Unlocking rice's genetic potential: big data-driven insights from population genomics

Yuhan Zhou[#]* and Ziqi Zhou[#]

State Key Laboratory of Rice Biology & Breeding, Zhejiang Provincial Key Laboratory of Crop Germplasm, The Advanced Seed Institute, Zhejiang University, Hangzhou 310058, China

[#] Authors contributed equally: Yuhan Zhou, Ziqi Zhou

* Corresponding author, E-mail: 12316095@zju.edu.cn

Abstract

Rice, a cornerstone of global food security, faces escalating demand amidst climate-driven agricultural challenges. By integrating data from over 10,000 cultivated and wild rice accessions, researchers have resolved structural variations, domestication trajectories, and adaptive alleles critical for stress resilience and yield improvement. From five major rice-growing regions in China, the newly expanded 6K genomic atlas revealed 3,131 quantitative trait loci associated with 53 phenotypes across 212 datasets under various environmental conditions. Platforms like RiceSuperPIRdb and RiceAltas integrate multidimensional datasets with visualization tools, bridging genetic discovery to breeding applications. Artificial intelligence (AI) models trained on 12,678 hybrid profiles predict heterosis with > 85% accuracy, while platforms such as AutoGP integrate genotype extraction, phenotypic extraction, and genomic selection (GS) models to halve breeding cycles. Yet, petabyte-scale data from initiatives challenge traditional storage and analysis frameworks, necessitating cloud-based architectures and interpretable AI to resolve nonlinear genotype-phenotype relationships. Single-cell multi-omics and 3D genome mapping now dissect spatiotemporal gene regulatory networks, transforming empirical breeding into a predictive science despite enduring bottlenecks in data standardization and equitable resource sharing. This synthesis underscores the pivotal role of population genomics in harnessing rice's evolutionary legacy to meet twenty-first-century food security imperatives.

Introduction

Global food security faces unprecedented challenges as the World Food Programme (WFP) projects that 343 million people across 74 countries will experience acute food insecurity by 2025. Rice, a staple crop supporting over half of the world's population, stands at the forefront of agricultural research[1]. Current estimates indicate a 130% surge in global rice consumption from 2010 levels by 2035, necessitating an additional 30 million tons annually to meet escalating demand[2].

The genus *Oryza* includes two cultivated rice species, Asian rice (*Oryza sativa*) and African rice (*O. glaberrima*), along with over 20 wild relatives[3], collectively representing a vast genetic repository shaped by divergent evolutionary histories. Maintained by the International Rice Research Institute (IRRI), the International Rice Genebank holds more than 132,000 available accessions as of December 2019 with *O. sativa* dominating agricultural systems through its subdivision into the *japonica* and *indica* subspecies. Together, these species represent an evolutionary mosaic of phylogenetic diversity, offering untapped reservoirs of stress tolerance and adaptive traits forged across diverse agroecological niches[4,5].

Advances in sequencing technologies and genome assembly algorithms have revolutionized rice genomics, enabling high-resolution exploration of its genetic architecture[6]. To date, researchers have generated chromosome-level or even complete genome assemblies for hundreds of rice varieties[7], while large-scale resequencing efforts have characterized genetic diversity across diverse germplasm collections. These endeavors have culminated in the development of pan-genomes spanning intervarietal and even intragenus boundaries, capturing structural variations (SVs)[8], single nucleotide polymorphisms (SNPs), and insertions/deletions (InDels)

at unprecedented scales. Such resources are critical for identifying rare alleles, beneficial natural variants, and novel haplotype combinations that underpin agronomic traits. Because of low sequencing depth, merely at 1×, structural resolution and causal gene identification remained elusive. Besides, sequencing data often show overlaps.

Concurrently, the exponential growth of rice genomic data has driven the creation of specialized platforms, including Rice Variation Map v2.0 (RiceVarMap2)[9,10], Rice Super Pan-genome Information Resource Database (RiceSuperPIRdb)[11], and the Rice Pan-genome Browser (RPAN)[12], which integrate multidimensional datasets with analytical tools to facilitate gene discovery and breeding applications. Modern plant breeding harnesses cutting-edge GS platforms like CropGS-Hub[13], BreedingAIDB[14], and Smart Breeding Platform[15]. These tools combine the power of machine learning (ML) and deep learning (DL) to analyze genetic and phenotypic data, significantly speeding up the breeding process.

However, rice population genomics now confronts dual frontiers, managing petabyte-scale data challenges and harnessing AI-driven insights across multi-omics layers. While cloud architectures address storage and computational bottlenecks posed by sequencing over 10,000 accessions, critical gaps persist in interpreting nonlinear genotype-environment interactions—a limitation exacerbated by the 'black box'[16] nature of DL models like AgroNT[17], despite their success in predicting cis-regulatory elements. Simultaneously, spatial genomics maps 3D chromatin interactions to resolve noncoding structural variants' roles in stress adaptation, yet scalable 3D omics technologies remain nascent[18]. Through federated learning frameworks and equitable data sharing, the integration of these domains will define the next decade of breeding innovation,

balancing computational scalability with biological interpretability to accelerate climate-resilient cultivar development.

This review synthesizes advances in rice population genomics, focusing on the integration of large-scale resequencing data and pan-genome frameworks, the interrogation of genomic diversity to decode domestication and adaptation mechanisms, and the translation of these insights into precision breeding strategies. We further explore challenges in data harmonization and the ethical implications of genomics-driven agriculture.

Current research progress in rice population genomics

The nascent era of rice population genomics (2010–2015) was characterized by pioneering yet depth-constrained studies, where seminal efforts laid the groundwork despite technical limitations[19]. In 2010, a pioneering study of 517 Chinese landraces established an integrative approach combining second-generation sequencing at merely 1× depth and genome-wide association studies (GWAS)[20], demonstrating its utility as a complementary strategy to classical biparental mapping for dissecting complex agronomic traits. By 2012, this methodology had been scaled to analyze 950 worldwide cultivars at the same depth, identifying 32 novel loci associated with flowering time and grain morphology[21], thereby expanding the scope of rice genetic architecture exploration from regional landraces to global diversity. Low sequencing depth underscored both the potential and constraints of early next-generation sequencing (NGS) frameworks, enabling broad variant surveys, yet structural resolution and causal gene identification remained elusive.

These foundational efforts culminated in 2018 with the landmark 3,000 Rice Genomes Project (3K-RG)[22], which resequenced 3,010 Asian cultivars to delineate subpopulation structures and provide evidence for multiple domestication origins. While this initiative cataloged 29 million SNPs and 90,000 structural variants (SVs), only 430 accessions exceeded 20× coverage, and reliance on short-read sequencing inherently limited SV detection, particularly in complex genomic regions[23], particularly those spanning complex genomic regions (Table 1).

