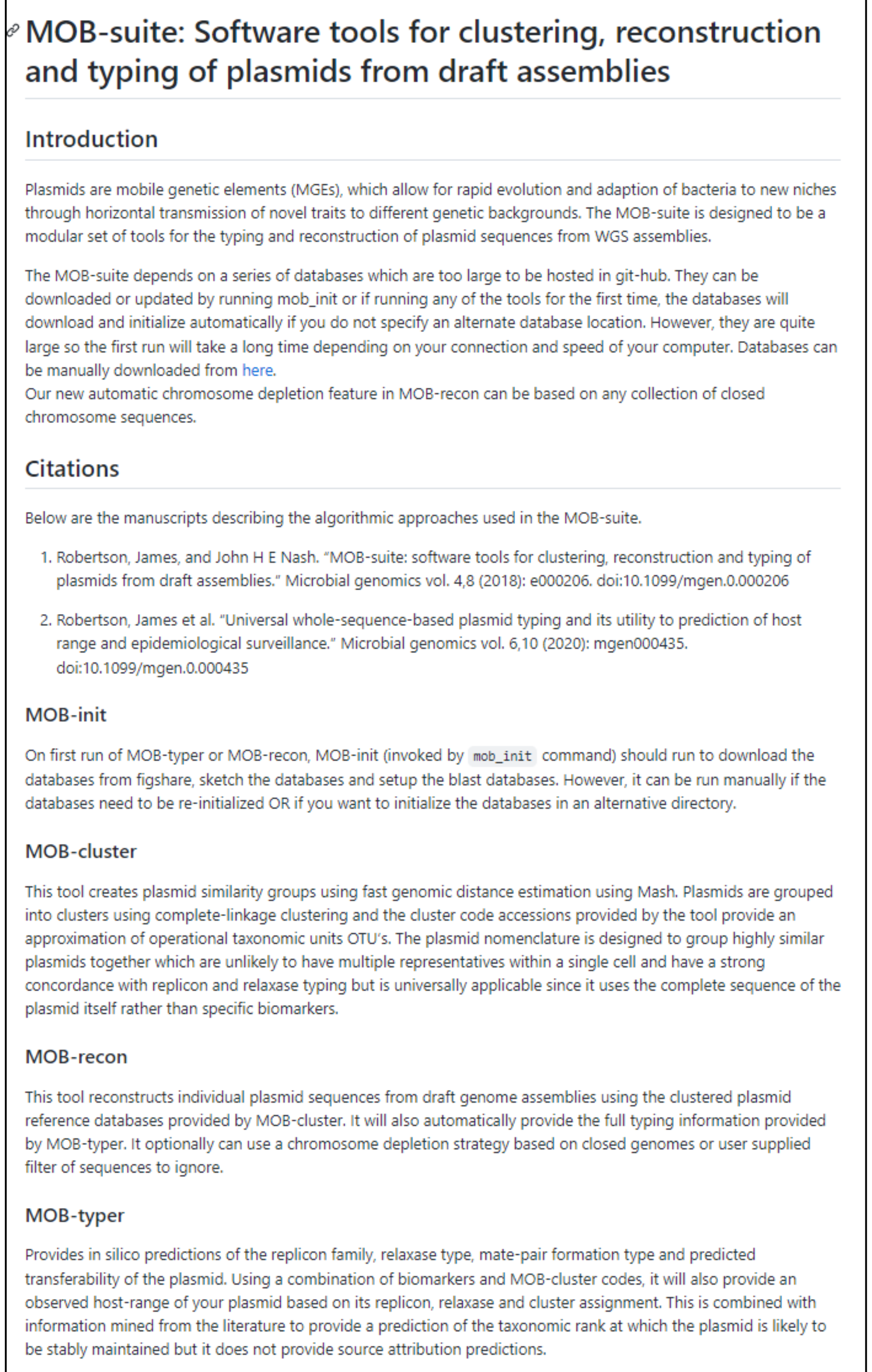
MICROBIAL GENOMICS

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

Annotation Methods

- Annotation refers to assign function to DNA sequences
- There are different annotation algorithms for protein-coding genes, tRNAs, rRNAs, other non-coding RNAs
- Prokka (http://www.vicbioinformatics.com/software.prokka.shtml) is an all-in-one wrapper for these tools

