Comparative microbial pangenomics to explore mobilome dynamics

Adelme Bazin, Guillaume Gautreau, Claudine Médigue, Alexandra Calteau, David Vallenet

Comparative microbial pangenomics to explore mobilome dynamics

Adelme Bazin[1], Guillaume Gautreau[1] , Claudine Médigue[1], Alexandra Calteau[1] and David Vallenet[1]

[1]LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, CNRS, Université d'Évry, Université Paris-Saclay, Evry, France.

Regions of genomic plasticity

Regions of genomic plasticity (RGP) are areas of the genome that encompass most of the strain variabilities within a genome including Genomic Islands (GI) such as pathogenicity islands, antibiotic resistance genes, environmental adaptations and secondary metabolites. For this reason they have raised a lot of attention from the scientific community, and numerous tools have been created to find them [1]. Most of them use sequence composition only and expect to differentiate them from the 'core' genome as they tend to not be fully adapted to the genome's nucleotide composition. With the advent of Next Generation Sequencing and the increase of available genomes, methods based on comparative genomics were developed to further improve RGP detection as those approaches can detect actual variations between genomes. However, as of today, no method is truly scalable to cope with the massive increase of available genomic data.

Pangenomics

For the last decade, pangenomics has provided new tools for researchers to estimate genomic diversity by partitioning gene families in terms of *core* and *accessory* genome [2]. The *core* genome consists in ubiquitous genes within the taxonomic group being studied and the *accessory* genome are the genes present in one or some individuals but not all. However, the content of the *core* genome is highly dependent on the number of genomes included in a study limiting the relevance of comparisons between studies. Moreover, the concept of *accessory* genome lacks subtlety as it gathers genes with a large range of frequencies.

Recently, a new tool named **PPanGGOLiN** (Gautreau *et al.,* in preparation) was developed to exploit gene neighborhood, gene frequency and population structure to classify gene families using a graph-based approach. 3 classes are obtained : p*ersistent*, which is a relaxed definition of the *core* genome, *shell* which corresponds to genes belonging to some individuals of the population and associated to environmental adaptations, and *cloud* which are genes present at very low frequencies.

