Comment

https://doi.org/10.1038/s41467-026-68624-9

On the use and misuse of pangenome and related terms

David Edwards

The growth of pangenomics has been rapid, as data and tools improve over time. However, care should be taken to avoid the misuse of pangenome and related terms, with clarity and precision taking precedent over fashion and trends.

With the growth of areas of science, new terminology often emerges to describe the field. This terminology is then sometimes expanded, mutated and overused as researchers join what may be considered a trend. This is exemplified by the growth in genomics, spawning a wide array of other "omic technologies", from the commonly used and usually accurate term "transcriptome", through to the less common "bibliome" (describing biological literature rather than biblical texts) and the probably tongue in cheek "shovelomics" (previously known as root physiology).

A similar process of terminology inflation has been occurring with the growth of pangenomics. A pangenome was originally defined as including a core genome containing genes or genomic regions present in all individuals, and a dispensable genome composed of genes or genomic regions absent from one or more individuals[1]. Pan, from the Greek word παν, as a prefix meaning "all" or "every", is logical when the study is an attempt to define all the genes or other genomic content within a group of individuals, usually a species, a set of related species, or in microbial studies even a diverse population of organisms. However, the term is increasingly being used to describe what would traditionally be considered comparative genomics or even the generation of genomic data from multiple individuals, without comparison or an attempt to define core and variable gene sets.

The terms pantranscriptome and panepigenome have also started to infiltrate scientific literature, sometimes with no clear definition or consideration of what "all" could mean in these examples. While it is possible to capture almost all genes and variable genomics regions in a pangenome, attempts to catalog all transcripts for an individual or group of individuals face major challenges due to the very low expression, and environment, tissue or developmental stage-specific expression of some genes. While transcript atlases are valuable resources in helping to determine the potential function of genes, describing them as pantranscriptomes has the potential to generate confusion, as pan in this case could be an attempt to classify all transcripts in an individual as well as all transcripts in a species. Pantranscriptomes, generated from individuals or across species may be better described as gene expression or transcriptome atlases given the second and tertiary information (tissue, environment and individual) associated with the data. Similarly, the cell and environment-specific variation in epigenetic marks makes capturing this data in its completeness an almost impossibility. If terminology is to be used to clarify rather than obfuscate, then the extended use of the pan prefix should be avoided and more precise terms, such as comparative epigenomics or epigenome atlas, should be applied.

In one of the early plant pangenome studies, Li et al.[2] describes the comparison of whole genome assemblies from seven individuals, and while it did find differences in genome composition, it did not attempt to describe all genome variation in soybean. A more comprehensive pangenome was constructed from low coverage whole genome sequencing for 1483 rice cultivars[3], assembling reads that did not match the reference to discover 1913 and 1120 novel genes in *indica* and *japonica* rice, though the low coverage prevented the allocation of variable genes to each individual. One of the first papers to predict the gene content of a plant species applied the iterative assembly approach, where reads from multiple individuals are mapped to a reference and unmapped reads assembled to identify genomic regions absent from the reference. Golicz et al.[4] suggested that they assembled nearly all genes for *Brassica oleracea*, though the predictions were conservative as very similar genes were classified by single representatives. The iterative mapping approach was developed at a time when whole genome DNA sequence data was still very expensive and computational resources more limited than today, and the method was unable to differentiate between genes with very high sequence identity such as recent tandem duplications that are common in some genomes. However, on a pragmatic level, the iterative mapping approach defines novel genes based on read mapping, supporting downstream applications such as SNP discovery and trait association that also often rely on read mapping. The approach is also independent of variation in genome assembly and annotation that may impact whole-genome comparison or gene clustering-based comparative analysis, and remains a valid approach, particularly in species that have limited genomic resources or for large populations, ideally in combination with graph pangenomes[5].

