

Feature Review

Open Access

Construction and Analysis of a Rice Pan-genome Reveals Structural Variation Hotspots Across Subspecies

Weijie Sun, Yaodong Liu ✉

Modern Agricultural Research Center, Cuixi Academy of Biotechnology, Zhuji, 311800, Zhejiang, China

✉ Corresponding email: yaodong.liu@cuixi.orgRice Genomics and Genetics, 2025, Vol.16, No.3 doi: [10.5376/rgg.2025.16.0015](https://doi.org/10.5376/rgg.2025.16.0015)

Received: 30 Apr., 2025

Accepted: 10 Jun., 2025

Published: 28 Jun., 2025

Copyright © 2025 Sun and Liu, This is an open access article published under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preferred citation for this article:

Sun W.J., and Liu Y.D., 2025, Construction and analysis of a rice pan-genome reveals structural variation hotspots across subspecies, Rice Genomics and Genetics, 16(3): 159-179 (doi: [10.5376/rgg.2025.16.0015](https://doi.org/10.5376/rgg.2025.16.0015))

Abstract Rice (*Oryza sativa*) is a staple cereal with immense global importance, yet a single reference genome cannot capture the full genetic diversity underlying key traits. Pan-genomics has emerged as a paradigm to characterize the “pan-genome”—the total genomic repertoire of a species—including core genes shared by all accessions and dispensable genes present in some but absent in others. Here, this study reviews the construction and analysis of rice pan-genomes and the insights they provide into structural variation hotspots across rice subspecies; outlines how the limitations of a single reference genome have driven the development of plant pan-genomics, enabling discovery of extensive genomic variation that was previously hidden; describes strategies for building rice pan-genomes, from early short-read sequencing approaches to recent long-read assemblies and graph-based genome models that integrate diverse accessions (*indica*, *japonica*, *aus*, *aromatic*, wild relatives). Major types of structural variation—insertions, deletions, inversions, translocations, copy number variations—are defined, and this study surveys computational tools for their detection; synthesizes findings on the distribution of structural variants (SVs) in the rice genome and identifies hotspots of variation specific to certain lineages. The functional impact of SVs is discussed, with case studies linking structural variants to agronomic traits (yield, stress tolerance, flowering time) and to gene presence/absence variation affecting gene families (e.g. disease resistance genes). Comparative pan-genome analyses across rice subspecies illuminate how evolutionary forces like domestication bottlenecks, introgression, and selection have shaped genomic differences between *indica* and *japonica* rice. Finally, this study highlights emerging applications of rice pan-genome research in germplasm utilization, genome-wide association studies, marker-assisted breeding, and de novo domestication, and discusses future prospects and challenges in integrating multi-omics data and developing pan-genomic resources for sustainable agriculture under climate change.

Keywords Rice pan-genome; Structural variation; Genetic diversity; Subspecies; Crop genomics

1 Introduction

Rice is a staple crop for over half the global population, making it central to food security. Advances in rice genomics have brought big improvements in traits like yield, quality, and disease resistance (Shang et al., 2022). One major milestone was the release of the first high-quality genome sequence of Asian rice, which opened the door to better gene discovery and breeding. But rice isn't a single, uniform crop. *Oryza sativa* includes two main subspecies—*indica* and *japonica*—as well as many subgroups adapted to different climates and growing conditions. Around the world, gene banks have collected more than 780 000 rice accessions. This huge range of genetic material holds the key to traits that can help rice cope with pests, poor soil, and climate stress. However, early studies focused heavily on a few elite varieties like Nipponbare (a *japonica* type). While these gave researchers a good foundation, it quickly became clear that one variety's genome doesn't tell the whole story—some important genes simply aren't present in every type. That realization shifted the focus toward pan-genomics, which looks at many genomes together. By comparing multiple cultivated and wild types, researchers are now finding new genes and structural differences that were missed before (Zhao et al., 2018). This broader view is helping scientists create better rice varieties faster—something that really matters in today's changing agricultural landscape.

Traditional genome analysis approaches have largely depended on a single reference genome per species. In rice, the use of one reference (e.g. Nipponbare) for read alignment and gene annotation has inherent limitations. A single reference genome does not represent the genetic diversity within the species, as evidenced by the high

degree of variation observed among different rice accessions. Many DNA sequences that are present in some rice varieties are absent or highly divergent in the reference genome. As a result, reference-biased analyses fail to detect structural variations or novel genes that do not align to the reference (Qin et al., 2021). For example, if a gene is missing in the reference but present in an *indica* landrace, short-read sequencing of that landrace would yield reads that go unmapped, leading to the gene's omission from analysis. This bias can skew SNP discovery, gene presence/absence calls, and trait mapping. Furthermore, single-reference approaches complicate the detection of large structural variants (SVs). Short-read sequencing data aligned to a reference often misses insertions, deletions, or rearrangements larger than the typical read length. Such SVs are a substantial source of genetic and phenotypic variation in rice populations (Vahedi et al., 2023). In summary, the single-reference paradigm inherently overlooks "hidden" variation outside the reference gene set. These limitations motivated the development of pan-genome strategies that integrate multiple genomes to more fully capture rice's genomic diversity.

The concept of a pan-genome—the complete set of genes or sequences present in all members of a species—was first introduced in bacterial genomics in 2005. In the ensuing years, pan-genomics has rapidly expanded to plant and animal studies. A plant pan-genome is typically composed of a core genome (genes present in every individual of the species) and a dispensable (or variable) genome (genes present in some individuals but missing in others). Early pan-genome analyses in plants began to reveal that a considerable portion of any given species' gene repertoire is dispensable. This was a paradigm shift from the assumption that a single reference could adequately represent a species. In rice, pan-genomic research gained momentum in the last decade as sequencing costs dropped and computational methods improved (Liu et al., 2021). Initial efforts involved comparing a few divergent cultivars, demonstrating that each new genome contributed novel genes absent from the reference. As more genomes were added, it became evident that the rice pan-genome is vast and still growing. Concurrently, pan-genome studies in other major crops (e.g. maize, soybean, brassicas) were yielding similar insights, underscoring that pan-genomics had "come of age" as a powerful framework in plant biology. The development of pan-genomics has thus been driven by the need to systematically catalog genetic variation at the species level. It represents a natural evolution of genomics from single-reference assemblies to comprehensive, population-scale genome resources.

This study provides a comprehensive overview of rice pan-genome research and to analyze how structural variations are distributed across different subspecies of rice; covers the concept and evolution of plant pan-genomes, emphasizing why they are necessary and how they have been applied in crop species; then focuses on strategies for constructing rice pan-genomes, including sequencing technologies and assembly approaches that enable integration of diverse rice genomes (*indica*, *japonica*, *aus*, *aromatic* and wild relatives). A major emphasis is placed on structural variations (SVs) uncovered by rice pan-genomes—their types, detection, distribution, and hotspots—and the functional implications of these SVs for agronomic traits and gene content. This study also presents comparative analyses that shed light on evolutionary divergence between rice subspecies. Four case studies illustrate key achievements in rice pan-genome research. Finally, this study discusses practical applications in breeding as well as current challenges and future perspectives for pan-genomic approaches in rice improvement and sustainable agriculture.

2 The Concept and Evolution of Plant Pan-genomes

2.1 Definition and structure of pan-genomes

A pan-genome encompasses the complete set of genomic elements (genes, regulatory sequences, structural variants, etc.) present in a species' population. It is typically defined in terms of two components: the core genome and the variable (or dispensable) genome. The core genome consists of genes found in all individuals of the species, representing functions that are presumably essential or ubiquitous. In contrast, the variable genome comprises genes or sequences that are present in some individuals and absent in others. These variable genes can include those gained or lost during evolution, often imparting specialized traits (for example, adaptation to specific stresses or environments) (Bayer et al., 2020). The pan-genome can be thought of as a union of all genes across all genomes of the species.

To build a pan-genome, scientists' sequence multiple individuals and bring all their genetic data together. This helps uncover genes that aren't in the usual reference genome—genes that might otherwise go unnoticed. Studies in rice and other crops have shown that using only one reference leaves out important parts of the genetic puzzle. By adding those missing pieces, the pan-genome gives a clearer picture of a species' full genetic makeup. What stands out is that many of these variable genes are linked to how plants handle stress from their surroundings. They may not be essential all the time, but they give plants the flexibility to adapt. That's what makes the pan-genome such a powerful tool for exploring biodiversity.

2.2 Core genome vs. dispensable genome

In plant genomes, including rice, the core genome comprises the genes that are consistently present in every accession examined. These genes typically encode fundamental cellular and developmental functions. The dispensable genome (also called the accessory genome) contains genes that are missing in one or more accessions. Dispensable genes often relate to environmental interactions, such as disease resistance, stress tolerance, or secondary metabolism, which may not be needed under all conditions. The balance between core and dispensable content can be quantified as more genomes are sequenced.

Early pan-genome research on Asian rice showed that the shared set of genes—the so-called core genome—makes up just over half of all genes found across different varieties. For instance, a study of 453 rice genomes identified about 12 770 gene families (roughly 53.5%) that were common to all samples. The rest varied from one variety to another. Similarly, the 3 000 Rice Genomes Project found more than 19 000 genes that weren't in the reference genome but were present in at least one of the varieties studied (Wang et al., 2023). This shows that rice genomes can differ a lot depending on the variety. Many of these variable genes appear in only certain groups and often reflect the plant's evolutionary history or responses to environmental pressures. These genes aren't just extras—they can be key to specific traits like disease resistance that help a plant thrive in particular conditions. Understanding which genes are common to all rice types and which are more specialized is crucial. It helps researchers see both the essentials for rice survival and the sources of diversity that allow different varieties to adapt.

2.3 Development and application of pan-genomes in major crops

Pan-genomic approaches have been successfully applied to many major crop species, leading to key biological and practical insights (Shi et al., 2022). In tomato (*Solanum lycopersicum*), a pan-genome analysis of 725 accessions uncovered dozens of novel genes, including a rare allele for fruit flavor that had been lost during domestication. In maize (*Zea mays*), initial pan-genome studies revealed that a significant fraction of genes is not shared among all lines, explaining heterosis and trait variation in hybrids. For soybean (*Glycine max*), pan-genomics helped identify structural variations and presence/absence variants linked to seed composition and stress responses. In wheat (*Triticum aestivum*)—a complex hexaploid crop—assembly of multiple genomes provided a wheat pan-genome that captures global variation from modern breeding; this work identified genomic regions and genes affected by selection in different breeding programs. Similarly, a pan-genome of barley (*Hordeum vulgare*) revealed “hidden” structural variants accumulated through decades of mutation breeding, some of which underlie agronomic traits.

