www.pcletters.com

PLANT AND CROP LETTERS

Plant and Crop Letters (2027), 9: 11-33 Review Article

Genomics and Bioinformatics for Next-Generation Crop Improvement: From Pangenomes to AI-Driven Precision Breeding

Hafiz Muhammad Faraz Azhar[1]*, Zarlish Attique[2] and Zeeshan Ahmad Abbasi[3]

[1]Department of Biology, Howard University, Washington, DC, USA
[2]Charles Sturt University, Australia
*Corresponding author: azhar@bison.howard.edu

Article History: 26-169 Received: 30 June 2026 Revised: 17 Aug 2026 Accepted: 03 Sep 2026 Published Online: 2027

Citation: Azhar HMF, Attique Z, and Abbasi ZA, 2027. Genomics and Bioinformatics for Next-Generation Crop Improvement: From Pangenomes to AI-Driven Precision Breeding. Plant and Crop Letters, 9: 11-33.
https://doi.org/10.65822/j.pcl/2027.02

This is an open-access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

ABSTRACT

Crop genomics has been based on single reference genomes for over last two decades, which have been used for the development of markers, genome-wide association studies and genomic selection, but which have underrepresented structural variation, gene content and haplotype diversity within species. This review explores the impact of the convergence of pangenomics, advanced bioinformatics, artificial intelligence and high dimensional biological datasets on crop improvement. This review follows the evolution of long-read and telomere-to-telomere genome assemblies to pangenomes and super-pangenomes, population and comparative genomics, haplotype based breeding, and genomic prediction for traits and environments. It also assesses the integration of transcriptomic, proteomic, metabolomic, epigenomic, single-cell and spatial omics data with genomic data, as well as high throughput phenotyping and enviromics. A special focus is on machine learning, deep learning, explainable AI, foundation models and pangenome-guided genome editing for gene discovery, phenotype prediction, sequence design and de novo domestication. Applications for yield stability, abiotic stress tolerance, nutrient use efficiency, disease resistance, and biofortification are critically evaluated. The persistent challenges of computational bottlenecks, unequal genomic resources and the disconnect between genomic discoveries and breeder ready tools are also explored. This review focuses on the shift towards an integrated, prediction-based approach to the design of optimal genotype-phenotype combinations under various production environments.

***Keywords*:** Pangenomics; Super-pangenome; Genomic selection; Haplotype-based breeding; Artificial intelligence; Machine learning; High-throughput phenotyping; Genetic diversity; Crop improvement

INTRODUCTION

For over two decades, crop genomics has relied on single reference genomes as the coordinate system for marker discovery, GWAS, genomic selection, and comparative analysis (Della Coletta et al., 2021). Pangenome studies now reveal the scale of this gap: rice pangenomes have identified over 90,000 structural variants and 10,000 novel protein-coding genes (SaiVenkat et al., 2026); barley pangenomes show extensive copy-number variation at loci controlling disease resistance, architecture, and trichome development (Jayakodi et al., 2024); and a 22-genome wheat pangenome uncovered around 1.98 million non-redundant structural variants and 140,000 gene families (Jayakodi et al., 2024). Single-reference genomics thus systematically underrepresents breeding-relevant variation. The challenge extends beyond variant discovery to prediction. Genotype × environment (G×E) interactions are important for yield, quality and stress tolerance, and genetic effects determined in one environment may not carry over to another (Napier et al., 2023). The inclusion of G×E terms in genomic prediction models increases cross-environment accuracy and the inclusion of environmental covariates in addition to genomic data further increases out-of-environment prediction (Cooper & Messina, 2023; Napier et al., 2023). The focus in modern breeding is thus no longer on the detection of genetic differences, but on the prediction of the performance of a particular genotype under different environments. These pressures are leading to a shift from reference-based to data-integrated breeding, which is pangenome enabled. Pangenomes are no longer just descriptive catalogues, but are now used for functional applications such as the discovery of causal structural variants, novel genes, and trait relevant alleles in barley, wheat, rice, and other crops

(Jiao et al., 2025; Nie et al., 2026). The field has moved from data generation to data integration, with recent advances in long-read sequencing, graph-based representation of genomes, high throughput phenotyping, multi-omics and machine learning allowing integration of genetic variation with phenotypic and environmental information (Zanini et al., 2022; Zhu et al., 2026).

AI and machine learning do not supplant quantitative-genetics techniques but serves as a computation level over the top of them, with the ability to pick up subtle patterns from the high-dimensionality of datasets from various stages of research from both genomic and phenomic and environmental data (Guo & Li, 2023). In a wheat study, the authors joined genomic and phenotypic plasticity descriptors together by employing machine-learning to build a model that better predicted out-of-environment response for a 2025 wheat crop while decreasing the strain on the field-testing of wheats (McBreen et al., 2025). But predicting gains are not the same as direct genetic gains and as for models, often greater complexity does not necessarily lead to greater breeding value which is equally important to identify proof of concept from true gain. Crop improvement therefore is shifting from a marker-centric paradigm, using a single reference to an integrated, prediction-driven one where pangenomes, population genomics, haplotypes, genomic prediction, enviromics, phenomics, multi-omics, and genome editing all play an important role. The goal is no longer to catalogue diversity but to identify which variants and haplotypes drive desirable phenotypes, predict their performance across target environments, and deploy them via selection or targeted editing, pangenome-informed editing being a direct outgrowth of this shift, as expanded genome resources expose target sequences and allelic configurations invisible to a single reference (Jayakodi et al., 2024). This review examines how pangenomics, bioinformatics, AI, and complementary high-dimensional datasets are converging to reshape crop improvement. Rather than cataloguing emerging technologies exhaustively, we critically evaluate their contribution to the progression from genome representation and variant discovery to causal inference, prediction, selection, and genome design, assessing where they yield measurable gains in biological understanding, prediction accuracy, selection efficiency, or resource efficiency, and where limitations still preclude routine deployment. This review argues for a transition from genome-assisted to predictive crop improvement, where the objective is not merely identifying superior alleles but designing optimal genotype-phenotype combinations for target production environments.

The Changing Architecture of Crop Genomics

Crop genomics has undergone a major shift from short read to long read and ultra-long read sequencing (Zanini et al., 2022). Short reads remain accurate and cost effective for population-scale genotyping, but their limited length struggles to resolve repetitive sequences, duplications, complex structural variants, and haplotypes, limitations especially consequential in the large, repetitive, heterozygous, or polyploid genomes typical of crops (Hajieghrari & Nejati-Jahromi, 2025). PacBio HiFi and Oxford Nanopore (ONT) ultra-long reads offer complementary strengths: HiFi delivers high nucleotide level accuracy, while ONT spans long repetitive regions to produce more contiguous assemblies (Chen et al., 2023). In rice, ONT ultra-long sequencing yielded markedly more contiguous assemblies with fewer repeat-associated errors, while HiFi assemblies had fewer small sequence errors, a complementarity that argues for integrating the two technologies rather than treating them as alternatives (Wang et al., 2025).

Combined with Hi-C and improved assembly algorithms, long-read sequencing has made chromosome-scale and telomere-to-telomere (T2T) assembly increasingly feasible in plants (Chen et al., 2023). The T2T genome of elite watermelon line G42 assembled all 11 chromosomes as single-contig pseudomolecules with all 22 telomeres and 11 centromeres characterized, yielding a 369.3-Mb genome with 24,205 predicted genes, a substantially more complete reference than prior fragmented assemblies (Deng et al., 2022). Building on this, Zhang et al., 2024 generated T2T assemblies for 27 accessions spanning all seven *Citrullus* species, producing a super pangenome that added ~399.2 Mb and 11,225 genes relative to the existing reference and revealed extensive structural variation across cultivated and wild genomes (Zhang et al., 2024). Critically, this resource moved beyond description: disease-resistance loci from *C. amarus* and *C. mucosospermus* were successfully introgressed into cultivated watermelon, directly demonstrating how complete assemblies enable utilization of wild genetic resources (Zhang et al., 2024).

Assembly algorithms are also lowering the computational barriers to near-T2T reconstruction. Hifiasm (UL) integrates HiFi with ONT ultra-long reads, optionally incorporating Hi-C or trio data, to generate haplotype-resolved assemblies (Wang et al., 2025). In benchmarks across human and plant genomes, it produced more contiguous assemblies than HiFi-only methods at substantially lower computational cost than Verkko, and improved haplotype contiguity and completeness in an autotetraploid potato dataset (Cheng et al., 2024). Near-T2T assembly can then be used, though it can require more Hi-C or genetic data to complete polyploidy phasing and performance relies on high accuracy long reads, which is still not common for large, highly polyploid crop genomes (Cheng et al., 2024).

This technological change is not merely a question of assembly statistics: this change would also result in recovery of previously fragmented, collapsed, and missing regions, which would improve detection of structural variants including insertions, deletions, duplication, inversions, translocations, and changes in gene presence or absence that alter gene dosage, order, regulatory landscapes, and gene presence or absence (Schreiber et al., 2024). These types of variants make a greater and greater contribution to variation in yield, quality, disease resistance, and response to stress. For wheat, using PacBio HiFi sequencing allowed indels to be robustly discovered even at relatively low depth, but was sensitive to aligner and variant-caller selection (Zhang et al., 2024). Longer contigs are therefore not in themselves the beneficial aspects in this context, but the increased resolution of variation which can

be retrieved from a longer contig, achievable through long-read sequencing, that allow it to be related to function and breeding traits (Zhang et al., 2024).

The complete, haplotype-resolved assemblies also allow a more wide-spread change in how comparative genomics is conducted; the genomes of one accession representing a species can be replaced by comparisons of multiple, high quality complete and flow-celled genomes to characterize gene content, gene organization, species haplotypes, repeats, and regulatory complexity, especially when comparing the genome of cultivated varieties to the genetically distant and structurally complex genome of landraces and wild relatives (Khan et al., 2020). These genome collections form the foundation for pangenomes and super-pangenomes, extending analysis from individual genome architecture to the distribution, evolutionary history, and breeding value of variation across populations and species (Schreiber et al., 2024). Table 1 synthesizes the preceding discussion by consolidating the successive genomic resources into a single comparative framework. It delineates each resource's primary sequence representation, the class of variation it captures, and its principal application in breeding. It further summarizes the key advantage, principal limitation, and breeding transition associated with each resource, situating single-reference genomes, pangenomes, and super-pangenomes along a common evolutionary trajectory.

Table 1: Evolution of genomic resources and their contribution to crop improvement

Genomic resource/approach	Primary representation	Major variation captured	Main breeding application	Key advantage	Principal limitation	Breeding transition enabled	References
Single reference genome	One linear genome	SNPs, small indels relative to reference	Marker development, GWAS, genomic selection	Simple coordinate system and established pipelines	Reference bias; incomplete representation of SVs, PAVs and haplotypes	Genome-assisted breeding	(McBreen et al., 2025)
Long-read assembly	High-contiguity individual genome	Large SVs, repetitive regions, complex loci	Genome characterization and variant discovery	Resolves regions inaccessible to short reads	Usually represents one genotype	High-resolution genome analysis	(Wang et al., 2025)
Telomere-to-telomere assembly	Near-complete chromosome sequence	Repeats, centromeres, complex structural variation	Complete genome annotation and allele discovery	Minimizes assembly gaps	High computational and sequencing requirements	Complete genome representation	(Deng et al., 2022)
Linear pangenome	Multiple genomes represented relative to a reference	SNPs, indels, SVs, PAVs and novel genes	Allele discovery and comparative genomics	Expands representational diversity	Linear representation can remain reference-biased	Pangenome-enabled breeding	(Zhou et al., 2022)
Graph pangenome	Multiple genomes represented as a variation graph	Alternative haplotypes, SVs and PAVs	Graph-based genotyping, GWAS and marker discovery	Represents alternative genomic paths	Greater computational complexity	Variation-aware breeding	(Zhou et al., 2022)
Haplotype-resolved pangenome	Phased chromosome-level haplotypes	Phase-specific variants and allele combinations	Haplotype selection and breeding by design	Preserves allelic combinations	Requires high-quality phased assemblies	Haplotype-based predictive breeding	(Meena et al., 2025)
Super-pangenome	Cultivated plus wild/related genomes	Wild-relative variation, novel genes and structural diversity	Pre-breeding, introgression and allele mining	Broadens the accessible breeding gene pool	Functional validation and transfer remain necessary	Diversity-enabled precision breeding	(Khan et al., 2020)

Beyond the Single Reference Genome: Pangenomes and Super-Pangenomes

Once multiple high quality genomes are available, crop genomics shifts from a single representative accession toward representing diversity across individuals, the central principle of the pangenome, which captures the non-redundant gene content and sequence variation across a defined group of genomes (Zanini et al., 2022). Pangenomes do not eliminate reference bias entirely, but substantially reduce it by incorporating sequences absent from the conventional reference (Raza et al., 2026). Genes are typically classified as core genes, which are present in nearly all sampled individuals, or variable genes, which are present in only a subset of individuals and form part of the variable genome (Napier et al., 2023). Core genes tend to be enriched for conserved cellular and metabolic functions, while variable genes are more often linked to environmental adaptation, stress response, and defense, though these patterns vary with species, sampling, and assembly quality (Napier et al., 2023).

