PANPROVA: PANgenomic PROkaryotic eVolution of full Assemblies

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Notes

OXFORD

Genome analysis

PANPROVA: PANgenomic PROkaryotic eVolution of full Assemblies

Vincenzo Bonnici [1,*], Rosalba Giugno [2]

[1]Department of Mathematical, Physical and Computer Sciences, University of Parma, Parma, 43124, Italy.
[2]Department of Computer Science, University of Verona, Verona, 37134, Italy.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Computational tools for pangenonic analysis have gained increasing interest over the past two decades in various applications such as evolutionary studies and vaccine development. Synthetic benchmarks are essential for the systematic evaluation of their performance. Currently, benchmarking tools represent a genome as a set of genetic sequences and fail to simulate the complete information of the genomes, which is essential for evaluating pangenomic detection between fragmented genomes.
Results: We present PANPROVA, a benchmark tool to simulate prokaryotic pangenomic evolution by evolving the complete genomic sequence of an ancestral isolate. In this way the possibility of operating in the pre-assembly phase is enabled. Gene set variations, sequence variation and horizontal acquisition from a pool of external genomes are the evolutionary features of the tool.
Availability and Implementation: PANPROVA is publicly available at `https://github.com/InfOmics/PANPROVA`.
Contact: vincenzo.bonnici@univr.it

1 Introduction

Computational pangenomic analysis aims at recognizing the sharing of biological information between living organisms (Tettelin and Medini, 2020). The focus of the analyses is to detect the presence of a given genetic family within a set of genomes (Tettelin *et al.*, 2005). Such an analysis is used in a wide range of applications, from studies of evolutionary behavior to vaccine development (Zhang *et al.*, 2019; Barrick *et al.*, 2009; Muzzi *et al.*, 2007). Comparing the performance of computational approaches for pangenomic analyses is a non-trivial task (Kim *et al.*, 2020; Bonnici *et al.*, 2021). Synthetic benchmarks are needed to: (i) trace the evolutionary steps of the simulated genomes, (ii) know the expected result of a pangenomic analyses of such genomes, and (iii) assess performance of the approaches comparing their output with the expected one.

Tools that simulate microbial whole genome evolution are already available Dalquen *et al.*, 2012, however, they don't produce populations showing expected pangenomic content. A main reason is the lack of procedures to simulate gene set variation to reflect horizontal gene transfer (HGT). Other methods evolve only a set of genes as representative of genomes and include HGT transfers (Bonnici *et al.*, 2021; Ferrés *et al.*, 2020; Bobay, 2020).

The above tools can not be applied when a pangenomic content must be retrieved from unassembled genomes. Recently, a plethora of fragment-level genomes are increasingly available and new pangenomic tools begin to address the problem of extracting biological information from them(Veras *et al.*, 2018; Gabrielaite and Marvig, 2020). In this perspective, there is a need of benchmarks obtained by simulating pangenomic processes in which the whole genomic information, rather than a gene set representation, is taken into account. Once whole genomes are produced, artificial fragmentation can be constructed.

We present PANPROVA, a computational tool for simulating prokaryotic pangenomic evolution by evolving the complete genetic sequences taking into account gene set variations, sequence alteration and horizontal acquisition. Unlike existing methods, PANPROVA applies variations at the DNA level with the advantage of evolving genetic sequences in a non-independent way and thus naturally preserving the overlap and proximity of genes.

2 Description

PANPROVA takes as input a full genomic sequence of an isolate and its gene sets annotated according to the GBK (GeneBank) file format. It generates a phylogenomic tree, rooted in the input genome, via a computational evolution. At each evolutionary step, a genome belonging to the tree is randomly selected to be the parent of a newly formed isolated. The genomic material is transmitted by allowing evolutionary variations according the user parameters.

Gene sequences are altered during vertical transmission.Single nucleotide modifications (insertions, deletions and substitutions) are randomly applied to the DNA sequence enclosed by the selected gene. Thus, the modification is automatically reflected to any other overlapping gene. The modifications never affect start and stop codons of any gene within the whole genomic sequence. The resultant transmitted gene set is subsequently modified, according to user-provided percentages, by simulating events of loss, duplication or acquisition of genetic annotation. Loss and acquisition are applied simultaneously and only after duplication events. Any gene set variation event affects the DNA sequence of the isolate preserving the other genes that are non involved in the event. In case of loss, only the portion of the selected gene that does not overlap with other annotations is removed from the genomic sequence. Instead, for what concerns duplication and acquisition, the new genetic sequence is inserted in a genomic location that is not covered by any current annotation.