Pan-genomes aren't just theoretical—they're useful tools for crop improvement. In rapeseed, comparing different ecotypes revealed genetic differences between spring and winter varieties, especially in traits like flowering time and glucosinolate levels. These insights help breeders target beneficial genes. Pan-genome data also support better genotyping tools, including presence/absence markers, which boost the accuracy of GWAS and genomic predictions in crops like rice and maize. Building crop pan-genomes has opened the door to understanding more of the genetic variation that exists—and has given breeders stronger resources to work with across many species.

2.4 Advances in computational methods and sequencing technologies

In recent years, plant pan-genomics has made great progress, largely thanks to advances in DNA sequencing and data analysis. Early studies depended on short-read methods like Illumina, which were precise but often missed big structural changes or repetitive regions in the genome. That changed with the arrival of long-read technologies

like PacBio and Oxford Nanopore. These tools read much longer stretches of DNA, helping researchers' piece together more complete genomes. They've made it possible to detect large insertions and deletions that used to go unnoticed. A good example is rice-scientists used long-read data to assemble over 30 genomes and found many hidden structural differences (Shang et al., 2022). Newer techniques, like HiFi reads and linked-read sequencing, along with scaffolding tools such as Hi-C, have made it much easier to build high-quality reference genomes for numerous plant lines. A standout example is the 251-genome rice pan-genome, which was built using high-coverage Nanopore sequencing and Hi-C, resulting in very large, high-contiguity assemblies.

Computational tools now let us compare multiple genomes more efficiently. Programs like MUMmer and Minimap2 are commonly used to spot structural differences and figure out which genes are missing or present. Recently, instead of using one fixed genome as a reference, researchers have started working with graph-based genome models. These graphs show the diversity across many genomes in a single structure. With tools like the VG toolkit, it's easier to find structural variants in plants. This shift is changing the way we study crop genomes-especially in species where genetic diversity plays a big role in breeding and adaptation.

3 Construction Strategies for Rice Pan-genomes

3.1 Rice as a model crop: diversity and genome complexity

Rice has long been a model for cereal genomics due to its relatively small genome (~390 Mb for the *japonica* subspecies) and its enormous agricultural importance. Despite its compact genome (especially compared to polyploid crops like wheat), rice exhibits tremendous genetic diversity. Asian cultivated rice (*Oryza sativa*) was domesticated from the wild progenitor *O. rufipogon* and consists of two major subspecies: *indica* (also called *O. sativa* subsp. *indica*) and *japonica* (*O. sativa* subsp. *japonica*). These subspecies further subdivide into distinct genetic subpopulations (such as *aus*, *aromatic/basmati*, *tropical japonica*, *temperate japonica*, etc.), which differ in their geographic origins and traits. *Indica* rices are typically grown in tropical regions and have broad genetic variation, whereas *japonica* rices are adapted to temperate climates and tend to be more genetically uniform. The divergence and partial reproductive isolation between *indica* and *japonica* (likely thousands of years ago) created deep structural variations and allelic differences between their gene pools. For instance, certain genomic segments are known to cause hybrid sterility when *indica* and *japonica* varieties are crossed, reflecting accumulated structural incompatibilities (Figure 1) (Wu et al., 2023).

Beyond *O. sativa*, the genus *Oryza* contains over 20 wild species with AA, BB, CC (and other) genome types. Some wild relatives (like *O. nivara* and *O. rufipogon*) readily cross with *O. sativa* and have contributed alleles for traits such as disease resistance and flood tolerance in breeding programs. The African cultivated rice (*O. glaberrima*) is another domesticated species, independently domesticated in West Africa, with a separate but overlapping gene pool. This rich tapestry of subspecies and wild species makes rice an ideal candidate for pan-genome analysis-the goal is to capture the full diversity from domestication, varietal group differentiation, and introgression from wild gene pools. Characterizing this diversity at the whole-genome level is essential to unlock novel alleles for crop improvement.

3.2 Sequencing strategies (e.g., short reads, long reads, HiFi, Hi-C)

Early rice pan-genome research mostly relied on short-read sequencing because it was cheap and got the job done. The 3 000 Rice Genomes Project is a classic example-researchers mapped short reads to a reference genome to find genetic differences. This method worked well for spotting small variations but wasn't great for assembling new genomes or detecting big structural changes. In recent years, long-read sequencing has stepped in to fill that gap. Technologies like PacBio's CLR and HiFi, as well as Oxford Nanopore, are better at reading complex or repetitive regions. One project used these tools to assemble genomes for 32 *O. sativa* and one *O. glaberrima*, showing a much more diverse pan-genome than expected. Another study used Nanopore to sequence over 250 rice genomes-both wild and cultivated-at high depth, creating a valuable resource for future work.

Combining long reads with scaffolding technologies greatly improves assembly contiguity. Hi-C sequencing, which captures information about chromatin contacts, has been used to order and orient contigs into

chromosome-level assemblies for each rice genome. This approach ensures that each assembled genome is of reference quality, facilitating accurate detection of large inversions or translocations. In addition, optical mapping and Bionano genome maps have sometimes been employed to resolve complex repeats or structural ambiguities. An emerging strategy is to use hybrid assembly pipelines that integrate Illumina, PacBio/Nanopore, and Hi-C data to maximize accuracy. In summary, the state-of-the-art in rice pan-genome sequencing is to generate high-depth long reads for each accession, polish the assemblies with short reads, and scaffold them with Hi-C, yielding multiple high-quality genomes for comparative analysis.

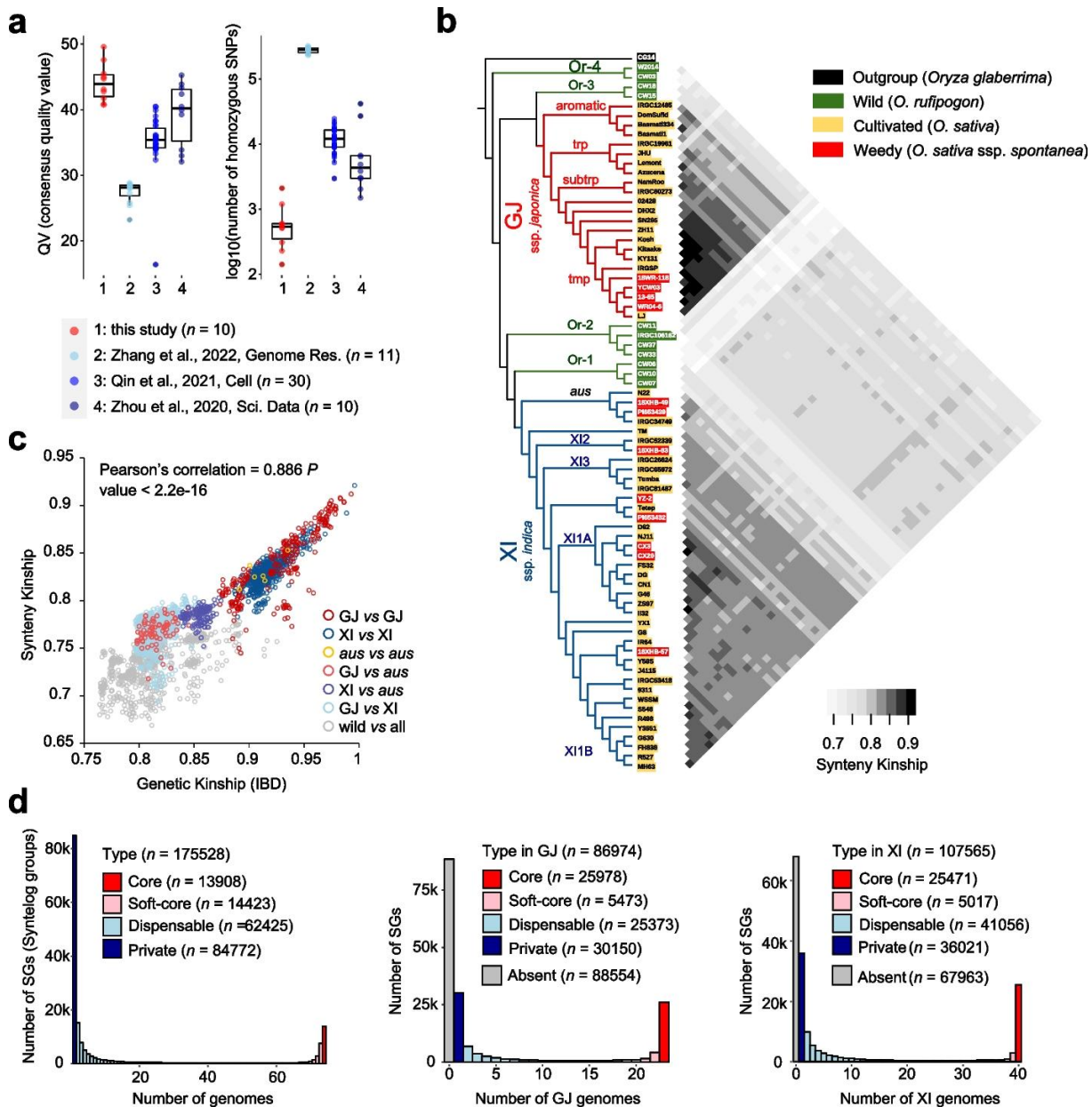


Figure 1 A syntelog-based pangenome of rice. a Consensus quality values and number of homozygous SNPs based on self-mapping for genomes in four pan-genome projects. b Phylogeny of rice genomes used in this study and their pairwise synteny kinship. c The relationship among rice genomes measured by genetic distance (IBD, identity-by-descent) based on SNPs and synteny based on gene orders. d Number of syntelog groups (SGs) represented in all 74 rice genomes versus the number of genomes. The subspecies pan-genome compositions for GJ and XI were extracted from the whole rice pan-genome. The numbers of core (present in the genomes of all accessions), soft-core (present in the genomes of > 90% accessions), dispensable (present in the genomes of less than 90% of all but at least two accessions), private (only present in a single genome), and absent SGs (identified by using all populations but absent in sub-population) were counted (Adopted from Wu et al., 2023)

3.3 Assembly pipelines and reference graph models

There are two primary strategies for constructing a rice pan-genome from multiple genomes: the de novo assembly approach and the reference-guided (iterative) approach. In the de novo approach, one generates an independent whole-genome assembly for each selected accession (using methods outlined above), and then aligns or merges these assemblies to identify the union of genomic sequences. Pipeline tools such as MUMmer4 or minimap2 can align assemblies to the reference, revealing insertions (segments present in a new assembly but not in reference) and other structural differences. The union of all unique sequences across the assemblies forms the pan-genome. For example, using 12 high-quality rice genomes as a starting point, researchers constructed a non-redundant sequence collection that served as a pan-genome reference. Subsequent diverse accessions can then be mapped onto this pan-genome reference to identify additional variants.