The advent of third-generation sequencing (TGS) technologies catalyzed a paradigm shift in rice genomics, enabling the *de novo* assembly of gap-free genomes[24–26] and the systematic exploration of SVs at population scales. In 2021, a study of 33 *O. sativa* and *O. glaberrima* accessions systematically identified over 170,000 SVs[27], linking these genomic rearrangements to environmental adaptation. This milestone laid the groundwork for the first comprehensive rice pangenome in 2022, which integrated 111 high-quality genomes, including wild relatives, and uncovered 879 Mb of non-redundant sequences enriched with stress-responsive genes[28]. Wild rice contributions, particularly long terminal repeat (LTR) retro-transposons, highlighted their role in shaping adaptive diversity.

Subsequently, the field witnessed a surge in super-pangenome initiatives. By 2023, a global consortium analyzed 10,548 accessions using multi-platform sequencing[11], cataloging rare alleles and establishing a framework for trait mining across agroecological zones. Concurrently, a synteny-based super-pangenome of 74 wild[29], cultivated, and weedy rice strains resolved introgression events and domestication signatures, demonstrating a single origin of *japonica* followed by adaptive hybridization. Wild rice genomes were further prioritized, and an analysis of 17 wild accessions revealed extensive subgenomic exchanges and regulatory variations that underpin stress resilience[30].

These resources have transitioned from structural catalogs to functional breeding tools. In 2024, haplotype-resolved assemblies of 75 Asian rice and wild relatives pinpointed large inversions (≥ 100 bp) as drivers of heterosis[31], while a 2025 analysis of 6,044 Chinese cultivars linked 3,131 QTLs to 53 agronomic traits[32], enabling the cloning of grain-length regulator *OsGL3.6*. Resequencing data in genomic databases often shows overlap, meaning some data is sequenced multiple times across different studies. Based on the information provided, it seems likely that there are 57 instances where the current 6,044 data duplicates earlier efforts compared to an initial set of 517 datasets. Additionally, other research efforts also exhibit similar overlaps, but these newer datasets have significantly deeper sequencing, which can improve accuracy and detail in digging target QTLs.

Integrated database for advancing rice population genomics

The rapid expansion of rice genomic resources has spurred the development of specialized databases that harmonize large-scale data storage with analytical functionalities (Table 2). Leading these efforts, RiceVarMap[9,10] serves as a cornerstone platform, integrating high-quality genotypes, GWAS linkages, and variant annotations for 4,726 cultivars, enabling trait dissection at unprecedented resolution. At the scale frontier, RiceSuperPIRdb[11] aggregates reference-grade assemblies from 10,904 wild and cultivated rice accessions, constructing a super pan-genome that captures structural diversity across subpopulations. Complementing these resources, RPAN[12] consolidates 3,010 accessions with pan-genome annotations, presence-absence variations (PAVs), and expression datasets, while RiceAtlas[47] bridges phenotypic and genomic heterogeneity across 3,733 germplasms through ML datasets. Notably, the Rice Gene Index[48] (RGI) pioneers multi-accession gene relational networks, mapping transcriptome dynamics across tissues and developmental stages, a critical resource for evolutionary and functional studies. These platforms collectively address three bottlenecks: standardization of heterogeneous data types (SNPs, SV, epigenetics), visualization of complex pan-genome architectures, and democratization of computational tools for non-specialists.

Emerging niche repositories further enrich this ecosystem. The Rice Genome Hub (RGH) provides evolutionary context through 32 reference genomes spanning 10 *Oryza* species, enabling comparative genomics at taxonomic depth. OryzaGenome[49,50] specializes in SNP discovery across 466 accessions from 19 wild relatives, featuring allele effect predictors for adaptive trait mining. For translational research, Rice Resource Center (RiceRc)[21] offers ancestral allele reconstruction across 33 cultivars, aiding the identification of domestication signatures and breeding targets. Future iterations must prioritize Findable, Accessible, Interoperable, Reusable (FAIR)[51] data principles and cloud-native architectures to enable federated queries across these distributed resources.

Modern plant breeding harnesses advanced GS platforms, such as CropGS-Hub[13], with its six ML models (BayesCπ, Bayesian Lasso, Bayesian Regression, GBLUP, RR-BLUP, and LightGBM)[52,53], BreedingAIDB[14] for boosting-type algorithms, and the Smart Breeding Platform[15] combining ML and DL to streamline trait optimization and breeding cycles. By integrating genotype-to-phenotype workflows, breeders can rapidly engineer climate-resilient crops, exemplified by drought-tolerant rice development. Starting with 500 rice accessions phenotyped for drought-related traits (e.g., root depth) and resequencing data uploaded to platforms like Smart Breeding

Table 1. Summary of landmark rice population genomics studies.