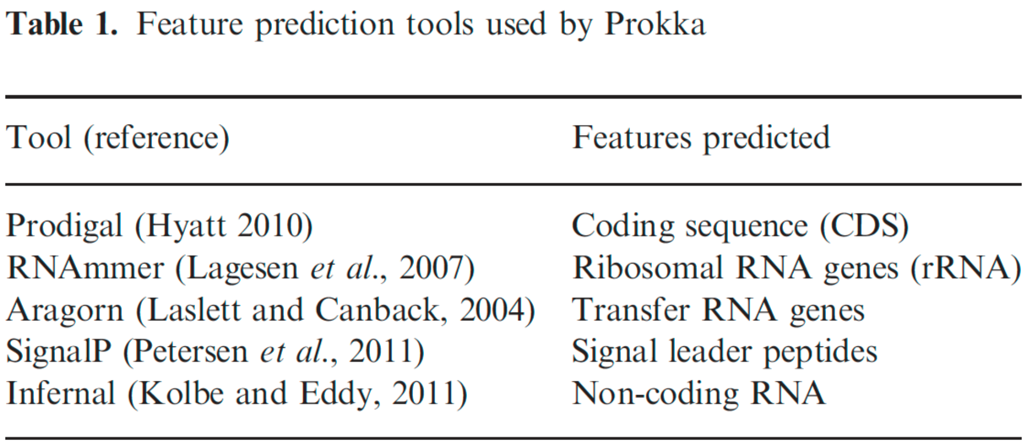
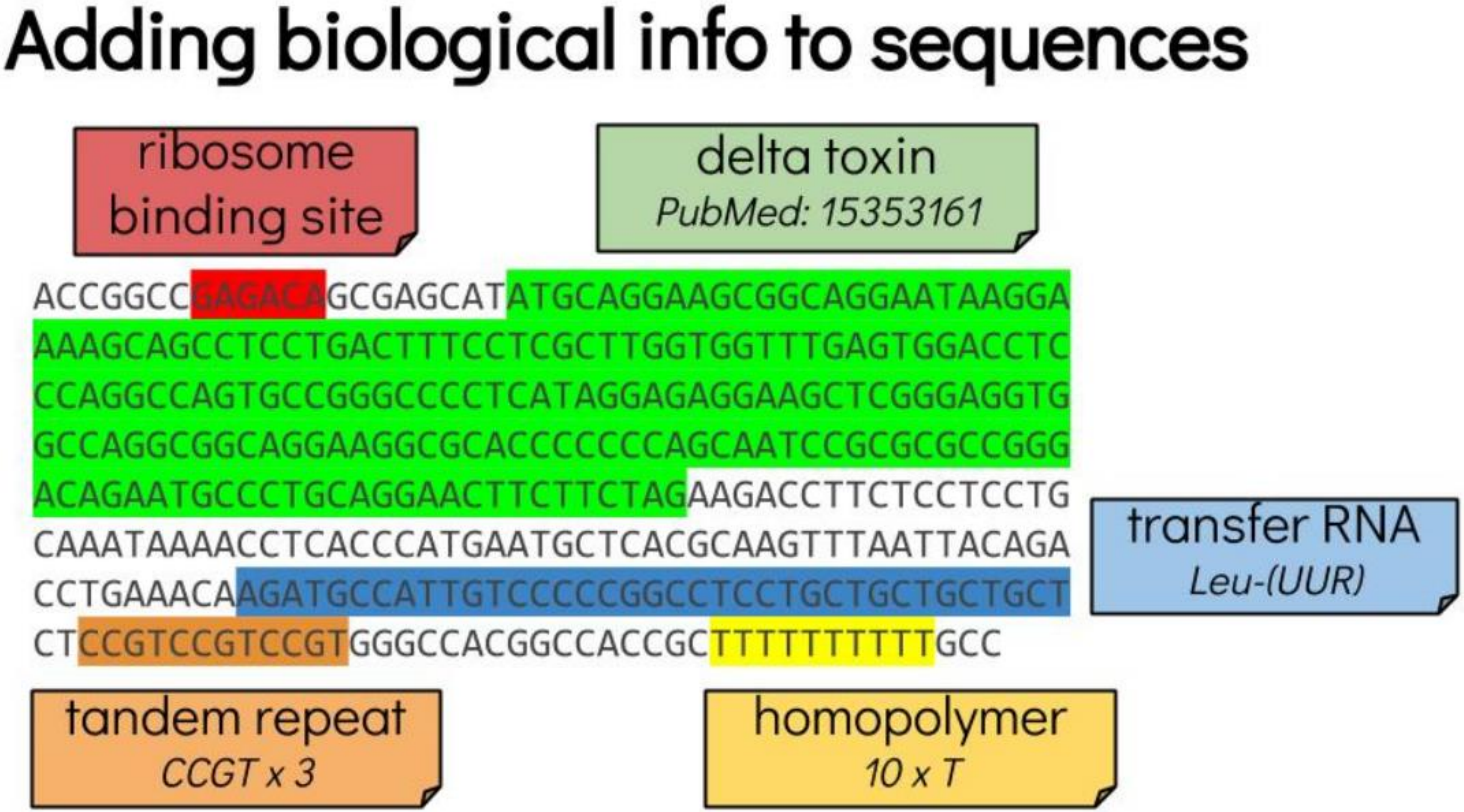


Table 1. Feature prediction tools used by Prokka

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

Then: annotate



Adding biological info to sequences

ribosome binding site

delta toxin
PubMed: 15353161

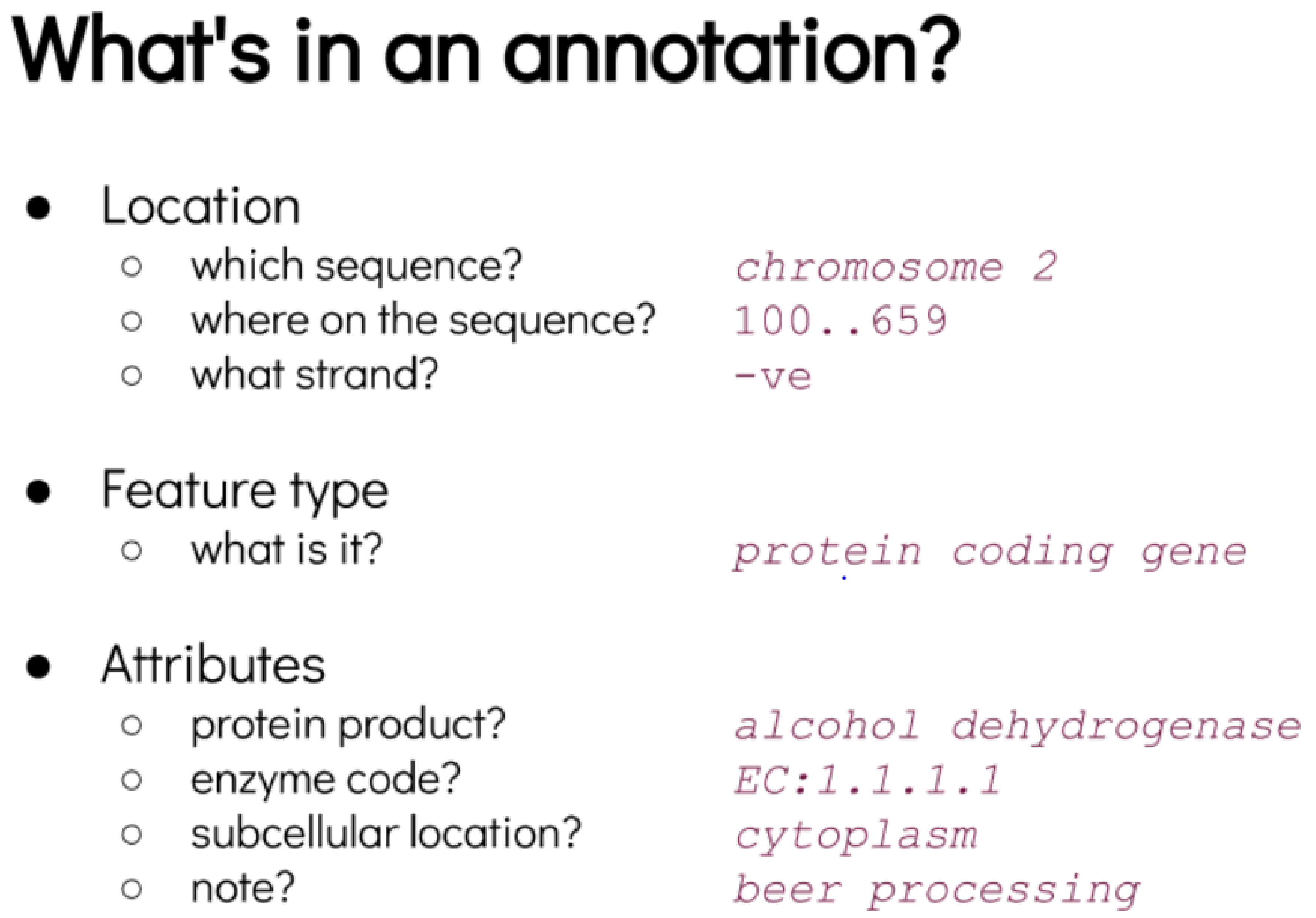
```
ACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGA
AAAGCAGCCTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTC
CCAGGCCAGTGCCGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTG
GCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGG
ACAGAATGCCCTGCAGGAACTTCTTCTAGAAGACCTTCTCCTCCTG
CAAATAAAACCTCACCCATGAATGCTCACGCAAGTTTAATTACAGA
CCTGAAACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCT
CTCCGTCCGTCCGTGGGCCACGGCCACCGCTTTTTTTTTTGCC
```

