Establishment of pangenome graphs for the analysis and monitoring of fungal plant pathogen populations

Nicolas Lapalu, Antoine Loth, Fabrice Legeai, Ludovic Duvaux, Elisabeth Fournier, Pierre Gladieux, Alice Feurtey, Cécile Lorrain, Marc-Henri Lebrun, Anne Genissel, et al.

Establishment of pangenome graphs for the analysis and monitoring of fungal plant pathogen populations

Nicolas Lapalu[1], Antoine Loth[1], Fabrice Legeai[2,3], Ludovic Duvaux[4], Elisabeth Fournier[5], Pierre Gladieux[5], Alice Feurtey[6], Cécile Lorrain[6], Marc-Henri Lebrun[1], Anne Genissel[1], Thierry C. Marcel[1]

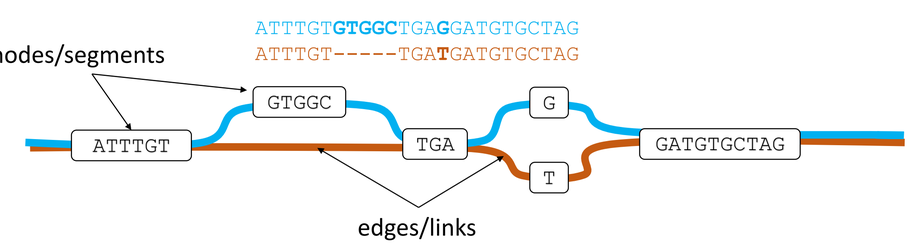
[1] Université Paris-Saclay, INRAE, UR BIOGER, Palaiseau, France
[2] IGEPP, INRAE, Institut Agro, University of Rennes, 35653 Le Rheu, France
[3] Inria, CNRS, IRISA, University of Rennes, 35000 Rennes, France
[4] BIOGECO UMR 1202 INRAE, Université Bordeaux, Cestas, France
[5] PHIM Plant Health Institute, Univ Montpellier, INRAE, CIRAD, Institut Agro, IRD, Montpellier, France
[6] Plant Pathology Group, Institute of Integrative Biology, ETH Zürich Universitätstrasse 2, 8006 Zürich, Switzerland

Corresponding author: nicolas.lapalu@inrae.fr

Scientific context and pangenome graph concepts

A single reference genome cannot represent all the variations present within a species, particularly when numerous mutation events (SNPs, InDels) and large structural rearrangements occur as a result of selective pressures or genome dynamics induced by transposable elements. Pangenome graphs (PGGs) have been developed to integrate population sequence diversity in a single data structure, providing a unique data source corresponding to many reference genomes. In the context of the analysis of populations of phytopathogenic fungi, our aim is to address the following questions:

- Do PGGs provide similar/better results than a reference genome for GWAS or GEA approaches ?
- What type of PGG is most appropriate for the different types of analysis ?
- How to modify, release and keep track of the different versions of a PGG ?

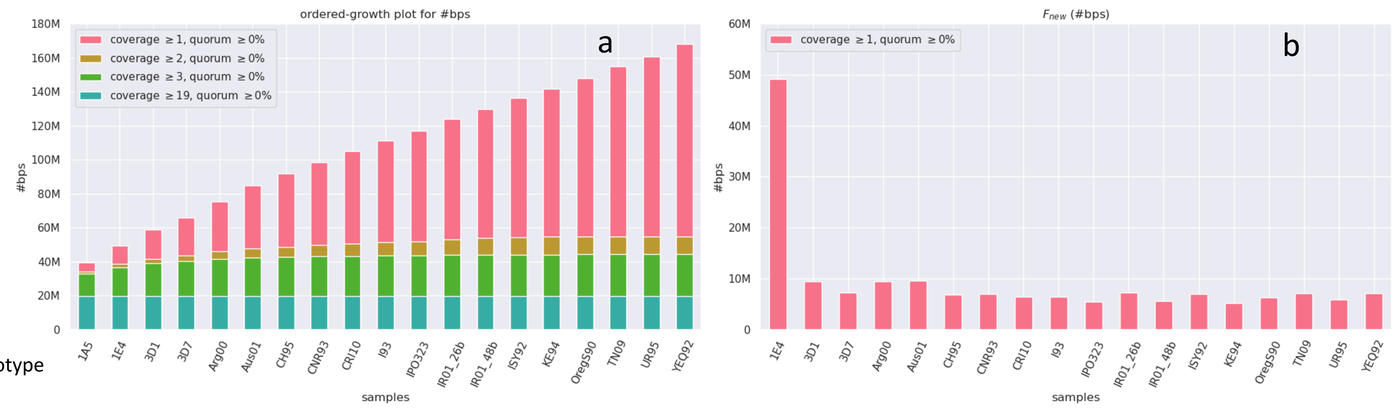


Zymoseptoria tritici as a case study

Based on the 19 genomes used to build the first *Z. tritici* pangenome (Badet *el al.*, 2020), we built PGGs with two different methods: Minigraph (export in rGFA with only large variants) and pggb (export in GFA at the SNP level). The complexity of the pggb graph increases with the number of genomes added (a), and the 19 genomes selected are not sufficient to reach the maximum diversity. In average, each strain contains 7Mb of unique sequences (b), most often found on the eight accessory chromosomes.

