Pan-Genomes: Why One Reference Genome Is No Longer Enough for Crop Improvement

Akash M Hosamani[1,2]*, Kavitha C[3,4] and Amasiddha Bellundagi[5]

[1]*Division of Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012, India*
[2]*The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi-110012, India*
[3]*Research Scholar, Department of Food Microbiology and Safety, CFT, VNMKV, Parbhani, Maharashtra, India*
[4]*Assistant Professor, Department of Food Processing Technology, CCSc, UAS Dharwad, Karnataka, India*
[5]*ICAR-Indian Institute of Millets Research (IIMR), Hyderabad-500030, India*

Corresponding Author
Akash M Hosamani
Email: akashmhosamani@gmail.com

OPEN ACCESS

Keywords

Pan-genome, Structural variation, Graph reference genome, Presence-absence variation, Crop improvement

How to cite this article:

Hosamani, A. M., Kavitha, C. and Bellundagi, A. 2026. Pan-Genomes: Why One Reference Genome Is No Longer Enough for Crop Improvement. *Vigyan Varta* 7 (09): 189-192.

ABSTRACT

For two decades crop genomics has run on a single reference genome per species, one plant standing in for millions. Long-read sequencing has shown that this convenient assumption discards a large share of a crop's genetic variation. A pan-genome, the combined gene content of many individuals of a species, reveals that only a fraction of genes occurs in every plant while the rest come and go between varieties: in potato, just a quarter of 51,401 pan-gene clusters are present in all 45 accessions examined. These presence-absence and structural variants underlie disease resistance, stress tolerance, yield and quality traits, yet they are largely invisible to a single linear reference. Graph pan-genomes, which hold every version of a region in one searchable structure, recover much of this hidden variation; in tomato they raised the average estimated heritability of 20,323 traits from 0.33 to 0.41. This article explains what a pan-genome is, how a graph reference differs from a linear one, and what it offers plant breeders working on crops from rice to chickpea.

INTRODUCTION:

A reference genome is a consensus sequence assembled from a single individual of a species: Nipponbare for rice, Heinz 1706 for tomato, DM for potato. That one sequence became the coordinate system for almost everything that followed, including gene annotation, SNP discovery, marker-assisted selection and genome-wide association studies. It was a practical triumph and, for a long time, the only affordable option.

It also carried a hidden assumption, namely that one plant's gene content fairly represents the whole species. Cheap long-read sequencing now lets researchers assemble dozens of genomes per crop, and the assumption has not survived. Two plants of the same crop differ not only at single bases but in entire genes present in one and absent in the other. The combined gene content of many individuals is called a pan-genome, and it is fast becoming the working reference for crop improvement.

ONE PLANT, ONE COORDINATE SYSTEM

In a routine genotyping experiment, short reads from breeding lines are aligned to the reference and the resulting variants become the markers used for mapping and selection. Any sequence absent from the reference plant has no coordinates to align to; its reads either fail to map or map spuriously elsewhere, and are then filtered out as noise. The variation is not simply missed, because the pipeline is structurally incapable of seeing it. If that blind spot holds a resistance gene or a transporter controlling a trait of interest, the breeder never learns it exists. The class of variation involved, presence-absence variation and structural variation more broadly, covering insertions, deletions, inversions, translocations and copy-number change, is both abundant and functionally important: the potato pan-genome catalogued 561,433 high-confidence structural variants, among them a 5.8 Mb inversion associated with carotenoid content (Tang *et al*., 2022), while the pearl millet graph pan-genome catalogued 424,085 structural variants (Yan *et al*., 2023).

CORE, SHELL AND PRIVATE GENES

A pan-genome is built by annotating many individuals, clustering their genes into families, and asking in how many individuals each family occurs. Core genes are present in every accession and typically carry housekeeping and essential developmental functions; soft-core genes occur in almost all; shell or dispensable genes in some accessions but not others; private genes in a single accession.

The proportions are the surprise. From 45 potato accessions, 51,401 pan-gene clusters were recovered, of which only 13,123 (25.5%) were present in all of them, while 5,743 (11.2%) were soft-core, 28,471 (55.4%) shell and 4,064 (7.9%) accession-specific (Tang *et al*., 2022). The rice super pan-genome, spanning 251 cultivated and wild Asian and African accessions, annotated 51,359 non-redundant genes of which 21,888 were core (Shang *et al*., 2022). Put plainly, most genes in a crop species are missing from at least one plant of that species. The size of the core also depends on the panel sampled and on each study's presence threshold, so these percentages describe the germplasm examined rather than the species outright.

How large the core is varies widely between crops and studies. Across 26 maize founder lines, 103,033 pan-genes were annotated of which 32,052 were core or near-core (Hufford *et al*., 2021), whereas a pan-genome of 89 pigeon pea accessions from India and the

Philippines contained 55,512 genes of which 48,067 were core, with the variable genes enriched for disease-response functions (Zhao *et al*., 2020). Part of that spread is biological and part methodological, since each study fixes its own presence threshold and panel breadth.

FROM A LINE TO A GRAPH

If genes come and go, the reference should not be a single string of letters. The graph pan-genome is now the standard answer. Shared segments form nodes on a common path, and wherever accessions differ the graph branches into alternative routes and rejoins afterwards, forming a bubble. A deletion, insertion, inversion or copy-number change becomes an explicit, addressable alternative rather than an absence. An individual is then genotyped by the route it takes through the graph, so reads from a gene missing in the old reference finally have somewhere to align, and its presence or absence becomes a scoreable marker like any SNP.