In contrast, the reference-guided iterative assembly approach starts with a reference genome and incrementally incorporates sequences from short reads of other accessions. Zhao et al. (2018) followed this strategy by taking 66 rice accessions (53 cultivated and 13 wild) and iteratively assembling contigs from unmapped reads, thereby building a composite reference that captured novel sequences absent in the original reference. This “map-to-pan” strategy was effective with short reads, though it can miss complex rearrangements. More recently, graph-based genome models have been introduced to represent a pan-genome. In a graph model, nodes represent genomic sequences (from any accession) and edges represent connections, allowing multiple alternative allelic sequences to coexist in one representation. Rice researchers have developed graph-based pan-genomes where all detected structural variants are embedded in a genome graph. This enables joint genotyping of variants across hundreds of accessions using tools like VG or PanGenie. Whether using aligned assemblies or genome graphs, the result is a pan-genome reference that replaces a single linear reference, providing a more comprehensive coordinate system to map reads and genetic data from diverse rice lines.

3.4 Inclusion of different rice subspecies (*indica*, *japonica*, *aus*, *aromatic*, *wild*)

When building a rice pan-genome, it's important to choose samples that truly represent the genetic variety of rice. The two main types—*indica* and *japonica*—must both be included because they each have unique genes. For example, popular varieties like IR64 (*indica*) and Nipponbare (*japonica*) can differ by millions of SNPs. Other groups, such as *aus* (a distinct *indica*-related type from South Asia) and aromatic types like Basmati, also carry special traits. One well-known example is the fragrance gene in Basmati, which isn't found in standard *indica* or *japonica*. To capture this kind of genetic diversity, researchers often include accessions from all these groups. That's exactly what was done in the 3K Rice Genomes project, which selected a wide mix: *indica*, *aus*, *tropical* and *temperate japonica*, and *aromatic* types. This kind of careful sampling helps ensure the pan-genome reflects the full range of rice diversity.

In addition, wild rice species have been incorporated to build an extended pan-genome (sometimes termed a “super-pangenome” when spanning multiple species). *O. rufipogon* (the Asian wild progenitor) contributes alleles that were lost or rare in cultivated rice, and including dozens of *O. rufipogon* genomes revealed thousands of wild-specific genes (Guo et al., 2025). For example, a recent study built a pan-genome from 129 wild *O. rufipogon* accessions and 16 cultivars, uncovering ~13 728 genes present only in wild rice and absent from domesticated rice. Other wild relatives like *O. nivara* or African *O. barthii* have further expanded the gene pool. By integrating *indica*, *japonica*, *aus*, *aromatic*, African rice, and wild *Oryza* species, researchers ensure that the pan-genome represents the full spectrum of rice genomic variation. This inclusive approach has highlighted, for instance, the much higher abundance of disease resistance genes in wild rice compared to cultivars (Guo et al., 2025), underlining the value of wild germplasm in enriching the pan-genome for crop improvement.

4 Structural Variations in the Rice Pan-genome

4.1 Types of structural variations: insertions, deletions, inversions, translocations, CNVs

Structural variations (SVs) are genomic differences that involve segments of DNA larger than about 50 base pairs. They contrast with single nucleotide polymorphisms (SNPs) and include a variety of mutation types. The major categories of SVs in rice and other organisms are:

- Insertions: segments of DNA present in one genome that are absent in another (often called presence/absence variants when comparing against a reference). An insertion could range from a few base pairs (e.g., transposon insertion) to thousands of base pairs harboring one or more genes.
- Deletions: the opposite of insertions, where a segment found in the reference is missing in another genome. Large deletions can knock out genes or regulatory regions, sometimes with phenotypic consequences. (Insertions and deletions are together often termed “indels,” especially when smaller than a few kb).
- Inversions: DNA segments that flip their direction without changing the actual gene content. Still, they may disrupt gene order and interfere with how genes are regulated or recombined.
- Translocations: happen when a piece of DNA shifts to a new location-either within the same chromosome or across different ones. In rice, such swaps between subspecies are rare, but a few have been found between wild and cultivated varieties.
- Copy Number Variations (CNVs): segments that are present in multiple copies (duplications) or have reduced copy (including complete deletion) in one genome relative to another. CNVs can involve gene duplications or deletions and are an important source of dosage variation for genes.

Presence/absence variants (PAVs)-essentially large insertions/deletions encompassing whole genes-are a particularly important class of SV in pan-genomes. These PAVs lead to genes that are entirely missing from some genomes but present in others, contributing heavily to the dispensable genome component. Overall, structural variants account for a substantial proportion of genetic differences among rice accessions and often have larger effect sizes on phenotype than SNPs, given that they can disrupt or duplicate entire genes.

4.2 Computational tools and pipelines for SV detection

To find structural differences in rice genomes, one practical way is to compare whole genome assemblies. When you have good-quality genome data from different rice varieties, you can line them up and spot the parts that don't match. Tools like MUMmer or nucmer help compare each genome to a standard one like Nipponbare, showing where pieces are missing or extra. This method has revealed thousands of insertions and deletions in past studies. It can also uncover more complex changes, like when genome sections are flipped or moved by checking for disrupted alignment patterns.

Another approach is read-based SV calling, which operates on sequencing reads mapped to a reference genome. Traditional SV callers for short reads (e.g., Pindel, DELLY, LUMPY) use patterns such as discordant read pairs or split reads to infer deletions, inversions, or duplications. However, short reads have limited power for complex or repetitive SVs. The advent of long reads improved this: long-read SV callers (such as *Sniffles* and PBHoney) can directly map long reads and identify SV signatures with higher sensitivity and specificity. Sedlazeck et al. (2018) introduced Sniffles, which leverages PacBio reads to accurately detect complex SVs, demonstrating far more insertions/deletions in a human genome than previously known. In rice, long-read data from multiple varieties have been similarly processed to identify tens of thousands of SVs that short-read analyses missed.

Beyond individual discovery, genotyping SVs across populations is crucial. Tools like *SVType* and graph-based genotypers use known SV coordinates to screen other accessions for presence/absence. The incorporation of SVs into a pan-genome graph enables mapping short reads from many accessions onto all alternate alleles, facilitating efficient genotyping. For example, Hickey et al. (2020) showed that the VG toolkit could genotype thousands of SVs in large panels using a variation graph. These tools and pipelines, combined in workflows, form the backbone of rice pan-genome projects, allowing researchers to systematically detect and compare SVs across hundreds or thousands of genomes.

4.3 Distribution and hotspots of SVs across rice subspecies

Structural variation in the rice genome isn't random. Studies of rice pan-genomes show that certain regions-especially those rich in transposable elements (TEs) or segmental duplications-tend to gather more structural changes (Lu et al., 2021). One common example is the pericentromeric area, where repetitive DNA is dense and insertions or deletions are more frequent. When researchers scan the genome using sliding windows, they often find that regions with more TEs also contain more novel insertions. This suggests TEs may be actively driving variation (Li et al., 2021). For instance, if a TE inserts itself into one rice variety but not another, that insertion appears as unique to that line. Over time, these insertions cluster in TE-rich zones, forming SV hotspots.

Additionally, different rice subspecies show lineage-specific SV patterns. Certain large structural variants have become fixed in *indica* but are absent in *japonica*, and vice versa. For instance, an extensive deletion overlapping the flowering regulatory gene *DTH8* is present only in *indica* populations, whereas a copy-number variant at the grain length gene *GL7* is found only in *japonica*. These variants mark subspecies differentiation and often map to QTLs underlying phenotypic differences (such as grain shape or flowering time) between *indica* and *japonica*. Some regions of the genome, such as those harboring clusters of NBS-LRR disease resistance genes, are particularly prone to presence/absence variation and show dramatic structural divergence among subpopulations. By contrast, other regions-often those with housekeeping genes-remain structurally conserved. Using pan-genome data, researchers have even constructed "inversion indexes" to identify large inversions distinguishing subpopulations. In one study, 1 769 non-redundant inversions (≥ 100 bp) were catalogued across Asian rice, collectively spanning ~29% of the reference genome sequence. These inversions and other SV hotspots provide clues to rice evolutionary history and may underlie some adaptation traits unique to specific lineages.

4.4 Relationship between SVs and genetic diversity

Structural variants (SVs) have a big impact on the genetic makeup and diversity of rice. Compared to single nucleotide polymorphisms (SNPs), SVs can remove or insert entire genes, revealing variations that SNPs might miss (Zhao et al., 2018). I find it especially intriguing how certain gene sequences show up in one group of rice but not in another, helping researchers tell subpopulations apart. For example, pan-genome studies show that some SVs are unique to groups like aus or aromatic rice, making them reliable indicators of those genetic lineages. When comparing wild and cultivated rice, more than 13 800 presence/absence variants have been identified that help distinguish between *indica* and *japonica* types. These differences highlight their separate domestication paths, including the genetic bottlenecks they each experienced. *Japonica* rice seems to have gone through a narrower bottleneck, which means it holds fewer rare variants compared to *indica*.

SVs also drive diversity by introducing new functions. Some rice types have genes that others don't-often due to gene duplication or gene flow (Li et al., 2021). These extra genes may help plants resist disease or adapt to tough growing conditions. SVs can even affect nearby regions. Large inversions, for instance, can reduce recombination, which causes more genetic differences to build up between populations. In short, SVs aren't just background noise in the genome. They often hold clues to traits we care about and help explain how rice varieties have developed over time. That's why newer studies tend to analyze SVs alongside SNPs to better understand rice evolution and diversity.

5 Functional Implications of SVs in Rice

5.1 Structural variations and agronomic traits (e.g., yield, stress tolerance, flowering time)

Structural variations in the rice genome can have major effects on agronomic traits by altering gene function or regulation. One clear example is a tandem duplication at the *GL7* locus in some *japonica* rice, which was shown to increase grain length and improve grain appearance; this ~17-kb duplication (absent in *indica*) boosts the expression of a positive regulator of grain length, resulting in longer grains. Another classic case involves the *Sub1A* gene for submergence tolerance: tolerant rice varieties (e.g., the landrace FR13A) have an extra copy of the *Sub1A* transcription factor gene (through duplication) that is not present in most intolerant varieties-this structural variation confers the ability to withstand prolonged flooding. Similarly, large insertions or deletions in promoter or

enhancer regions can modulate trait expression. For instance, the rice pan-genome analysis by Gao et al. (2025) identified structural variants associated with grain weight differences: one variety group had an ~126 bp insertion upstream of the gene *SHAT1*, altering its expression and contributing to different grain shattering tendencies (Shang et al., 2022). Independent selection of such upstream insertions in Asian vs. African rice suggests their role in domestication of seed non-shattering.