Rice type	Number of accessions	Sequencing method	Key results	Sparking point	Ref.	Published year
129 *O. rufipogon* and 16 *O. sativa*	145	ONT (103×), PacBio HiFi (24×) and NGS	69,531 pan-genes, with 28,907 core genes and 13,728 wild-rice-specific genes	Supports Asian rice's monophyletic origin from *Or-IIIa* (*japonica* ancestor)	[33]	2025
Chinese cultivars	6,044	NGS (31×)	3,131 QTLs for 53 traits; cloned *OsGL3.6* regulating grain length	Unprecedented QTL scale with direct breeding applications	[32]	2025
Asian cultivars	18,421	ONT (69×) and NGS (100×) on founders and NGS (0.3×) on F$_2$	1,207 QTLs for 16 traits; validated *OsMADS22* and *OsFTL1*	Hybrid sequencing resolves QTL-phenotype causality	[34]	2024
73 Asian rice and two wild relatives	75	Pacbio HiFi (18×)	1,769 large inversions (≥ 100 bp) in pan-genome	First pan-genome landscape defining and validating inversion dynamics in rice	[31]	2024
Inbred lines	12	PacBio HiFi (30×)	47,490 gene families, 489,713 genes and heterosis of 17 agronomic traits	SV-heterosis mechanism bridge structural and regulatory evolution	[35]	2024
Wild rice	17	PacBio HiFi (60×) and NGS (100×)	101,723 gene families including 63,881 gene families absent in cultivated	Loss of some resistance genes in cultivated rice during domestication and artificial selection	[30]	2024
2,839 hybrid cultivars and 9,839 F$_2$	12,678	NGS (F$_1$ 35×; F$_2$ 0.2×)	5.2 million SNPs, 1.7 million InDels and 22,555 SV	Heterosis decoded and developed a genomic model	[36]	2023
11 wild, 51 cultivated and 12 weedy rice	74	PacBio HiFi (32×)	175,528 syntelog groups in a syntelog-based pangenome with putative introgression regions	Constructed syntelog-based pangenome	[29]	2023
Northeast Asian cultivars	546	NGS (10×)	111 known QTLs genotyped and novel GWAS-QTLs	Summarized a rice breeding principle	[37]	2023
Indian rice landraces	108	ONT (20×) and NGS (67×)	7415 and 131 differentially expressed transcripts and miRNAs	Identified miRNAs and their target genes in regulating anthocyanin biosynthesis	[38]	2022
Asian and African rice	251	ONT (100×) and NGS (65×)	1.52 Gb non-redundant DNA sequences in pan-genome	Revealed extensive SVs associated with grain weight and gene PAV	[39]	2022
O. sativa and wild relatives	111	ONT (68×) and NGS (69×)	879 Mb novel sequences and 19,000 novel genes in the rice pan-genome	New pan-genome construction method for long-read data	[28]	2022
32 *O. sativa* and 1 *O. glaberrima*	33	PacBio (60×), and NGS (20×)	171,072 SVs and 25,549 gCNVs underlying environmental adaptation	SV-adaptation nexus across species boundaries	[27]	2021
299 papers published from 1995 to 2020	/	/	562 alleles in 225 QTGs and 348 QTNs	A comprehensive QTN map and genome navigation systemin of rice	[40]	2021
Chinese cultivars	1,275	NGS (7×)	143 association loci, including three new genes, control heading date or amylose content	GWAS-powered gene prioritization framework	[41]	2020
53 *O. sativa* and 13 *O. rufipogon*	66	NGS (115×)	23 million sequence variants include many known QTNs	Variant-to-QTN translation bridges genomics and breeding	[42]	2018
Asian cultivars	3,010	NGS (18×)	29 million SNPs, 2.4 million indels, and over 90,000 SVs	Domestication model challenger via population structure	[22]	2018
10,074 F$_2$ lines from 17 hybrid rice crosses	10,074	NGS (0.2×)	A dense genotype map with 347,803 recombination events; 74 QTLs for seven yield traits	Dominance theory revived in hybrid rice	[43]	2016
Hybrid rice (F$_1$)	1,495	NGS (2×)	38 agronomic traits and 130 associated loci	Hybrid trait architecture blueprint	[44]	2015
Asian cultivars and landraces	1,479	NGS (3×)	6,428,770 SNPs total and 200 selected regions between *IndI* and *IndII*	Breeding signatures between *IndI* and *IndII* including loci associated with agronomic traits	[45]	2015
446 *O. rufipogon* and 1,083 cultivated varieties	1,529	NGS (2×)	5,037,497 non-singleton SNPs and 55 selective sweeps	Domestication geography redefined	[46]	2012
Worldwide cultivars	950	NGS (1×)	32 new loci associated with flowering time and with ten grain-related traits	Extend to a larger and more diverse sample of rice varieties	[21]	2012
Chinese landraces	517	NGS (1×)	GWAS of 14 traits via 3.6M SNPs in 517 *indica* landraces	Complex trait genomics and GWAS on crop pioneer	[20]	2010

Platform, automated SNP calling, and GWAS pinpoint loci linked to drought adaptation, while GS models predict progeny performance, enabling strategic crosses and early-generation selection of high-scoring individuals. Iterative genomic filtering across generations concentrates favorable alleles, minimizing field trialing of low-potential lines and compressing breeding timelines by ~50%, with elite drought-tolerant varieties advancing to multi-environment testing within 3–4 years. Such platforms democratize precision breeding, merging computational power with biological insight to address global agricultural challenges.

Rice population genomics promotes crop breeding

Harnessing genetic diversity to uncover agronomic trait genes

Advances in rice population genomics have systematically decoded the genetic architecture of agronomic traits through large-scale sequencing and multi-omics integration[59]. By analyzing global germplasm spanning wild, landrace, and modern cultivated rice, researchers identified key loci governing critical traits (Fig. 1). In 2016,

Table 2. Databases for rice population genomics data storage with online visualization function.

Database	Number of cultivars	Description	Ref.	URL
RiceVarMap	4,726	High quality and complete genotype data; Comprehensive annotations of genomic variations; Phenotype data and GWAS results and tools	[9,10]	https://ricevarmap.ncpgr.cn/
RiceSuperPIRdb	356 (TGS), 10,548 (NGS)	The super pan-genome based on reference-quality *de novo* long-read assemblies accessions	[11]	http://ricesuperpir.com/
RGI	16	Establishes gene relationships between different accessions and contains transcriptomes across different accessions and tissues	[48]	https://riceome.hzau.edu.cn/
RPAN	3,010	Consolidates 3,010 accessions, pan-genome annotations, PAVs, expression data, and analytical tools	[12]	https://cgm.sjtu.edu.cn/3kricedb/
RiceAtlas	3,733	Integrates large-scale rice phenotypic, genomic variation, and germplasm information	[47]	http://60.30.67.242:18076/#/home
Rice Genome Hub	32	Reference genomes of 10 species in *Oryza*	/	https://rice-genome-hub.southgreen.fr/
OryzaGenome	466	A collection of short reads from 217 accessions covering 19 *Oryza* species and 446 *O. rufipogon* SNP viewer	[49,50]	http://viewer.shigen.info/oryzagenome21detail/
RiceRc	33	Integrates a genome browser and a BLAST function and includes the ancestral allele data for SNP/Indels	[21]	http://ricerc.sicau.edu.cn/
RAP-DB	685	Facilitates a comprehensive analysis of the genome structure and function of rice on the basis of the annotation	[54,55]	https://rapdb.dna.affrc.go.jp/
MBKbase-rice	137,796	Integrates rice germplasm information, reference genomes with gene loci, phenotypic and gene expression data	[56]	https://www.mbkbase.org/rice
Rice3KGS	3K	LGBMY GS model with a dataset of 56 trait phenotypes and seven genotype datasets.	[57]	http://101.201.107.228:8002/#/Rice3KGS/home/
AutoGP	/	A dataset of 143,477 rice G2P paired data for 41traits	[58]	http://autogp.hzau.edu.cn
CropGS-Hub	2,578	Over 224 billion genotype data and 434 thousand phenotype data generated from >30 000 individuals belonging to 7 major crop species	[13]	https://iagr.genomics.cn/CropGS/#/
BreedingAIDB	415	129,449 rice genome-to-phenotype paired data	[14]	http://ibi.zju.edu.cn/BreedingAIDB/
Smart Breeding Platform	/	Various statistical analyses, GWAS and GS using both classical ML and DL models	[15]	https://sbp.ibreed.cn/#/