A pangenome's value depends as much on how variation is represented as on which genomes are included. The first generation of models involved various linear or reference-based assemblies, that is a primary reference sequence and an appending non-reference sequence expanding sequence space but still having a linear dominant coordinate system (Zhao et al., 2026). Both isoforms are translated into a graph-based pangenome, which enables reads to be mapped against various allelic states instead of just a single reference to enhance the discovery and genotyping of variants in divergent sequences (Cochetel & Cantu, 2026). A tomato graph pangenome composed of 838 tomato and 19 million variants increased the average estimated heritability of 20,323 molecular traits from 0.33 to 0.41, with the majority of the increase attributable to SVs, providing direct evidence that pangenome representation not only enhances genomic characterization but also downstream genetic analysis (Zhou et al., 2022).

Additional layers to add onto this are the haplotype-resolved pangenomes that maintain phase information between variants within the same chromosome, which is also important in breeding, as the impact of a variant could differ based on the genetic context of other variants in a chromosomal region, effects that are not detected by SNP-level analysis alone (Lian et al., 2025; Schreiber et al., 2024). Superior combination of alleles could be obtained by haplotype-aware analysis, but is challenging in highly heterozygous and polyploid crops to accurately phase alleles over long regions for genotype-phenotype interpretation (Cheng et al., 2024). Haplotype-aware pangenomics therefore is not a change from SNP-centric pangenomics, but rather a change in the genomic context of the modelling of variation (Vorbrugg et al., 2024) .

Most important extension for breeding is the super-pangenome concept that moves beyond a single species to the CWRs and related taxa (Raza et al., 2023). This is particularly useful, as domestication and breeding have altered the frequencies of alleles in genes, they have reduced the diversity at a few genetic places while maintaining a lot of diversity at other places, but did not simply remove the wild variation (Raza et al., 2026). Super-pangenomes provide a systematic framework for identifying and prioritizing this diversity for pre-breeding (Khan et al., 2020). The Citrullus super-pangenome illustrates this: 27 T2T assemblies spanning all seven Citrullus species added ~399.2 Mb and 11,225 genes relative to the cultivated watermelon reference, revealing domestication-associated structural variation and wild-derived gene families absent from the cultivated pool (Sun et al., 2026; Zhang et al., 2024). The study went beyond cataloguing multi disease resistance loci from *C. amarus* and *C. mucosospermus* were introgressed into cultivated *C. lanatus*, demonstrating the pathway from genome diversity to super-pangenome to candidate variation to pre-breeding.

Comparative population genomics extends this value beyond novel-sequence discovery. Analysis of wild and domesticated maize and rice revealed convergent selection at orthologous loci, for example, KRN2 in maize and its rice ortholog OsKRN2, both negative regulators of grain number under convergent selection (Jabeen et al., 2026; Wang et al., 2025). Knockout increased grain yield by 10% in maize and 8% in rice with no apparent major trade-offs, and genome-wide analysis identified 490 orthologous gene pairs under convergent selection across the two species evolution (Jabeen et al., 2026; Wang et al., 2025). Though based on comparative population genomics rather than a pangenome, this shows how expanded sampling across wild and domesticated gene pools can reveal conserved genetic solutions and candidate breeding targets (Raza et al., 2023).

Ultimately, super-pangenomes breeding value depends on connection to functional validation and deployment: a catalogue of novel genes or variants has limited use unless linked to phenotype, validated, and transferred into elite backgrounds (Nie et al., 2026). These resources can identify candidates for marker-assisted selection, haplotype-based selection, genomic prediction, introgression, and targeted editing, including by revealing alternative gene copies, promoter configurations, and natural alleles absent from the reference (Issah et al., 2026). But these applications still depend on accurate annotation, functional validation, transformation efficiency, and locus-specific biology (Issah et al., 2026). Pangenomes are therefore best understood not as autonomous breeding technologies but as genomic infrastructure that expands the search space for breeding decisions.

Bioinformatics Pipelines: Converting Raw Sequence into Breeding Knowledge

Producing a high quality genome assembly or pangenome is not enough, the breeding value is completely dependent on the ability to apply a computing pipeline to make sequence data interpretable (Yao et al., 2020). The initial steps in this pipeline involve assembly and annotation, which involve assembling reads into contigs, scaffolds, and then structurally and functionally annotated to identify genes, regulatory elements, and repetitive content. The decisions at this stage are then important and guide nearly all downstream analyses (Yao et al., 2020). For example, a marker discovery study in wheat, a large and highly polyploid genome, identified BWA-MEM for sequence alignment and SAMtools/mpileup for variant calling as an optimal combination. For wheat and potentially other large polyploid crops, the study suggested that extensive preprocessing should not be pursued if it substantially increases analytical complexity without providing a clear improvement in downstream performance (Wu et al., 2019). A helpful reminder that pipelines that do well on human and/or model organism data are not automatically suitable for crop genomes that have high ploidy and are rich in repeats (Wu et al., 2019; Yao et al., 2020).

In a more comprehensive benchmarking of aligners (SOAP2, Bowtie2, BWA-MEM) and variant callers (GATK HaplotypeCaller vs. SAMtools mpileup) on plant data, BWA-MEM consistently was able to align the highest number of reads with the highest accuracy while variant-caller precision and recall were obscured by genetic diversity, coverage and genome complexity (Wu et al., 2019). Thus, pipeline benchmarking should not be considered a cross taxa comparison that needs to be done only once, but should ideally be repeated on each crop, since the differences

in heterozygosity, ploidy and repeat structure will influence the results of the benchmarking runs (Yao et al., 2020; Zhang et al., 2024).

Variants, whether they are SNPs, indels, copy number variants, or structural variants, must be classified upon being detected, and each type of change after detection must be supported by different detection algorithms and different confidence thresholds, and each will entail different functional implications (Abdelwahab & Torkamaneh, 2025). Functional annotation then provides the context of the variants: associating variants with gene models, predicted effects on proteins and to regulatory areas then makes a raw variant call something much more useful: a plausible candidate in the form of a difference between the breeding and reference plants, instead of just a particular point of difference (Mahood et al., 2020). In recent years, AI tools for variant calling and interpretation have been able to assist this step and work to reduce accuracy, efficiency and scale issues that have traditionally been addressed using statistical tools, with its evaluation side by side with those methods being a very active methods research area (Sheikh et al., 2024).

As pangenomes and graph-based representations have become standard, computational infrastructure has scaled accordingly from comparing a handful of linear genomes to handling dozens or hundreds of accessions (Zhou et al., 2022). Dedicated graph-genome frameworks for building, indexing, mapping, and visualizing pangenome graphs have emerged to meet this demand, increasingly run at scale on cloud infrastructure (Zhou et al., 2022). This reflects a genuine shift in what breeding bioinformatics must handle: population-scale, multi-genome analysis as the routine expectation rather than single-genome workflows.

Databases and visualization platforms that store variant calls, annotations, and pangenome structure in breeder-accessible, query able formats are essential for translating pipeline output into line-advancement and crossing decisions (Naithani et al., 2023). Version-controlled pipelines with documented parameters and provenance tracking are now recognized best practice, since undocumented analysis choices are a major source of irreproducibility at genomic scale (Yang & Zhang, 2025). Finally, the scale of modern sequencing and pangenome data often terabytes across hundreds of resequenced accessions has made cloud and high-performance computing a practical necessity rather than a convenience, shifting the primary bottleneck in many programs from data generation to computational capacity and bioinformatics expertise (Raza et al., 2026).

Mining Crop Diversity Through Population Genomics

With large numbers of resequenced genomes across wild, landrace, and cultivated accessions, analytical focus shifts from describing individual genomes to reconstructing how a species diversity is structured and has moved over time (Raza et al., 2023). Population structure analysis reveals how variation is partitioned across geographic regions, subspecies, and breeding pools information that directly shapes GWAS and genomic prediction design, since unaccounted-for population structure is a leading source of false-positive associations in crop genetics (Khan et al., 2020; Shen et al., 2026).

Comparing wild and domesticated genomes within this framework identifies selective sweeps regions where diversity was sharply reduced as favorable alleles rose to high frequency during domestication or breeding (Schreiber et al., 2024). Resequencing studies show the scale of this footprint, comparisons of wild and cultivated rice have identified dozens of sweep regions spanning several percent of the genome, and whole-genome resequencing in maize found a comparable proportion under selection during domestication (Nie et al., 2026). These signals are not always straightforward to interpret, since domestication and later trait improvement reduce diversity in similar ways, and multiple, geographically distinct domestication events can produce overlapping, confounding sweep signals population-genomic inferences about domestication history should therefore be read cautiously rather than as a single clean narrative (Zanini et al., 2022; Zhao et al., 2026).

It has always been difficult to fully isolate domestication, and clearly, there are genomic signatures of gene flow between wild and cultivated plants (Shen et al., 2026). Grapevine genomics suggests that the domestication event has been single, and machine-learning based identification of soft sweeps reveals loci for synthesis of aromatic compounds as the targets for ongoing introgression highlighting not so much introgression as a historical background, but as an active, traceable source of trait-relevant variation (Lorenzo & Cervantes, 2026). Legumes show homology with similar signatures; genomic scans of the scarlet runner bean revealed introgression with introgressed chromosomal windows exhibiting both increased introgression flow from sympatric wild bean populations and introgression signals within cultivated populations, hinting an introgression response to local selection (Pratap et al., 2021; Bellucci et al., 2023). Not all introgression is good, as it may be the import of stress-tolerance or quality alleles, and alternatively, it may be the import of variants that are deleterious and have been purged over time, a dichotomy that population genomics increasingly can tell apart (Pratap et al., 2021).

Analysis of sweep detection and introgression mapping play a key role in the inference of rare and favorable alleles from whole-genome scans and may be used as tools to prioritize eventual functional follow-up in identified sweep regions or wild-derived introgression, prior to causal gene discovery (He et al., 2019). Finally, because the diversity that has been excluded from elite germplasm through intensive selection is most likely to be found in wild relatives and landraces, population-genomic characterization is needed to prioritize accessions most likely to contain such truly novel and non-redundant alleles (Wambugu & Henry, 2022).

Collectively, these tools now function less as retrospective evolutionary description and more as a forward-looking targeting mechanism for pre-breeding (Raza et al., 2023), sweep and introgression maps indicate which wild and landrace regions are most likely to harbor untapped, breeding-relevant variation, letting pre-breeding programs

prioritize introgression targets on population-genomic evidence rather than trait screening alone narrowing an enormous wild germplasm search space to a manageable set of high-probability genomic regions (Khan et al., 2020).

From Association to Causality: Finding Genes Behind Complex Traits

Identifying the genes underlying complex agronomic traits has traditionally relied on two complementary strategies: linkage-based mapping within controlled populations and association mapping across diverse germplasm (Roy et al., 2026). QTL-seq and bulked segregant analysis exemplify the linkage-based approach, sequencing pools of individuals with contrasting phenotypes from a segregating population and flagging regions with skewed allele frequency between pools as likely to harbor the causal locus, valued for speed and relative insensitivity to population structure, though limited to variation segregating within the biparental cross used (Wang et al., 2023; Baek et al., 2026).

GWAS instead exploits historical recombination across large, diverse panels, testing genome-wide markers against a trait across far more alleles than a single cross could capture (Scott et al., 2021). A GWAS on a large international wheat panel illustrates this scale, detecting several hundred QTLs across a dozen agronomic traits and multiple environments, with a meaningful subset resolved narrowly enough to implicate previously cloned genes and mapped QTLs as strong candidates (Varshney et al., 2021). Multi-locus GWAS methods, which model several associated loci jointly rather than independently, have become standard for this polygenic architecture, since they improve power and reduce false positives and negatives when a trait is controlled by many small-effect loci as most agronomic traits are (Kumar et al., 2024).