Structural variants can also be linked to flowering time. A notable deletion in the flowering repressor gene *DTH8* (also known as *Ghd8*) is present in some *indica* cultivars but not in *japonica*, leading to earlier flowering under long-day conditions; breeders have utilized this deletion allele for adapting rice to temperate climates. Furthermore, presence/absence of entire gene clusters, such as photoperiod-sensitivity genes or hormone regulators, can create qualitative trait differences. A striking example outside of cultivated rice is the *SNORKEL1* and *SNORKEL2* genes from deepwater wild rice: these genes (encoding ethylene response factors) are absent in most cultivars, but when present as an introgression, they allow rice plants to rapidly elongate internodes under flooding. This trait is vital for survival in deepwater environments. Overall, structural variation provides a rich source of phenotypic variation in rice-by changing gene dosage, creating novel chimeric genes, or modifying gene regulation, SVs underlie many quantitative trait loci (QTLs) that breeders have historically selected for improved yield, plant architecture, or stress responses.

5.2 SVs affecting gene presence/absence and gene family evolution

One of the most significant consequences of structural variation in a pan-genomic context is gene presence/absence variation (PAV). Because of large insertions and deletions, some genes are completely missing in certain rice genomes while present in others. This leads to differences in gene content that can influence phenotypes and adaptation. Many of these PAVs involve genes in multigene families that are known to evolve rapidly. For example, plant disease resistance genes (notably the NBS-LRR class) often occur in clusters that are subject to duplication and deletion. As a result, any given rice variety typically has a unique repertoire of NBS-LRR genes-some genes in these families are present in one cultivar but absent in another (Shang et al., 2022). A species-wide study in the model plant *Arabidopsis thaliana* found dozens of immune receptor genes present only in certain accessions, revealing extreme cases of presence/absence variation in a plant's immune gene repertoire. Similarly in rice, pan-genome analysis has identified entire clusters of defense-related genes that are part of the dispensable genome. These PAVs drive gene family evolution: through duplication (a type of CNV), new gene copies can arise and diverge in function; conversely, through deletion, some lineages lose certain gene family members.

An illustrative case is the *LRR-RLK* gene family (Leucine-Rich Repeat Receptor-Like Kinases) involved in pathogen recognition-different rice subpopulations have gained or lost specific members of this family via structural variations over time, reflecting local adaptation to pathogen pressure (Zhao et al., 2018). Presence/absence polymorphisms also extend to genes controlling metabolic profiles, such as fragrance: the well-known fragrance allele in aromatic rice is caused by an 8 bp deletion in the *BADH2* gene (technically a small indel), which in homozygous form leads to fragrance. While a small mutation, it underscores how loss-of-function via deletion can disseminate as a favorable trait. On a larger scale, the rice pan-genome has revealed that each new genome added contributes on average several hundred “novel” genes that were absent from prior references (Gao et al., 2025). Many of these novel genes belong to expanded gene families or are duplicated copies that acquired new functions. Thus, structural variation and PAV are intimately linked with gene family evolution, constantly shaping and reshaping the gene content of rice populations.

5.3 Case studies of SVs linked to domestication and adaptation

Structural variants have played conspicuous roles in rice domestication-the process by which wild rice was transformed into cultivated rice-and in subsequent adaptation to diverse agricultural environments. One hallmark of domestication in cereals is loss of seed shattering (to facilitate harvesting). In rice, a key domestication gene is *SHATTERING 1* (*SHAT1*), a regulator of seed detachment. A comparative analysis between Asian and African rice

revealed different structural variants upstream of *SHAT1* were selected in each domestication event. Asian cultivated rice (both *indica* and *japonica*) carries a specific ~126 bp insertion about 4.5 kb upstream of *SHAT1*, while African cultivated rice evolved a different ~70 bp insertion ~3.5 kb upstream of the orthologous gene (Shang et al., 2022). These insertions, which likely alter regulatory elements, contributed to the loss of shattering in domesticated rice—a prime example of parallel domestication using distinct SVs.

Another case involves adaptation to flooding. Deepwater rice varieties, which grow in flood-prone areas, evolved the ability to elongate when submerged. The discovery of the *SNORKEL1/2* genes (absent in standard cultivars) in certain traditional rice indicated that introgression of these genes (a structural gain) gave plants a novel adaptation—a dramatic elongation response to rising water. This structural gain from wild rice was likely selected by farmers in regions with deep floods. Domestication also often entailed selective sweeps around beneficial structural variants. For instance, a well-known quantitative trait locus for rice plant architecture, *PROG1*, underwent a causative 2-bp deletion in coding sequence (not a large SV, but a mutational variant) that changed plant growth from prostrate (in wild rice) to erect (in domesticated rice). While that example is a small indel, structural changes such as copy-number variation at the *Fragrance* gene (*badh2*) locus or deletion of dormancy genes have been identified as contributors to domestication syndrome traits (loss of seed dormancy, white pericarp, etc.).

Looking at rice evolution, researchers have found that certain structural changes in the genome—like insertions, deletions, or duplications—are closely tied to domestication. These changes often show up repeatedly in breeding programs, suggesting they've been under strong selection pressure. For example, removing part of a gene that controls starch breakdown helps rice sprout better in flooded fields, which is great for direct seeding. Whether adapting rice for upland or lowland fields, or tropical versus temperate regions, structural variations—like SNPs—have clearly played a major role in shaping the crop we rely on today.

5.4 Regulatory roles of SVs in gene expression networks

Structural variants can influence gene expression in myriad ways, acting as a form of structural regulation. Inversions, for example, can disrupt the local synteny and place genes in new regulatory contexts or prevent their recombination with certain regulatory alleles. If a transcription factor gene is caught in an inversion, its expression might change due to altered chromatin environment or the breaking of linkage with distant enhancers. In rice, large inversions identified in the pan-genome have been correlated with expression differences for genes inside the inverted regions. This suggests some inversions may underlie eQTL (expression quantitative trait loci) by modifying how genes contact enhancers or insulators in the 3D genome architecture.

Copy number variations (duplications) can directly alter gene expression levels by gene dosage. A rice variety carrying a duplication of a transcription factor gene will often express that factor at higher levels, potentially amplifying its downstream effects (Qin et al., 2021). A case in point is the green revolution semi-dwarf gene *Sd1* (though the classic semi-dwarf allele is a loss-of-function SNP, one can imagine a duplication of a growth repressor gene leading to a similar dwarf phenotype via increased dosage). CNVs can also foster neofunctionalization: duplicated gene copies may diverge—one maintaining the original function, another taking on a new expression pattern—thereby rewiring networks.

Another regulatory impact of SVs comes from transposable element (TE) insertions in regulatory regions. Rice genomes contain many TEs, and new TE insertions can bring regulatory motifs that either enhance or repress nearby gene expression. For example, if a TE carrying a strong promoter inserts 5' of a gene, it can cause overexpression of that gene. In maize, the famous *tb1* gene controlling apical dominance was upregulated by an adjacent transposon insertion; analogous phenomena are likely present in rice as well (such as hopping of MITEs near stress-responsive genes contributing stress inducibility). In pan-genome data, many structural insertions are in non-coding regions and may represent novel regulatory sequences. A recent study demonstrated that presence/absence of a ~366 bp promoter insertion in foxtail millet (*SiGW3* gene) led to expression variation and differences in grain weight—a concept transferable to rice, where promoter indels can modulate traits like grain size

and panicle architecture. In conclusion, structural variations often act as large-effect cis-regulatory mutations: by changing gene copy number, disrupting chromosomal context, or introducing new regulatory DNA, SVs rewire gene expression networks in ways that can produce significant phenotypic outcomes.

6 Comparative Pan-genome Analysis Across Rice Subspecies

6.1 Core and variable genome components across subspecies

Comparative pan-genome analysis allows us to quantify how much of the rice genome is shared between subspecies (core) and how much is unique to each lineage (variable). When examining *indica* and *japonica* rice, researchers find a substantial core genome common to both, alongside significant subspecies-specific content. For example, a recent pan-genome reference constructed from both Asian and African rice reported about 28 900 core genes that were found in all accessions examined (Guo et al., 2025). Beyond this core, thousands of genes were restricted to particular groups—around 10 101 genes were specific to Asian rice (*indica/japonica*) and not present in African rice, and conversely about 1 259 genes appeared unique to African rice. Focusing within *O. sativa*, *indica* and *japonica* share the majority of their genes, but each has lost or gained some genes relative to the other. The 3K pan-genome analysis, for instance, showed that *japonica* varieties lack some genes that are present in *indica* (and vice versa), which correlates with their independent domestication bottlenecks and breeding histories.

The aus subpopulation (considered an early-diverging group of *indica*) and aromatic group (Basmati-type) also contribute unique genes to the pan-genome. *Aus* rice was found to carry novel alleles and even novel genes not observed in mainstream *indica*, reflecting its separate evolutionary trajectory. Meanwhile, aromatic rices, which include the basmati and sadri types, have some genomic segments more similar to *japonica* (due to historical introgression) but also unique content. Wild relatives dramatically expand the variable genome: incorporating *O. rufipogon* adds many wild-specific genes (e.g., ~13k genes unique to wild in one study), most of which are absent from all cultivated lines. These wild-specific genes include those related to stress tolerances or life history traits (like seed shattering or perennial growth) that were not retained during domestication. In summary, the core genome forms the backbone of rice's biological functions, while the variable genome—differing across subspecies and populations—provides the genomic basis for the diverse phenotypes and local adaptations observed in rice.

6.2 SV enrichment in lineage-specific regions

When comparing genomes of different rice subspecies, certain genomic regions emerge as enriched for structural variation unique to one lineage. These often correspond to regions that experienced differential selection or drift after the *indica-japonica* split. For example, genomic regions around known reproductive isolation genes show lineage-specific SVs. The S5 locus on chromosome 6, implicated in hybrid sterility between *indica* and *japonica*, contains an inversion and a small deletion in *indica* relative to *japonica* that cause meiotic failure in hybrids. Such structural differences act as barriers maintaining lineage separation. More broadly, a pan-genome inversion analysis identified nearly 1 769 inversions across Asian rice, some of which are fixed in one subpopulation but absent in others. These inversions are essentially lineage-specific markers; they can suppress recombination locally, contributing to subpopulation structure.

Researchers have also observed that some chromosomes harbor clusters of subspecies-specific SVs. A striking example is on chromosome 11 and 12 where *indica* carries private large deletions affecting seed dormancy and shattering genes, whereas *japonica* carries ancestral versions of those loci. In the rice super-pan-genome, an analysis of SV hotspot distribution found that certain deletions or copy-number changes were exclusively present in the *indica* subgroup (denoted *O. sativa indica*, Osi) but completely absent in the *japonica* subgroup (Osj), and vice versa. For instance, Shang et al. (2022) noted a large deletion in the flowering gene *DTH8* only in *indica* accessions, and a copy-number expansion of the grain length gene *GL7* only in *japonica* accessions. These lineage-specific SVs often align with QTLs that differ between subspecies—e.g., *DTH8* deletion contributes to early flowering in some *indica*, *GL7* duplication contributes to longer grains in *japonica*. Hence, structural variation is not evenly spread: it tends to cluster in genomic neighborhoods that have undergone separate evolutionary paths in *indica* versus *japonica* (or other subgroups), representing the genetic signatures of lineage divergence.

