Genome analysis

PanExplorer: a web-based tool for exploratory analysis and visualization of bacterial pan-genomes

Alexis Dereeper 1,2,* , Marilyne Summo 2,3 and Damien F. Meyer 1

1CIRAD, UMR ASTRE, F-97170 Petit-Bourg, France, 2CIRAD, INRAE, IRD, F-34398 Montpellier, France and 3CIRAD, UMR AGAP, F-34398 Montpellier, France

*To whom correspondence should be addressed.

Abstract

Motivation: As pan-genome approaches are largely employed for bacterial comparative genomics and evolution analyses, but still difficult to be carried out by non-bioinformatician biologists, there is a need for an innovative tool facilitating the exploration of bacterial pan-genomes.

Results: PanExplorer is a web application providing various genomic analyses and reports, giving intuitive views that enable a better understanding of bacterial pan-genomes. As an example, we produced the pan-genome for 121 Anaplasmataceae strains (including 30 Ehrlichia, 15 Anaplasmataceae, 68 Wolbachia).

Availability and implementation: PanExplorer is written in Perl CGI and relies on several JavaScript libraries for visualization (heatmap.js, MauveViewer, CircosJS). It is freely available at http://panexplorer.southgreen.fr. The source code has been released in a GitHub repository https://github.com/SouthGreenPlatform/PanExplorer. A documentation section is available on PanExplorer website.

Contact: alexis.dereeper@ird.fr

1 Introduction

In the past decade, the pan-genome concept has been largely employed to investigate the bacterial comparative genomics and evolution analyses (Hyun et al., 2022; Jonkheer et al., 2021; Vernikos, 2020). Many programs have been developed for this purpose such as Roary (Page et al., 2015) or more recently PanACoTA (Perrin et al., 2021), and a need is still present for the efficient storage, exploitation and visualization of data derived from pan-genome analyses. To address this question, we developed a web-based application, PanExplorer, which performs online pan-genome analysis and displays resulting information as a comprehensive and easy-to-use solution, through several modules facilitating the exploration of gene clusters and interpretation of data. Several web applications have been recently published for this purpose. PanX (Ding et al., 2018) is a sophisticated web application for browsing among clusters after pre-computed pan-genome analyses, but it does not allow matrix visualization, synteny exploration nor genome import. On the other hand, PGAWeb (Chen et al., 2018), PGAPX (Zhao et al., 2018) and PanWeb (Pantoja et al., 2017) are web servers for running online the PGAP software (Zhao et al., 2012) on users Genbank files, but output files and images cannot be manipulated and browsed interactively. Thus, in contrast to the existing web-based tools for exploring pan-genomes, PanExplorer brings together in a single web application a whole set of data representation modules that allow to view pan-genomic information from different angles and offers the possibility to submit online a customized selection of bacteria strains if completely assembled and annotated.

2 The PanExplorer application

The application takes a list of GenBank identifiers as input and the server will execute successively: (i) retrieval of the corresponding published complete genome sequences and their annotations, (ii) gene clustering and pan-genome analysis using PGAP (Zhao et al., 2012), Roary (Page et al., 2015) or PanACoTA (Perrin et al., 2021), (iii) attribution of the Clusters of Orthologous Groups (COG) functional categories using RPSBlast against the COG database (Tatusov et al., 2000) and (iv) GC content calculation using SkewIT (Lu et al., 2020) (Fig. 1A). The PanExplorer application is developed in Perl CGI, web interactivity is allowed by external JavaScript libraries: heatmap.js (https://github.com/nicgirault/heatmap), MauveViewer (https://github.com/PATRIC3/mauve-viewer), CircosJS (https://github.com/nicgirault/circosJS), MSAViewer (Yachdav et al., 2016), phylotree.js (Shank et al., 2018) and D3 (Bostock et al., 2011).

3 Interactive data exploration

3.1 Pan-genome visualization

As a presence/absence variation (PAV) matrix (Fig. 1B) using the heatmap.js javascript library. This overview allows to easily identify
and distinguish core-genes (present in all strains), dispensable genes (genes from the accessory genome) and strain-specific genes. Furthermore, this section allows a particular cluster of interest to be selected for further investigation and displayed in detail (Fig. 1B). This section plots the distribution of COG function categories for each analyzed genome and highlights potentially over- or under-represented functional category in given strains. In addition, it also reports in a table list COG and COG categories assigned for each gene cluster of the core-genome.

3.2 Circos
The physical map of core-genes and strain-specific genes can be displayed as a circular genomic representation (Krzywinski et al., 2009) (Fig. 1B), for each genome taken independently. Each gene is colorized according to its COG category. It also plots the GC skew values, in sliding windows along the genome.