Association mapping has also moved beyond simple SNP-to-phenotype testing. eQTL and expression-based mapping link variants to gene expression rather than phenotype directly, helping distinguish regulatory from coding-sequence effects and revealing which gene's expression drives an association (Wang et al., 2025). This is particularly powerful combined with pangenome resources: a study integrating eQTL and GWAS on a rice super-pangenome uncovered key salt-tolerance regulators that conventional single-reference GWAS had missed, showing how expression evidence can prioritize among candidate genes within the same associated region (Vikas et al., 2022; Vorbrugg et al., 2024). Haplotype-based association analysis takes a related approach, testing co-inherited variant combinations rather than individual SNPs, increasing power when a functional effect depends on a specific combination rather than any single polymorphism, an increasingly common strategy as full haplotype data has become more accessible (Kumar et al., 2024).

Integrating pangenome resources into GWAS like pangenome-GWAS is among the field's more consequential recent shifts. Mapping reads directly to a graph-based pangenome rather than a linear reference allows structural variants and presence or absence variation to be genotyped alongside SNPs within the same framework (Shen et al., 2026). A GWAS meta-analysis using a graph-based rice pangenome from several hundred accessions demonstrated improved gene-mining efficiency over single-reference approaches, capturing variants a linear reference simply could not represent (Raza et al., 2026). This capacity to test structural variants as first-class association targets like structural-variant GWAS which is increasingly necessary, given how much functionally important variation (large insertions, deletions, copy-number changes) is missed by SNP-only panels built on a single reference (Issah et al., 2026). Recent wheat pangenome work illustrates this directly: a pangenome built from Chinese cultivars revealed structural variants shaped by dietary, cultural, and environmental pressures that a single reference had not captured (Jiao et al., 2025).

Statistical association, however sharply resolved, is never proof of causality, and the pipeline's final stages address exactly that gap. Fine mapping narrows a broad associated interval to a small number of strong candidates using recombination breakpoints, linkage disequilibrium decay, and expression or functional annotation evidence (Issah et al., 2026). Candidate-gene prioritization then applies further filters, tissue-specific expression, known pathways, sequence conservation, presence or absence patterns tied to the phenotype, a to rank candidates before functional validation (Deng et al., 2022; Khan et al., 2020). That validation, via gene editing, transgenic complementation, or independent segregating populations, remains the only way to confirm a causal gene, and its scarcity relative to the volume of published GWAS hits remains a persistent bottleneck in translating association genomics into breeding-ready targets (Kumar et al., 2024).

Haplotype-Based Breeding: Moving Beyond SNP-by-SNP Selection

Selecting on individual SNPs one at a time has an inherent limitation: complex traits are rarely governed by single markers acting independently, and SNP-by-SNP selection risks breaking apart co-inherited variant combinations that together produce a favorable phenotype while retaining unlinked, less useful ones (Sun et al., 2024). Haplotype-based breeding addresses this by treating the haplotype block variants inherited together due to limited recombination within a chromosomal segment, rather than the single marker, as the functional unit of selection (Meena et al., 2025). This approach has gained traction because it captures allele combinations and their interactions, which marker-by-marker selection tends to overlook, enabling more deliberate design of superior trait combinations (Bhat et al., 2021).

Determining which haplotypes are worth deploying requires linking haplotype identity directly to phenotype analysis. A study mining superior drought-responsive haplotypes in rice illustrates the scale now achievable: whole-genome resequencing of nearly 400 landrace and breeding-line genotypes, combined with candidate-gene association analysis, identified dozens of marker-trait associations for yield under drought stress, from which superior haplotypes at several loci were confirmed to correlate with higher drought-yield (Singh et al., 2024). Genotypes carrying these

haplotypes were proposed as donor material for drought-tolerant breeding programs, illustrating a direct pipeline from discovery to donor selection (Singh et al., 2024). A comparable study in pigeonpea similarly mined drought-responsive haplotypes across breeding lines, landraces, and wild species, finding landraces and wild accessions harbored substantially more distinct haplotypes than elite lines, consistent with the broader pattern of intensive selection narrowing haplotype diversity in elite germplasm (Sinha et al., 2020).

Once validated against phenotype, favorable haplotypes can be tracked through a breeding pipeline via haplotype-specific markers, much like conventional marker-assisted selection but at the haplotype level (Meena et al., 2025). This offers a practical advantage beyond improved trait prediction: tracking co-inherited blocks rather than isolated markers can help circumvent linkage drag, where a desirable allele remains linked to nearby undesirable alleles across generations of backcrossing (Bhat et al., 2021). Extending this logic, deliberately stacking multiple validated superior haplotypes across genomic regions and traits within a single line, designing elite haplotype combinations which is a more proactive application than simply selecting existing haplotypes, shifting haplotype breeding from a selection tool toward a design tool (Zhu et al., 2026).

Climate relevance has become an explicit focus of recent literature, driven by the recognition that climate resilience typically depends on multiple interacting loci governing overlapping stress responses, unlike many single-gene traits (Cooper & Messina, 2023). Haplotype-based pangenomics, genome collections representing each haplotype across multiple strains which has been proposed as a blueprint for identifying and deploying climate-adaptive variation, on the reasoning that such traits are better captured by co-inherited stress-responsive allele combinations than by isolated markers (Meena et al., 2025). This connects directly to the pangenome and super-pangenome resources discussed earlier as haplotype-resolved pangenomes become available for more crops, the practical bottleneck is shifting from haplotype discovery itself toward the harder problem of validating which candidate haplotypes actually confer resilience under real, multi-stress field conditions rather than the controlled conditions in which they were first characterized (Zhu et al., 2026).

Genomic Selection Enters the Predictive Breeding Era

Conventional genomic prediction rests on a simple premise: genome-wide markers, rather than a handful of major-effect loci, estimate an individual's cumulative genetic value by relating marker genotypes in a training population to observed phenotypes, then applying that trained model to predict performance in untested, genotyped individuals before field phenotyping (Jiao et al., 2025). This distinguishes genomic selection from marker-assisted selection and GWAS, which emphasize major-effect loci; genomic selection instead captures the cumulative contribution of many small-effect loci, the typical architecture of complex traits like yield, stress tolerance, and disease resistance (Bhat et al., 2021; Wang et al., 2026).

Since the early developments of GBLUP and Bayesian regression methods (Bayes A, Bayes B, Bayesian LASSO) statistical machinery has diversified. Genomic breeding value prediction based on a genomic relationship matrix is straightforward and simple (Ma et al., 2024), whereas a Bayesian approach allows for varying marker effects, under different priors regarding their size and distribution, and usually outperforms genomic relationship matrix methods in the case of a few moderate and large-effect loci and a strong polygenetic background (McGowan et al., 2021). When both are applied together with machine learning to predict both yield and quality in cereals and forages, there has been no consistent trend that the one can uniformly outperform the other; instead, model selection should be based on the particular type of data set and the trait in question, and should be made on an ad hoc basis and supported by experimental evidence (Mohd Ali et al., 2026).

One of the main strands of recent research is that prediction can move outside of the scope of individual traits and individual environments (He et al., 2019). Multi-trait genomic prediction exploits inter-trait correlation to improve prediction accuracy by using more than a single trait, or to predict a trait with low heritability or a non-directly measurable trait, by using some other correlated response with a higher heritability (Zhou et al., 2022). Gain is reported to be significant, and reasonably similar, with traits of white clover forage improved by about 25-100% when compared to the single-trait models, and wheat end-use quality improved by about 5-8% when compared to single-trait GBLUP/Bayesian baselines (Ehoche et al., 2025).

Multi-environment genomic prediction addresses a complementary problem, explicitly modeling genotype-by-environment and genotype-by-year interactions since a genotype's relative ranking frequently shifts across sites and seasons (Yan & Wang, 2026). Combining both extensions into multi-trait multi-environment (MTME) models has become common in cereal breeding, as the two information sources are largely complementary rather than redundant (Puglisi et al., 2026). Work on advanced winter wheat lines across ten site-year environments shows the practical relevance directly as accuracy advantages from multivariate models on diverse, unrelated panels do not automatically transfer to the more closely related material in advanced breeding trials, so MTME performance needs validation in the specific breeding context of deployment rather than assumption from panel-level studies (Singh et al., 2024; Z. Zhang et al., 2024). Incorporating G×E directly into prediction models has consistently been shown to improve predictive ability over models ignoring environmental interaction, though the benefit shrinks or disappears in low-heritability environments where signal-to-noise is already poor (Napier et al., 2023).

Pangenome-informed genomic selection extends this framework by incorporating structural variants and presence/absence variation genotyped against a pangenome rather than relying solely on single-reference SNPs (Bayer et al., 2022). This is not a marginal refinement: pangenome-based marker sets have increased heritability

estimates and improved prediction accuracy in multiple crops relative to single-reference genotyping, since structural variants captured only via pangenome genotyping often explain variance that SNP-only models leave as missing heritability (Puglisi et al., 2026).

Genomic prediction rarely operates in isolation. Integrated frameworks combining pedigree records, genomic markers, and high-throughput phenomic measurements deliver greater prediction accuracy and higher genetic gain per unit time, at reduced cost, relative to marker data alone (McGowan et al., 2021). Sensor- and image-derived phenomic data are valuable both as an additional predictor layer and as a way to extend selection into traits and growth stages too costly or slow to phenotype manually at scale; greenhouse or early-stage phenotyping integrated with genomic prediction has enabled earlier selection decisions, even at moderate accuracy, when combined with techniques that shorten generation interval (Guo & Li, 2023; Issah et al., 2026). Table 2 integrates the analytical stages discussed above into a unified bioinformatics-to-breeding pipeline. It specifies, for each stage, the requisite input data, the major computational task performed, and the resulting biological output, spanning genome assembly through precision breeding. It further indicates the specific breeding decision enabled at each stage, thereby illustrating the sequential conversion of raw sequence data into actionable breeding value.

Table 2: Integrated bioinformatics-to-breeding pipeline

Analytical stage	Input data	Major computational task	Biological output	Breeding decision enabled	References
Genome assembly	Long/short sequencing reads	Assembly and polishing	Chromosome-scale genome	Reference construction	(Kopalli et al., 2025)
Genome annotation	Assembled genome	Gene and regulatory annotation	Gene models and functional elements	Candidate gene identification	(Zanini et al., 2022)
Variant discovery	Reads + reference genome	SNP/indel/SV calling	Genomic variation	Marker development	(Yao et al., 2020)
Pangenome construction	Multiple assemblies	Genome comparison and graph construction	Core/accessory genome and PAVs	Expanded allele discovery	(Cochetel & Cantu, 2026)
Population genomics	Population genotypes	Diversity, structure and selection analysis	Population-specific variation	Donor and allele selection	(Adak et al., 2024)
Association analysis	Genotypes + phenotypes	GWAS/QTL analysis	Trait-associated loci	Candidate region identification	(Kopalli et al., 2025)
Fine mapping	Dense markers + phenotypes	Recombination and association analysis	Candidate genes/variants	Target prioritization	(Yao et al., 2020)
Multi-omics	Genomic + transcriptomic/proteomic/metabolomic data	Cross-layer integration	Regulatory and molecular mechanisms	Functional candidate prioritization	(Ali et al., 2026)
Haplotype analysis	Phased genotypes	Haplotype construction and association	Superior allele combinations	Haplotype selection	(Meena et al., 2025)
Genomic prediction	Genome-wide markers + phenotypes	Prediction model training	Genomic estimated breeding values	Early selection	(Adak et al., 2024)
G×E prediction	Genotypes + multi-environment phenotypes	Multi-environment modelling	Environment-specific breeding values	Target-environment selection	(Cooper et al., 2022)
Phenomics integration	Genomics + HTP data	Image/sensor feature extraction and prediction	Quantitative phenotypic descriptors	Earlier and higher-throughput selection	(Ali et al., 2026)
AI-assisted prioritization	Genomic, phenomic and environmental data	ML/DL/XAI	Candidate features and predictions	Decision support	(Abdelwahab & Torkamaneh, 2025)
Functional validation	Candidate genes/variants	Experimental validation	Causal evidence	Validated breeding targets	(Salotti et al., 2022)
Precision breeding	Validated targets + editing systems	Targeted genome modification	Designed allele/genotype	Direct deployment	(Jabeen et al., 2026)

AI and Machine Learning as the New Bioinformatics Layer

If genomic selection marked the first major statistical shift in breeding decision-making, the current wave of machine learning and deep learning is arguably a larger second one, applied across nearly every stage of the genomics to breeding pipeline (Cooper & Messina, 2023). In gene and allele discovery, machine learning builds on the association and fine-mapping methods discussed earlier but adds the capacity to detect non-linear and interaction effects between markers that conventional statistical tests cannot capture, a natural complement to, rather than replacement for, GWAS and QTL-mapping (Wang et al., 2026).

