PanExplorer2: Explore multi-scale genetic markers derived from pangenome graph for interactive comparative genomics

Easily explore the pan-genomic landscape of your favorite microbes.

Bayram Boukhari[1], Damien Meyer[4], Alvaro Perez[1], Ian Quibod[1], Sébastien Cunnac[1], Ralf Koebnik[1] and Alexis Dereeper[1,2,*]
(1) PHIM Plant Health Institute, Univ Montpellier, IRD, CIRAD, INRAE, Institut Agro, Montpellier, France
(2) French Institute of Bioinformatics (IFB) - South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, F-34398 Montpellier, France
(4) CIRAD, UMR ASTRE, 97170, Petit-Bourg, Guadeloupe, France
* email: alexis.dereeper@ird.fr

https://panexplorer.southgreen.fr/

1) Import your genomes
Import your selection of microbial genomes.
Genomes must be completely assembled and annotated.
Input can be:
- ✓ a list of **Genbank identifiers** (public genomes)
- ✓ a collection of **Genbank files** (private genomes)
- ✓ a collection **Fasta+GFF files** (eukaryotes)

Import Data

.gb

INPUT: List of Genbank ids or Genbank files or fasta+GFF files

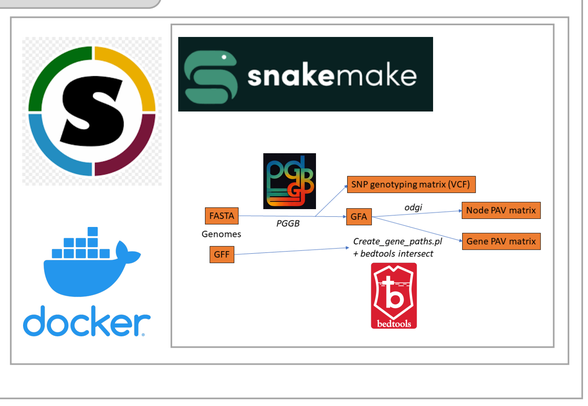
Options:
- *Minimim percentage identity*
- *Pangenome construction software (PGGB, Panacota, roary, orthofinder)*

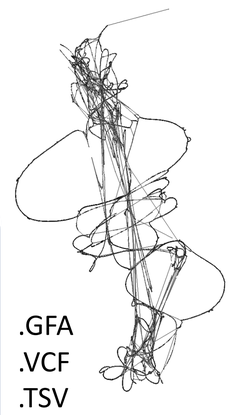
Galaxy Bioblend API

2) Process the PanExplorer workflow
The workflow consists of a Snakemake workflow that executes successively:
- ✓ Retrieval of the corresponding gene sequences
- ✓ Average Nucleotide Identity (ANI)
- ✓ Gene clustering and pan-genome analysis using PGGB (1), Roary (2) or PanACoTA (3),
- ✓ Attribution of COG functional categories and KEGG metabolic pathways
- ✓ GC content calculation using SkewIT (4)
- ✓ Hierarchical clustering, Heatmap, Upset diagram using R

PanExplorer workflow

Galaxy wrapper





.GFA
.VCF
.TSV

https://github.com/SouthGreenPlatform/PanExplorer_workflow (Snakemake workflow)

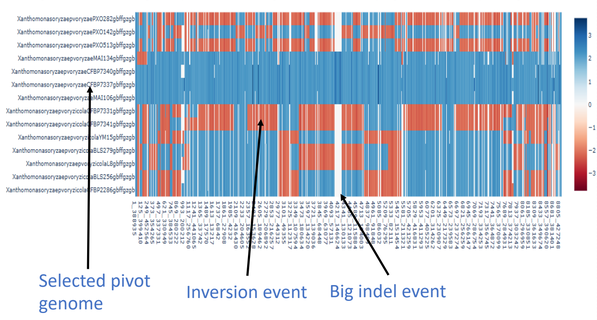
docker://dereeper/panexplorer_workflow:v2