GWAS using a high-density rice array (HDRA) of 700,000 SNPs identified *OsSPL13* as a key gene simultaneously enhancing grain length, panicle grain number, and yield[60], while subsequent work in 2018 revealed *OsSAP16* as a critical regulator of low-temperature germination, expanding the molecular toolkit for stress-resilient rice breeding[61]. Recent advances in GWAS have integrated multi-omics approaches, including transcriptome-wide (TWAS), proteome-wide (PWAS), and metabolome-wide (MWAS) association analyses—to pinpoint target traits for precision breeding. For instance, a landmark study combining GWAS with TWAS and methylome profiling deciphered the genetic architecture underlying rice glycemic index (GI), identifying *OsAmy3D*[62] as a key amylase gene regulating starch digestion rates. Similarly, volatilome-based GWAS revealed *OsWRKY19* and *OsNAC021*[63] as master regulators of 2-acetyl-1-pyrroline biosynthesis, the compound responsible for aromatic fragrance in premium rice varieties. This shift from simplex to multi-omics and data-driven strategies has significantly accelerated the development of high-yielding, stress-resilient varieties.

While single-gene discoveries provide foundational insights, the phenotypic manifestation of traits is often shaped by intricate epistatic interactions and genetic background dependencies. A 2024 analysis of 18,421 rice lines uncovered 70 significant locus pairs with additive effects[34], illustrating how genes like the semi-dwarf *Sd1* modulate eight distinct traits through context-dependent interactions. Similarly, *Ghd8* was shown to influence grain protein content only in specific haplotypes of *Ehd1*, underscoring the limitations of reductionist approaches. Population genomics addresses this complexity by mapping genome-wide interaction networks, empowering breeders to select germplasm with compatible genetic backgrounds. This paradigm ensures stable trait expression, mitigating risks of allele dilution or pleiotropic trade-offs in

diverse environments. While integrative pan-genome, GWAS, and GS approaches may not directly identify novel genes, they systematically uncover pleiotropic functions in previously characterized loci. For example, *Ghd7*[64], initially recognized as a heading date regulator, has been shown through population allele mining to coordinate yield-related traits, including plant height and panicle architecture, demonstrating how multi-trait optimization emerges from revisiting 'known' genes.

While traditional analyses primarily rely on population SNPs or small InDels, TGS of rice populations now enables comprehensive characterization of SVs, including gene presence/absence polymorphisms that are often undetectable with NGS. The first TGS-based pan-genome, distinct from earlier approaches like the 3K-RG project that mapped short reads to reference genomes, was constructed by comparing 12 high-quality *de novo* genome assemblies generated from PacBio long-read data (> 100× coverage). Subsequent TGS pan-genomes, such as the 33-TGSRG and 251-TGSRG, have revealed SVs with direct agronomic relevance. For example, a 987-bp LTR insertion upstream of *Os06g13470*[27] was linked to premature leaf senescence, while SV-based GWAS identified a 4-kb insertion disrupting *LOC_Os06g04820*[39] function, reducing grain length. These advancements underscore the critical role of SVs in bridging genomic diversity to trait improvement.

Decoding population structure for elite germplasm utilization

Classified into nine subpopulations, including four *indica* clusters and three *japonica* clusters, the 3K-RG accessions bring more resolution to the within-species diversity of large numbers of SNPs, InDels, and SVs. This granular stratification not only clarifies the monophyletic origin of Asian rice but also pinpoints underutilized rare alleles in peripheral groups, such as drought-responsive *OsNAC6*[65] haplotypes unique to
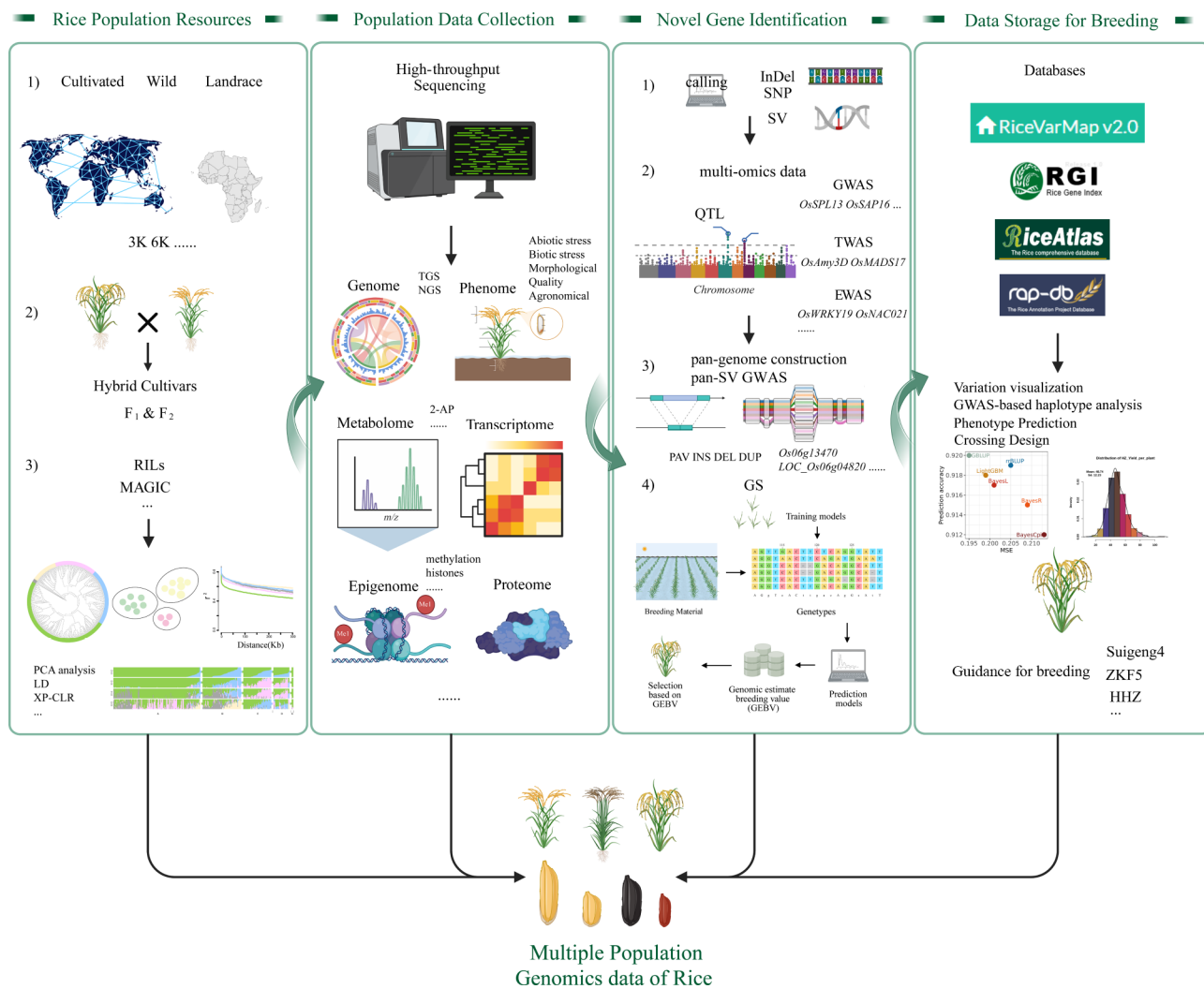
Fig. 1 The integrative framework of population genomics and multi-omics approaches for advancing rice breeding. Key components include: (1) Rice population resources, encompassing cultivated varieties, wild relatives, landraces, hybrid cultivars (F_1/F_2), recombinant inbred lines (RILs), multi-parent advanced generation inter-cross (MAGIC) populations, and so on. (2) High-throughput data collection across genomic, phenomic, transcriptomic, epigenomic, proteomic, and metabolomic layers. (3) Novel gene discovery strategies, leveraging pan-genome analyses (SNPs, InDels, structural variations), genome-wide association studies (GWAS), and genomic selection (GS). (4) Data storage and breeding applications featuring databases (RiceVarMapv2.0, RGI) that guide cultivar development (e.g., ZKF5, HHZ, Sujgeng4). Collectively, this big data-driven paradigm accelerates the identification of agronomically vital traits and informs precision breeding strategies to enhance global rice productivity and resilience.