Deep learning's contribution to phenotype prediction has followed a different trajectory, driven by architectures capable of modeling complex, non-additive genetic effects from sequence and marker data (Abdelwahab & Torkamaneh, 2025; Wang et al., 2026). Crop former, a deep learning framework combining convolutional neural networks with self-attention, illustrates both the promise and a recurring limitation (Wang et al., 2025). It showed strong performance predicting phenotypic traits across five major crops, but its developers noted persistent poor interpretability and comparatively low robustness relative to simpler statistical models, a tension between predictive power and interpretability that the next generation of tools has tried to address directly (Guo & Li, 2023).

Explainable AI (XAI) emerged to solve exactly this problem. Because deep neural networks interact non-linearly across millions of parameters, reconstructing their internal decision-making is inherently difficult (Pelosi et al., 2025); XAI methods clarify which input features drive a given prediction, aiding detection of bias or confounding and translating model output into usable biological insight rather than an unexplained score (Alqudah & Moussavi, 2025). Concrete applications illustrate this: explanations from a random forest model identified specific genomic regions including a seed-development gene as drivers of predicted almond shelling traits (Novielli et al., 2024), while a deep learning model on tomato ripening-stage expression data used gradient-based explanation to decode regulatory-element variation, with ~96% of flagged transcription-factor-binding motifs subsequently validated experimentally (Akagi et al., 2022). Beyond scientific insight, explainability supports faster development of usable genetic markers and builds breeder confidence in AI-based decisions arguably a precondition for real-world adoption regardless of raw predictive accuracy.

The most recent, structurally different development is the emergence of foundation models and large biological models trained on massive, often cross-species genomic or proteomic datasets rather than a single crop or trait, then fine-tuned for downstream tasks. DNA language models such as AgroNT, trained on large edible-plant genomic corpora (Mendoza-Revilla et al., 2024), and cross-species DNA foundation models now modeling plant genomes at single-nucleotide resolution across diverse taxa (Zhai et al., 2025), extend a paradigm first developed for human genomics into plant-specific contexts.

This same paradigm underlies generative AI in sequence and protein design. DNA language models and diffusion-based generative models can design realistic regulatory DNA sequences and predict features like promoter strength directly from sequence (Sheikh et al., 2024), while protein language models support rational protein design with function, expression, and codon usage optimized computationally before wet-lab synthesis (Yang & Zhang, 2025). This marks a meaningful shift from earlier ML applications built to predict or classify existing variation, toward generative models that propose novel sequences and designs directly.

Beyond sequence-level modeling, knowledge graphs are being built to organize relationships between genes, traits, and environmental variables across large bodies of literature and experimental data, providing a structured way to query accumulated knowledge that no single dataset can represent alone (Cochetel & Cantu, 2026). Recent examples include large-scale plant literature knowledge graphs mined from tens of thousands of published articles, making dispersed gene-function and trait-association findings queryable without manual review of every candidate gene (Zhou et al., 2022).

Taken together, these developments are increasingly framed as components of integrated AI-assisted breeding decision-support systems rather than isolated tools. Multimodal foundation models processing natural language alongside genomic sequence and other biological data are being proposed as infrastructure spanning the full breeding chain from phenotype-gene association and environmental adaptability analysis to germplasm data mining positioning AI not as a standalone tool but as an integrative layer connecting genomics, phenomics, and genome editing into a coherent decision-support framework (Bayer et al., 2022). This framing is itself significant: it implies these tools eventual usefulness will be judged less by individual benchmark accuracy and more by how well they interoperate with the genomic, phenomic, and editing pipelines already described.

Panomics: Connecting Genotype to Phenotype

Everything discussed so far has operated primarily at the DNA sequence level genomes, pangenomes, variants, and models predicting phenotype directly from genotype (Kumar et al., 2024). Yet the genome is a static blueprint, and much of the biological distance between genotype and phenotype is filled by intermediate molecular layers that genomic data alone cannot capture (Novielli et al., 2024). This is the basic motivation for panomics: integrating genomics, transcriptomics, metabolomics, and other omics platforms allows researchers to systematically characterize the biological elements that actually give rise to a trait, rather than inferring them indirectly from statistical association (Singh et al., 2024).

Transcriptomics, typically via RNA sequencing, captures which genes are actively expressed and at what level in a given tissue and condition in which a dynamic readout static sequence cannot provide, since a structurally intact gene contributes nothing to a phenotype if it isn't expressed under relevant conditions (Bergen et al., 2020). Proteomics extends this further, characterizing protein abundance, sequence, and post-translational modification directly via mass spectrometry, important because transcript abundance does not always correlate cleanly with functional protein abundance (Guo et al., 2026). Metabolomics represents the biochemical phenotype most directly, since metabolite concentrations sit closest to observable physiological outcomes; integrating it with genomics and transcriptomics has proven particularly powerful for dissecting mechanisms underlying stress resistance and quality traits (Ćavar Zeljković et al., 2026).

Epigenomics adds a regulatory dimension sequence alone cannot explain: chemical modifications to DNA and histones like methylation, acetylation, that alter expression without changing sequence, profiled via whole-genome bisulfite sequencing and ATAC-seq (Ohishi & Au Yeung, 2026). This matters for breeding because heritable epigenetic states can influence trait expression independently of genetic variation at a locus, meaning DNA-sequence-only prediction models may miss a portion of heritable variation (Kumar et al., 2024). Ionome and microbiome data extend the framework beyond the plant itself: ionomic profiling characterizes mineral nutrients and trace elements relevant to nutrient-use efficiency and biofortification (Alzate Zuluaga et al., 2024), while microbiome data captures root- and tissue-associated microbial communities that influence nutrient uptake, stress tolerance, and disease resistance, variation outside the plant genome that still shapes realized phenotype (Winters et al., 2020).

Machine learning has become the practical backbone connecting these disparate layers, since the dimensionality and heterogeneity of combined multi-omics data, spanning millions of genomic variants alongside transcript, protein, and metabolite measurements which is generally intractable for classical statistical integration alone (Guo & Li, 2023; Alqudah & Moussavi, 2025). Recent applications show the payoff: a machine learning framework integrating metabolomic and transcriptomic data helped unveil molecular mechanisms of salt tolerance, and metabolic-marker-assisted genomic prediction models improved accuracy in hybrid breeding relative to genomic data alone (Mahood et al., 2020). However, recent multi-omics work in perennial crops offers an important caveat: the added predictive value of combining omics layers is highly context-dependent, shaped by tissue type, trait genetic architecture, and genuine biological complementarity between layers, multi-omics integration should be validated empirically as a trait- and context-specific tool, not assumed to outperform single-layer prediction universally (Sheikh et al., 2024; H. Wang et al., 2025).

Single-Cell and Spatial Omics: Resolving Traits at Cellular Resolution

All omics approaches discussed so far share a common limitation. They are typically applied to bulk tissue, averaging molecular signals across thousands of cells of different types (Cochetel & Cantu, 2026). This can obscure decisive information, since a stress response or developmental process is often driven by a small, specific subset of cells that bulk sequencing dilutes into an undifferentiated average (Cooper & Messina, 2023). Single-cell RNA sequencing was developed to resolve this, profiling expression at individual-cell resolution; first applied in plants in 2017 via isolated rice mesophyll protoplasts, establishing feasibility despite recovering data from only a small number of cells (Bao et al., 2025). Single-nucleus sequencing offers a related approach better suited to plant tissues that resist protoplast isolation, profiling RNA from isolated nuclei rather than requiring intact dissociated cells (Nobori, 2025).

Single-cell resolution enables reconstruction of cell-type-specific regulatory networks, genes and regulatory elements whose activity differs systematically across cell types within a tissue, rather than assuming a uniform tissue-wide program (Phua et al., 2026). This is a conceptual shift from the bulk, tissue-averaged panomics networks discussed earlier: cell-resolved networks make it possible to identify which specific cell types actually drive a trait-associated expression signature (Ćavar Zeljković et al., 2026).

Spatial transcriptomics extends single-cell resolution by preserving each measurement's physical location within tissue architecture information single-cell approaches, by dissociating tissue, necessarily lose (Phua et al., 2026). This matters considerably in plants, where strongly layered tissue architecture makes spatial context decisive for interpreting transcriptional changes; spatial methods preserve the coordinates needed to infer intercellular communication, developmental boundary effects, and the spatial organization of stress or defense responses (Jabeen et al., 2026). Recent platform advances have moved spatial capture from coarse, multi-cell barcoded spots toward near single-cell resolution, substantially increasing achievable granularity (Ehoche et al., 2025). Spatial metabolomics and proteomics extend the same logic to metabolites and proteins, though these remain less developed in plants than spatial transcriptomics, constrained by limited antibody availability and plant-specific sample-preparation challenges (Wang et al., 2026).

Mapping stress responses within tissues is where single-cell and spatial approaches have shown their clearest breeding-relevant value (Cooper & Messina, 2023). A study combining single-cell RNA sequencing and spatial transcriptomics in rice roots (soil vs. gel conditions) found substantial expression changes concentrated in outer root cell types, linked to nutrient homeostasis, cell wall integrity, and defense, differences bulk root transcriptomics would likely have obscured (Phua et al., 2026). Disease mapping provides an equally direct example: a combined spatial and single-nucleus study of soybean infected with Asian soybean rust identified two distinct cell states in the defense response (Wang et al., 2026), and capturing both host and pathogen transcriptomes within the same spatial experiment allowed distinguishing expression within infected tissue regions from surrounding uninfected areas, a distinction bulk RNA-seq could not resolve (Hu et al., 2025).

The broader breeding relevance lies less in what has already been mapped than in what becomes newly measurable for traits that have historically resisted mechanistic dissection via bulk approaches (Ali et al., 2026). Root architecture, which depends on coordinated but distinct behavior of specific root cell types; disease resistance, where localized immune response depends on cells in direct pathogen contact mounting a local rather than systemic defense (Ali et al., 2026); reproductive development, which unfolds through cell-type-specific programs in floral tissue; and abiotic stress adaptation, where cell-specific responses to heat, drought, or salinity have proven more mechanistically informative than whole-tissue averages (Li et al., 2026). Single-cell and spatial omics in plants nonetheless remain considerably less mature than in animal systems, constrained by isolation difficulties, metabolite interference during

processing, and comparatively limited plant genomic reference resources, the breeding applications above should be read as a promising trajectory rather than tools already validated at breeding-program scale (Guo et al., 2026).

Genomics Meets Phenomics and Enviromics

Genomic prediction is only as good as the phenotypic data used to train it, and phenotyping has historically been the slower, more labor-intensive half of the genotype-to-phenotype pipeline (Diakite et al., 2026). High-throughput phenotyping (HTP) closes this gap, using imaging, spectral sensing, and automated platforms to measure traits non-destructively across large populations at a scale manual scoring cannot match (Ali et al., 2026). Field-based HTP platforms capture plant performance directly under realistic agronomic conditions, providing more accurate insight into G×E interactions than data from controlled-environment facilities (Elliott et al., 2022).

UAV and drone-based sensing has become one of the most widely adopted HTP modalities, scaling well across large trials while remaining cheaper and faster to deploy than ground-based phenomobiles (Ćavar Zeljković et al., 2026). A winter wheat study combining UAV-assisted HTP with deep learning showed that integrating UAV-derived traits into phenomic and multi-trait genomic prediction models improved accuracy for grain yield, test weight, and protein content over genomic data alone (Tanin et al., 2026). Similar gains appear in maize, where drone-derived vegetation indices summarized across the season via functional principal component analysis enhanced phenomic and genomic prediction of yield and plant height across years and environments, though matching biomarker measurement time points consistently across years and environments remains a real obstacle for models intended to generalize across seasons (Adak et al., 2024).