transfer RNA
Leu-(UUR)

tandem repeat
CCGT x 3

homopolymer
10 x T

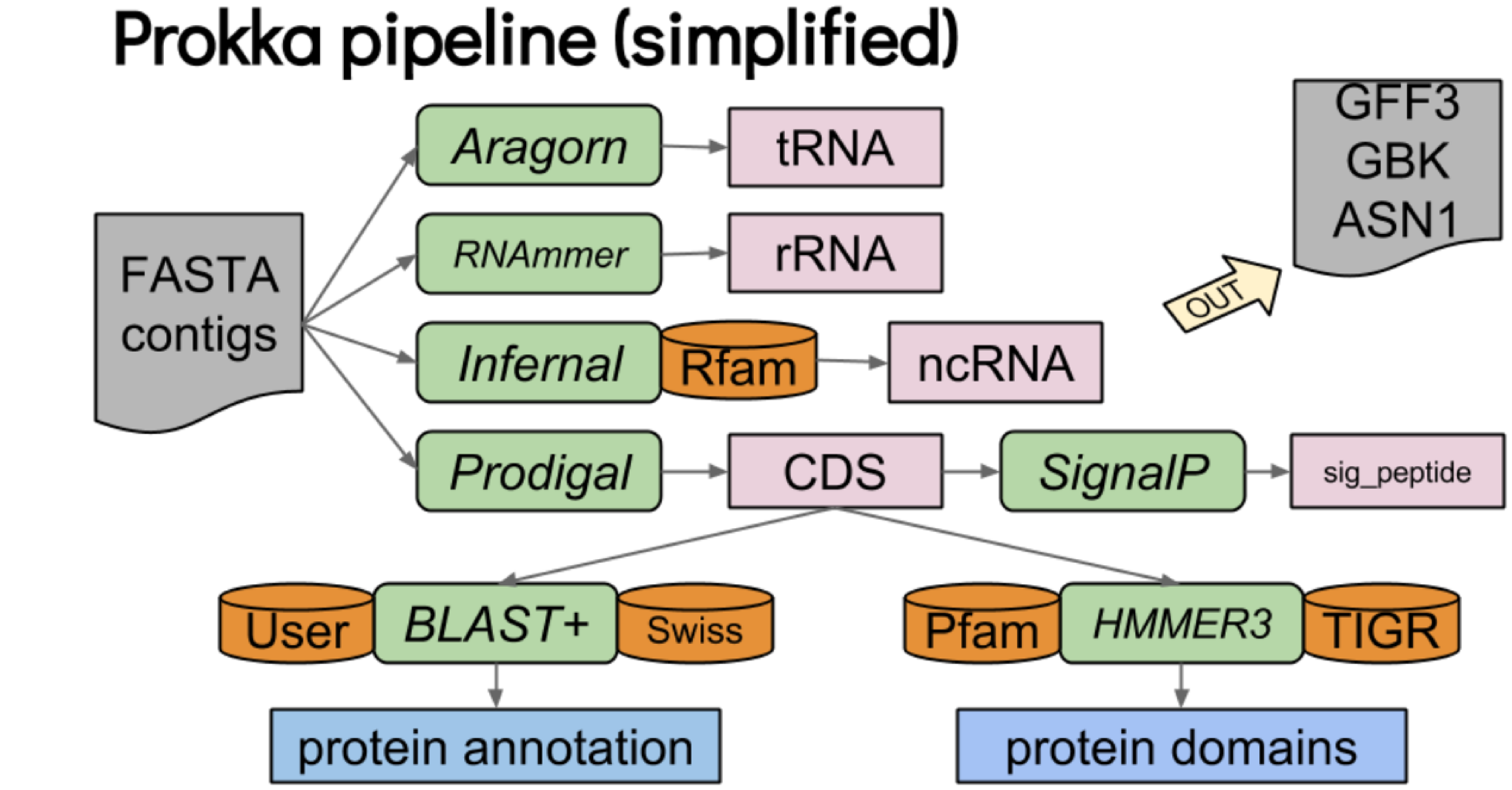
Annotation



What's in an annotation?

