



POPULAR ARTICLE

Beyond the reference genome how pangenomics is accelerating plant breeding?

T. Vijay¹, A.L. Chavan², M.S. Hanif³, N. Asif⁴, R. G. Roy⁵

1 Department of Genetics and Plant Breeding, Junagadh Agricultural University, Junagadh 362001, Gujarat, India

2 Department of Genetics and Plant Breeding, Keladi Shivappa Nayaka University of Agricultural and Horticultural Sciences, Shivamogga 577204, Karnataka, India

3 Centre for Agricultural Resources Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Shijiazhuang 050021, China

4 College of Fisheries and Life Sciences, Shanghai Ocean University, Shanghai 201306, China

5 Department of Genetics and Plant Breeding, Tanta University, Sri Ganganagar 335002, Rajasthan, India

Corresponding authors email: thakurdr72@gmail.com

Manuscript received: August 4, 2026; Decision on manuscript, August 30, 2026; Manuscript accepted: September 10, 2026

Introduction

When the first crop reference genomes were published rice in 2002, maize in 2009, the full bread wheat genome not until 2018, given its sheer size each one was treated as a kind of master blueprint for the species. Every subsequent study aligned its data back to that single reference sequence, as if one individual's DNA could stand in for an entire species' genetic diversity. It turned out to be a useful fiction, but a fiction nonetheless. A single reference genome, however carefully assembled, captures only the DNA of one plant. Any gene, structural variant, or stretch of sequence absent from that particular individual simply does not exist as far as the reference is concerned even if it is common, and agronomically important, in the wider gene pool. Pangenomics arose directly out of this realization, and over the last few years it has moved from a niche methodological

debate to a tool reshaping how breeders think about genetic diversity (Jayakodi, *et al.*, 2025).

The blind spot of a single reference

Imagine trying to describe an entire language using only one person's vocabulary. You would capture the common words, certainly, but you would miss regional dialects, technical jargon, and slang all real, all in use, just not spoken by your one reference speaker. A single reference genome has the same blind spot. Genes involved in disease resistance, stress tolerance and specialty quality traits are disproportionately likely to be "dispensable" present in some individuals or populations and absent in others precisely because these are the genes shaped by local adaptation and selection pressure. Studies across many crops have found that a substantial share of species genes in some cases a third or more



fall into this dispensable category, missing from any single reference assembly (He, *et al.*, 2025). For a breeder searching for a resistance gene that simply does not exist in

the one individual chosen as the reference, conventional reference-based analysis will never find it, no matter how deeply the population is sequenced.

Fig 1: Representation of the pangenome

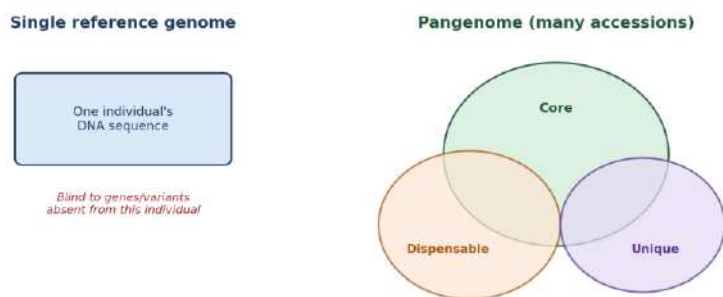


Figure 1. A single reference genome captures only one individual's DNA. A pangenome adds the dispensable and unique sequence found across many accessions - often where disease-resistance and stress-tolerance genes hide.

What a pangenome actually is?

A pangenome is built by sequencing and assembling the genomes of many individuals from across a species and often its wild relatives too rather than relying on one. The resulting resource is typically described in three layers: a core genome of genes and sequences present in essentially all individuals; a dispensable (or variable) genome present in some individuals but not others; and a unique genome found in only a single accession studied so far. Together, these layers form a far more complete catalogue of a species' genetic potential than any single assembly could offer. Some research groups have taken the idea a step further with "super-pangenomes," which extend the concept across an entire genus, including wild relatives, capturing genetic variation that domestication bottlenecks may have narrowed out of cultivated germplasm entirely (He, *et al.*, 2025).

The technology behind the shift

Pangenomics depends on assembling many

complete, high-quality genomes rather than simply calling variants against one reference a much heavier computational and sequencing lift than conventional resequencing. Several technological threads have converged to make this practical. Long-read sequencing platforms, producing reads of many kilobases, resolve repetitive regions and structural variants that short-read sequencing routinely misses or misassembles. Graph-based pangenome tools represent the pangenome not as a single linear sequence but as a graph, where different accessions' variants form alternate paths a structural shift that avoids the reference bias baked into linear alignment. Falling sequencing costs have made it economically realistic to sequence dozens or hundreds of individuals from a single species to genome-assembly quality, rather than just one. And multi-omics integration, layering transcriptomic and other functional data onto pangenome assemblies, is helping researchers move from simply cataloguing variation to understanding what that variation actually does (Pangenomics combined with artificial intelligence and precision breeding, Current Opinion in Plant Biology, 2025).