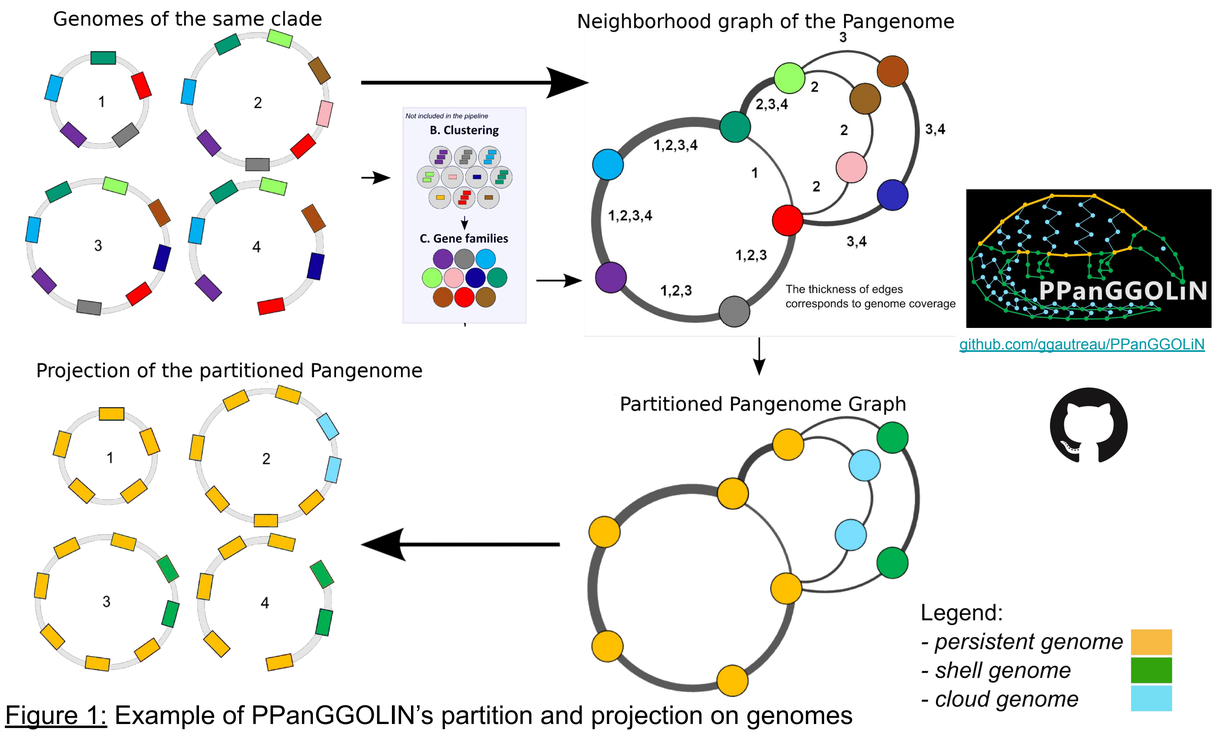


Figure 1: Example of PPanGGOLIN's partition and projection on genomes

Detecting RGP through pangenomics

RGP have a high turnover in terms of gene content. At the scale of the pangenome, we expect them to be made in majority of *shell* and *cloud* gene families. Thus, we can easily extract RGP by searching through the genomes for regions mostly *variable*. We detect them by projecting the pangenome partition on a genome and annotating genes with the partition of their corresponding gene families. Then we use a global scoring approach named **panRGP** where we try to maximize a score penalized by *persistent* genes and improved by *variable* genes to extract potential RGP in each genomes. Its simplicity makes it easily scalable to thousands of genomes as long as the pangenomic analysis scales as well.

To evaluate this approach, we tested it on a curated literature-based dataset from [1]. We built the pangenome of each of the involved species using all the genomes from genbank that have less than 1000 contigs and L90 < 100. We used the GI predictions available from [1] for the other tools.

Tool	MCC	F1-score	Accuracy	Precision	Recall
panRGP	**0.833**	**0.912**	**0.909**	**1.0**	**0.839**
IslandViewer 4	0.71	0.826	0.832	0.996	0.705
IslandPath-DIMOB	0.573	0.696	0.736	0.995	0.535
Islander	0.383	0.442	0.595	**1.0**	0.284
SIGI-HMM	0.332	0.372	0.55	**1.0**	0.228