- Location
  - which sequence? `chromosome 2`
  - where on the sequence? `100..659`
  - what strand? `-ve`
- Feature type
  - what is it? `protein coding gene`
- Attributes
  - protein product? `alcohol dehydrogenase`
  - enzyme code? `EC:1.1.1.1`
  - subcellular location? `cytoplasm`
  - note? `beer processing`

Prokka

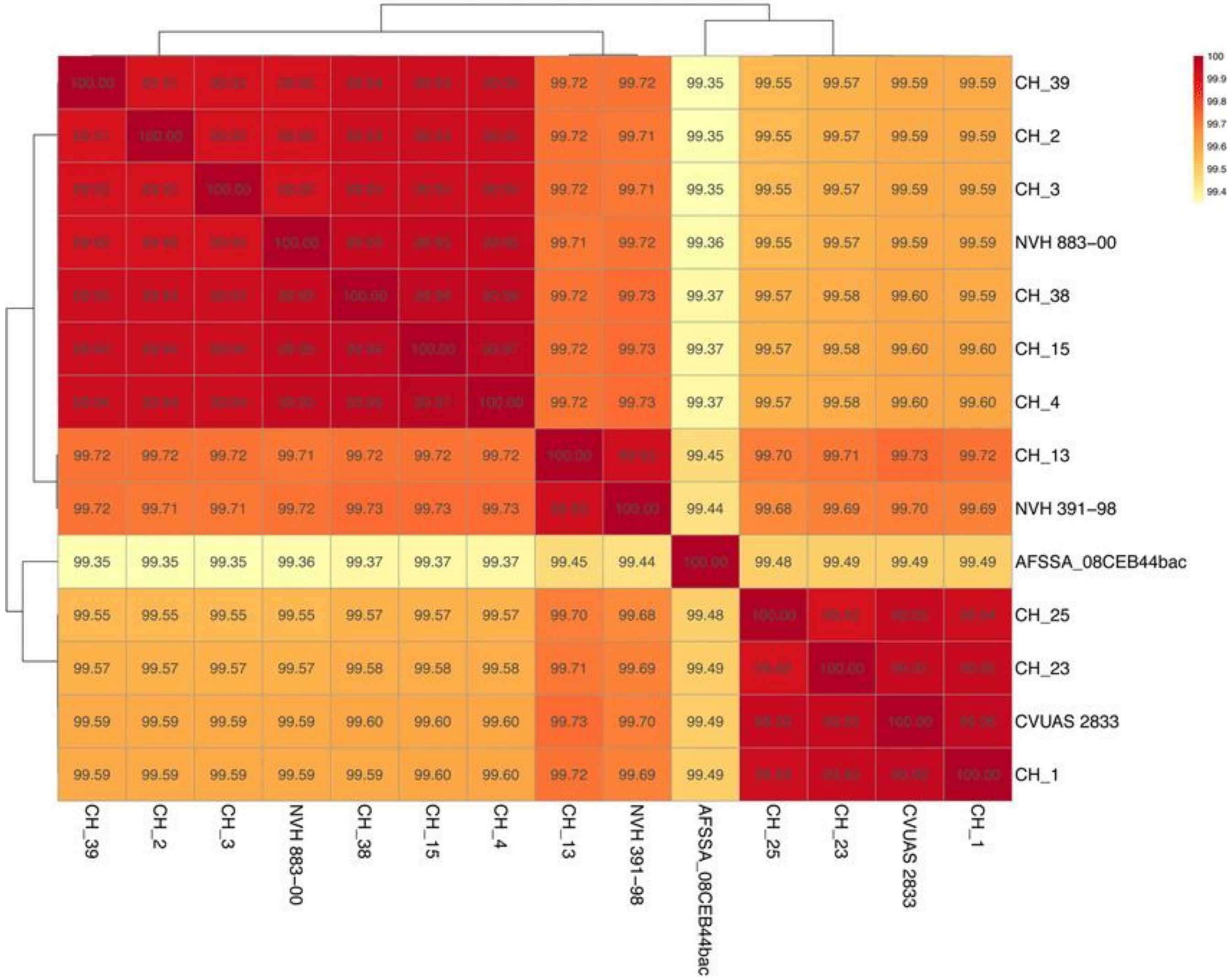

Prokka pipeline (simplified)
FASTA contigs
Aragorn
tRNA
RNAmmer
rRNA
Infernal
Rfam
ncRNA
Prodigal
CDS
SignalP
sig_peptide
GFF3
GBK
ASN1
OUT
User
BLAST+
Swiss
protein annotation
Pfam
HMMER3
TIGR
protein domains