the *aus* subpopulation. Leveraging inbreeding coefficient profiles and homozygous genomic blocks, breeders can now avoid parental combinations with excessive genetic similarity (> 85% identity-by-descent), mitigating inbreeding depression risks in hybrid programs. Population differentiation analyses such as F_{ST} further dissected ecological adaptation mechanisms, a photoperiod sensitivity gene divergently selected in low-latitude (high-temperature) versus high-latitude (short-day) ecotypes.

A comprehensive study of 6,044 Chinese rice accessions delineated five major growing regions: Southern China (SC), Central China (CC, subdivided into CC-I and CC-J), Southwest China (SW, with SW-I and SW-J), Northern China (NC), and Northeast China (NE)[32]. This diversity panel uncovered distinct genomic backgrounds associated with regional adaptation, including localized selection for key agronomic traits such as heading date, biotic/abiotic stress resistance, and grain morphology—critical factors in matching crop performance to environmental constraints and consumer demands.

Leveraging big data to build predictive breeding models

Building on these genomic insights, the integration of AI has ushered in a transformative era of predictive breeding[66]. GS models, trained on vast datasets of hybrid and inbred lines[67], now forecast trait performance with unprecedented accuracy. Furthermore, DL-based models have shown competitive or even better capability than traditional models. According to model architectures and various development strategies, three advancements can be categorized: CNN-based architectures (DeepGS[68], DNNGP[69], Foundations[70], and SoyDNGP[71]); transformer-based architectures (GPformer[72] and Cropformer[73]; automated frameworks (Auto-GS[74] and AutoGP[58]), and others (TrG2P[75] and DEM[76]).

In 2024, a maize breeding initiative enhanced yield optimization across 64,620 potential hybrid candidates through ML[77], which has since been extensively applied to wheat[78] and soybean[71] breeding programs, as well as resistance gene research in *Spartina alterniflora*[79]. In addition, DeepCCR[70] , tailored to Chinese rice population data, demonstrates robust capability in predicting phenotypic

traits and identifying ecological regions optimal for cultivation. The growing availability of multi-omics datasets in rice research is poised to further advance the development of refined ML frameworks. Beyond trait prediction, AI-driven tools decode nonlinear relationships between structural variations and phenotypes, enabling the deployment of rare alleles, such as anthocyanin biosynthesis regulators from stress-adapted landraces, into elite backgrounds. This synergy of genomics and computational intelligence not only refines directional breeding but also pioneers climate-smart crop design, where adaptive traits are engineered through iterative modeling.

Future perspectives in rice population genomics

Data-driven challenges

Despite substantial progress, significant challenges persist in this field. The foremost challenge in crop genomics lies in managing the exponentially growing data lifecycle, from collection and sequencing to storage, transmission, and interpretation. Establishing comprehensive germplasm repositories faces inherent logistical barriers, particularly for acquiring rare wild rice varieties and landraces essential for biodiversity preservation. Modern sequencing efforts now routinely generate petabyte-scale datasets, exemplified by recent studies producing 60.78 Tb of 150-bp paired-end reads from 5,164 rice accessions[32]. Storing such *Oryza* genomic resources would require over 300 3TB hard drives (occupying about 15 m²), while transferring these datasets at 50 Mb/s bandwidth demands > 200 d—challenges dwarfed by polyploid species like *Brassica napus* (AACC, 1.1 Gb genome size)[80] and *Triticum aestivum* (AABBDD, 16 Gb genome size)[81], where sequencing complexity escalates exponentially. Traditional analysis frameworks, reliant on local data transfers and on-premise computation, falter under such scales. Cloud-based architectures address these limitations by enabling direct data access and scalable computational resources. A key innovation lies in the automated management of intermediate files, which are systematically purged during virtual machine termination. This approach not only reduces physical infrastructure burdens but also standardizes data hygiene, enhancing reproducibility across studies.

Current pan-genomic studies, encompassing thousands of rice accessions, have cataloged vast structural and allelic variations linked to stress resilience, grain quality, and yield. Integrating transcriptomic, epigenomic, and metabolomic datasets will enable precise annotation of cryptic genes and their interaction networks, moving beyond static variant catalogs to dynamic functional insights. Emerging single-cell sequencing technologies further promise to resolve spatiotemporal gene expression patterns, unraveling the genetic regulation of complex traits such as root architecture or drought-responsive metabolic pathways.

