Bioinformatics

https://doi.org/10.1038/s41592-024-02402-7

Constructing and personalizing population pangenome graphs

Rayan Chikhi, Yoann Dufresne & Paul Medvedev

Pangenome graphs signify a new frontier in genome representation. Recent advances in constructing and personalizing them mark progress in this area.

The identification of genomic variants is a core bioinformatics challenge with a wide range of impactful biological applications such as disease diagnosis, cancer therapy and genetic diversity studies. A recent paper[1] once again highlighted the biological significance of using a pangenome reference, rather than a linear one, to genotype genomic variants. The algorithmic bioinformatic community was early in identifying the promise of pangenomes[2] and has been developing the methodological underpinnings of today's pangenomic breakthroughs for many years. Now that the advantages of pangenome references are more widely recognized, a new array of algorithmic and statistical challenges are opening up[3]. Two papers in this issue of *Nature Methods* tackle two of those challenges: how to construct a general reference that is not biased toward a single individual[4] and how to downsample from a general reference to improve the genotyping of a specific sample[5].

In the first paper, Garrison et al.[4] present PanGenome Graph Builder (PGGB), a method for constructing pangenome graphs for any species, also described in ref. 1 (Fig. 1, left). By performing an all-versus-all alignment of a set of input genomes, PGGB creates a pangenome graph that is not biased toward any particular haplotype, unlike previous methods. PGGB offers a comprehensive representation of variation, ranging from single-nucleotide polymorphisms to complex structural variants, and is scalable to hundreds of genomes. The PGGB tool suite enables researchers to identify variation, measure conservation and detect recombination events with improved precision. A strength of this work is its focus on usability for biologists and bioinformaticians alike[6]. The software is designed in a modular manner, with each module's interface well documented and accompanied by examples. The integration of a pangenome visualization module is particularly relevant for the biological interpretation of the graphs.

In the second, Sirén et al.[5] propose a method for personalizing a general pangenome graph, such as the one generated by PGGB or Minigraph-Cactus[7] (Fig. 1, right). On one hand, including as many individuals as possible is important for making a pangenome graph representative; on the other, it can complicate the genotyping of a single individual by creating read mapping ambiguities. Thus, when genotyping a single individual, having too many distant individuals included in a pangenome graph goes from being a likely advantage to a potential burden. Implemented as part of the Giraffe tool[8], the method of Sirén et al. samples haplotypes by searching for *k*-mers that are unique in the pangenome graph and present in the sequenced reads. Distant haplotypes are thus removed from the graph, and read mapping and

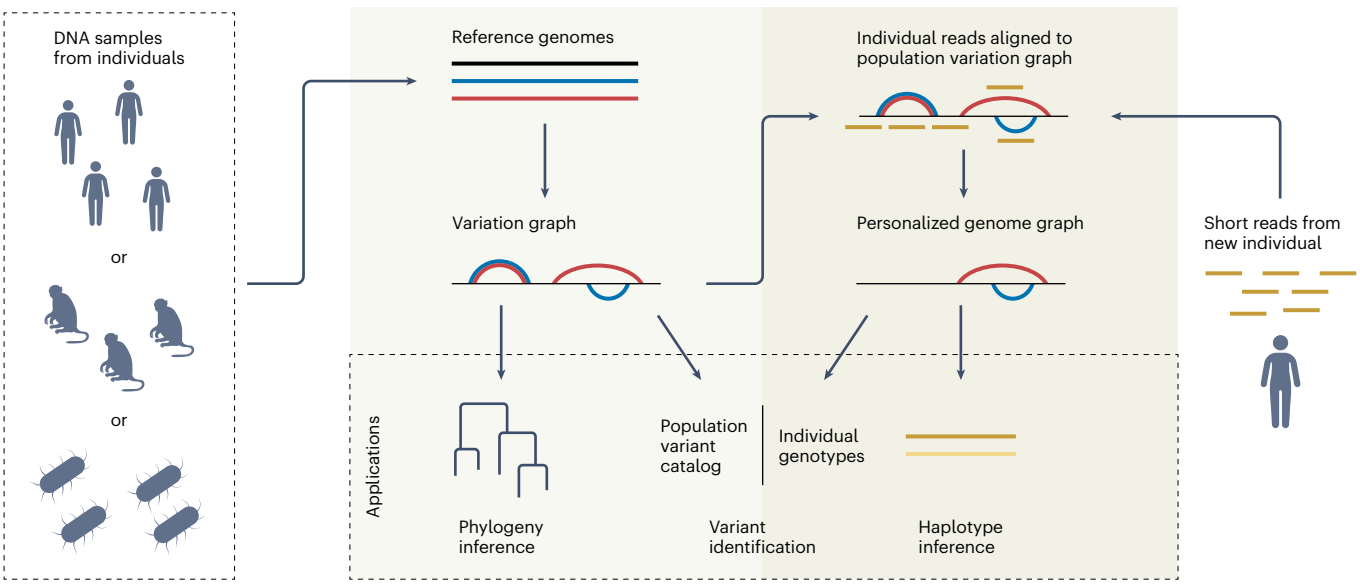

Fig. 1 | Workflow for the construction and personalization of pangenome graphs. At left, a set of reference genomes is transformed into a population pangenome graph that records haplotypes of the entire population[4]. At right, the population pangenome graph is personalized using sequencing data from a new individual, enabling improved genotyping of the individual[5].