4) 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)*

5) Pan-genome and Gene clustering

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…

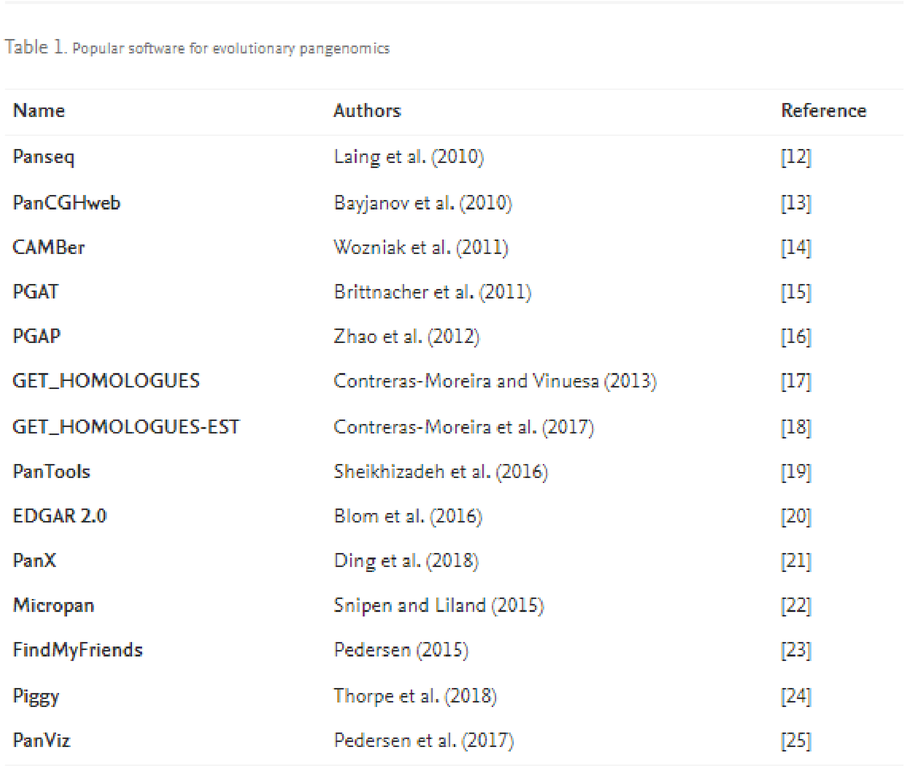
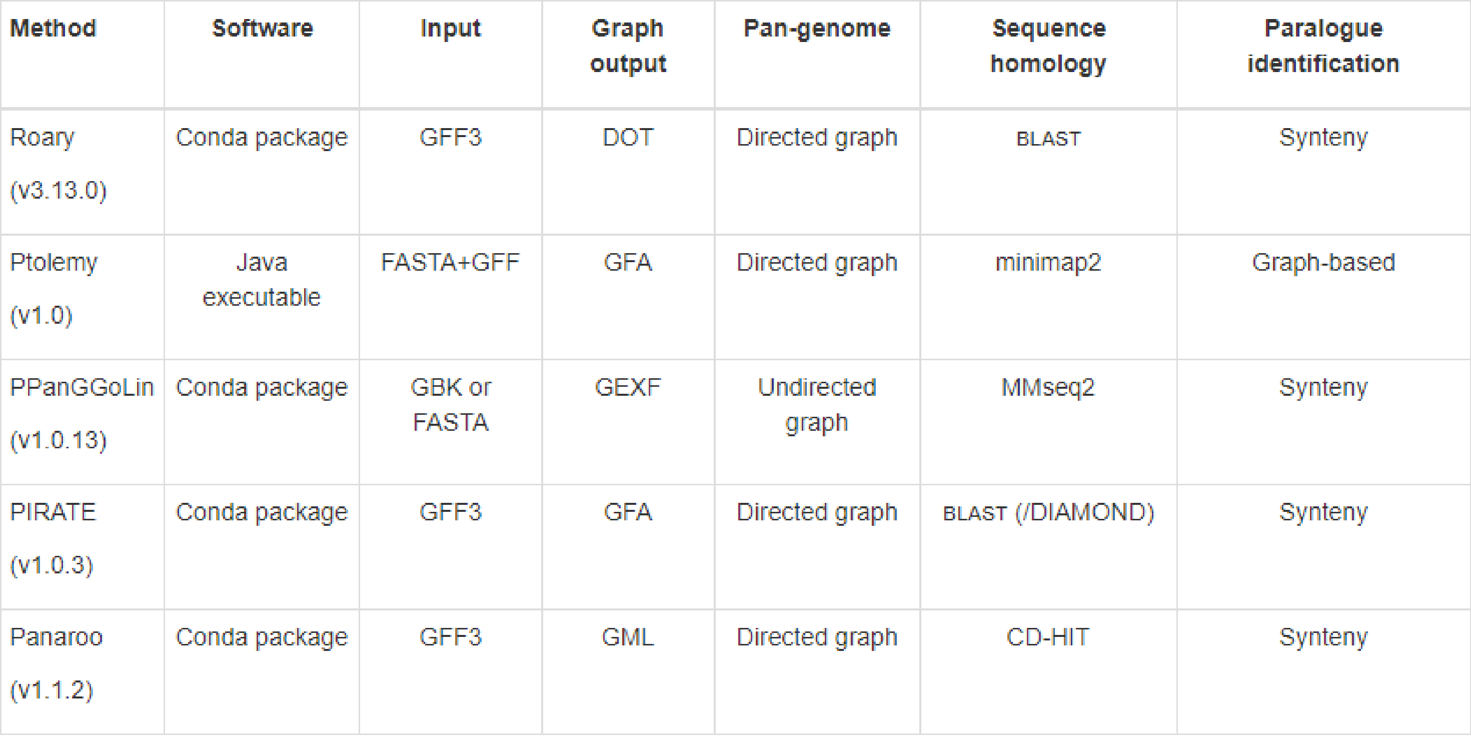
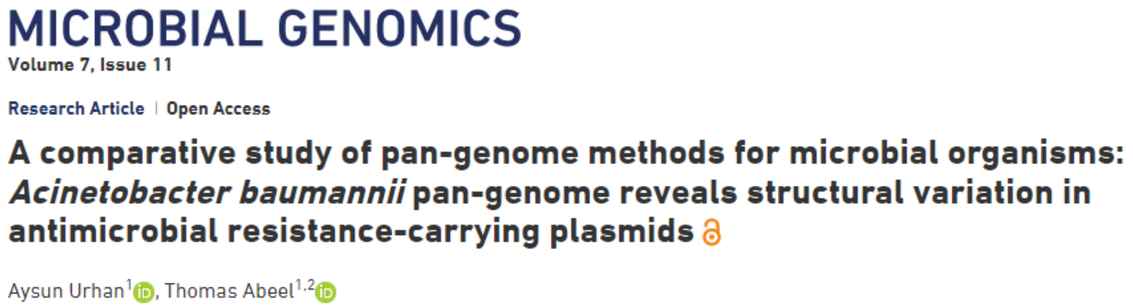


Table 1. Popular software for evolutionary pangenomics

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]