Table 1: Quality metrics for each tool on a literature dataset

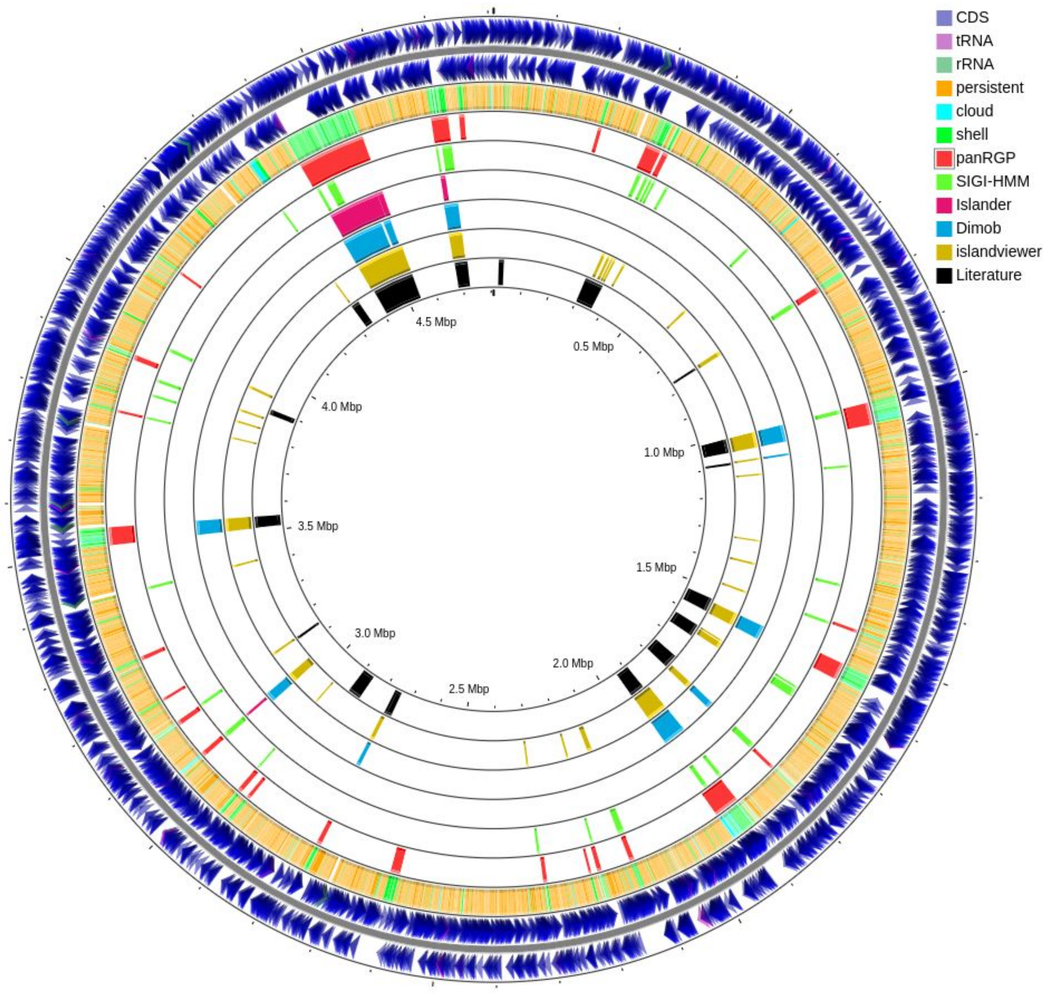


Figure 2: RGP predictions in *Salmonella enterica subsp. enterica serovar Typhi* str. CT18 (NC_003198.1) From inside out the different circles are : Literature dataset, IslandViewer 4, IslandPath-Dimob, Islander, SIGI-HMM, panRGP, PPanGGOLiN partitions, forward strand CDS, reverse strand CDS. Made with [5]

Modularizing regions and detecting co-localized and co-occurring gene sets

Genes involved in the same function tend to be co-localized in genomes [3]. Functions encoded by colocalized genes can be transferred in a single horizontal gene transfer event. If the function provides a fitness advantage it will remain active in the genomes and the genes will be kept by the organisms. Therefore we expect gene families that are in co-localized RGP and co-occurrent in the pangenome to be functionally linked. We use a cover Itemset Mining algorithm [4] using a Jaccard distance to extract gene sets that match defined criteria.