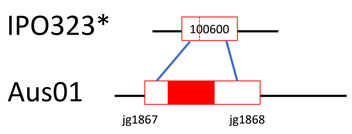
	Minigraph*	pggb
Nodes/Segments	70,187	4,131,367
Edges/Links	99,436	5,599,460
Components	21	21
Average segment length (bp)	2092	40.6
Largest component (Mbp)	18.6	20.2
Total length (Mbp)	146.8	168.0
Size GFA file (Mb)	149	635

*Minigraph export larger variation in rGFA (no export of haplotype paths) requiring sequence mapping to restore haplotype

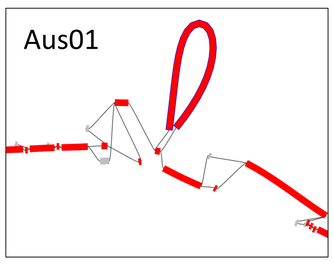


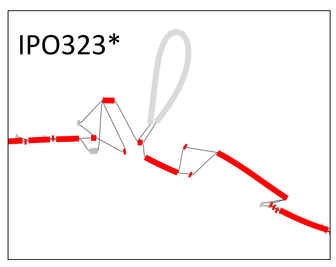
Previously detected events from Z. tritici pangenome were easily retrieved from PGG

1) Singleton GH43 detected in Aus01 (gene:jg1867) corresponds to IPO323 100600 gene split in two parts by a genomic insert (bubble in graph).



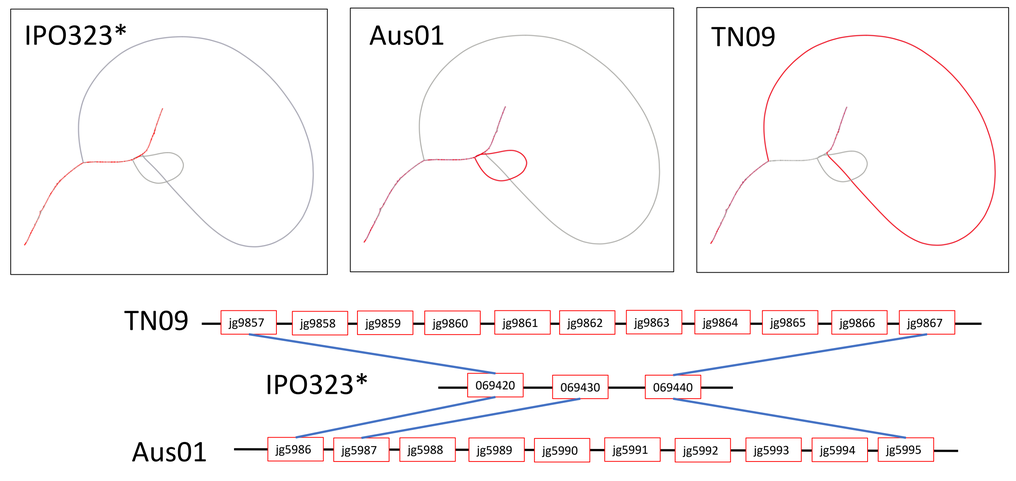
*IPO323 and other strains



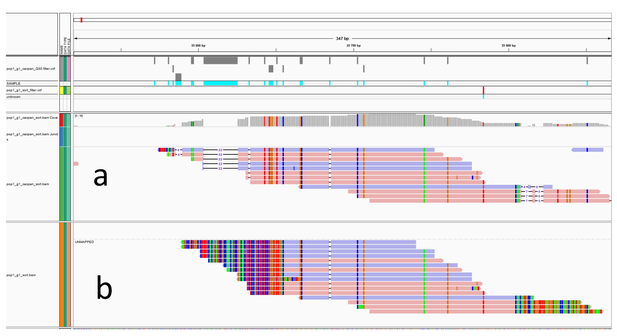


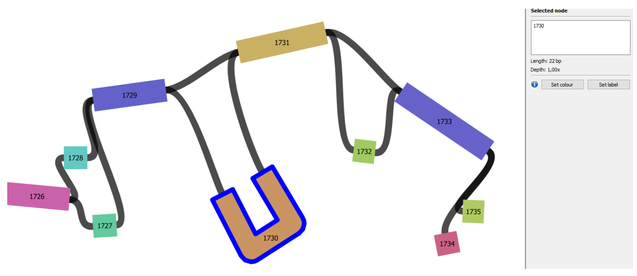
2) Sole indole cluster (jg9859-jg9862) specific to TN09 is highlighted in the PGG by a large bubble (51kb) between IPO323 069420 and 069440 genes.