Image-based trait extraction underlies much of this sensing, converting multispectral, thermal, or LiDAR imagery into quantitative measurements like canopy cover, biomass, or stress indices. Ground-based systems like tractor-mounted sensors and phenomobiles, use the same sensor types at higher spatial resolution and proximity, trading scalability for precision relative to aerial platforms (Adak et al., 2024). Machine vision and AI-based image analysis increasingly sit at the center of this pipeline, detecting and quantifying stress signals from raw imagery that would otherwise require slow, subjective manual scoring (Guo & Li, 2023).

Phenotypic data alone does not fully explain genotype performance without accounting for environment, motivating enviromics: systematic quantification of weather, soil, and management as structured, analyzable covariates rather than an unmeasured categorical trial label (Guo & Li, 2023). Genomics, phenomics, and enviromics together form what has been described as the G–P–E triangle in plant breeding; while genomic and phenomic data have each received substantial methodological attention, the interaction between phenomics and enviromics remains comparatively underexplored (Li et al., 2026). An identified gap now framed as phenomics–enviromics crosstalk" deserving its own coordinated data strategy (Li et al., 2026).

This enviromic layer becomes essential once G×E interaction is treated not as a statistical nuisance but as a genuine deployment problem (Liang et al., 2026): breeders must rank genotypes across a target population of environments differing in weather, soils, management, and stress timing, and for traits like yield, flowering time, and stress adaptation, this interaction determines whether a model's ranking holds up under commercial deployment across variable conditions (Yan & Wang, 2026). Modeling genotype × environment × management (G×E×M) interactions extends this further by incorporating management decisions like planting density, irrigation timing, input regimes as a third interacting factor, since the optimal genotype for an environment can itself depend on the management under which it is grown, a distinction pure G×E models cannot capture (Cooper et al., 2022).

Linking genomic, phenomic, and environmental data into a single jointly analyzable framework is the practical culmination of this integration (Adak et al., 2024). The principal bottleneck constraining multi-environment genomic prediction is no longer marker density but the availability and quality of structured environmental covariate data and the sophistication of models integrating all three data types jointly (Gogna et al., 2026). The eventual payoff is predicting genotype performance under future climate conditions not yet observed in a program's historical trial network, extrapolating trained genomic-enviromic-phenomic models to project performance under warmer or more variable conditions (Liang et al., 2026). This capability depends directly on the quality of historical G×E×M training data and remains an active area of methodological development rather than a mature, routinely deployed tool.

From Gene Discovery to Genome Design: CRISPR and Precision Editing

Once a candidate gene is identified and functionally validated through the association and fine-mapping pipeline described earlier, the natural next step is to modify it directly rather than wait for the corresponding allele to arise via natural variation, introgression, or crossing (Bayer et al., 2022; Bergen et al., 2020). This is the logic connecting gene discovery to genome design, and it depends first on bioinformatic identification of editing targets, requiring precise sequence-level knowledge of the target locus, its regulatory context, and any polymorphisms affecting guide binding, drawing directly on the high-resolution genomic and pangenomic data described earlier (Khan et al., 2020).

CRISPR/Cas systems remain the dominant editing platform, though the toolkit has diversified well beyond Cas9. Compact nucleases including Cas9, Cas12, Cas13, and newer ultra-compact variants offer a range of PAM compatibilities, editing outcomes, and delivery constraints suited to different targets and transformation systems (Kumar et al., 2025). Base editing and prime editing have extended what's achievable at the sequence level: prime editing functions as a search-and-replace mechanism introducing precise substitutions, insertions, or deletions without a double-strand break, and has been used in crops to replace a 57–186 bp sequence with a defined 72 bp target (Hillary

& Ceasar, 2022). However, prime editing in plants faces a persistent bottleneck of low, variable editing efficiency across species and target sites identified by a recent systematic review as the major factor still constraining broader adoption (Hillary & Ceasar, 2022).

Multiplex genome editing addresses a different limitation: most agronomically important traits are polygenic, poorly matched to editing one gene at a time (Lin et al., 2026). Multiplex approaches allow simultaneous modification of many loci in a single transformation event (Lin et al., 2026). Recent demonstrations include multiplex editing of a dozen agronomic genes in rice combined with rational crossing to select plants with improved heading time and architecture without sacrificing yield or disease resistance, and a related effort improving seven agronomic traits simultaneously in a different rice line while preserving salt tolerance (Wei et al., 2024). Newer Cas12a-based systems push this further, with engineered orthologs enabling stable, simultaneous editing of six to eight loci within a single vector across multiple crop species (Cheng et al., 2026). The same multiplex logic extends into regulatory regions, using CRISPR activation and interference to modulate expression directly, and targeted modification of promoters and other regulatory elements, extending editing into the tunable, quantitative variation already discussed in the context of eQTL mapping (Cheng et al., 2026; Wang et al., 2025).

Pangenome-guided genome editing connects directly back to the pangenome resources introduced earlier, since accurate guide RNA design and off-target prediction depend on knowing the full range of sequence variation at and around a target locus across a breeding population, not just a single reference genome (Tay Fernandez et al., 2022). Variant-aware off-target prediction tools now explicitly incorporate pangenome data, since relying on a single linear reference risks missing true off-target sites present only in non-reference haplotypes, or flagging false concerns at sites a given breeding line doesn't actually contain (Wei et al., 2024).

Perhaps the most conceptually striking application is de novo domestication: using genome editing to install a defined set of domestication traits like reduced shattering, altered architecture, non-dormancy, directly into a wild or semi-domesticated species, bypassing the generations of selection historical domestication required (Zeibig et al., 2024). A published roadmap for de novo domestication of wild allotetraploid rice illustrates this concretely, integrating genome sequencing, transformation protocols, and multiplex editing to target domestication genes while deliberately preserving the wild species existing stress-tolerance and disease-resistance traits (Yu et al., 2021), expanding the usable genetic pool by domesticating wild relatives directly, rather than relying solely on introgression into an already domesticated background, as the pre-breeding and super-pangenome approaches discussed earlier generally do (Yu et al., 2021).

Trait-Focused Applications in Crop Improvement

Pangenomes, GWAS, haplotype breeding, genomic selection, multi-omics, and precision editing ultimately serve as infrastructure for specific, agronomically consequential traits (Khan et al., 2020). Yield and yield stability remain the central target of most breeding programs, but the framing has shifted from maximizing performance under narrow, favorable conditions toward stabilizing performance across the broader, variable conditions a variety will actually encounter commercially, precisely why multi-environment, G×E-aware genomic prediction has become as central to yield breeding as raw prediction accuracy (Ali et al., 2026; Baek et al., 2026).

Drought, heat, and salinity tolerance are frequently discussed together because they share physiological mechanisms, co-occur under field conditions, and increasingly interact in ways that complicate breeding for any single stress (Ali et al., 2026). Combined abiotic stresses salinity alongside heat or nutrient imbalance, for instance present a genuinely harder challenge than any isolated stress, since plants must withstand compounded physiological disruption; recent literature argues breeding and panomics strategies should target realistic stress combinations rather than treating each as a separate objective (Cooper & Messina, 2023). Genomics-assisted breeding has nonetheless delivered concrete field outcomes: introgressing characterized salt-tolerance genes into locally adapted Bangladeshi wheat lines increased grain yield by 10–20% on high-salinity soils, a direct pathway from gene discovery to farmer-relevant gain (Ali et al., 2026). Pangenome resources have added further value by revealing stress-responsive genes absent from single references altogether, pangenome analysis of radish identified a previously uncharacterized gene carrying diverse stress-responsive regulatory elements that single-reference scanning had missed (Liu et al., 2026).

Nutrient-use efficiency has attracted growing attention as fertilizer costs and nutrient-runoff consequences become harder to ignore, with genomic tools increasingly framed as enabling lower-input crops that maintain productivity while reducing both costs and runoff (Ali et al., 2026). Disease and insect resistance remains one of the most mature application areas, since resistance loci are often amenable to clean marker-trait association, drawing directly on the GWAS, haplotype-based selection, and multiplex-editing approaches for stacking resistance genes described earlier (Jabeen et al., 2026).

Nutritional biofortification illustrates a trait category where breeding intent, target trait, and delivery technology are especially tightly linked (Kumar et al., 2025). Breeding strategies have progressed from phenotypic selection to genomic and marker-assisted methods, successfully raising iron, zinc, and provitamin A content in rice, cassava, beans, and wheat (Tessema et al., 2026). Several hundred biofortified varieties across a dozen staple crops had been released in over forty countries, reaching well over a hundred million consumers including many smallholder households, a scale of deployment that stands out from the more experimental technologies discussed elsewhere, having moved well past proof-of-concept into large-scale, multi-country adoption (Gaikwad et al., 2020). CRISPR-based editing has extended this further, with precise edits targeting anti-nutrient biosynthesis genes reducing phytic

acid content by up to 25% in soybean, wheat, and rice, improving mineral bioavailability that phytic acid would otherwise limit (Cheng et al., 2026; Jabeen et al., 2026).

Quality and post-harvest traits like grain or fruit composition, processing characteristics, shelf life, sensory attributes, have historically received less genomic attention than yield or major stress traits, largely because they are harder and more expensive to phenotype at scale (Wu et al., 2026). The high-throughput and image-based phenotyping approaches discussed earlier are increasingly closing this gap, making quality traits considerably more tractable for genomic selection and GWAS than a few years ago (Baek et al., 2026). Climate resilience functions less as a discrete trait than as an integrative framing drawing together drought, heat, and disease resistance under a shared objective of stable performance as conditions shift unpredictably (Cooper & Messina, 2023).

The Neglected Frontier: Genomics of Orphan and Underutilized Crops

Nearly every advance, from T2T assembly through pangenome-informed genomic selection to AI-assisted precision editing, has been developed and validated overwhelmingly in a small handful of globally dominant staples: wheat, rice, maize, and a few others with large breeding programs and correspondingly large research investment (Ćavar Zeljković et al., 2026; Cheng et al., 2024). This distribution of genomic resources is neither accidental nor biologically justified; it reflects decades of concentrated funding in a narrow set of crops, leaving orphan and underutilized crops, often critical to food and nutritional security where they are grown without comparable genomic infrastructure (Hossain et al., 2021). Grown across Africa, Asia, and South America, orphan crops play vital local and cumulative global roles yet have received comparatively little research investment relative to their actual importance (Zafar et al., 2024).

Pangenomics is particularly well-suited to closing this gap, since orphan crops rarely have decades of accumulated single-reference infrastructure to fall back on, and lack the extensive marker development, association panels, and pedigree records that compensate for single-reference limitations in major crops (Cochetel & Cantu, 2026). Recent reviews argue pangenomics lets orphan-crop research bypass the sequential, decades-long developmental path major crops originally followed, since pangenome and multi-omics tools developed for major crops can now be applied far more rapidly than they were originally built, allowing researchers to leapfrog stages of infrastructure-building major crop programs worked through slowly (Kumar et al., 2026).

Genomic resource development for orphan crops has consequently accelerated, with dedicated assembly and pangenome projects now underway across previously under-sequenced species (Malik & Chaudhary, 2019). Recent efforts have characterized little millet's tetraploid genome structure and the genetic basis of its micronutrient content, resolved groundnut genome alongside population genetics and seed-coat-color candidate genes, and resequenced, revealing genetic redundancy and candidate genes for grain color, size, and metabolite content (Vetriventhan et al., 2020). This wave mirrors, at smaller scale and with newer technology, the resource-building major crops underwent over a much longer timeframe, compressed considerably by long-read sequencing and pangenome-construction methods unavailable when major crop references were first assembled (Cochetel & Cantu, 2026).

Wild germplasm carries particular weight for orphan crops, which have undergone comparatively limited formal domestication, so wild and landrace material retains a larger share of original genetic diversity not yet eroded by intensive selection (Khan et al., 2020). Work characterizing smallholder-maintained cowpea varieties in Southern Africa illustrates this directly, identifying climate-adaptation traits in farmer-maintained material that formal breeding programs, working with a narrower set of elite lines, had not yet captured (Issah et al., 2026).