Integrating AI and GS with multiscale genomic data

AI and GS models are poised to revolutionize rice breeding paradigms. The rapid evolution of AI technology has established large language models (LLMs) as a pivotal, transformative force within the AI domain. The rapid development of plant-specialized LLMs, including Genomic Pre-trained Network (GPN)[82], GPN-MSA (GPN with multiple-sequence alignment)[83], AgroNT (1 billion parameters)[17], FloraBERT[84], and Plant DNA LLMs (PDLLM)[85] are revolutionizing genomic interpretation through their capacity to decode complex biological patterns. Among these, AgroNT stands out by establishing new benchmarks in predicting cis-regulatory elements (promoters/terminators), tissue-specific transcriptional dynamics, and functional variant prioritization. This computational capacity aligns strategically with the explosive growth of rice population genomics data, where traditional methods struggle to scale. By pre-training on terabase-scale genomic sequences, these LLMs systematically uncover nonlinear relationships and epigenetic dependencies that govern trait formation, a capability critical for annotating millions of rice genomic variants with functional impact scores. Such genome-to-phenome predictive frameworks could transform genetic resource utilization, enabling targeted screening of agriculturally valuable alleles while reducing experimental validation costs. However, significant challenges impede the effective application of LLMs in rice genomics. The field's characteristic 'big data' challenges, which include enormous data volumes, high dimensionality, and intricate biological complexity, create unprecedented computational and storage demands. Moreover, the inherent opacity of DL architectures presents a critical bottleneck. While these models demonstrate impressive predictive accuracy across multiple tasks, their decision-making processes remain largely opaque 'black boxes', severely limiting biological interpretability and hindering researcher trust. This interpretability crisis persists despite model performance improvements, creating a fundamental barrier to translating computational predictions into actionable biological insights.

GS, initially developed for livestock breeding in 2001, has transformed plant breeding through methodologies such as genomic best linear unbiased prediction (GBLUP) and ridge regression BLUP (rrBLUP). Platforms like AutoGP[58] harness large-scale rice genotype-to-phenotype (G2P) datasets, comprising 143,477 accessions and 41 agronomic traits of rice, to refine breeding value prediction. Multi-species frameworks such as CropGS-Hub[13] integrate ML with genomic data, optimizing crossbreeding strategies and allele prioritization in key crops. Future advancements require overcoming critical challenges: cost-effective genotyping arrays for underrepresented germplasm, high-throughput phenotyping systems with micron-level precision, computational efficiency in model training, and adaptive parameter optimization to address nonlinear genotype-environment interactions. Addressing these hurdles will catalyze next-generation breeding frameworks, enabling the systematic development of climate-resilient cultivars through predictive genotype-environment models.

Blueprints of population spatial genomics

The integration of three-dimensional chromatin architecture and population-scale genomic data is transforming rice genomics by bridging structural, functional, and evolutionary insights[86]. Spatial chromatin interaction maps enable functional annotation of noncoding structural variants, revealing their roles in disrupting regulatory circuits at domain boundaries or driving adaptive chromatin states linked to stress resilience. Evolutionary constraints identified through conserved chromatin hubs and plasticity hotspots offer critical design principles for breeding, distinguishing immutable genomic regions from dynamic targets for trait optimization. Overcoming barriers in scalable 3D omics technologies, multidimensional data integration, and equitable data sharing will unlock 'structural intelligence' breeding, precisely editing spatial interaction modules to engineer climate-resilient rice while minimizing unintended pleiotropic effects. This paradigm shift promises to accelerate the development of adaptive crop varieties tailored to future agricultural challenges.

Looking ahead, the integration of these frontiers will catalyze a shift from reactive to anticipatory crop improvement.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: Zhou Y; draft manuscript preparation: Zhou Y;

writing - review and editing: Zhou Z. Both authors reviewed the draft manuscript and approved the final version of the manuscript.

Data availability

Data sharing is not applicable to this article as no datasets were generated or analyzed in this review.

Conflict of interest

The authors declare that they have no conflict of interest.

Dates

References

1. Song S, Tian D, Zhang Z, Hu S, Yu J. 2018. Rice genomics: over the past two decades and into the future. *Genomics, Proteomics & Bioinformatics* 16:397−404

2. Muthayya S, Sugimoto JD, Montgomery S, Maberly GF. 2014. An overview of global rice production, supply, trade, and consumption. *Annals of the New York Academy of Sciences* 1324:7−14

3. Huang C, Chen Z, Liang C. 2021. *Oryza* pan-genomics: a new foundation for future rice research and improvement. *Crop Journal* 9:622−32

4. Izawa T, Shimamoto K. 1996. Becoming a model plant: The importance of rice to plant science. *Trends in Plant Science* 1:95−99

5. Wang C, Han B. 2022. Twenty years of rice genomics research: From sequencing and functional genomics to quantitative genomics. *Molecular Plant* 15:593−619

6. Li H, Durbin R. 2024. Genome assembly in the telomere-to-telomere era. *Nature Reviews Genetics* 25:658−70

7. Shirasawa K, Harada D, Hirakawa H, Isobe S, Kole C. 2021. Chromosome-level *de novo* genome assemblies of over 100 plant species. *Breeding Science* 71:117−24

8. Saxena RK, Edwards D, Varshney RK. 2014. Structural variations in plant genomes. *Briefings in Functional Genomics* 13:296−307

9. Zhao H, Li J, Yang L, Qin G, Xia C, et al. 2021. An inferred functional impact map of genetic variants in rice. *Molecular Plant* 14:1584−99

10. Zhao H, Yao W, Ouyang Y, Yang W, Wang G, et al. 2015. RiceVarMap: a comprehensive database of rice genomic variations. *Nucleic Acids Research* 43:D1018−D1022

11. Wang T, He W, Li X, Zhang C, He H, et al. 2023. A rice variation map derived from 10548 rice accessions reveals the importance of rare variants. *Nucleic Acids Research* 51:10924−33

12. Sun C, Hu Z, Zheng T, Lu K, Zhao Y, et al. 2017. RPAN: rice pan-genome browser for ~3000 rice genomes. *Nucleic Acids Research* 45:597−605

13. Chen J, Tan C, Zhu M, Zhang C, Wang Z, et al. 2024. CropGS-Hub: a comprehensive database of genotype and phenotype resources for genomic prediction in major crops. *Nucleic Acids Research* 52:D1519−D1529

14. Shen Z, Shen E, Yang K, Fan Z, Zhu QH, et al. 2024. BreedingAIDB: a database integrating crop genome-to-phenotype paired data with machine learning tools applicable to breeding. *Plant Communications* 5:100894