6.3 Evolutionary forces shaping the rice pan-genome

Several evolutionary forces have shaped the composition and variation of the rice pan-genome. Natural selection during domestication and diversification has had a profound impact. Domestication imposed bottlenecks and strong selection for certain traits, which in turn fixed some structural variants and eliminated others. For instance, as rice was domesticated from *O. rufipogon*, alleles conferring non-shattering, reduced dormancy, and erect growth were favored—many of these traits, as discussed, involved structural changes such as insertions or deletions at key loci. This directional selection reduced diversity around those loci in cultivated rice relative to wild rice, effectively making wild-specific haplotypes (and their SVs) disappear from the cultivated gene pool. The relatively larger bottleneck in *japonica* (a narrower genetic base) compared to *indica* means *japonica* lost more genomic variation (including SVs) during domestication (Guo et al., 2025). This is reflected in the pan-genome by fewer private alleles and SVs in *japonica* and a smaller effective pan-gene set for *japonica* alone, as opposed to *indica* which retained or acquired more variation via introgression.

Speaking of introgression, gene flow between different rice populations has also shaped the pan-genome. *Indica* rice is thought to have picked up domestication alleles through hybridization with *japonica* (for example, the *sd1* dwarf allele was originally *japonica* and transferred to *indica* breeding lines). Such introgressions are structural events at the population level—large chromosomal segments (containing multiple genes and variants) moved between gene pools. Introgression with wild relatives (either deliberate in breeding or natural in sympatric growth) has introduced new structural variants into cultivated rice, such as the aforementioned *Sub1A* locus for flood tolerance from wild *O. rufipogon*.

Another force is transposable element activity and genome turnover. Rice genomes have high transposon content, and bursts of transposition can generate new insertions that differentiate lineages. Over evolutionary time, active transposons in one subpopulation but not another will lead to accumulation of private insertions, expanding the pan-genome. Similarly, gene duplication and divergence (a form of mutation and selection combined) contribute to the pan-genome as new gene copies arise (e.g., disease resistance gene duplications in response to pathogen pressure). If those duplicates confer advantage, they may be retained in some populations (under positive selection) but could be lost in others (neutral loss or lack of selection). Genetic drift, especially in small farmer-maintained landrace populations, can also result in random loss of genes or fixation of structural peculiarities in certain lineages without adaptive reason.

6.4 Insights into *indica*–*japonica* divergence and hybridization

Through comparing pan-genomes, we can better understand how *indica* and *japonica* rice diverged and how hybridization played a role in shaping their genomes. Genomic studies show that both types of rice—*indica* and *japonica*—likely came from the same original domestication event of *Oryza rufipogon*. *Indica* seems to have formed later, by mixing with domesticated *japonica*. The pan-genome supports this single-origin idea by showing that key domestication genes—like those affecting seed shattering, pericarp color, and plant structure—are mostly the same in both types (Lu et al., 2022). This suggests that *japonica* passed these useful traits to early *indica* through hybridization. For instance, both types carry the non-shattering *sh4* allele and the white-pericarp *Rc* mutation, which are nearly fixed. These shared traits point to a common history, rather than separate domestication paths. So, rather than being independently domesticated, *indica* likely inherited key domestication traits from *japonica* or a related early domesticate.

However, after their split, *indica* and *japonica* pursued largely separate evolutionary paths, accumulating distinct sets of structural variants. *Indica*–*japonica* comparative analyses show about 13 853 presence/absence variants differentiating the two groups, many of which can be traced back to divergence between their wild progenitor populations or post-domestication selection (Kou et al., 2025). These include differences such as the aforementioned DTH8 deletion (*indica*-specific) and GL7 duplication (*japonica*-specific), as well as numerous small indels in regulatory regions. Hybrid sterility loci like *S5* on chromosome 6 illustrate how certain SVs reinforce divergence: *S5* involves a gene complex where *indica* and *japonica* have incompatible alleles (including

a minor indel in one of the HSA genes), preventing full fertility in hybrids. Pan-genome analysis has pinpointed such incompatibility regions and even guided the discovery of “neutral” alleles (so-called wide-compatibility genes) that breeders use to enable *indica-japonica* crosses.

Some rice varieties clearly show traces of hybridization in their genomes. For example, aromatic or Basmati rice is a genetic blend of both *indica* and *japonica* types. Data from the rice pan-genome reveals that while aromatic rice mostly carries *indica* background, it also contains sections inherited from *japonica* and has its own unique gene variations. These special genetic combinations-including chromosomal rearrangements and introgressed blocks-reflect how human breeding created a distinct subgroup. In recent breeding work, parts of the *japonica* genome have been introduced into *indica* varieties and vice versa, forming genomes that are a patchwork of both. This kind of mosaic structure can be difficult to spot using a single reference genome. But with a pan-genome approach, we can track where each DNA segment comes from by looking at variations and haplotypes tied to specific lineages. In short, the rice pan-genome maps out both ancient crossings and modern breeding efforts that have shaped today’s diverse rice types.

7 Case Studies in Rice Pan-genome Research

7.1 Case 1: the 3 000 rice genomes project

The 3,000 Rice Genomes Project (3K RGP), published in 2018, was one of the first major efforts to explore rice pan-genomics at scale. It focused on sequencing over 3,000 rice varieties from Asia and Africa using low-coverage, short-read methods. Although these genomes weren’t fully assembled from scratch, the project still uncovered an incredible amount of genetic variation. Researchers identified more than 29 million SNPs and hundreds of thousands of small insertions and deletions, all by comparing the sequences to the Nipponbare reference genome. They also looked into larger structural differences by analyzing read depth and assembling unmapped reads. One of the key outcomes was a draft version of the rice pan-genome. This included core genes shared by all varieties and others that appeared in only some. About 12 000 gene families were consistently present across all accessions, while nearly half of all gene families were found to vary between varieties-a clear sign of rice’s genomic diversity.

But the project didn’t just produce data. It shed light on rice population structure, revealing nine distinct subgroups within cultivated rice. It also enabled genome-wide association studies for traits like grain size, pericarp color, and disease resistance. Importantly, the team made their findings publicly available, including a user-friendly rice pan-genome browser (RPN) for tracking gene presence or absence. Despite its limitations in capturing large structural variants, the 3K RGP paved the way for more advanced pan-genome work. It revealed how much valuable genetic information had been missed by relying on a single reference genome and underscored the importance of sequencing rice more deeply and broadly. For many researchers, it was a wake-up call-and a strong foundation for what came next.

7.2 Case 2: graph-based pan-genome

A representative case of rice pan-genome research is the work by Song et al. (2021), who used a graph-based genome approach to better understand complex traits. Their team brought together 12 different rice genomes-spanning *indica*, *japonica*, and wild types-into one variation graph. This setup made it possible to align sequencing data from over 400 rice lines to a more comprehensive reference, instead of relying on just one genome like Nipponbare. This broader reference uncovered a wider range of genetic variation, including structural changes and presence/absence variants (PAVs) that typical single-reference methods often miss. One of the most significant findings was a new QTL linked to grain weight, tied to a gene the authors called *qGW candidate*. This gene wasn’t detectable when using a standard linear reference, but it became visible within the graph-based framework.

Beyond discovering traits, this method also sharpened the accuracy of variant detection, especially in repetitive or hard-to-map regions. By offering multiple alignment paths, the graph genome reduced the bias that comes from forcing data to fit a single template. Overall, this study highlights how graph-based pan-genomes can move rice

genomics forward-not only by capturing more genetic variation, but also by turning that variation into meaningful insights. It's a good example of how using multiple genomes together can make a real difference in trait analysis, especially in crops like rice where population diversity is high.

7.3 Case 3: pan-genome study in wild rice

Wild rice carries a treasure trove of genes that modern cultivated rice has either lost or never had. A notable example comes from a 2025 study by Guo and his team (Figure 2) (Guo et al., 2025). They assembled a pan-genome from 145 high-quality rice genomes-129 from wild *Oryza rufipogon* and 16 from cultivated *O. sativa*. Compared to the traditional Nipponbare reference genome, this new dataset added 3.87 gigabases of fresh sequences, mostly from wild rice. Much of it came from repeated DNA regions like those near centromeres and telomeres, along with duplicated genes. In total, over 69 000 genes were annotated. About 29 000 were shared across all samples (core genes), while roughly 13 700 were unique to wild rice. Many of these wild-specific genes are tied to stress resistance-useful traits for dealing with pests, diseases, or harsh environments. One striking detail: wild rice contains far more disease resistance (R) genes than cultivated varieties. This suggests that domestication may have unintentionally weeded out valuable genes or failed to tap into them at all. The study also highlighted key genes linked to deep root systems and perennial growth-traits that have largely disappeared in today's annual rice crops.

Another important takeaway came from comparing the wild and cultivated genomes. The patterns of variation supported the idea that Asian rice was domesticated just once, with the two main subspecies, *indica* and *japonica*, diverging later. The team identified nearly 14 000 structural differences between these two types, helping to explain their evolutionary paths. Interestingly, *japonica* showed signs of a stronger genetic bottleneck, having lost more wild traits than *indica*. Today, this wild rice pan-genome serves as a valuable tool for breeders. It offers a treasure trove of genes-especially those for stress tolerance and disease resistance-that could be used to improve modern rice through crossbreeding or gene editing. This case clearly shows the importance of keeping wild relatives in the picture when working to expand and improve crop diversity.

7.4 Case 4: application in marker-assisted selection

One practical outcome of rice pan-genome research is the development of more effective genotyping tools. A notable example is the Rice Pan-genome Genotyping Array (RPGA), a high-density SNP array designed to detect not just common variants but also those missing from the reference genome (Nipponbare) but found in other rice varieties. Daware et al. (2023) compiled roughly 80 000 markers for this array, including probes aimed at dispensable genomic regions uncovered through pan-genome studies. Using this tool, researchers genotyped a diverse set of rice accessions. The results were quite revealing. When performing GWAS with the RPGA, they identified 42 QTLs associated with grain size and weight-8 of which had been completely missed using only the reference genome. One particularly interesting locus involved a gene absent in Nipponbare: a WD40 repeat-containing protein on chromosome 7. This gene, found only in some rice lines, was linked to longer grain length and was confirmed through QTL mapping to have a real impact on the trait.

Beyond discovery, the RPGA also proved useful for practical breeding tasks. It successfully distinguished population structures, tested hybridity in crosses, and helped build dense linkage maps. For breeders using marker-assisted selection (MAS), having markers tied to presence/absence variants is especially valuable. For example, if a beneficial gene from a wild rice donor is missing in elite lines, the RPGA can track its introgression during backcrossing. Overall, this case shows how pan-genomic knowledge can be translated into breeding tools. Instead of relying only on a single reference, breeders can now tap into a broader range of genetic diversity-including structural variants-to make more informed selection decisions and improve rice varieties more effectively.