WHAT IT DELIVERS FOR BREEDING

Recovering missing heritability. In tomato, a graph pan-genome built from 838 genomes, including 32 new reference-quality assemblies, catalogued more than 19 million variants. Across 20,323 gene-expression and metabolite traits the average estimated heritability rose from 0.33 with a single linear reference to 0.41 with the graph, a 24% gain attributed largely to structural variants resolving incomplete linkage disequilibrium, and two new genes potentially contributing to soluble solid content were identified (Zhou *et al*., 2022).

Finding trait genes a linear reference hides. The rice super pan-genome uncovered grain-weight-associated structural variants that act by altering the expression of nearby genes, together with variants linked to submergence tolerance, seed shattering and plant architecture (Shang *et al*., 2022). The barley pangenome, from 76 long-read genomes, yielded novel powdery mildew resistance alleles, copy-number variation at a branching regulator and an expansion of starch-cleaving enzyme genes in malting barleys (Jayakodi *et al*., 2024), and multiple wheat genomes resolved rearrangements and wild-relative introgressions, including the Sm1 insect-resistance region (Walkowiak *et al*., 2020).

Triaging germplasm. The chickpea variation map, based on sequencing 3,366 genomes (3,171 cultivated and 195 wild accessions), identified superior haplotypes for improvement-related traits in landraces that can be introgressed into elite lines, and located deleterious mutations in elite germplasm that could be purged by genomics-assisted breeding or gene editing (Varshney *et al*., 2021). For a breeder this is a catalogue indicating which gene-bank accession carries the allele worth crossing in. In soybean, structural variants from a graph pan-genome were genotyped across 2,898 accessions and linked to candidate genes for agronomically important traits (Liu *et al*., 2020).

WHERE INDIAN CROPS STAND

The pulses and millets central to Indian agriculture are already covered. Pigeon pea has a pan-genome with 225 trait-associated SNPs across nine agronomic traits (Zhao *et al*., 2020); chickpea has the 3,366-genome variation map with its landrace haplotypes (Varshney *et al*., 2021); and in pearl millet, pangenomic analysis tied an expansion of RWP-RK transcription factors and endoplasmic-reticulum-related genes to heat tolerance, a trait that matters more with every passing summer (Yan *et al*., 2023). For Indian breeding programmes the resources exist; the bottleneck is now trained people who can query them.

LIMITATIONS

Pan-genomes are not free. They require many high-quality, ideally chromosome-scale assemblies and consistent annotation, since presence-absence calls are only as reliable as the gene models behind them. Graph-based tools are younger and less standardised than the linear alignment pipelines most laboratories already run, and they demand more computing and bioinformatics skill. A catalogued structural variant also remains a hypothesis until field phenotyping and multi-location testing confirm that it behaves as the association suggests.

CONCLUSION

The single reference genome was the scaffold that let crop genomics begin, and it is now the limiting factor. Pan-genomes, and graph references in particular, change the question from how a line differs from one reference plant to which version of each region that line carries, which is much closer to what a breeder needs to know. As sequencing costs fall further, the expectation for a major crop will not be a reference genome but a reference population, updated as new germplasm is assembled and queried routinely from breeding programmes. For students entering crop bioinformatics, fluency with pan-genome and graph-based methods is moving from a specialist skill to a basic requirement.

REFERENCES

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K., & Dawe, R. K. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, *373* (6555), 655–662.

Jayakodi, M., Lu, Q., Pidon, H., Rabanus-Wallace, M. T., & Bayer, M. (2024). Structural variation in the pangenome of wild and domesticated barley. *Nature*, *636*(8043), 654–662.

Liu, Y., Du, H., Li, P., Shen, Y., & Peng, H. (2020). Pan-genome of wild and cultivated soybeans. *Cell*, *182*(1), 162–176.

Shang, L., Li, X., He, H., Yang, Q., & Wang, W. (2022). A super pan-genomic landscape of rice. *Cell Research*, *32*(10), 878–896.

Tang, D., Jia, Y., Zhang, J., Li, H., & Cheng, L. (2022). Genome evolution and diversity of wild and cultivated potatoes. *Nature*, *606*(7914), 535–541.

Varshney, R. K., Roorkiwal, M., Sun, S., Bajaj, P., & Chitikineni, A. (2021). A chickpea genetic variation map based on the sequencing of 3,366 genomes. *Nature*, *599*(7886), 622–627.

Walkowiak, S., Gao, L., Monat, C., Haberer, G., & Kassa, M. T. (2020). Multiple wheat genomes reveal global variation in modern breeding. *Nature*, *588*(7837), 277–283.

Yan, H., Sun, M., Zhang, Z., Jin, Y., & Zhang, A. (2023). Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. *Nature Genetics*, *55*(3), 507–518.

Zhao, J., Bayer, P. E., Ruperao, P., Saxena, R. K., & Khan, A. W. (2020). Trait associations in the pangenome of pigeon pea *(Cajanus cajan)*. *Plant Biotechnology Journal*, *18*(9), 1946–1954.

Zhou, Y., Zhang, Z., Bao, Z., Li, H., & Lyu, Y. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, *606*(7914), 527–534.