At the same locus, Aus01 contains a specifc insert of 11kb with 7 predicted genes (jg5988-jg5994) between IPO323 069430 and 069440 genes.



The genetic diversity of PGGs improves mapping of short reads for variant detection

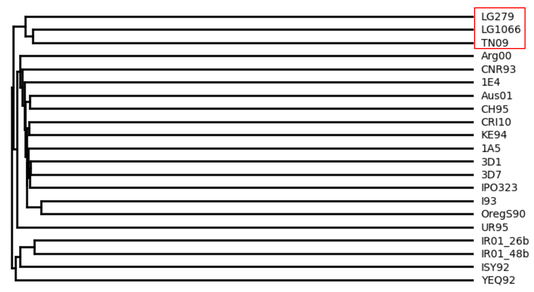




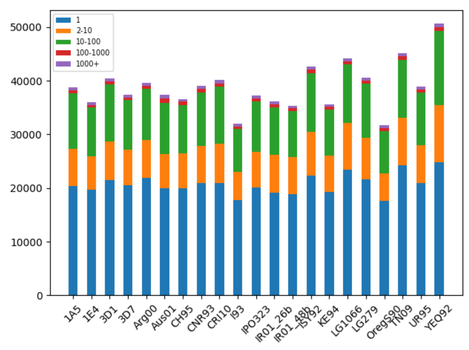
Mapping (GraphAligner) of short reads on the PGG (a) from field-collected strains of *Z. tritici,* detected deletions of 4bp and 22bp (segment 1732,1730 respectively) not present in mapping (bwa) on the IPO323 reference genome (b). The variants detected from the PGG can be exported in VCF format with any strain (contained in the graph) as reference sequence, increasing the number and reliability of detected variants.

PGG topology as a proxy to describe strains similarity

UPGMA tree of pair-wise distances: 1-(#common bases) of the 13 core chromosomes

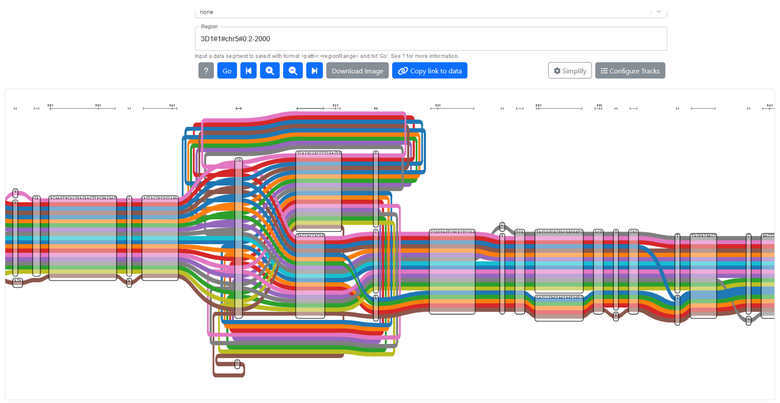


Number of specific variants / size / strains



After the addition to the PGG of two new *Z. tritici* strains (LG1066 and LG279) isolated from durum wheat, clustering of shared bases in pair-wise manner, highlighted the structural proximity with TN09, another durum wheat strain. In terms of the number of variants, the SNPs correspond to the same proportion of the sum of larger variants, highlighting the PGG approaches for characterising all sizes of InDels and structural variations.

Visualization and exploitation of PGGs



PGGs are subject to change as new data becomes available. Implementation of the FAIR principles is essential for sharing with the scientific community as new versions are released. We also believe that visualization will be essential, just as the genome browser is for linear genomes. Currently, projecting genome annotations (gene, TE) onto a graph in an understandable way remains a challenge.

Conclusion and future work

Pangenome graphs are relatively easy to build with different level of complexity, depending on the intended use. For haploid and small genomes such as *Z. tritici*, PGGs with full sequence variations remain small. We plan to build a PGG suitable for variant detection and subsequent GWAS analyses with all *Z. tritici* genomes and field populations available. To facilitate the access for end-users, a web portal will be developed with dedicated browsers and tools to explore specific loci to track allele variations (e.g. inserts in the MFS1 gene promoter) in field populations. The same work will be carried out on *Pyricularia oryzae* as another usecase (with host adaptation considerations) to define and propose a common framework for exploiting PGGs for plant pathogenic fungi.

Badet, T., Oggenfuss, U., Abraham, L., McDonald, B. A., and Croll, D. 2020. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen *Zymoseptoria tritici*. BMC Biol. 18

Funding: **GraPanPhy** (IB INRAE SPE 2024) & **PANANNOT** (SAPI INRAE)
Acknowledgments: Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12), BIOGER for the support of Saclay Plant Sciences-SPS (ANR-17-EUR-0007).