plotly | Dash

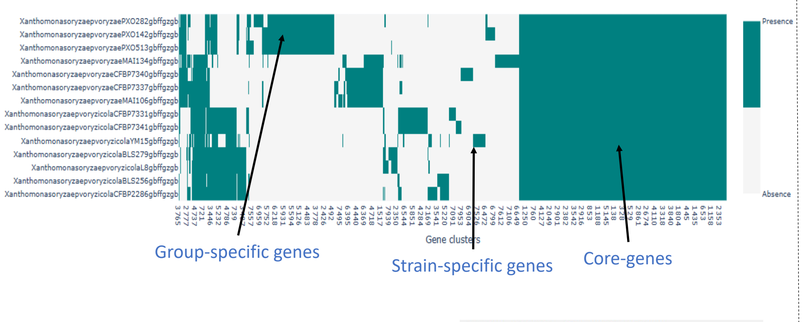
OUTPUTS/VISUALIZATIONS

The web application allows interactive data exploration at different levels

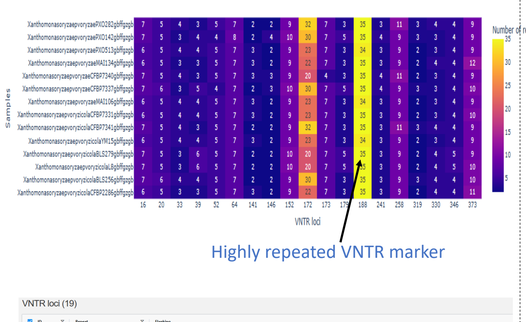
Pangenome-graph segments



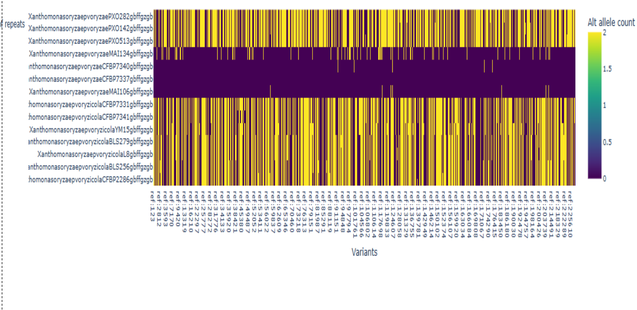
Genes Presence/Absence Variation (PAV)



Repeats

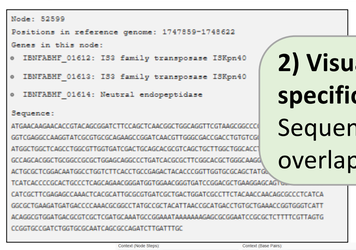


Core-SNPs



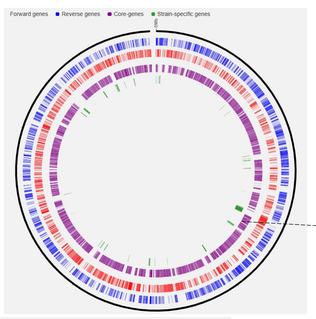
1) Pan-genome graph overview:
- ✓ Representation of presence/absence of segments respecting a minimum size
- ✓ Allows the identification of large structural variations: indels, inversion events

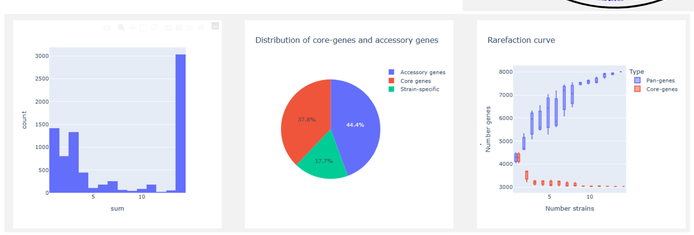
2) Visual inspection of a specific segment.
Sequence, gene overlapping



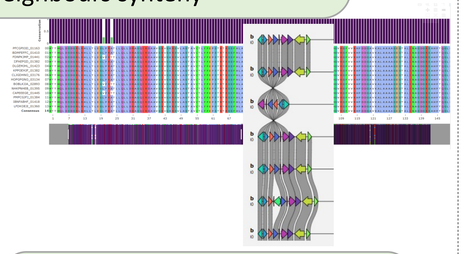
3) Subgraph extraction and visualization using various viewers:
- ✓ Bandage, ODGI, VG

4) Physical map of core-genes and strain-specific genes can be displayed as a circular genomic representation (**Circos**), for each genome taken independently