This combination of accelerated genomic development and preserved wild diversity underlies genomics-enabled crop diversification as a deliberate food-systems strategy, not a minor regional curiosity (Raza et al., 2023). Underutilized crops are increasingly framed as essential to diversifying food systems toward greater security, since reliance on a small handful of dominant staples carries structural vulnerability to any stress, disease, or climate shock affecting them simultaneously across major production regions (SaiVenkat et al., 2026). This is close to the central justification for orphan-crop genomic research: many orphan crops are inherently well-adapted to low-input and marginal environments where major staples perform poorly, and integrating pangenomic, multi-omics, and genome-editing approaches into their breeding offers a genuine path toward repositioning them as crops of legitimate agricultural importance (Malik & Chaudhary, 2019). Supporting sustainable diversification of global food production rather than continued near-total reliance on the narrow set of staples this review has otherwise focused on.

The Bioinformatics Bottleneck

Nearly every advance in modern crop genomics like pangenomes, multi-omics networks, AI-driven prediction, precision editing, depends on a computational infrastructure that has struggled to keep pace with the data these methods generate (Bayer et al., 2022). The problem begins with raw data volume: genomic, transcriptomic, phenomic, and environmental datasets now routinely span many terabytes per project. Even a simple operation illustrates the scaling challenge: aligning resequencing data from a large rice diversity panel against a single linear reference can generate tens of terabytes of output before any pangenome graph construction or multi-omics integration is layered on top (Cheng et al., 2026; Nie et al., 2026).

Storage and compute demands scale accordingly. Graph-based pangenome methods generally improve genotyping accuracy over single-reference alignment (Bayer et al., 2022), but they typically require substantially more compute, a constraint that disproportionately limits scalability for the large, complex genomes typical of crops,

as opposed to the smaller, better-resourced genomes on which these tools were originally developed and benchmarked (Cooper & Messina, 2023). Most pangenome tools were built and tested on human genome data, and their performance on plant genomes, which are frequently larger, more repetitive, and more heterozygous, remains poorly characterized (Kumar et al., 2024). Polyploidy and repeat density compound this gap: accurate variant genotyping in repeat-enriched regions such as centromeres and disease-resistance gene clusters is inherently difficult, and because many structural variants arise from transposable-element activity, identical or near-identical repeats routinely create multiple candidate alignment positions or erase the unique sequence signatures a pangenome graph depends on (Liu et al., 2026; Yao et al., 2020). This problem is measurably more severe in large, repeat-dense, often polyploid crop genomes than in the genomes on which these tools were first developed.

Reference and ascertainment bias is a subtler but equally consequential problem: even graph-based approaches designed to reduce single-reference bias still depend heavily on which accessions were included during construction (Nie et al., 2026), and a pangenome built predominantly from elite, geographically limited material simply won't represent variation in unsequenced wild relatives or landraces, however sophisticated the graph representation (Vorbrugg et al., 2024). Data standardization compounds this at the tooling level, graph-based bioinformatics currently lacks mature, universally adopted standards for data formats, interfaces, and APIs that exist for linear genome representations, and competing graph formats have caused tool fragmentation and interoperability difficulties between research groups (Santangelo et al., 2024).

Reproducibility concerns run through this gap as well. Graph-based genotyping workflows often lack standardized benchmarking datasets and consistent version control for the underlying graphs, making it harder than it should be to independently reproduce a genotyping result across groups or software versions (Du et al., 2024); reproducible SV genotyping workflows, stable graph versioning, and standardized metadata remain priority areas still requiring development (Kopalli et al., 2025). This connects to a concern raised earlier regarding AI-based prediction: model overfitting and transferability remain persistent problems, since deep learning models trained on one crop, population, or environment frequently fail to generalize to a different genetic background or context meaning reported accuracy on training data offers limited assurance of performance on genuinely novel material (Tsega & Mullualem, 2026). AI explainability, discussed earlier, remains unresolved for the same underlying reason: the black-box nature of complex deep learning architectures limits how confidently breeders can trust and act on predictions, particularly when they conflict with field-based monitoring (Adak et al., 2024).

Perhaps the most consequential bottleneck, however, is the broader translational gap between genomic research output and actual breeding-program use (Guo et al., 2026). Breeders rely on established selection criteria and pipelines built around familiar marker types, and introducing complex structural-variation or presence or absence data from pangenomic analysis demands a genuine shift in practice one requiring substantial communication between genomic researchers and breeders rather than simply handing over a new dataset (Schreiber et al., 2024). The technical complexity of large-scale pangenomic datasets often exceeds the in-house bioinformatics expertise breeding programs traditionally maintain (De Filippis, 2017), which is why collaborative training programs, genuinely user-friendly tools, and decision-support platforms tailored to real breeding scenarios rather than research-oriented workflows have been identified as critical, currently underdeveloped needs (Simeunović et al., 2025). This is the central argument increasingly emerging across the pangenomics literature. The major challenge is no longer building more complete and sophisticated pangenomes, but translating already-available genomic resources into interoperable, reproducible, breeder-usable tools (Nie et al., 2026), a conclusion that reframes much of the technical progress catalogued earlier as necessary but, on its own, insufficient for realizing pangenomics full breeding potential. Table 3 consolidates the major bottlenecks constraining the translation of pangenomic and AI-driven genomic resources into breeder-usable tools. For each challenge, it specifies the current consequence, its significance for breeding decision-making, and the priority solution required for resolution. It additionally indicates the current readiness level of each proposed solution, ranging from increasingly mature to major, unresolved gaps.

Democratizing Genomics-Assisted Breeding

The genomic tools and pipelines described throughout this review have limited practical value if accessible only to well-resourced breeding programs, making democratized access an increasingly urgent goal in its own right (Tanin et al., 2026). Low-cost sequencing has done much of the initial work: as costs have fallen, low-coverage whole-genome skim sequencing has emerged as an attractive genotyping strategy for genomic selection where dense, high-cost marker arrays are unaffordable (Ehoche et al., 2025). Recent work implementing whole-genome prediction at breeding-program scale using ultra-low-coverage skim sequencing as low as 0.01-0.05x per sample, via commercial lab products and open-source imputation software (Novielli et al., 2024), was developed specifically to make genomic selection tractable for emerging and minor crops, where programs rarely have access to the well-funded, high-density platforms major crop programs take for granted (Simeunović et al., 2025).

Open genomic databases provide the second essential pillar, since tools are only as useful as the reference data and variant catalogues researchers can access (Riaz et al., 2021). International nucleotide sequence archives operate as open-access, globally shared infrastructure so sequence data generated anywhere can be reused freely elsewhere, built around FAIR principles findable, accessible, interoperable, and reusable rather than locked within the originating institution's systems (Gaffney et al., 2020). In plant breeding specifically, applying FAIR principles to large multi-institutional datasets has been proposed to convert fragmented, institution-specific silos into genuinely reusable,

cross-program resources extending the interoperability concerns raised earlier from a technical infrastructure problem into a broader data-governance challenge (Bhat et al., 2021).

Table 3: Major bottlenecks and priorities for breeder-ready predictive genomics

Challenge	Current consequence	Why it matters for breeding	Priority solution	Readiness	References
Reference bias	Missing alleles and structural variation	Important breeding variation may remain undetected	Pangenomes and graph genomes	Increasingly mature	(McGowan et al., 2021)
Polyploid complexity	Difficult assembly, genotyping and phasing	Reduces confidence in variant and haplotype calls	T2T, long reads and graph approaches	Developing	(Liu et al., 2026)
Computational burden	Large data-processing requirements	Limits routine analysis	Optimized pipelines, cloud computing and scalable algorithms	Developing	(Damos et al., 2026)
Interoperability	Heterogeneous data formats	Prevents integrated analyses	Standardized data structures and workflows	Developing	(De Filippis, 2017)
Reproducibility	Variable analytical outcomes	Weakens cross-study comparability	FAIR principles and standardized pipelines	Developing	(Wang et al., 2026)
Unequal genomic resources	Limited resources in minor/orphan crops	Restricts adoption outside major crops	Open databases, shared resources and cost-efficient sequencing	Major gap	(Zhai et al., 2025)
Breeder usability	Complex outputs difficult to interpret	Slows practical adoption	User-friendly platforms and decision-support tools	Major gap	(Nie et al., 2026)
Phenotyping bottleneck	Limited high-quality phenotype data	Restricts prediction accuracy	HTP and automated phenotyping	Rapidly developing	(Altaf et al., 2025)
Environmental data limitations	Weak G×E prediction	Reduces deployment accuracy	Structured enviromics and environmental covariates	Developing	(Cooper et al., 2022)
AI interpretability	Black-box predictions	Limits biological confidence	XAI and transparent model development	Developing	(Abdelwahab & Torkamaneh, 2025)
Benchmark-to-field gap	High prediction accuracy may not translate into gain	Breeding value depends on realized genetic gain	Field validation and selection-response assessment	Major gap	(Hossain et al., 2021)
Functional validation	Candidate variants remain uncertain	Discovery does not equal causality	Experimental validation and editing	Developing	(Puglisi et al., 2026)
Transformation constraints	Limits deployment of validated targets	Restricts precision breeding	Improved transformation/editing systems	Developing	(Kumar et al., 2025)
Autonomous breeding	Limited field validation	May be premature for routine deployment	Integrated robotics, AI and controlled validation	Prospective	(Altaf et al., 2025)

Cloud-based bioinformatics has become a practical necessity for extending analysis capacity to programs that could never justify their own high-performance computing infrastructure (Ali & Goh, 2026). It enable computationally demanding pangenome, variant-calling, and prediction pipelines run on shared, rented infrastructure rather than requiring every program to acquire dedicated hardware, directly addressing the compute and expertise barrier identified earlier as limiting pangenome adoption (Liu et al., 2026). User-friendly breeding platforms extend this logic further, translating raw genomic and pangenomic output into decision-support interfaces breeders without extensive bioinformatics training can navigate directly, addressing the breeder-friendly interface gap discussed as an unresolved bottleneck (Tanin et al., 2026; Zhu et al., 2026).

Capacity building in developing countries addresses a dimension infrastructure and open databases alone cannot solve, since usable tools and open data still require trained personnel to apply them locally (Ugwu et al., 2025). Coordinated training illustrates the scale achievable when organized well: one continental open-science genomics initiative reported organizing dozens of workshops in a single year, training several hundred researchers across dozens of countries in genomics and bioinformatics explicitly aimed at building regional research capacity rather than leaving countries dependent on external expertise (Diakite et al., 2026; H. Li et al., 2026).

Future Perspective: Toward Predictive and Autonomous Crop Design

The trajectory traced across this review from single-reference genomes toward pangenomes, marker-by-marker selection toward haplotype and genomic prediction, and single-omics description toward AI-integrated multi-omics points toward a broader shift already visible in forward-looking literature: from marker-assisted to genuinely prediction-based breeding, where decisions about crosses and line advancement are increasingly driven by computational prediction rather than direct phenotypic evaluation of every candidate. This is not hypothetical; it is

already underway in the genomic and phenomic prediction pipelines discussed earlier, and best understood as an extrapolation of trends already operating today rather than an entirely novel departure.

Digital twins of crops represent one concrete manifestation of this shift, extending the concept of a virtual, continuously updated replica established in engineering and increasingly in medicine into plant breeding. A digital twin is a virtual replica capturing key agronomic development processes like growth response to irrigation, fertilization, pest pressure that can be simulated computationally before any intervention is tested in the field; open, FAIR-principled soil-plant digital twin infrastructure integrating multi-source data through formal data assimilation has recently been proposed to support this kind of reproducible virtual crop modeling. Digital twins of entire breeding programs are also being explored, extending the concept from simulating a single plant's physiology to simulating a program's overall structure and capacity to adapt to shifting objectives.

Genotype-to-phenotype foundation models extend the foundation-model paradigm discussed earlier toward the prediction problem underlying genomic selection itself. Newer transformer-based architectures modeling long-range dependencies across genomic sequence are being applied directly to genotype-to-phenotype prediction in crops, motivated by the recognition that genetic and environmental heterogeneity jointly challenge the stability of genotype-phenotype associations demanding robust, better-generalizing methods that transfer across environments and germplasm rather than overfitting to a single training population, echoing the transferability concerns raised earlier regarding the bioinformatics bottleneck.