variant calling are then performed on the resulting personalized graph. The method demonstrates that, with a minimal increase in runtime, more variants are correctly detected, allowing for more precise downstream analysis. This approach enhances the accuracy of structural variant genotyping, bringing short-read genotyping performance close to that of long-read methods—a longstanding challenge in genomics.

The two new papers reinforce a bifurcation in what a pangenome graph can mean for the biological community. On the one hand, a population pangenome graph faithfully represents the complete variation of haplotypes present in some population of interest. It could be used to detect population polymorphisms, look for signals of selection or reconstruct phylogenies. The paper by Garrison et al. addresses one important aspect of such a graph, which is that it should not be biased toward any single individual within a population. On the other hand, an application-specific pangenome graph reflects only some of the population variability and is instead tailored toward maximizing the performance of application-specific downstream tools. The paper by Sirén et al. proposes such a graph to improve the accuracy of individual genotyping. Importantly, their graph no longer accurately represents the whole population, as it removes sequences that can confound the mapping algorithm used for genotyping. Personalized medicine may drive many such application-specific pangenome graphs in the future (for example, for detection of novel splice sites in a tumor sample). The two papers contribute to what we believe should be a clear decoupling of these 'population' pangenome graphs from 'application-specific' pangenome graphs. We note that such a decoupling is not specific to references that are pangenomes and was already relevant for linear ones[9].

We are now in a formative period, and pangenome tools developed in the next few years will set the bar for what biological studies will come to expect. It will become increasingly difficult to integrate algorithmic improvements into biological applications once the early mainstream tools become entrenched. Therefore, we think it is a good time to reinforce the standard that the output produced by a bioinformatics tool should be well defined. In particular, the pangenome produced by a tool must be defined as precisely as possible. When a user has a pangenome reference in their hands, they should know what properties it has with respect to the haplotype sequences that were used to generate it. If not, then the biological results obtained with respect to this reference become technically meaningless or, even worse, may be ascribed meaning that is inaccurate. Along the same lines, pangenome tool developers should, in our opinion, strive to minimize ad hoc heuristics and arbitrary parameters, which lead to hard-to-predict artifacts of downstream tools. Such artifacts may mislead the biomedical interpretations of the results: for example, what may look like an enrichment of variants in a genomic region may just be an artifact of an ad hoc graph construction approach.

Rayan Chikhi [1] ✉, **Yoann Dufresne** [1] **& Paul Medvedev** [2]
[1]Institut Pasteur, Université Paris Cité, Sequence Bioinformatics Unit, Paris, France. [2]The Pennsylvania State University, State College, PA, USA.
✉e-mail: rayan.chikhi@pasteur.fr

Published online: 21 October 2024

References

1. Liao, W.-W. et al. *Nature* **617**, 312–324 (2023).
2. Computational Pan-Genomics Consortium. *Brief. Bioinform.* **19**, 118–135 (2018).
3. Baaijens, J. A. et al. *Nat. Comput.* **21**, 81–108 (2022).
4. Garrison, E. et al. *Nat. Methods* https://doi.org/10.1038/s41592-024-02430-3 (2024).
5. Sirén, J. et al. *Nat. Methods* https://doi.org/10.1038/s41592-024-02407-2 (2024).
6. Andreace, F., Lechat, P., Dufresne, Y. & Chikhi, R. *Genome Biol.* **24**, 274 (2023).
7. Hickey, G. et al. *Nat. Biotechnol.* **42**, 663–673 (2024).
8. Sirén, J. et al. *Science* **374**, abg8871 (2021).
9. Schröder, J., Girirajan, S., Papenfuss, A. T. & Medvedev, P. *PLoS One* **10**, e0136771 (2015).

Acknowledgements

R.C. was supported by Agence Nationale de la Recherche (ANR) grants ANR-22-CE45-0007, ANR-19-CE45-0008, PIA/ANR16-CONV-0005, ANR-19-P3IA-0001 and ANR-21-CE46-0012-03 and Horizon Europe grants 872539, 956229, 101047160 and 101088572. P.M. was supported by US National Science Foundation grant DBI2138585 and National Institute of General Medical Sciences award R01GM146462.

Competing interests

The authors declare no competing interests.