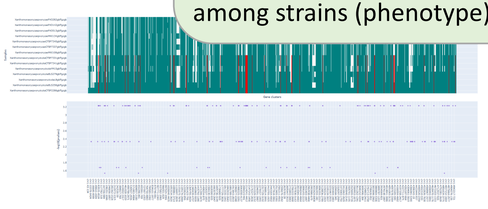




5) Visual inspection of a specific cluster.
Cluster composition, Sequence alignement, haplotypes, neighbours synteny



6) Pan-GWAS (association studies)
A dedicated module allows to search for clusters that respect particular pattern of gene presence among strains (phenotype)

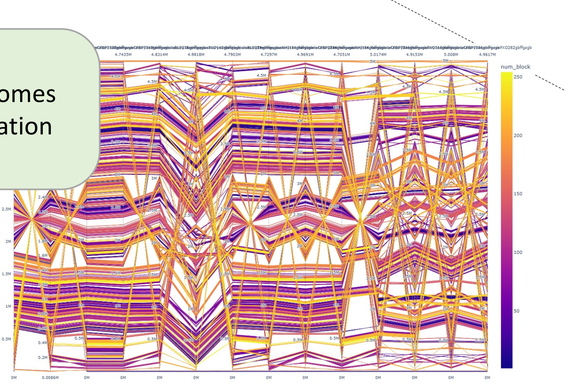


Scoary
microbial pan-GWAS

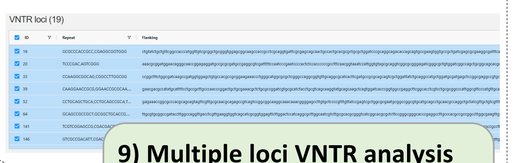
7) COG functions and KEGG metabolic pathways enrichment

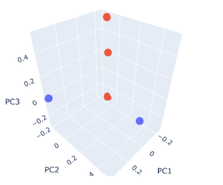
COG Term	Odds ratio	P value	FDR
[R] COG1159 GTPase	17.9941972920696632	0.000126624389057 4	0.0039615346576548
[Q] COG3386 Gluconolactonase	16.7880897913843775	0.000248841401244 4	0.0067418062110341
[K] COG3710 DNA-binding winged-HTH domains	16.7880897913843775	0.000248841401244 4	0.0067418062110341
[R] COG0546 Predicted phosphatases	15.5829145772954322	0.000489366186457	0.0111830842208784
[M] COG0438 Glycosyltransferase	14.3786707882534 78	0.000962348029094	0.0173222645236935
[R] COG1011 Predicted hydrolase (HAD superfamily)	14.3786707882534 78	0.000962348029094	0.0173222645236935
[R] COG1100 GTPase SAR1 and related small G proteins	14.3786707882534 78	0.000962348029094	0.0173222645236935
[R] COG1160 Predicted GTPases	14.3786707882534 78	0.000962348029094	0.0173222645236935

8) Synteny analysis.
The conservation of gene order between genomes can be investigated using graphical representation (HivePlot, Mauve viewer, Clinker)



9) Multiple loci VNTR analysis (MLVA) analysis





10) Genetic diversity and population structure analyses.
SNP-based phylogenetic tree and PCA (using plink) and population structure (with sNMF)

References
(1) Garrison E. et al. (2024) Building pangenome graphs. Nature Methods, 28, 416–418.
(2) Page A.J. et al. (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics, 31, 3691–3693.
(3) Perrin A. et al. (2021) PanACoTA: a modular tool for massive microbial comparative genomics. NAR Genomics Bioinform., 3, lqaa106.
(4) Lu J. et al. (2020) SkewIT: the skew index test for large-scale GC skew analysis of bacterial genomes. PLoS Comput. Biol., 16, e1008439.
(5) Darling A.C. et al. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, 14, 1394-1403.

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE
SouthGreen bioinformatics platform
cirad LA RECHERCHE AGRONOMIQUE POUR LE DÉVELOPPEMENT
IRD Institut de Recherche pour le Développement FRANCE
PHIM Plant Health Institute Montpellier
i-Trop bioinformatics

https://github.com/SouthGreenPlatform/PanExplorer (Web application)