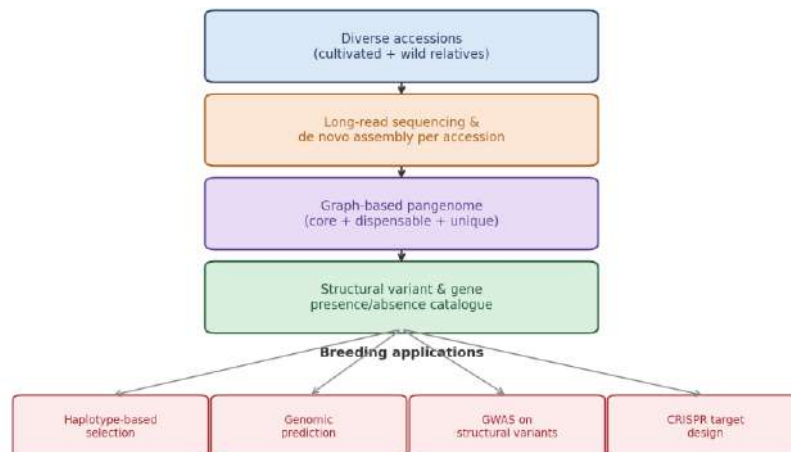


Figure 2. From diverse accessions to a graph-based pangenome, feeding directly into breeding decisions that a single reference genome could not support.

What pangenomes are revealing that references miss?

The payoff of this heavier investment has been substantial. A 2025 pangenome study in rice assembled 145 genomes spanning both wild *Oryza rufipogon* and cultivated *Oryza sativa* accessions, uncovering nearly four gigabases of sequence entirely absent from the standard Nipponbare reference, along with roughly fourteen thousand genes specific to the wild populations sampled many involved in disease resistance that had been narrowed out of the cultivated gene pool during domestication (Guo *et al.*, 2025). Tens of thousands of structural variants, including presence/absence variation and inversions, also distinguished indica and japonica rice subspecies in that same study, shedding light on how each subpopulation adapted to its own agricultural history. Similar patterns a wealth of structural variation, much of it linked to agronomically important traits, sitting outside the reference have now been documented across a wide range of crops, from cereals to cucurbits, reinforcing the idea that pangenomes reveal hidden structural variation underlying trait diversity that reference-based genome-wide association studies were simply never built to see (Frontiers editorial on structural variants in

plant pangenomics, 2025; cucurbit pangenome review, 2025).

From catalogue to crop improvement

Cataloguing variation is only useful if it feeds back into breeding decisions, and pangenomics is increasingly doing exactly that, in several concrete ways. Haplotype-based selection lets breeders track entire haplotype blocks sets of variants inherited together rather than single markers, giving a richer and more predictive unit of selection. Incorporating pangenome-derived structural variants alongside conventional SNPs has been shown to improve the accuracy of genomic prediction models, particularly for traits where the causal variation lies in presence/absence variants invisible to standard SNP arrays (Current Opinion in Plant Biology review, 2025). Knowing the full variant landscape at a locus also lets researchers design sharper genome-editing targets, accounting for variation across germplasm rather than targeting a sequence that might not even be present in the line being edited. Pangenome-wide association studies can implicate structural variants and dispensable genes that conventional genome-wide association studies, anchored to a single reference, would never see. And pangenomics



are revealing how transposable elements have reshaped genome structure across accessions, downstream effects on gene expression and trait variation (Jayakodi, *et al.*, 2025). For orphan and underutilized crops many of which never had the R and D investment that produced deep reference resources for maize or rice, pangenomics is proving especially valuable, offering a faster route to characterizing genetic diversity in species that are otherwise data-poor (Hu *et al.*, 2025).

Challenges that remain

Pangenomics is not without its growing pains. Assembling and comparing dozens of complete genomes generates enormous datasets that strain storage, computation, and the bioinformatics skills of many breeding programs, particularly in the public sector and in developing countries growing crops that matter locally but attract less global research investment. Graph-based representations, while more accurate, are also more complex to analyze than a simple linear reference, and standardized formats and tools are still maturing (Hu, Scheben, Wang, Li, Batley and Edwards, 2024). Deciding how many and which accessions to include in a pangenome enough to capture real diversity without ballooning costs indefinitely remains an open question for every new species tackled.

Looking forward

The trajectory is nonetheless clear. As sequencing costs keep falling and graph-based computational tools mature, pangenomics is moving from a resource available to only a handful of well-funded model crops toward something achievable for a much wider range of species, including many with real regional and smallholder importance. Combined with artificial intelligence for variant interpretation

and with genomic selection for prediction, pangenomes are shifting plant breeding's starting point from "what does this one reference genome say?" to "what does the full diversity of this species make possible?" (Pangenomics combined with artificial intelligence and precision breeding, 2025). That shift matters. Every dispensable gene sitting outside a reference genome is a trait breeders could not previously see, let alone select for. Pangenomics is, in a very real sense, handing plant breeding back the diversity that a single reference genome had quietly hidden from view.

References

1. Guo, D., Li, Y., Lu, H., Zhao, Y., Kurata, N., Wei, X., et al. 2025. A pangenome reference of wild and cultivated rice. *Nature*, 642(8068): 662–671. doi:10.1038/s41586-025-08883-6.
2. He, W., Li, X., Qian, Q. and Shang, L. 2025. The developments and prospects of plant super-pangenomes: Demands, approaches, and applications. *Plant Commun.*, 6(2): 101230. doi:10.1016/j.xplc.2024.101230.
3. Hu, H., Scheben, A., Wang, J., Li, F., Li, C., Edwards, D. and Zhao, J. 2024. Unravelling inversions: Technological advances, challenges, and potential impact on crop breeding. *Plant Biotechnol. J.*, 22(3): 544–554. doi:10.1111/pbi.14224.
4. Hu, H., Zhao, J., Thomas, W. J. W., Batley, J. and Edwards, D. 2025. The role of pangenomics in orphan crop improvement. *Nat. Commun.*, 16(1): 118. doi:10.1038/s41467-024-55260-4.
5. Jayakodi, M., Shim, H. and Mascher, M. 2025. What are we learning from plant pangenomes? *Annu. Rev. Plant Biol.*, 76(1): 663–686. doi:10.1146/annurev-arplant-090823-015358.