Because horizontal gene transfer has a high impact on the pangenomic content of microbial species, an ad hoc procedure for gene acquisition is proposed. A pool of genetic material to be acquired is taken into account. At each gene acquisition event, a gene is randomly chosen from the pool and it is added to the given genome. When the pool is empty, random sequences are created following a uniform nucleotide distribution. The set of genes composing the pool is provided by the user. Alternatively, a set of *non-redundant* genetic sequences is retrieved by a user-defined set of genomes. Two genetic sequences are defined *non-redundant* if their sequence similarity is lower that a user-defined threshold. PANPROVA applies the generalized Jaccard distance defined in Bonnici *et al.*, 2018 to compare sequences by means of their k-mer content. An initial version of the pool is composed of genes belonging to the ancestral isolate. Then, it is enlarged by iteratively scanning the user-provided genomes.

The final product of the simulation is a set of isolates from which an artificial fragmentation can be constructed. Parental relationships and relationships between their genes are also provided. PANPROVA is intended to be used on Linux systems, under MIT license. The core components of PANPROVA are developed in C++ for reducing computational requirements, and accessory procedures are developed in Python3 and as Bash scripts.

3 Results

We tested PANPROVA by generating a population of 1,000 genomes starting from a *Mycoplasma genitalium (G37)* ancestor having 580,076 nucleotides and 476 genes. The genetic content of other bacterial species, such as *Bacillus subtilis (168) and Campylobacter jejuni (NCTC11168)*, was extracted for composing an HGT pool of 21,846 unredundant genes. All the genomes were downloaded from NCBI. Default parameters were used: 0.5 probability of gene alteration, 0.01 for nucleotide alteration, 0.001 for gene duplication, 0.9 and of 0.1 for capturing or deleting genes, and the value 1% for gene set variation.

The generated genomes tend to increase their size (with an average of 508 genes and 61k nucleotides), mainly due to the ration between the probability of acquiring and deleting a gene. Figure 1 shows that the generated pangenomic distribution, reporting the number of genes present in a given number of genomes, follows an expected U-shape curve. Extraction of the unredundant collection may takes hours because hundreds

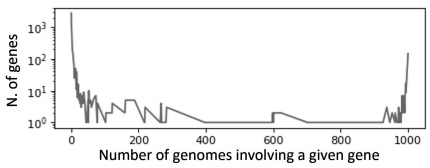

Fig. 1. Pangenomic distribution of the generated population of 1,000 genomes.

of thousands of genes are compared, but the pool can be reused multiple times. Simulation of 1,000 isolates only takes a few minutes. Further experiments were conducted by using an *Escherichia coli* genome as ancestor, and similar results were obtained (not shown here). Details of the experiments are available through the GitHub repository of PANPROVA.

4 Conclusions and future directions

We presented PANPROVA, the first available tool for building benchmarks for pangenomic analyses from full genomes. It generates full assemblies showing expected pangenomic contents. Because the whole genome structure is preserved, it enables synthetic analyses and simulating of genome fragmentation. Future directions regard the extension of user customization capability such as defining non-uniform nucleotide variation probabilities, enabling gene fusion events and recombinant variation of the genomic sequences.

Funding

Authors were partly supported by GNCS-INDAM, JPND 2019-466-037.

References

Barrick, J. E. *et al.* (2009). Genome evolution and adaptation in a long-term experiment with escherichia coli. *Nature*, **461**(7268), 1243–1247.

Bobay, L.-M. (2020). Coresimul: a forward-in-time simulator of genome evolution for prokaryotes modeling homologous recombination. *BMC bioinformatics*, **21**(1), 1–7.

Bonnici, V. *et al.* (2018). Pandelos: A dictionary-based method for pan-genome content discovery. *BMC bioinformatics*, **19**(15), 47–59.

Bonnici, V. *et al.* (2021). Challenges in gene-oriented approaches for pangenome content discovery. *Briefings in Bioinformatics*, **22**(3), bbaa198.

Dalquen, D. A. *et al.* (2012). Alf—a simulation framework for genome evolution. *Molecular biology and evolution*, **29**(4), 1115–1123.

Ferrés, I. *et al.* (2020). simurg: simulate bacterial pangenomes in r. *Bioinformatics*, **36**(4), 1273–1274.

Gabrielaite, M. and Marvig, R. L. (2020). Genapi: a tool for gene absence-presence identification in fragmented bacterial genome sequences. *BMC bioinformatics*, **21**(1), 1–8.

Kim, Y. *et al.* (2020). Current status of pan-genome analysis for pathogenic bacteria. *Current opinion in biotechnology*, **63**, 54–62.

Muzzi, A. *et al.* (2007). The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials. *Drug discovery today*, **12**(11-12), 429–439.

Tettelin, H. and Medini, D. (2020). *The pangenome: Diversity, dynamics and evolution of genomes*. Springer Nature. ISBN: 978-3-030-38283-4.

Tettelin, H. *et al.* (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.

Veras, A. *et al.* (2018). Pan4draft: a computational tool to improve the accuracy of pan-genomic analysis using draft genomes. *Scientific reports*, **8**(1), 1–8.

Zhang, B. *et al.* (2019). The poplar pangenome provides insights into the evolutionary history of the genus. *Communications biology*, **2**(1), 1–8.