There has been a rapid growth and evolution of pangenome construction methods in recent years with improvements derived from advances in DNA sequencing, particularly the increased accuracy of long-read sequencing and the development of graph-based pangenome construction methods[6]. The construction of high-quality reference assemblies using long-read sequence data permits detailed genomic comparisons and the construction of pangenome graphs with greater accuracy than previous short-read genome assemblies. The development of these graph pangenomes provides the ability to distinguish between very similar genes and even tandem duplications that are especially common in some genomes and would previously collapse during pangenome assembly. This has highlighted the presence of a much larger number of variable genes in species than previously considered. However, graph-based pangenome methods are still in their infancy and may fail to capture the full spectrum of structural variation within a population, potentially leading to errors in downstream analyses. Careful evaluation and parameter optimization are therefore necessary, particularly for large and/or complex

genomes. Nevertheless, the improved contiguity and completeness of genic information provided by high-quality genome assemblies now enable more accurate comparison of gene composition between individuals, and pangenome graphs provide a superior reference for trait association compared to single-reference assemblies.

One approach that has been applied for the construction of gene-based pangenomes is clustering predicted coding sequences using tools such as CD-HIT, orthoMCL and OrthoFinder[7–9]. However, these tools were not designed for pangenome construction and instead aim to produce clusters of similar genes derived from related species for comparative gene family analysis. Clustering has been applied as a step towards gene-based pangenome construction using tools such as Roary[10]. While crude clustering approaches (CD-hit, orthoMCL) are intrinsically problematic, tools that incorporate phylogenetic and synteny information when clustering genes can produce more informative clusters. However, the results vary, even for relatively small bacterial genomes, and are highly contingent on the clustering method, parameters applied and even computation resource allocation[11,12], with the selection of a representative sequence from each cluster introducing further bias. Even with the limitations of the approach, clustering-based analysis of genomes are a valuable tool for comparative genome analysis, particularly where data or computing resources are limited or where there has been major structural rearrangements between individuals that limit the benefit of synteny-based orthologue resolution, and in some cases they may outperform graph-based methods for complex genomes with limited synteny. The final approach selected should be driven by the end use of the analysis, with gene clustering providing information on orthologue presence/absence between individuals, whereas graph pangenomes, even with their current limitations, capture more heritability in trait association studies compared to single references.

As pangenome methods continue to develop, care should be taken in the approach used, with the methods selected based on the final application of the pangenome, the variation between individuals and the type and quality of the data generated. Microbes, plants and vertebrates demonstrate differences in genome structure and within population variation that may also suggest differences in analysis approaches. However, a common thread should be the robustness of the analysis to ensure that subsequent applications of the pangenomes are valid and reproducible.

David Edwards
Centre for Applied Bioinformatics and School of Biological Sciences, University of Western Australia, Perth, WA, Australia.
e-mail: dave.edwards@uwa.edu.au

Received: 4 August 2025; Accepted: 12 January 2026;
Published online: 19 January 2026

References

1. Tettelin, H. et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial pan-genome. *Proc. Natl. Acad. Sci.* **102**, 13950–13955 (2005).
2. Li, Y.-H. et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052 (2014).
3. Yao, W. et al. Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol.* **16**, 187 (2015).
4. Golicz, A. A. et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* **7**, 13390 (2016).
5. Hu, H. et al. Plant pangenomics, current practice and future direction. *Agric. Commun.* **2**, 100039 (2024).
6. Hu, H., Li, R., Zhao, J., Batley, J. & Edwards, D. Technological development and advances for constructing and analyzing plant pangenomes. *Genome Biol. Evol.* **16**, evae081 (2024).
7. Li, L., Stoeckert, C. J. Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178–2189 (2003).
8. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
9. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
10. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
11. Manzano-Morales, S., Liu, Y., González-Bodí, S., Huerta-Cepas, J. & Iranzo, J. Comparison of gene clustering criteria reveals intrinsic uncertainty in pangenome analyses. *Genome Biol.* **24**, 250 (2023).
12. Dimonaco, N. J. PyamilySeq: Exposing the fragility of conventional gene (re)clustering and pangenomic inference methods. *NAR Genom. Bioinform.* **8**, lqaf198 (2026).

Author contributions

D.E. wrote the article.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to David Edwards.

Peer review information *Nature Communications* thanks Zhangjun Fei and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.