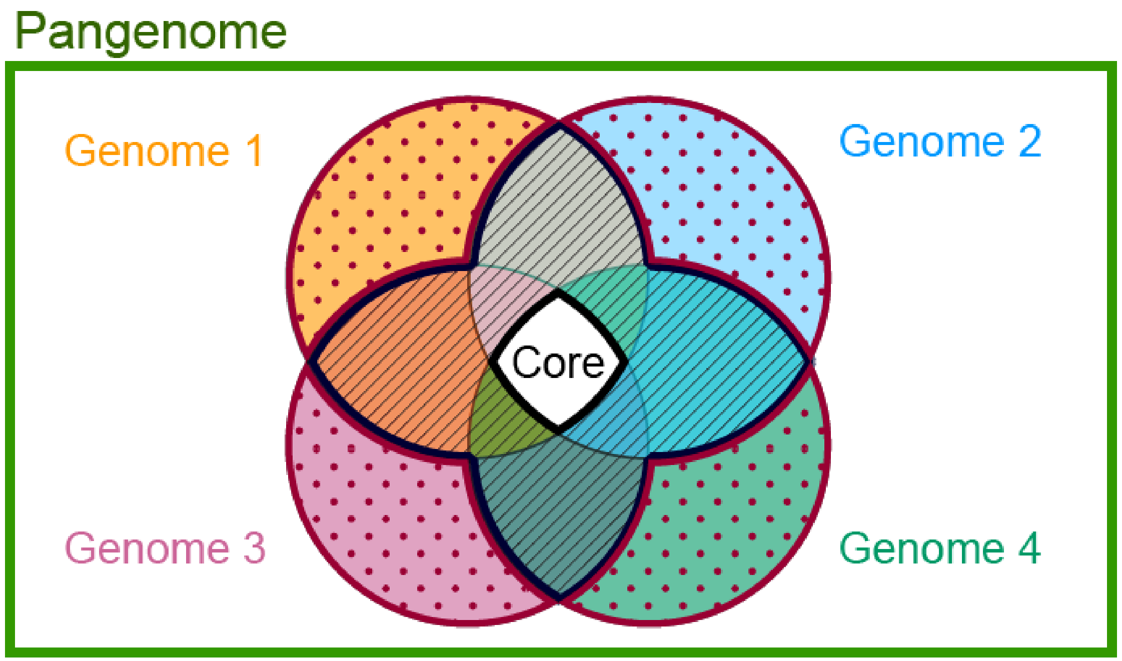
Method	Software	Input	Graph output	Pan-genome	Sequence homology	Paralogue identification
Roary (v3.13.0)	Conda package	GFF3	DOT	Directed graph	BLAST	Synteny
Ptolemy (v1.0)	Java executable	FASTA+GFF	GFA	Directed graph	minimap2	Graph-based
PPanGGoLin (v1.0.13)	Conda package	GBK or FASTA	GEXF	Undirected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

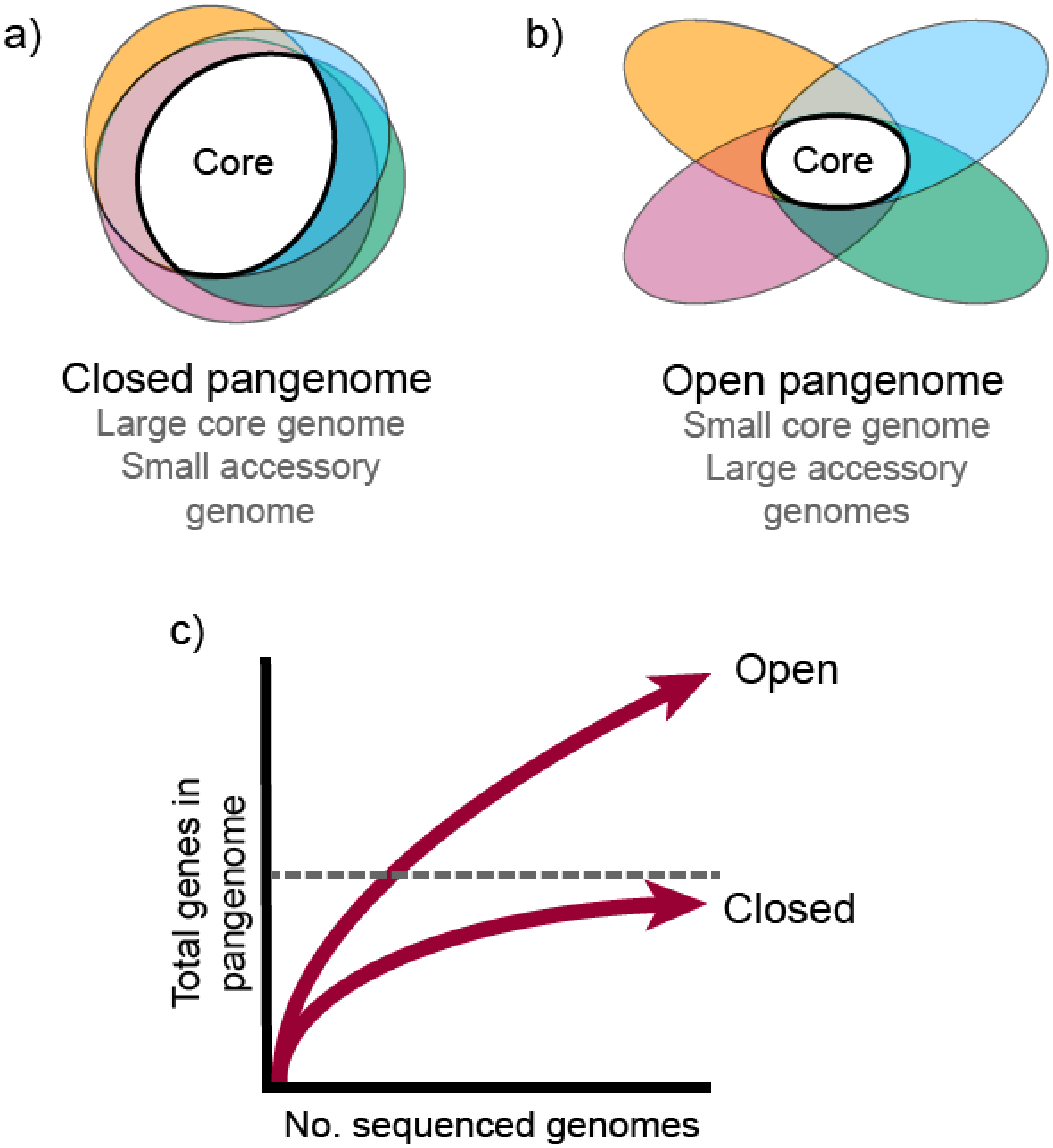
Aysun Urhan[1], Thomas Abeel[1,2]