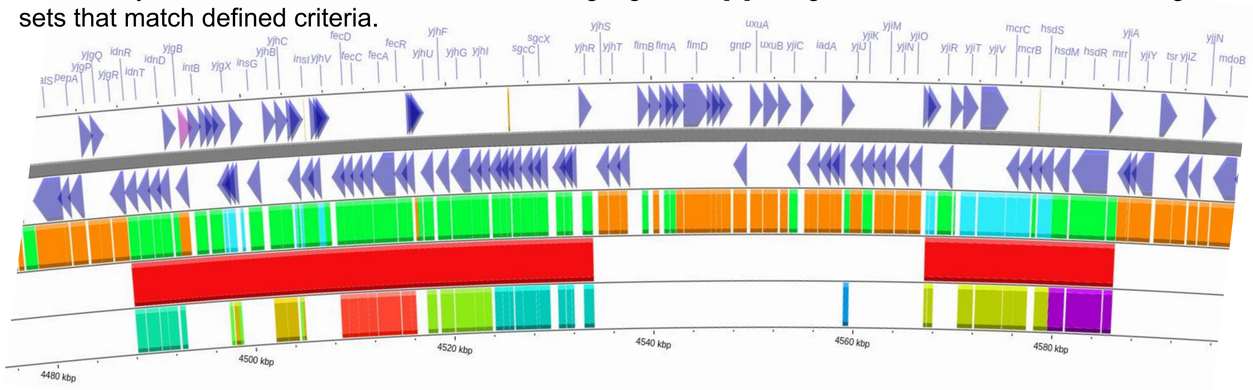
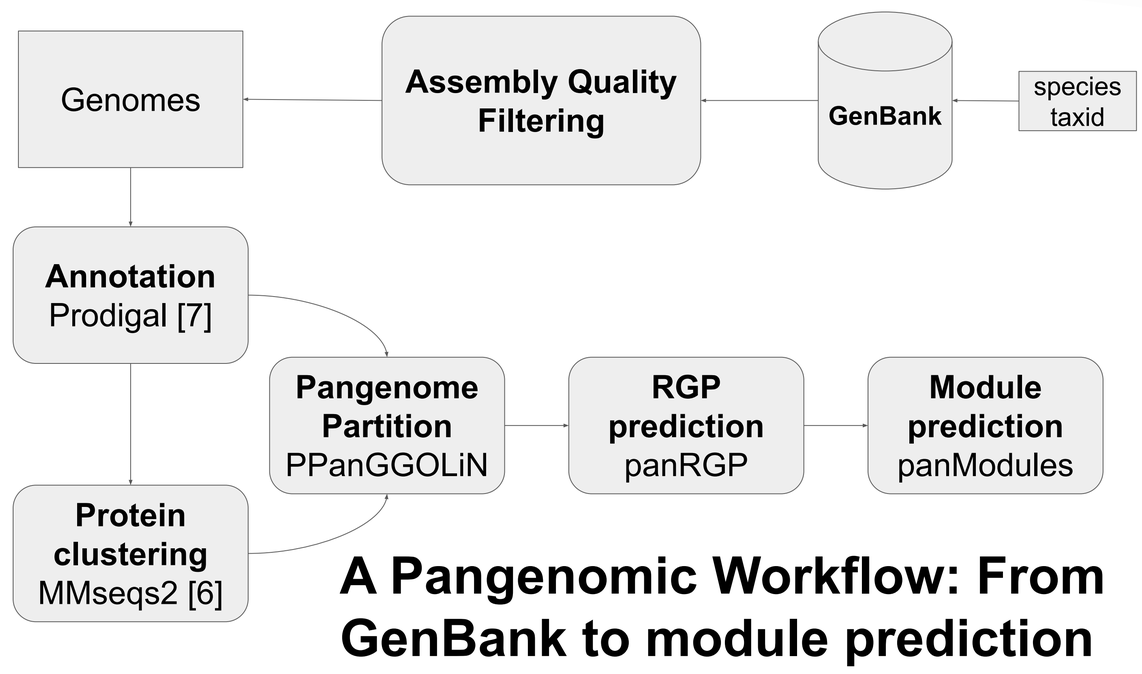


Figure 3: Example of Module predictions in RGP of *Escherichia coli* K-12. From up to down : gene annotation, forward strand CDS, reverse strand CDS, PPanGGOLiN partitions, panRGP predictions, module predictions. Made using [5]



References

[1] Bertelli, C., Tilley, K. E., & Brinkman, F. S. (2018). Microbial genomic island discovery, visualization and analysis. *Briefings in bioinformatics*.

[2] Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., ... & DeBoy, R. T. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, *102*(39), 13950-13955.

[3] Overbeek, R., Fonstein, M., D'souza, M., Pusch, G. D., & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences*, *96*(6), 2896-2901.

[4] Segond, M., & Borgelt, C. (2011). Item set mining based on cover similarity. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 493-505). Springer, Berlin, Heidelberg.

[5] Grant, J. R., & Stothard, P. (2008). The CGView Server: a comparative genomics tool for circular genomes. *Nucleic acids research*, *36*, W181-W184.

[6] Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, *35*(11), 1026.

[7] Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, *11*(1), 119.