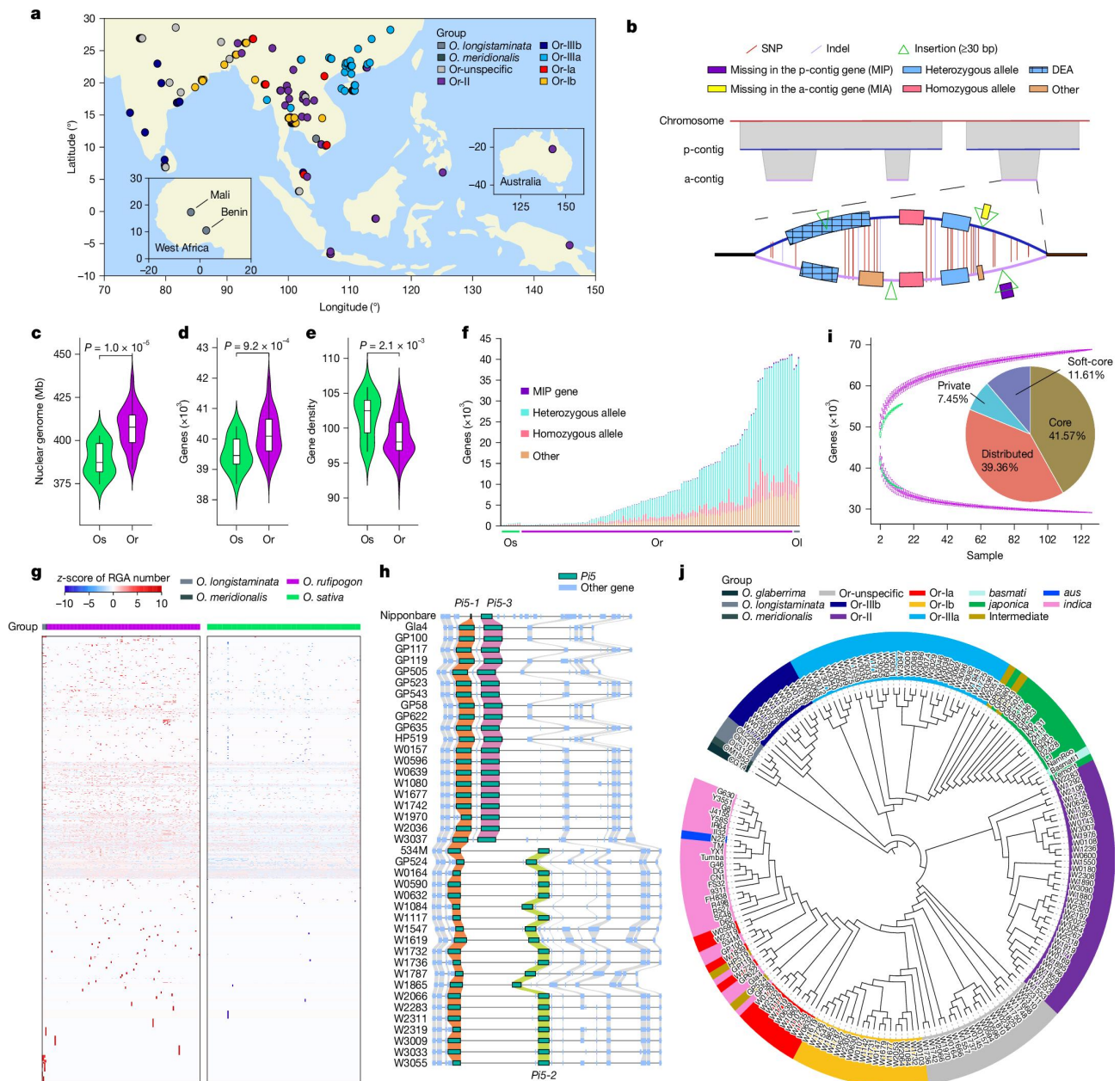


Figure 2 Pangenome analysis of 149 representative wild and cultivated rice accessions (Adopted from Guo et al., 2025)

Image captions: a, Geographical distribution of 133 wild rice accessions, with dots coloured by group. Inset maps show enlarged views of West Africa and Australia. Positive values indicate east longitude ($^{\circ}$ E) and north latitude ($^{\circ}$ N). b, Schematic representation of p-contigs and a-contigs, showing gene models and variant information from HiFi genome assemblies. c–e, Violin plots comparing *O. sativa* (Os; $n = 16$) and *O. rufipogon* (Or; $n = 129$) varieties for nuclear genome size (c), gene number (d) and gene density (gene number per Mb) (e). The box edges mark the 25% and 75% quartiles and the central line marks the median. Whiskers extend to 1.5 times the interquartile range (IQR). Statistical significance was determined by two-sided Wilcoxon test; P values are shown at the top. f, Distribution of genes on a-contigs across 133 HiFi genomes, including *O. longistaminata* (Ol). g, Heat map showing RGA loci that have a higher average copy number in wild rice than in cultivated rice, normalized using a row-wise z-score method. Each row indicates an RGA locus, and each column indicates a variety. h, Local synteny plot of *Pi5* loci (chr. 9: 9.65–9.79 Mb in the Nipponbare reference genome) from 41 accessions. *Pi5* genes (*Pi5-1*, *Pi5-2* and *Pi5-3*) are shown in green with collinearity represented by lines of different colours; adjacent genes are shown in blue with grey collinearity lines. i, Accumulation curve of the number of pan-genes (uptrend lines) and core genes (downtrend lines) across *O. sativa* (green; $n = 16$) and *O. rufipogon* (purple; $n = 129$) varieties. The pie chart shows the gene proportions in the combined wild–cultivated pangenome ($n = 145$). j, Phylogenetic tree of the 180 *Oryza* varieties, on the basis of single-copy genes. Diverse groups are denoted by colour-coded labels (Adopted from Guo et al., 2025)

8 Applications of Rice Pan-genome Research

8.1 Germplasm utilization and genome-wide association studies (GWAS)

Rice pan-genome resources greatly enhance the utilization of germplasm collections and improve the power of genome-wide association studies. Traditional GWAS in rice often struggled with “missing heritability,” partly because they ignored structural variants and presence/absence variation. By providing a more complete set of markers, including those from non-reference sequence, pan-genome-based GWAS can capture previously untagged genetic effects. As described in Case 4, using a pan-genome genotyping array led to the discovery of multiple novel QTLs for grain traits that single-reference SNP chips failed to detect. Similarly, researchers performing GWAS with a graph-based pan-genome approach were able to pinpoint trait-associated structural variants (like an insertion affecting plant height) that were hidden to linear reference analysis. Thus, integrating pan-genomic variants into GWAS increases QTL detection power and can explain additional phenotypic variance (closing some of the “missing heritability” gap) (Yang et al., 2025).

Beyond GWAS, pan-genomes improve germplasm characterization. For instance, breeders and gene bank managers can use pan-genome data to more thoroughly genotype diverse landraces and wild accessions. Each accession can be characterized not just by SNPs relative to Nipponbare, but by its unique gene content. This helps identify accessions carrying novel genes of interest—for example, a particular landrace might be the only one (among those sequenced) harboring a wild-derived disease resistance gene. Such information directs breeders to germplasm that should be tapped for specific traits. In practice, national and international gene banks are beginning to integrate pan-genomic markers into their characterization protocols. The International Rice Research Institute (IRRI), for example, now has access to variations identified by 3K RGP and subsequent pan-genome studies to guide germplasm mining.

Pan-genome data makes it possible to bring together results from different studies, even when they used different reference genomes or analysis methods. For example, a recent meta-GWAS on rice used a graph-based pan-genome to combine six separate datasets. This led to the identification of 156 QTLs related to traits like yield-116 of which weren’t found in individual studies. Many of these were linked to structural variations or presence/absence markers only visible through the pan-genome.

8.2 Marker development and genomic selection

The rise of rice pan-genome studies has really changed how we develop molecular markers for breeding. Instead of relying only on SNPs, breeders now have access to a broader range of genetic variations—like structural variants and new genes tied to important traits (Daware et al., 2022). These can be turned into practical tools, such as PCR markers. Take disease resistance, for instance: if researchers find that deleting a specific gene makes a plant resistant to rice blast, it’s straightforward to design a simple InDel marker to test whether breeding lines carry that deletion. This is better than using a nearby SNP because it targets the actual cause of resistance.

Genomic selection isn’t just riding on the success of marker-assisted breeding—it’s growing with it. With the rise of pan-genome data, we now catch details that used to slip through the cracks, like gene deletions or duplications. These subtle variations matter, especially when trying to improve complex traits such as yield or stress resistance. By factoring in this broader range of genetic markers, our prediction models become more accurate and useful in real breeding scenarios. One tool I find particularly promising is the Practical Haplotype Graph (PHG). Instead of sequencing the whole genome every time—which costs a lot—we can now sample a small portion, and PHG helps fill in the blanks using a reference built from pan-genome data. It’s efficient, cost-saving, and still leverages full genetic diversity.

8.3 Implications for *de novo* domestication and gene editing

The rice pan-genome provides a blueprint for *de novo* domestication—the idea of taking wild species or unadapted germplasm and rapidly domesticating them (or improving them) using modern tools. By comparing genomes of domesticated rice with wild relatives, pan-genome studies pinpoint which genes and structural variants were critical in domestication (Shang et al., 2022). For example, pan-genome analysis confirms that wild rice contains

certain alleles (and gene presences) that make it less suitable for agriculture, such as shattering or dormancy alleles, but it also contains many beneficial alleles absent in cultivars, like stress tolerance genes. Using CRISPR/Cas9 gene editing, it is now feasible to introduce domestication-related mutations into wild rice in a designed way. The pan-genome guides this by listing all key differences: e.g., loss-of-function of *sh4* and *progl* for non-shattering and erect growth, respectively; deletion in *Rc* for white pericarp; perhaps a semi-dwarf allele in *Sd1*; and so forth. Recent work on de novo domestication of wild allotetraploid rice successfully edited a suite of genes to create a phenotype approaching cultivated rice in a single generation. The pan-genome ensures that while we are modifying those known domestication genes, we retain the wild rice's novel genetic content (such as additional disease resistance genes or high nutrient content genes) that we ultimately want in the new crop.

Pan-genomic data offers new direction for improving traits in existing rice varieties. Instead of relying only on a single reference genome, pan-genomes show a broader picture, revealing genes that may have been missed before. For example, if a wild rice species has a unique gene that helps it resist a certain disease-and this gene isn't found in any cultivated types-we could introduce it into elite lines using gene editing or transformation. In other cases, genome-wide association studies (GWAS) based on pan-genomes might find that a key promoter sequence is missing in high-yielding lines. That missing piece can then be added through precise editing to boost performance. One recent example used a graph-based pan-genome to identify two new QTLs for grain size. Researchers confirmed these by knocking out the genes with CRISPR/Cas9, showing their clear effect on grain shape. This approach-from pan-genome data to editing and trait validation-shows how structural variants can be efficiently turned into breeding targets. It speeds up the entire process of bringing useful traits from wild relatives into cultivated rice.