The genuinely novel element here is the explicit push to integrate AI, pangenomic, phenomic, and enviromic data streams within a single coordinated modeling framework rather than treating each as a separate pipeline feeding breeding decisions independently. This integration underlies in silico crossing and progeny prediction, where the expected genotype and phenotype of a hypothetical cross can be simulated computationally before it is physically made, letting breeders evaluate far more candidate crosses than could realistically be attempted in the field. Established breeding simulation software already enables forward- and backward-in-time simulation of haplotypes and phenotypes across an entire simulated program, testing crossing and selection schemes and modeling genomic prediction throughout before committing physical resources. Virtual breeding populations generated this way can improve resource allocation, evaluate large numbers of potential recombinants entirely in silico, and augment limited real phenotypic data with the larger sample sizes needed to train prediction models extending a program's practical reach beyond what was physically measured.

AI-guided allele stacking builds on the haplotype and multiplex-editing concepts introduced earlier, using predictive models to identify which combinations of favorable alleles or haplotypes across multiple loci are most likely to produce a desired phenotype when stacked together, rather than relying on breeders to manually reason through individual locus effects. This feeds directly into predicting optimal genotypes for specific target environments, extending the multi-environment and enviromic frameworks discussed earlier toward a prospective use case: rather than only predicting how existing genotypes will perform in a given environment, these models can in principle identify which allele combination would perform optimally in a defined target environment, then work backward toward the crosses or edits needed to assemble it.

Taken to its logical conclusion, this combination of digital twins, foundation models, in silico crossing, and environment-specific optimization points toward autonomous breeding pipelines, where computational systems handle a substantial share of candidate generation, prediction, and prioritization with limited direct human intervention at each decision point. This remains considerably more prospective than the genomic selection, pangenomics, or CRISPR applications discussed earlier, most of which already have documented field deployments and measurable outcomes; autonomous breeding pipelines represent an extrapolation of developing component technologies rather than an established, field-validated approach. This trajectory is nonetheless increasingly framed in recent literature as a shift toward "breeding by design" rather than breeding by selection where the breeder's role shifts from selecting the best performers among what chance crossing and conventional breeding produce, toward computationally specifying a target genotype and assembling it directly through the pangenomic, editing, and predictive tools this review has described. Whether this shift delivers on its promise will depend less on any single technology discussed in isolation, and considerably more on whether the translational, interoperability, and breeder-usability gaps identified in earlier sections can be closed at the pace the underlying computational and genomic technologies are advancing.

CONCLUSION

This review has synthesized eighteen interconnected areas spanning genome assembly, pangenomics, population genetics, association mapping, genomic prediction, artificial intelligence, multi-omics, phenomics and enviromics, and precision genome editing. Rather than representing isolated methodological advances, these areas constitute sequential and complementary components of a single underlying problem: converting genetic diversity into predictable, deployable breeding value under variable field conditions. Integration across these components, rather than any individual method, distinguishes the current period from earlier phases of genomics-assisted breeding. Pangenomes expand the representable genetic search space; population and comparative genomics localize where selection and introgression have already acted within it; association mapping and haplotype analysis connect specific configurations of that variation to phenotype; genomic, multi-trait, and multi-environment prediction models estimate performance

prior to field evaluation; multi-omics and cell-resolved data provide the mechanistic layer linking genotype to phenotype; and genome editing supplies a direct route from validated target to deployed genotype. Artificial intelligence increasingly functions as connective infrastructure across these stages rather than as a tool confined to any single one, a pattern evident from variant interpretation through explainable prediction to guide-RNA design.

Several limitations qualify this progress. Genomic resources remain concentrated in a small number of major staple crops; computational infrastructure for pangenome-scale analysis continues to lag behind the data volumes it must process; and the translational gap between genomic research output and breeder-usable tools constitutes, by the field's own assessment, a more consequential constraint than remaining limitations in sequencing or modeling technology. Predictive accuracy in benchmark studies does not translate automatically into genetic gain under applied breeding conditions, and several prospective applications discussed in this review, including digital twins, genotype-to-phenotype foundation models, in silico crossing, and autonomous breeding pipelines, remain extrapolations of developing component technologies rather than field-validated approaches.

These findings collectively support the central argument of this review: crop improvement is transitioning from a paradigm centered on cataloguing genetic differences relative to a single reference toward one centered on predicting and designing genotype–phenotype combinations for defined production environments. Progress toward this objective will depend less on further gains in any individual technology than on resolving the interoperability, reproducibility, and breeder-usability constraints identified throughout this review, and on extending genomic infrastructure currently concentrated in major crops to the broader range of species on which global food and nutritional security depend.

DECLARATIONS

Funding: This research received no external funding.

Acknowledgment: The authors acknowledge their respective institutions for providing academic and research support.

Conflicts of Interest: The authors declare no conflicts of interest.

Data Availability: No new data were generated or analyzed in this review article. All information discussed is derived from previously published literature cited in the manuscript.

Ethics Statement: Ethical approval was not required for this review article, as it did not involve human participants, animals, or experimental research.

Author's Contribution: Hafiz Muhammad Faraz Azhar: Conceptualization, literature review, writing—original draft, and writing—review and editing. Zarlish Attique: Literature review, data curation, and writing—review and editing. Zeeshan Ahmad Abbasi: Literature review, validation, and writing—review and editing. All authors read and approved the final manuscript.

Generative AI Statement: The authors declare that generative artificial intelligence tools were not used in the writing, preparation, or generation of scientific content for this manuscript.

Publisher's Note: All opinions, analyses, and conclusions presented in this publication are those of the authors alone. They do not necessarily correspond to the positions of the publisher, editorial team, reviewers, or affiliated institutions. The publisher neither certifies nor endorses any products, procedures, or claims discussed and accepts no liability arising from their use.

REFERENCES

Abdelwahab, O., & Torkamaneh, D. (2025). Artificial intelligence in variant calling: A review. Frontiers in Bioinformatics, 5, 1574359. https://doi.org/10.3389/fbinf.2025.1574359

Adak, A., DeSalvio, A. J., Arik, M. A., & Murray, S. C. (2024). Field-based high-throughput phenotyping enhances phenomic and genomic predictions for grain yield and plant height across years in maize. G3: Genes|Genomes|Genetics, 14(7), jkae092. https://doi.org/10.1093/g3journal/jkae092

Akagi, T., Masuda, K., Kuwada, E., Takeshita, K., Kawakatsu, T., Ariizumi, T., Kubo, Y., Ushijima, K., & Uchida, S. (2022). Genome-wide cis-decoding for expression design in tomato using cistrome data and explainable deep learning. The Plant Cell, 34(6), 2174–2187. https://doi.org/10.1093/plcell/koac079

Ali, B., Khan, Z., Imin, N., Janda, T., & Gholizadeh, F. (2026). Multi-omics dissection of drought stress responses in crops: From molecular regulatory networks to climate-resilient breeding applications. International Journal of Molecular Sciences, 27(11), 5008. https://doi.org/10.3390/ijms27115008

Ali, B., Rehman, A., Latef, A. A. H. A., Imran, M., Javed, M. A., Wise, N., & Imin, N. (2026). Engineering abiotic stress tolerance in cereal crops: Current advances and future directions. Cereal Research Communications. https://doi.org/10.1007/s42976-026-00804-1

Ali, M., Sharma, R., Parkirti, Singh, A., Sharma, A., Rana, H., Singh, A. D., Singh, J., & Ohri, P. (2026). Validation methods for plant genomic prediction. In N. Zencirci, F. S. Baloch, J.-Y. Gou, V. Mladenov, S. Fragkostefanakis, M. da Silva Lopes, E. Habyarimana, H. Ulukan, & A. Kaplan Evlice (Eds.), CABI books (pp. 216–230). CABI. https://doi.org/10.1079/9781836991328.0011

Ali, M., Wang, Z., Guo, Q., Wang, Y., Cai, Y., Du, J., Pi, E., Ding, P., & Shen, J. (2026). Mapping plant cell-type-specific responses to environmental stresses. Trends in Plant Science, 31(8), 1030–1056. https://doi.org/10.1016/j.tplants.2026.03.003

Ali, S., & Goh, H.-H. (2026). Multi-omics integration for abiotic stress acclimation in tropical and underutilized plants. Phyton-International Journal of Experimental Botany, 95(6). https://doi.org/10.32604/phyton.2026.084657

Alqudah, A. M., & Moussavi, Z. (2025). Bridging signal intelligence and clinical insight: A comprehensive review of feature engineering, model interpretability, and machine learning in biomedical signal analysis. Applied Sciences, 15(22), 12036. https://doi.org/10.3390/app152212036

Altaf, M. T., Ali, A., Mortazvi, P., Liaqat, W., Tatar, M., Sucur, R., Bedir, M., Javed, J., Gou, J.-Y., Nadeem, M. A., Alsaleh, A., Aktas, H., & Baloch, F. S. (2025). Combination of high-throughput phenotyping and genomics in precise wheat breeding. In N. Zencirci, F. S. Baloch, J.-Y. Gou, V. Mladenov, S. Fragkostefanakis, M. da Silva Lopes, E. Habyarimana, H. Ulukan, & A. Kaplan Evlice (Eds.), Empowering wheat cultivation with GIS, digital approaches and artificial intelligence (pp. 105–124). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-99954-3_6

Alzate Zuluaga, M. Y., Fattorini, R., Cesco, S., & Pii, Y. (2024). Plant-microbe interactions in the rhizosphere for smarter and more sustainable crop fertilization: The case of PGPR-based biofertilizers. Frontiers in Microbiology, 15, 1440978. https://doi.org/10.3389/fmicb.2024.1440978

Baek, I., Bhatt, J., Lim, S., Lee, D., Jang, J. H., Cohen, S. P., Lovelace, A. H., Kim, M. S., Meinhardt, L. W., Park, S., & Ahn, E. (2026). A GWAS–machine learning framework reveals protein-synthesis pathway signals for yield in *Theobroma cacao* after population-structure correction. Scientific Reports, 16, 13840. https://doi.org/10.1038/s41598-026-42273-w

Bao, H., Zhou, F., Ma, G., & Zheng, G. (2025). A systematic review of single-cell RNA sequencing applications and insights in *Oryza sativa* response to environmental stresses. Molecular Biology Reports, 52(1), 739. https://doi.org/10.1007/s11033-025-10836-1

Bayer, P. E., Petereit, J., Durant, É., Monat, C., Rouard, M., Hu, H., Chapman, B., Li, C., Cheng, S., Batley, J., & Edwards, D. (2022). Wheat Panache: A pangenome graph database representing presence–absence variation across sixteen bread wheat genomes. The Plant Genome, 15(3), e20221. https://doi.org/10.1002/tpg2.20221

Bellucci, E., Benazzo, A., Xu, C., Bitocchi, E., Rodriguez, M., Alseekh, S., Di Vittori, V., Gioia, T., Neumann, K., Cortinovis, G., Frascarelli, G., Murube, E., Trucchi, E., Nanni, L., Ariani, A., Logozzo, G., Shin, J. H., Liu, C., Jiang, L., ... Papa, R. (2023). Selection and adaptive introgression guided the complex evolutionary history of the European common bean. Nature Communications, 14(1), 1908. https://doi.org/10.1038/s41467-023-37332-z

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., & Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nature Biotechnology, 38(12), 1408–1414. https://doi.org/10.1038/s41587-020-0591-3

Bhat, J. A., Yu, D., Bohra, A., Ganie, S. A., & Varshney, R. K. (2021). Features and applications of haplotypes in crop breeding. Communications Biology, 4(1), 1266. https://doi.org/10.1038/s42003-021-02782-y

Ćavar Zeljković, S., Saeed, F., Šamec, D., & Chaudhry, U. K. (2026). Methodological advances in transcriptomics and metabolomics for assessing crop stress resilience. Physiologia Plantarum, 178(1), e70717. https://doi.org/10.1111/ppl.70717

Chen, P., Sun, Z., Wang, J., Liu, X., Bai, Y., Chen, J., Liu, A., Qiao, F., Chen, Y., Yuan, C., Sha, J., Zhang, J., Xu, L.-Q., & Li, J. (2023). Portable nanopore-sequencing technology: Trends in development and applications. Frontiers in Microbiology, 14, 1043967. https://doi.org/10.3389/fmicb.2023.1043967

Cheng, H., Asri, M., Lucas, J., Koren, S., & Li, H. (2024). Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. Nature Methods, 21(6), 967–970. https://doi.org/10.1038/s41592-024-02269-8