⁙ Cloud genome

▨ Shell genome

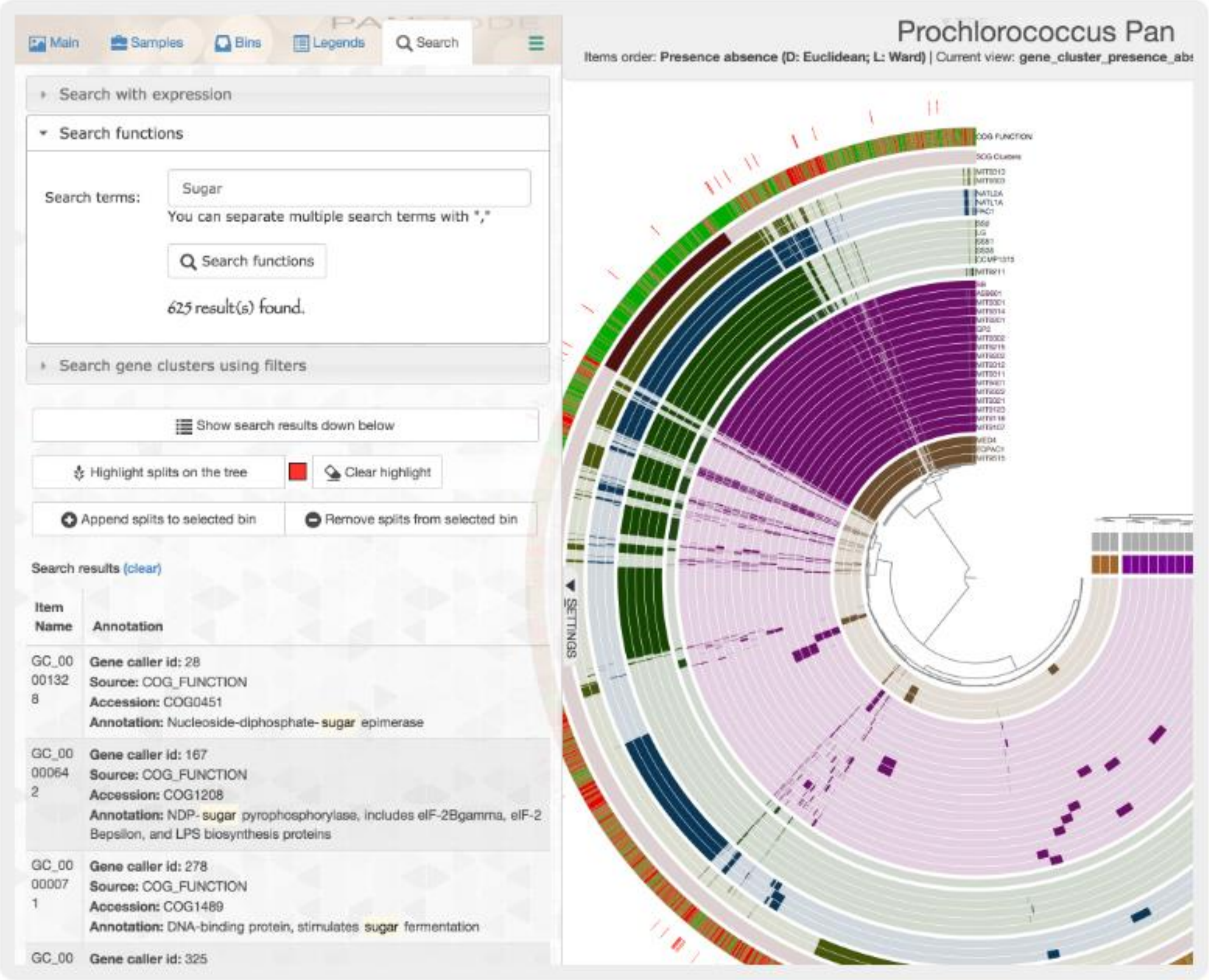
Accessory genome
=
Dispensable genome



Closed pangenome
Large core genome
Small accessory genome

Open pangenome
Small core genome
Large accessory genomes

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.

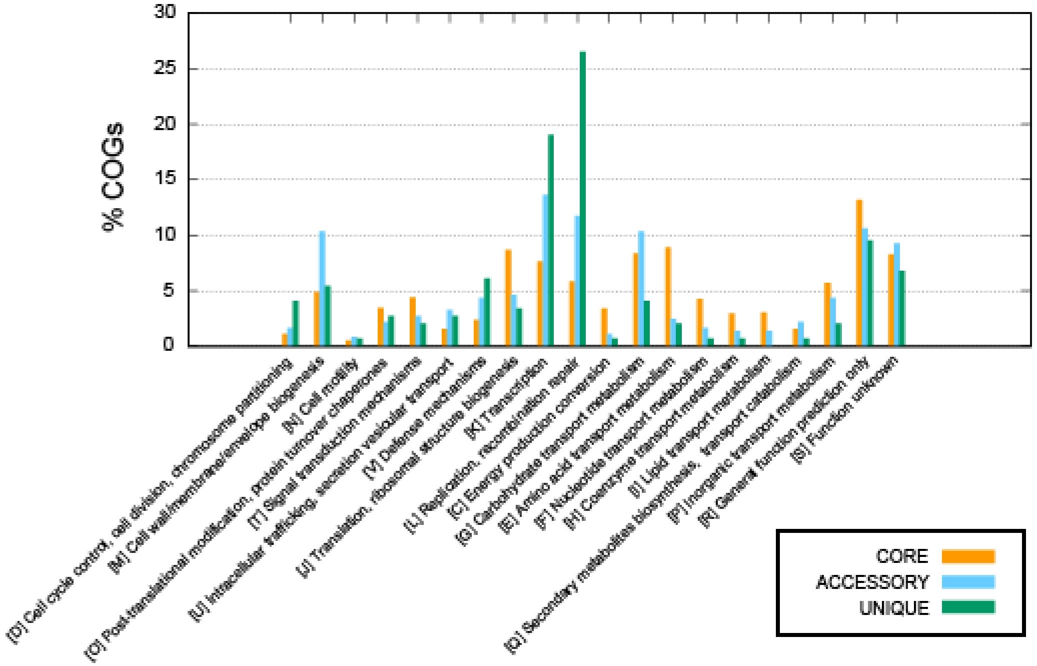
Search results (clear)