3.3 Synteny
The conservation of gene order between genomes can be investigated using two graphical representations (Fig. 1B). A Hive Plot built with d3.js (Bostock et al., 2011) is displayed for a global macro-synteny overview, allowing to detect rearrangements between strains (translocation, duplication or inversion) while a Mauve viewer (Darling et al., 2004) allows the zoom in depth until giving access to the gene information. For both representations, each cluster of the core-genome is materialized as a link between genomes.

Comparison is possible after the selection of three genomes among those available in the project.

3.4 Visual inspection of a specific cluster
PanExplorer offers the possibility to focus on any cluster of the pan-genome, and get access to its composition in terms of protein or DNA sequences, to their alignment using Muscle (Edgar et al., 2004) and subsequent gene-based phylogeny (NJ distance tree) or colorized minimum spanning network (Fig. 1B). This examination can be reached from any cluster tables, outputted either from the cluster or gene search panel, or directly after clicking on a gene from Circos or pan-genome matrix overview.

3.5 Dynamic reevaluation using a subset of genomes
The application allows to reduce the PAV matrix after the selection of a subset of strains from the original collection of genomes (e.g. for focusing on one species or one particular phenotype), which finally results to a different list of genes defined as core-genes.

4 Conclusions
PanExplorer offers broad-spectrum of tools for easily exploring pan-genomes for scientists without programming skills (automatization of process, no installation is required). Thanks to a connection to NCBI ftp server, the application guarantees an up-to-date availability of public genomes, to be analyzed on-the-fly, and represents a versatile tool for genome exploration filling a need for bacteriologist
community. Among perspectives and further development, new functionalities might be implemented shortly such as online pan-Genome Wide Association Studies (pan-GWAS) based on Scoary software (Brynildsrud et al., 2016) or COG statistical enrichment studies. By allowing visualization of high dimensional data, PanExplorer can assist investigations of pan-genome dynamics of various bacterial lifestyles. It combines efficiently in-depth search of individual genes and their mutations as well as broad examination of pan-genomes. With routine publication of bacterial genomes, PanExplorer is now a web application of choice to assist microbiological research for a better control of bacterial infectious diseases.

Acknowledgements
This work was technically supported by the CIRAD HPC Data Center, IRD itrop HPC, and web servers of the South Green Bioinformatics platform (https://www.southgreen.fr/).

Funding
This work was a part of the MALIN project funded by the European Union on the Guadeloupe Region under the European Research and Development Funds (ERDF) 2014–2020 program [2018-FED-1084].

Conflict of Interest: none declared.

References
Bostock,M. et al. (2011) D3: data-driven documents. IEEE Trans. Visual. Comput. Graph., 17, 2301–2309.
Brynildsrud,O. et al. (2016) Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol., 17, 238.
Chen,X. et al. (2018) PGAWS-web: a web server for bacterial Pan-Genome analysis. Front. Microbiol., 9, 1910.
Darling,A.C. et al. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res., 14, 1394–1403.
Ding,W. et al. (2018) panX: pan-genome analysis and exploration. Nucleic Acids Res., 46, e5.
Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res., 32, 1792–1797.
Hyan,J.C. et al. (2022) Comparative pangenomics: analysis of 12 microbial pathogen pan-genomes reveals conserved global structures of genetic and functional diversity. BMC Genomics, 23, 7.
Jonkheer,E.M. et al. (2021) The Pectobacterium pangenome, with a focus on Pectobacterium brasiliense, shows a robust core and extensive exchange of genes from a shared gene pool. BMC Genomics, 22, 265.
Krzynowksi,M. et al. (2009) Circos: an information aesthetic for comparative genomics. Genome Res., 19, 1639–1645.
Lu,J. et al. (2020) SkewIT: the skew index test for large-scale GC-skew analysis of bacterial genomes. PLoS Comput. Biol., 16, e1008439.
Page,A.J. et al. (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics, 31, 3691–3693.
Pantoja,Y. et al. (2017) PanWeb: a web interface for pan-genomic analysis. PLoS One, 12, e0178154.
Perrin,A. et al. (2021) PanACoTA: a modular tool for massive microbial comparative genomics. NAR Genomics Bioinform., 3, lgaa106.
Shank,S.D. et al. (2018) phylotree.js—a JavaScript library for application development and interactive data visualization in phylogenetics. BMC Bioinformatics, 19, 276.
Tatusov,R.L. et al. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res., 28, 33–36.
Vernikos,G.S. (2020) A review of pangenome tools and recent studies. In: Tettelin,H. and Medini,D. (eds.), The Pangenome. Springer, Cham, pp. 89–112.
Yachdav,G. et al. (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. Bioinformatics, 32, 3501–3503.
Zhao,Y. et al. (2012) PGAP: pan-genomes analysis pipeline. Bioinformatics, 28, 416–418.
Zhao,Y. et al. (2018) PGAP-X: extension on pan-genome analysis pipeline. BMC Genomics, 19, 36.