15. Li H, Li X, Zhang P, Feng Y, Mi J, et al. 2024. Smart Breeding Platform: a web-based tool for high-throughput population genetics, phenomics, and genomic selection. *Molecular Plant* 17:677−81

16. Yang L, Wang H, Zou M, Chai H, Xia Z. 2025. Artificial intelligence-driven plant bio-genomics research: a new era. *Tropical Plants* 4:e015

17. Mendoza-Revilla J, Trop E, Gonzalez L, Roller M, Dalla-Torre H, et al. 2024. A foundational large language model for edible plant genomes. *Communications Biology* 7:835

18. Liang Z, Zhang Q, Ji C, Hu G, Zhang P, et al. 2021. Reorganization of the 3D chromatin architecture of rice genomes during heat stress. *BMC Biology* 19:53

19. Jing CY, Zhang FM, Wang XH, Wang MX, Zhou L, et al. 2023. Multiple domestications of Asian rice. *Nature Plants* 9:1221−35

20. Huang X, Wei X, Sang T, Zhao Q, Feng Q, et al. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics* 42:961−67

21. Huang X, Zhao Y, Wei X, Li C, Wang A, et al. 2012. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genetics* 44:32−39

22. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, et al. 2018. Genomic variation in 3, 010 diverse accessions of Asian cultivated rice. *Nature* 557:43−49

23. Cameron DL, Di Stefano L, Papenfuss AT. 2019. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications* 10:3240

24. Song JM, Xie WZ, Wang S, Guo YX, Koo DH, et al. 2021. Two gap-free reference genomes and a global view of the centromere architecture in rice. *Molecular Plant* 14:1757−67

25. Shang L, He W, Wang T, Yang Y, Xu Q, et al. 2023. A complete assembly of the rice Nipponbare reference genome. *Molecular Plant* 16:1232−36

26. Zhang Y, Fu J, Wang K, Han X, Yan T, et al. 2022. The telomere-to-telomere gap-free genome of four rice parents reveals SV and PAV patterns in hybrid rice breeding. *Plant Biotechnology Journal* 20:1642−44

27. Qin P, Lu H, Du H, Wang H, Chen W, et al. 2021. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 184:3542−3558.e16

28. Zhang F, Xue H, Dong X, Li M, Zheng X, et al. 2022. Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. *Genome Research* 32:853−63

29. Wu D, Xie L, Sun Y, Huang Y, Jia L, et al. 2023. A syntelog-based pan-genome provides insights into rice domestication and de-domestication. *Genome Biology* 24:179

30. Long W, He Q, Wang Y, Wang Y, Wang J, et al. 2024. Genome evolution and diversity of wild and cultivated rice species. *Nature Communications* 15:9994

31. Zhou Y, Yu Z, Chebotarov D, Chougule K, Lu Z, et al. 2023. Pan-genome inversion index reveals evolutionary insights into the subpopulation structure of Asian rice. *Nature Communications* 14:1567

32. Ma X, Wang H, Yan S, Zhou C, Zhou K, et al. 2025. Large-scale genomic and phenomic analyses of modern cultivars empower future rice breeding design. *Molecular Plant* 18:651−68

33. Guo D, Li Y, Lu H, Zhao Y, Kurata N, et al. 2025. A pangenome reference of wild and cultivated rice. *Nature* 642:662−71

34. Wei X, Chen M, Zhang Q, Gong J, Liu J, et al. 2024. Genomic investigation of 18, 421 lines reveals the genetic architecture of rice. *Science* 385:eadm8762

35. Dan Z, Chen Y, Huang W. 2025. Structural variations contribute to genetic diversity and heterosis in rice. *bioRxiv* Preprint

36. Gu Z, Gong J, Zhu Z, Li Z, Feng Q, et al. 2023. Structure and function of rice hybrid genomes reveal genetic basis and optimal performance of heterosis. *Nature Genetics* 55:1745−56

37. Chen Z, Bu Q, Liu G, Wang M, Wang H, et al. 2023. Genomic decoding of breeding history to guide breeding-by-design in rice. *National Science Review* 10:nwad029

38. Singh PK, Rawal HC, Panda AK, Roy J, Mondal TK, et al. 2022. Pan-genomic, transcriptomic, and miRNA analyses to decipher genetic diversity and anthocyanin pathway genes among the traditional rice landraces. *Genomics* 114:110436

39. Shang L, Li X, He H, Yuan Q, Song Y, et al. 2022. A super pan-genomic landscape of rice. *Cell Research* 32:878−96

40. Wei X, Qiu J, Yong K, Fan J, Zhang Q, et al. 2021. A quantitative genomics map of rice provides genetic insights and guides breeding. *Nature Genetics* 53:243−53

41. Li X, Chen Z, Zhang G, Lu H, Qin P, et al. 2020. Analysis of genetic architecture and favorable allele usage of agronomic traits in a large collection of Chinese rice accessions. *Science China Life Sciences* 63:1688−702

42. Zhao Q, Feng Q, Lu H, Li Y, Wang A, et al. 2018. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics* 50:278−84

43. Huang X, Yang S, Gong J, Zhao Q, Feng Q, et al. 2016. Genomic architecture of heterosis for yield traits in rice. *Nature* 537:629−33

44. Huang X, Yang S, Gong J, Zhao Y, Feng Q, et al. 2015. Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nature Communications* 6:6258

45. Xie W, Wang G, Yuan M, Yao W, Lyu K, et al. 2015. Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proceedings of the National Academy of Sciences of the United States of America* 112:E5411−E5419

46. Huang X, Kurata N, Wei X, Wang ZX, Wang A, et al. 2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490:497−501

47. Laborte AG, Gutierrez MA, Balanza JG, Saito K, Zwart SJ, et al. 2017. RiceAtlas, a spatial database of global rice calendars and production. *Scientific Data* 4:170074

48. Yu Z, Chen Y, Zhou Y, Zhang Y, Li M, et al. 2023. Rice Gene Index: a comprehensive pan-genome database for comparative and functional genomics of Asian rice. *Molecular Plant* 16:798−801

49. Ohyanagi H, Ebata T, Huang X, Gong H, Fujita M, et al. 2016. OryzaGenome: genome diversity database of wild *Oryza* species. *Plant and Cell Physiology* 57:e1

50. Kajiya-Kanegae H, Ohyanagi H, Ebata T, Tanizawa Y, Onogi A, et al. 2021. OryzaGenome2.1: database of diverse genotypes in wild *Oryza* species. *Rice* 14:24

51. Reiser L, Harper L, Freeling M, Han B, Luan S. 2018. FAIR: a call to make published data more findable, accessible, interoperable, and reusable. *Molecular Plant* 11:1105−108