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

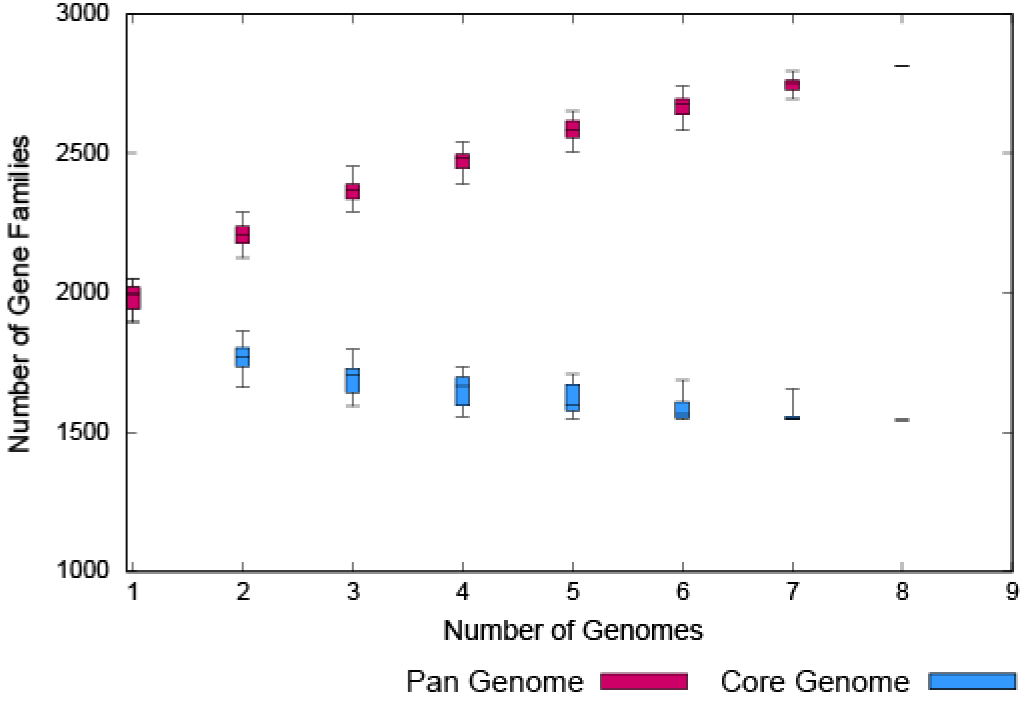
BPGA (Bacterial Pan Genome Analysis tool)

Streptococcus agalactiae

COG Distribution



Pan and Core Genome Plot



6) 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

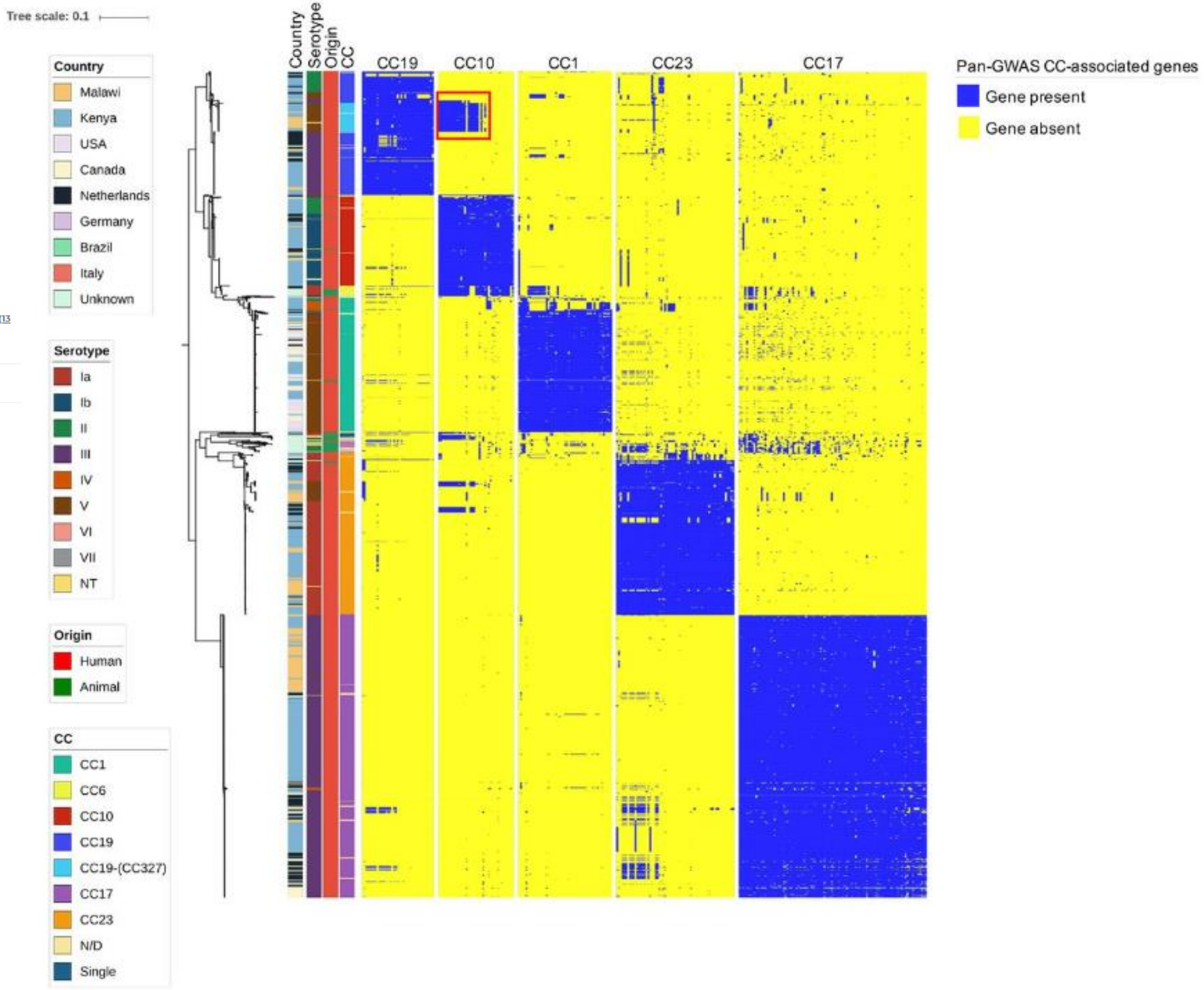


FIG 2 Core genome-based population structure of GBS. The phylogenetic tree is annotated with 4 colored strips representing the clonal complex, the country of isolation, the origin, and the serotype of each strain. The three binary heatmaps represent the presence (blue) or absence (yellow) of the genes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

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.

SouthGreen bioinformatics platform

Merci pour votre attention !

SUIVEZ NOUS SUR TWITTER !

South Green : @green_bioinfo

I-Trop : @ItropBioinfo

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”