8.4 Integration into breeding pipelines and seed industry

As rice pan-genome data become more accessible, they are being integrated into breeding pipelines and even seed industry practices. Modern rice breeding increasingly uses decision support tools that incorporate genomic information at various stages-from parental selection to line advancement. Pan-genome databases and browsers (e.g., the RFGB-Rice Functional & Genomic Breeding platform-which includes pan-genomic info) allow breeders to check if a parent line possesses certain presence/absence alleles or structural variants of interest. For example, if a breeder wants to improve a popular rice cultivar by adding a gene that is only present in aus rice, pan-genome resources will identify which aus accessions carry that gene and what markers tag it. This guides the choice of donor parent. After making the cross, breeders can use markers (based on that gene's presence) to track the introgression in progeny, an approach that has been made more efficient with the advent of pan-genome-informed arrays (like the 80K SNP-array which captures pan-genome variation) (Daware et al., 2022). This ensures the desired genomic segment is retained while background genome is recovered.

Pan-genome research plays a key role in keeping genetic diversity broad within breeding programs. It helps breeders see how much variation in a species is actually captured by current elite lines. If essential parts of the pan-genome-such as genes from wild relatives-are missing, targeted efforts like introgression or pre-breeding can bring in useful traits, especially for stress tolerance or new disease threats. Seed companies also value pan-genomic markers for identifying and protecting their varieties. A complete set of presence/absence markers allows for high-precision genetic fingerprinting, even among closely related lines. This strengthens seed purity checks and safeguards intellectual property by linking each variety to a distinct genomic signature.

On the industry scale, pan-genome data enable the creation of customized breeding panels-subsets of diverse lines that maximize pan-genome coverage. For instance, the discovery that only a few accessions harbor a given rare allele could prompt including those accessions in a breeding consortium to ensure that allele isn't lost. In summary, integration of pan-genomes into breeding is making selection more precise and comprehensive: breeders can now select not just for known genes and SNPs, but for the presence of entire genomic regions that were previously outside their awareness. As a result, the seed industry is moving toward more data-driven breeding decisions, leveraging the full genetic potential outlined by the rice pan-genome.

9 Challenges and Future Perspectives

9.1 Technical limitations: sequencing depth, assembly errors, SV annotation

Although there has been good progress in rice pan-genome studies, several technical hurdles remain. One key issue is the depth and quality of sequencing needed to uncover rare variants. To fully detect structural variants-especially those that occur infrequently-deep long-read sequencing of many individuals is ideal. But this approach is expensive. Many projects try to balance cost by sequencing a small, diverse subset deeply and the rest at lower coverage. This helps reduce costs but may miss low-frequency SVs. In the 3K Rice Genome Project, for example, the lower coverage meant that only common variants were detected reliably, and many rare insertions or deletions were probably missed (Zhang et al., 2022). As studies expand to tens of thousands of genomes, maintaining accuracy will be tough. New algorithms that work well with low-depth data will be essential.

Another challenge is the reliability of assemblies. Even with better long-read tools, misassemblies in repetitive regions or collapsed duplications are still common. These errors can lead to false SV calls. Advances like Hi-C scaffolding and trio binning are improving things, but manual checking is still needed to reduce errors (Shang et al., 2023). Interpreting SVs is no small task. We can now identify huge numbers of them, but figuring out which ones matter biologically is tricky. SVs in coding regions are easier to assess, but many lie in regulatory or non-coding regions, where their effects are hard to predict. Plus, current formats like VCF struggle to represent complex SVs, making annotation and analysis more difficult.

Finally, pan-genome updating and maintenance is technical work: as more genomes are added, computing the “incremental” pan-genome without starting from scratch is non-trivial. Efficient algorithms are needed to merge new assemblies into existing pan-genome graphs or alignment maps. In summary, generating a flawless and exhaustive rice pan-genome is still constrained by sequencing resources, assembly/algorithmic accuracy, and bioinformatic frameworks for SV interpretation (Qin et al., 2021). Overcoming these technical limitations will be crucial to fully realize the benefits of pan-genomics.

9.2 Integration with transcriptome, epigenome, and phenome data

The power of a rice pan-genome can be greatly amplified by integrating it with other layers of genomic and phenotypic data. One future direction is the development of pan-transcriptomes-catalogs of all transcripts expressed across different rice varieties and conditions. Different rice lines may express alternative splicing variants or even lineage-specific genes (from the dispensable genome) under certain conditions. By mapping RNA-seq data from diverse varieties to a pan-genome reference, researchers could identify novel transcripts originating from sequences absent in the standard reference (Woldegiorgis et al., 2022). Some initial studies have begun constructing such cross-variety transcriptome comparisons, revealing, for example, that certain stress-induced transcripts in wild rice have no counterpart in cultivated rice. Integrating transcriptomic data will help pinpoint which variable genes are actually functional (transcribed) and under what circumstances, linking the structural presence of a gene to a biochemical function.

Similarly, incorporating epigenomic data (like DNA methylation, histone modification profiles) in a pan-genomic context is an emerging frontier. It is known that transposable element activation and silencing can vary between rice strains; these epigenetic differences could influence gene expression and might correlate with structural variations (e.g., TEs near variable genes might be silenced in some strains and active in others) (Li et al., 2025). A pan-epigenome approach would track how epigenetic marks differ on core vs. dispensable genomic regions in different genetic backgrounds. This could provide insight into regulation of newly introgressed DNA or domestication-related chromatin changes.

On the phenotype side, bridging the gap between pan-genome genotype and the phenome (the set of phenotypes) is the ultimate goal. This will involve large panels of diverse lines grown in multiple environments with extensive trait measurements (phenomics), and analyzing these in conjunction with pan-genome variants. Approaches like GWAS and machine learning can link complex combinations of structural variants to traits. For instance, if a subset of pan-genome presence/absence variants consistently correlates with drought tolerance (supported by both genomics and transcriptomics under drought stress), it strengthens the causal inference.

One particular integration challenge is dealing with gene presence/absence in gene expression studies. If a gene is missing in some lines, traditional differential expression analysis must account for that (treat missing genes appropriately rather than as zero expression). New computational methods are needed to handle such scenarios seamlessly. However, as these integrative analyses mature, we expect a more holistic understanding: not just which genes exist in the pan-genome, but which are turned on or off, methylated or not, and ultimately how they drive complex traits across the myriad contexts in which rice is grown.

9.3 From graph-based pan-genomes to pan-transcriptomes

Graph-based representations of the rice pan-genome are likely to serve as a foundation for next-generation multi-omics integration. By encoding all genomic variants in a graph, we set the stage for aligning not only DNA reads but also RNA transcripts and other sequence-based data to the same structure. A pan-transcriptome can be built by aligning RNA-seq reads from diverse rice varieties to a pan-genome graph rather than a single reference (Woldegiorgis et al., 2022). This approach would allow discovery of transcripts arising from sequences unique to certain subpopulations. For example, if an *indica*-specific gene is expressed under heat stress, its mRNA reads will align to the *indica* branch of the pan-genome graph and be correctly assembled into an *indica*-specific transcript, rather than being lost or misaligned when using a *japonica* reference. Early efforts in model plant systems hint at the promise of this approach—for instance, graph-based read mapping has improved the detection of allele-specific expression and splice variants in *Arabidopsis*. In rice, a graph-based pan-transcriptome could reveal new isoforms that are present only in, say, aus rice, or alternative splicing patterns that differ between *indica* and *japonica* due to structural variants affecting splice sites.

Another frontier is constructing pan-metabolic networks or pan-proteomes, extending the concept beyond the genome. A pan-genome graph can be annotated with protein-coding and non-coding elements, and one could overlay proteomic data (peptide mass spectra) to detect peptides from subspecies-specific genes (Shrestha et al., 2024). Similarly, one could map chromatin immunoprecipitation (ChIP-seq) data for transcription factors or histone marks onto the pan-genome graph to see how regulatory landscape differs across genomes.

All of this will require robust graph bioinformatics tools. The graph approach is computing-intensive—the rice genome is already ~390 Mb, and a graph representing hundreds of genomes will be larger. Algorithms for mapping RNA or DNA reads to a variation graph (such as VG, GraphAligner, or Minigraph) are improving, but handling the sheer data volumes of population-scale transcriptomes is still challenging. Moreover, visualizing and interpreting results on a graph is non-trivial for biologists. User-friendly interfaces will be needed to query, for example, “show expression of this gene across all varieties, including those where the gene is absent or truncated.” Despite these challenges, the concept of pan-transcriptomics is on the horizon. As graph-based pan-genomes become standard, it’s natural that all sequence-based assays—RNA-seq, ATAC-seq, methyl-seq—will be re-analyzed in a graph context to maximize discovery. This will lead to richer functional catalogs, identifying not just the static presence of genes, but their dynamic usage across the species.

9.4 Prospects for pan-genomes in sustainable agriculture and climate adaptation

Pan-genomics is poised to play a key role in breeding crops that can withstand the challenges of climate change and ensure sustainable agriculture (Shang et al., 2022). By capturing the full genetic diversity of rice, the pan-genome provides a reservoir of traits that can be tapped for adaptation. For instance, climate change is leading to more erratic weather patterns—droughts, floods, extreme temperatures—and pan-genome analyses have identified many stress-response genes that were left behind during domestication. These include heat shock factors, dehydration-responsive element-binding proteins, and various osmoprotectant biosynthesis genes present in wild or traditional varieties but absent in modern high-yield cultivars. Armed with this knowledge, breeders can re-introduce such genes into elite backgrounds to create climate-resilient varieties. Indeed, efforts are underway to breed “climate-smart” rice, such as flood-tolerant varieties (*Sub1* introgression lines) and drought-tolerant varieties, and pan-genomic data accelerates the identification of new tolerance genes and the markers to track them.

Beyond abiotic stresses, sustainable agriculture also calls for improved resistance to pests and diseases, potentially reducing the need for chemical inputs. The rice pan-genome contains a vast arsenal of resistance (R) genes and novel allelic variants—far more than any single variety possesses. By exploring this diversity, breeders can pyramid multiple R genes against evolving pathogens, or discover broad-spectrum resistance genes that were hidden in unsequenced germplasm. For example, a pan-genome search might find a wild rice gene that confers resistance to a virulent new rice blast fungus strain; gene editing could then rapidly deploy this resistance into susceptible but high-performing cultivars.

Pan-genomics also dovetails with sustainable practices like varietal diversification. Instead of monocultures of a single variety, farmers might plant mixed varietal populations to mitigate risk. Pan-genome analysis helps select varieties that are genetically distinct (carrying complementary resistance genes, for instance) to maximize the benefits of such diversification.