52. VanRaden PM. 2008. Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91:4414−23

53. Ahmadi Z, Kesbi FG. 2021. Assessing the performance of a novel method for genomic selectio: rrBLUP-method6. *Journal of Genetics* 100:24

54. Sakai H, Lee SS, Tanaka T, Numa H, Kim J, et al. 2013. Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant and Cell Physiology* 54:e6

55. Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, et al. 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6:4

56. Peng H, Wang K, Chen Z, Cao Y, Gao Q, et al. 2020. MBKbase for rice: an integrated omics knowledgebase for molecular breeding in rice. *Nucleic Acids Research* 48:D1085−D1092

57. Zhang C, Cui Y, Yuan C, Zhang F, Zhu Q, et al. 2025. Rice3kGS: a powerful web platform and database for large-scale genome selection. *Plant Communications* 00:101369 (inpress)

58. Wu H, Han R, Zhao L, Liu M, Chen H, et al. 2025. AutoGP: an intelligent breeding platform for enhancing maize genomic selection. *Plant Communications* 6:101240

59. Chen R, Deng Y, Ding Y, Guo J, Qiu J, et al. 2022. Rice functional genomics: decades' efforts and roads ahead. *Science China Life Sciences* 65:33−92

60. Si L, Chen J, Huang X, Gong H, Luo J, et al. 2016. OsSPL13 controls grain size in cultivated rice. *Nature Genetics* 48:447−56

61. Wang X, Zou B, Shao Q, Cui Y, Lu S, et al. 2018. Natural variation reveals that OsSAP16 controls low-temperature germination in rice. *Journal of Experimental Botany* 69:413−21

62. Anacleto R, Badoni S, Parween S, Butardo VM Jr, Misra G, et al. 2019. Integrating a genome-wide association study with a large-scale transcriptome analysis to predict genetic regions influencing the glycaemic index and texture in rice. *Plant Biotechnology Journal* 17:1261−75

63. Li Y, Miao Y, Yuan H, Huang F, Sun M, et al. 2024. Volatilome-based GWAS identifies OsWRKY19 and OsNAC021 as key regulators of rice aroma. *Molecular Plant* 17:1866−82

64. Xue W, Xing Y, Weng X, Zhao Y, Tang W, et al. 2008. Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. *Nature Genetics* 40:761−67

65. Lee DK, Chung PJ, Jeong JS, Jang G, Bang SW, et al. 2017. The rice OsNAC6 transcription factor orchestrates multiple molecular mechanisms involving root structural adaptions and nicotianamine biosynthesis for drought tolerance. *Plant Biotechnology Journal* 15:754−64

66. Farooq MA, Gao S, Hassan MA, Huang Z, Rasheed A, et al. 2024. Artificial intelligence in plant breeding. *Trends in Genetics* 40:891−908

67. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, et al. 2017. Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science* 22:961−75

68. Ma W, Qiu Z, Song J, Li J, Cheng Q, et al. 2018. A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* 248:1307−18

69. Wang K, Abid MA, Rasheed A, Crossa J, Hearne S, et al. 2023. DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Molecular Plant* 16:279−93

70. Ma X, Wang H, Wu S, Han B, Cui D, et al. 2024. DeepCCR: large-scale genomics-based deep learning method for improving rice breeding. *Plant Biotechnology Journal* 22:2691−93

71. Gao P, Zhao H, Luo Z, Lin Y, Feng W, et al. 2023. SoyDNGP: a web-accessible deep learning framework for genomic prediction in soybean breeding. *Briefings in Bioinformatics* 24:bbad349

72. Wu C, Zhang Y, Ying Z, Li L, Wang J, et al. 2024. A transformer-based genomic prediction method fused with knowledge-guided module. *Briefings in Bioinformatics* 25:bbad438

73. Wang H, Yan S, Wang W, Chen Y, Hong J, et al. 2025. Cropformer: an interpretable deep learning framework for crop genomic prediction. *Plant Communications* 6:101223

74. He K, Yu T, Gao S, Chen S, Li L, et al. 2025. Leveraging automated machine learning for environmental data-driven genetic analysis and genomic prediction in maize hybrids. *Advanced Science* 12:e2412423

75. Li J, Zhang D, Yang F, Zhang Q, Pan S, et al. 2024. TrG2P: a transfer-learning-based tool integrating multi-trait data for accurate prediction of crop yield. *Plant Communications* 5:100975

76. Ren Y, Wu C, Zhou H, Hu X, Miao Z. 2024. Dual-extraction modeling: a multi-modal deep-learning architecture for phenotypic prediction and functional gene mining of complex traits. *Plant Communications* 5:101002

77. Xiang Y, Xia C, Li L, Wei R, Rong T, et al. 2024. Genomic prediction of yield-related traits and genome-based establishment of heterotic pattern in maize hybrid breeding of Southwest China. *Frontier in Plant Science* 15:1441555

78. Montesinos-López A, Crespo-Herrera L, Dreisigacker S, Gerard G, Vitale P, et al. 2024. Deep learning methods improve genomic prediction of wheat breeding. *Frontiers in Plant Science* 15:1324090

79. Chen S, Du T, Huang Z, He K, Yang M, et al. 2024. The Spartina alterniflora genome sequence provides insights into the salt-tolerance mechanisms of exo-recretohalophytes. *Plant Biotechnology Journal* 22:2558−74

80. Song JM, Guan Z, Hu J, Guo C, Yang Z, et al. 2020. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nature Plants* 6:34−45

81. Cavalet-Giorsa E, González-Muñoz A, Athiyannan N, Holden S, Salhi A, et al. 2024. Origin and evolution of the bread wheat D genome. *Nature* 633:848−55

82. Benegas G, Batra SS, Song YS. 2023. DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences of the United States of America* 120:e2311219120

83. Benegas G, Albors C, Aw AJ, Ye C, Song YS. 2025. A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*

84. Levy B, Xu Z, Zhao L, Kremling K, Altman R, et al. 2022. FloraBERT: cross-species transfer learning withattention-based neural networks for gene-expression prediction. *Research Square* Preprint

85. Liu G, Chen L, Wu Y, Han Y, Bao Y, et al. 2025. PDLLMs: a group of tailored DNA large language models for analyzing plant genomes. *Molecular Plant* 18:175−78

86. Zhang Y, Chen G, Deng L, Gao B, Yang J, et al. 2023. Integrated 3D genome, epigenome and transcriptome analyses reveal transcriptional coordination of circadian rhythm in rice. *Nucleic Acids Research* 51:9001−18