Finally, as crop scientists aim for future-proofing crops, they are looking at traits like nitrogen use efficiency, carbon sequestration (root biomass traits), and allelopathy for weed suppression. Many of these traits haven't been heavily selected in modern breeding but exist in traditional or wild varieties. The pan-genome offers a systematic way to mine genes related to these traits. For example, if deeper roots for drought resilience and carbon sequestration are desired, pan-genome GWAS might identify novel root development regulators present in upland landraces (He et al., 2024).

Acknowledgments

We sincerely thank Dr. Qian for providing valuable comments and suggestions during the writing of this paper, which were instrumental in improving its quality.

Conflict of Interest Disclosure

The authors affirm that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Bayer P., Golicz A., Scheben A., Batley J., and Edwards D., 2020, Plant pan-genomes are the new reference, *Nature Plants*, 6(8): 914-920.
<https://doi.org/10.1038/s41477-020-0733-0>
- Daware A., Malik A., Srivastava R., Das D., Ellur R.K., Singh A.K., Tyagi A.K., and Parida S.K., 2023, Rice pangenome genotyping array: an efficient genotyping solution for pangenome-based accelerated genetic improvement in rice, *The Plant Journal*, 113(1): 26-46.
<https://doi.org/10.1111/tpj.16028>
- Guo D., Li Y., Lu H., Zhao Y., Kurata N., Wei X., Wang A., Wang Y., Zhan Q., Fan D., Zhou C., Tian Q., Weng Q., Feng Q., Huang T., Zhang L., Gu Z., Wang C., Wang Z., Wang Z., Huang X., Zhao Q., and Han B., 2025, A pangenome reference of wild and cultivated rice, *Nature*, 642(7930): 662-671.
<https://doi.org/10.1038/s41586-025-08883-6>
- He H., Leng Y., Cao X., Zhu Y., Li X., Yuan Q., Zhang B., He W., Wei H., Liu X., Xu Q., Guo M., Zhang H., Yang L., Lv Y., Wang X., Shi C., Zhang Z., Chen W., Zhang B., Wang T., Yu X., Qian H., Zhang Q., Dai X., Liu C., Cui Y., Wang Y., Zheng X., Xiong G., Zhou Y., Qian Q., and Shang L., 2024, The pan-tandem repeat map highlights multiallelic variants underlying gene expression and agronomic traits in rice, *Nature Communications*, 15(1): 7291.
<https://doi.org/10.1038/s41467-024-51854-0>
- Hickey G., Heller D., Monlong J., Sibbesen J.A., Sirén J., Eizenga J., Paten B., 2020, Genotyping structural variants in pangenome graphs using the vg toolkit, *Genome Biology*, 21(1): 35.
<https://doi.org/10.1186/s13059-020-1941-7>
- Kou Y., Liao Y., Toivainen T., Lv Y., Tian X., Emerson J., Gaut B., Zhou Y., and Purugganan M., 2020, Evolutionary genomics of structural variation in Asian rice (*Oryza sativa*) domestication, *Molecular Biology and Evolution*, 37(12): 3507-3524.
<https://doi.org/10.1093/molbev/msaa185>
- Li K., Jiang W., Hui Y., Kong M., Feng L., Gao L., Li P., and Lu S., 2021, Gapless *indica* rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution, *Molecular plant*, 14(10): 1745-1756.
<https://doi.org/10.1016/j.molp.2021.06.017>
- Li X., Dai X., He H., Chen W., Qian Q., Shang L., Guo L., and He W., 2025, Uncovering the breeding contribution of transposable elements from landraces to improved varieties through pan-genome-wide analysis in rice, *Frontiers in Plant Science*, 16: 1573546.
<https://doi.org/10.3389/fpls.2025.1573546>
- Liu C., Peng P., Li W., Ye C., Zhang S., Wang R., Li D., Guan S., Zhang L., Huang X., Guo Z., Guo J., Long Y., Li L., Pan G., Tian B., and Xiao J., 2021, Deciphering variation of 239 elite *japonica* rice genomes for whole genome sequences-enabled breeding, *Genomics*, 113(5): 3083-3091.
<https://doi.org/10.1016/j.ygeno.2021.07.002>

- Lu Y., Wang J., Chen B., Mo S., Lian L., Luo Y., Ding D., Ding Y., Cao Q., Li Y., Li Y., Liu G., Hou Q., Cheng T., Wei J., Zhang Y., Chen G., Song C., Hu Q., Sun S., Fan G., Wang Y., Liu Z., Song B., Zhu J., Li H., and Jiang L., 2021, A donor-DNA-free CRISPR/Cas-based approach to gene knock-up in rice, *Nature Plants*, 7(11): 1445-1452.
<https://doi.org/10.1038/s41477-021-01019-4>
- Lu Y., Xu Y., and Li N., 2022, Early domestication history of asian rice revealed by mutations and genome-wide analysis of gene genealogies, *Rice*, 15(1): 11.
<https://doi.org/10.1186/s12284-022-00556-6>
- Qin P., Lu H., Du H., Wang H., Chen W., Chen Z., He Q., Ou S., Zhang H., Li X., Li X., Li Y., Liao Y., Gao Q., Tu B., Yuan H., Ma B., Wang Y., Qian Y., Fan S., Li W., Wang J., He M., Yin J., Li T., Jiang N., Chen X., Liang C., and Li S., 2021, Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations, *Cell*, 184: 3542-3558.
<https://doi.org/10.1016/j.cell.2021.04.046>
- Sedlazeck F.J., Rescheneder P., Smolka M., Fang H., Nattestad M., von Haeseler A., and Schatz M.C., 2018, Accurate detection of complex structural variations using single-molecule sequencing, *Nature Methods*, 15(6): 461-468.
<https://doi.org/10.1038/s41592-018-0001-7>
- Shang L., He W., Wang T., Yang Y., Xu Q., Zhao X., Yang L., Zhang H., Li X., Lv Y., Chen W., Cao S., Wang X., Zhang B., Liu X., Yu X., He H., Wei H., Leng Y., Shi C., Guo M., Zhang Z., Zhang B., Yuan Q., Qian H., Cao X., Cui Y., Zhang Q., Dai X., Liu C., Guo L., Zhou Y., Zheng X., Ruan J., Cheng Z., Pan W., and Qian Q., 2023, A complete assembly of the rice Nipponbare reference genome, *Molecular Plant*, 16(8): 1232-1236.
<https://doi.org/10.1016/j.molp.2023.08.003>
- Shang L., Li L., He H., Yuan Q., Song Y., Wei Z., Lin H., Hu M., Zhao F., Zhang C., Li Y., Gao H., Wang T., Liu X., Zhang H., Zhang Y., Cao S., Yu X., Zhang Y., Tan Y., Qin M., Ai C., Yang Y., Zhang B., Hu Z., Wang H., Lv Y., Wang Y., Ma J., Lu H., Wu Z., Liu S., Sun Z., Zhang H., Wang Y., Gao L., Li Z., Zhou Y., Li J., Zhu Z., Xiong G., Ruan J., and Qian Q., 2022, A super pan-genomic landscape of rice, *Cell Research*, 32(10): 878-896.
<https://doi.org/10.1038/s41422-022-00685-z>
- Shi J., Tian Z., Lai J., and Huang X., 2022, Plant pan-genomics and its applications, *Molecular plant*, 16(1): 168-186.
<https://doi.org/10.1016/j.molp.2022.12.009>
- Shrestha A., Gonzales M., Ong P., Larmande P., Lee H., Jeung J., Kohli A., Chebotarov D., Mauleon R., Lee J., and McNally K., 2024, RicePilaf: a post-GWAS/QTL dashboard to integrate pangenomic, coexpression, regulatory, epigenomic, ontology, pathway, and text-mining information to provide functional insights into rice QTLs and GWAS loci, *GigaScience*, 13: giae013.
<https://doi.org/10.1093/gigascience/giae013>
- Song J.M., Guan Z., Hu J., Guo C., Zhang H., Wang S., Liu D., Wang B., Lu S.P., Zhou R., Xie W., Cheng Y., Zhang Y., Liu K., Yang Q., Chen L.L., and Guo L., 2020, Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*, *Nature Plants*, 6(1): 34-45.
<https://doi.org/10.1038/s41477-019-0577-7>
- Vahedi S., Momen M., Mousavi S., Banabazi M., Hasanvandi M., Bhatta M., Roudbar A., and Ardestani S., 2023, Population genetic analysis and scans for adaptation and contemporary selection footprints provide genomic insight into aus, *indica* and *japonica* rice cultivars diversification, *Journal of Genetics*, 102: 1-14.
<https://doi.org/10.1007/s12041-023-01440-y>
- Wang J., Yang W., Zhang S., Hu T., Yuan H., Dong J., Chen L., Ma Y., Yang T., Zhou L., Chen J., Liu B., Li C., Edwards D., and Zhao J., 2023, A pangenome analysis pipeline provides insights into functional gene identification in rice, *Genome Biology*, 24(1): 19.
<https://doi.org/10.1186/s13059-023-02861-9>
- Woldegiorgis S., Wu T., Gao L., Huang Y., Zheng Y., Qiu F., Xu S., Tao H., Harrison A., Liu W., and He H., 2022, Identification of heat-tolerant genes in non-reference sequences in rice by integrating pan-genome, transcriptomics, and QTLs, *Genes*, 13(8): 1353.
<https://doi.org/10.3390/genes13081353>
- Wu D., Xie L., Sun Y., Huang Y., Jia L., Dong C., Shen E., Ye C., Qian Q., and Fan L., 2023, A syntelog-based pan-genome provides insights into rice domestication and de-domestication, *Genome Biology*, 24(1): 179.
<https://doi.org/10.1186/s13059-023-03017-5>
- Yang L., He W., Zhu Y., Lv Y., Li Y., Zhang Q., Liu Y., Zhang Z., Wang T., Wei H., Cao X., Cui Y., Zhang B., Chen W., He H., Wang X., Chen D., Liu C., Shi C., Liu X., Xu Q., Yuan Q., Yu X., Qian H., Li X., Zhang B., Zhang H., Leng Y., Zhang Z., Dai X., Guo M., Jia J., Qian Q., and Shang L., 2025, GWAS meta-analysis using a graph-based pan-genome enhanced gene mining efficiency for agronomic traits in rice, *Nature Communications*, 16(1): 3171.
<https://doi.org/10.1038/s41467-025-58081-1>
- Zhao Q., Feng Q., Lu H., Li Y., Wang A., Tian Q., Zhan Q., Lu Y., Zhang L., Huang T., Wang Y., Fan D., Zhao Y., Wang Z., Zhou C., Chen J., Zhu C., Li W., Weng Q., Xu Q., Wang Z., Wei X., Han B., and Huang X., 2018, Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice, *Nature Genetics*, 50(2): 278-284.
<https://doi.org/10.1038/s41588-018-0041-z>

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.