Cheng, Y., Li, G., Zhou, M., Mandlik, R., Wang, D., & Qi, Y. (2026). Enhancing CRISPR-Cas12a base editing in plants with LbCas12a variants and introns. Journal of Integrative Plant Biology, 68(9), 3243–3259. https://doi.org/10.1111/jipb.70249

Cochetel, N., & Cantu, D. (2026). From structural pangenomes to functional panomics in plants. Molecular Biology and Evolution, 43(4), msag087. https://doi.org/10.1093/molbev/msag087

Cooper, M., & Messina, C. D. (2023). Breeding crops for drought-affected environments and improved climate resilience. The Plant Cell, 35(1), 162–186. https://doi.org/10.1093/plcell/koac321

Cooper, M., Messina, C. D., Tang, T., Gho, C., Powell, O. M., Podlich, D. W., Technow, F., & Hammer, G. L. (2022). Predicting genotype × environment × management (G × E × M) interactions for the design of crop improvement strategies. In I. M. M. van Eeuwijk & D. J. de Koning (Eds.), Plant breeding reviews (Vol. 44, pp. 467–585). Wiley. https://doi.org/10.1002/9781119874157.ch8

Damos, P., Papathanasiou, F., Ampas, V., Kyriakidis, T., & Louta, M. (2026). Hybrid computational framework for climate-resilient plant disease forecasting via stochastic weather simulation and dynamic risk modeling. Ecological Informatics, 97, 103898. https://doi.org/10.1016/j.ecoinf.2026.103898

De Filippis, L. F. (2017). Plant bioinformatics: Next generation sequencing approaches. In K. R. Hakeem, A. Malik, F. Vardar-Sukan, & M. Ozturk (Eds.), Plant bioinformatics: Decoding the phyta (pp. 1–106). Springer International Publishing. https://doi.org/10.1007/978-3-319-67156-7_1

Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B., & Hirsch, C. N. (2021). How the pan-genome is changing crop genomics and improvement. Genome Biology, 22(1), 3. https://doi.org/10.1186/s13059-020-02224-8

Deng, Y., Liu, S., Zhang, Y., Tan, J., Li, X., Chu, X., Xu, B., Tian, Y., Sun, Y., Li, B., Xu, Y., Deng, X. W., He, H., & Zhang, X. (2022). A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. Molecular Plant, 15(8), 1268–1284. https://doi.org/10.1016/j.molp.2022.06.010

Diakite, S., Norman, P. E., Kamara, L., Pehlivan, N., & Zargar, M. (2026). Genetic enhancement of root, tuber and cereal crops via pangenomics, multi-omics integration and AI-driven prediction. Frontiers in Plant Science, 17, 1793924. https://doi.org/10.3389/fpls.2026.1793924

Du, Z.-Z., He, J.-B., & Jiao, W.-B. (2024). A comprehensive benchmark of graph-based genetic variant genotyping algorithms on plant genomes for creating an accurate ensemble pipeline. Genome Biology, 25, 91. https://doi.org/10.1186/s13059-024-03239-1

Ehoche, O. G., Arojju, S. K., Jahufer, M. Z. Z., Jauregui, R., Larking, A. C., Cousins, G., Tate, J. A., Lockhart, P. J., & Griffiths, A. G. (2025). Genomic selection shows improved expected genetic gain over phenotypic selection of agronomic traits in allotetraploid white clover. Theoretical and Applied Genetics, 138(1), 34. https://doi.org/10.1007/s00122-025-04819-w

Elliott, K., Berry, J. C., Kim, H., & Bart, R. S. (2022). A comparison of ImageJ and machine learning based image analysis methods to measure cassava bacterial blight disease severity. Plant Methods, 18(1), 86. https://doi.org/10.1186/s13007-022-00906-x

Gaffney, J., Tibebu, R., Bart, R., Beyene, G., Girma, D., Kane, N. A., Mace, E. S., Mockler, T., Nickson, T. E., Taylor, N., & Zastrow-Hayes, G. (2020). Open access to genetic sequence data maximizes value to scientists, farmers, and society. Global Food Security, 26, 100411. https://doi.org/10.1016/j.gfs.2020.100411

Gaikwad, K. B., Rani, S., Kumar, M., Gupta, V., Babu, P. H., Bainsla, N. K., & Yadav, R. (2020). Enhancing the nutritional quality of major food crops through conventional and genomics-assisted breeding. Frontiers in Nutrition, 7, 533453. https://doi.org/10.3389/fnut.2020.533453

Gogna, A., Kamali, B., Wimmer, V., Schmidt, R. H., Rezaei, E. E., Eckhoff, W. M., Reif, J. C., & Zhao, Y. (2026). Predicting enviromically adapted varieties with big data. Genome Biology, 27, 3. https://doi.org/10.1186/s13059-025-03914-x

Guo, T., & Li, X. (2023). Machine learning for predicting phenotype from genotype and environment. Current Opinion in Biotechnology, 79, 102853. https://doi.org/10.1016/j.copbio.2022.102853

Guo, Z., Yan, X., & Liang, J. (2026). Proteome–transcriptome discordance in rice under drought is modulated by post-translational modifications with functional consequences for photosynthesis and energy metabolism. Plants, 15(10), 1559. https://doi.org/10.3390/plants15101559

Hajieghrari, B., & Nejati-Jahromi, S. (2025). Next generation sequencing and beyond: A review of genomic sequencing methods. Functional & Integrative Genomics, 25(1), 242. https://doi.org/10.1007/s10142-025-01724-9

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., Forrest, K., Fritz, A., Hucl, P., Wiebe, K., Knox, R., Cuthbert, R., Pozniak, C., Akhunova, A., Morrell, P. L., Davies, J. P., Webb, S. R., Spangenberg, G., Hayes, B., ... Akhunov, E. (2019). Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nature Genetics, 51(5), 896–904. https://doi.org/10.1038/s41588-019-0382-2

Hillary, V. E., & Ceasar, S. A. (2022). Prime editing in plants and mammalian cells: Mechanism, achievements, limitations, and future prospects. BioEssays, 44(9), 2200032. https://doi.org/10.1002/bies.202200032

Hossain, A., Islam, M. T., Maitra, S., Majumder, D., Garai, S., Mondal, M., Ahmed, A., Roy, A., Skalicky, M., Brestic, M., & Islam, T. (2021). Neglected and underutilized crop species: Are they future smart crops in fighting poverty, hunger and malnutrition under changing climate? In S. M. Zargar, A. Masi, & R. K. Salgotra (Eds.), Neglected and underutilized crops: Towards nutritional security and sustainability (pp. 1–50). Springer Singapore. https://doi.org/10.1007/978-981-16-3876-3_1

Hu, Y., Schaefer, R., Rendleman, M., Slattery, A., Cramer, A., Nahiyan, A., Breitweiser, L. A., Shah, M., Kaehler, E., Yao, C., Bowling, A. J., Crow, J., May, G., Tabor, G., Thatcher, S., Uppalapati, S. R., Muppirala, U., & Deschamps, S. (2025). Spatial and single-cell transcriptomics capture two distinct cell states in soybean defense response to Phakopsora pachyrhizi infection. Frontiers in Plant Science, 16, 1637176. https://doi.org/10.3389/fpls.2025.1637176

Issah, A. K., Wang, Y., Azupio, S., Zhang, H., Chu, J., Sayibu, W., Xie, Q., & Jiang, X. (2026). Natural genetic variation in calcium sensor genes as a novel resource for abiotic stress tolerance in crops. Frontiers in Plant Science, 17, 1747177. https://doi.org/10.3389/fpls.2026.1747177

Issah, A. R., Agyare, R. Y., Suraj, M. M., Burton, G. P., Ryan, P., Idun, I., Gutaker, R. M., Chater, C., Adjebeng-Danquah, J., & Banful, B. K. (2026). Preserving crop genetic diversity through traditional seed systems: Insights from farmer-saved fonio (Digitaria exilis (Kippist) Stapf) landraces in Northern Ghana. Genetic Resources and Crop Evolution, 73(5), 194. https://doi.org/10.1007/s10722-026-02824-6

Jabeen, N., Niazi, A., Rashid, M. A., & Tabasum, S. (2026). CRISPR Cas based precision genome editing for enhancing disease resistance and yield in major crops. Research Consortium Archive, 4(2), 3677–3691. https://doi.org/10.5281/zenodo.21869835

Jayakodi, M., Lu, Q., Pidon, H., Rabanus-Wallace, M. T., Bayer, M., Lux, T., Guo, Y., Jaegle, B., Badea, A., Bekele, W., Brar, G. S., Braune, K., Bunk, B., Chalmers, K. J., Chapman, B., Jørgensen, M. E., Feng, J.-W., Feser, M., Fiebig, A., ... Stein, N. (2024). Structural variation in the pangenome of wild and domesticated barley. Nature, 636(8043), 654–662. https://doi.org/10.1038/s41586-024-08187-1

Jiao, C., Xie, X., Hao, C., Chen, L., Xie, Y., Garg, V., Zhao, L., Wang, Z., Zhang, Y., Li, T., Fu, J., Chitikineni, A., Hou, J., Liu, H., Dwivedi, G., Liu, X., Jia, J., Mao, L., Wang, X., ... Zhang, X. (2025). Pan-genome bridges wheat structural variations with habitat and breeding. Nature, 637(8045), 384–393. https://doi.org/10.1038/s41586-024-08277-0

Khan, A. W., Garg, V., Roorkiwal, M., Golicz, A. A., Edwards, D., & Varshney, R. K. (2020). Super-pangenome by integrating the wild side of a species for accelerated crop improvement. Trends in Plant Science, 25(2), 148–158. https://doi.org/10.1016/j.tplants.2019.10.012

Kopalli, V., Arslan, K., Morales-Díaz, N., Zanini, S. F., & Golicz, A. A. (2025). Toward a standardized framework for pangenome graph evaluation: Assessing crop plant pangenome variation graph construction from multiple assemblies. GigaScience, 14, giaf121. https://doi.org/10.1093/gigascience/giaf121

Kumar, A., Muthuramalingam, P., Verma, L., Kumar, R., Kumar, N., Misra, J., Ravi, K., Shin, H., & Ramesh, M. (2026). Modern genomic and omics-based technologies for millet breeding and genetic improvement. Frontiers in Plant Science, 17, 1782766. https://doi.org/10.3389/fpls.2026.1782766

Kumar, R., Das, S. P., Choudhury, B. U., Kumar, A., Prakash, N. R., Verma, R., Chakraborti, M., Devi, A. G., Bhattacharjee, B., Das, R., Das, B., Devi, H. L., Das, B., Rawat, S., & Mishra, V. K. (2024). Advances in genomic tools for plant breeding: Harnessing DNA molecular markers, genomic selection, and genome editing. Biological Research, 57(1), 80. https://doi.org/10.1186/s40659-024-00562-6

Kumar, U., Dwivedi, D., & Das, U. (2025). Advancements in CRISPR-mediated multiplex genome editing: Transforming plant breeding for crop improvement and polygenic trait engineering. Biotechnology Journal, 20(11), e70148. https://doi.org/10.1002/biot.70148

Li, H., Gao, S., Gebrewahid, T. W., Li, W.-X., Wang, J., Han, Z., Reynolds, M. P., Araus, J. L., Hearne, S., & Xu, Y. (2026). Omics-driven plant breeding through phenomics-enviromics crosstalk. Nature Communications, 17(1), 5408. https://doi.org/10.1038/s41467-026-73097-x

Li, Y., Vigil, J., Pradhan, R., Zhu, J., & Libault, M. (2026). Integrating single-cell and spatial multi-omics to decode plant–microbe interactions at cellular resolution. Microorganisms, 14(2), 380. https://doi.org/10.3390/microorganisms14020380

Lian, Q., Jiao, W.-B., & Wang, Y. (2025). Designing better crops with phased pangenomes. Molecular Plant, 18(11), 1806–1808. https://doi.org/10.1016/j.molp.2025.08.014

Liang, X., Yu, S., Ju, Y., Wang, Y., & Yin, D. (2026). Integrating envirotyping and phenomics for AI-enabled multi-environment genomic prediction in crop breeding. Agronomy, 16(10), 1019. https://doi.org/10.3390/agronomy16101019

Lin, J., Hazaisi, H., Guan, Y., & Bai, M. (2026). Multiplex gene editing drives revolution in crop breeding: Overlaid editing of multiple genes and customization of complex traits. Advanced Biotechnology, 4(1), 5. https://doi.org/10.1007/s44307-026-00099-7

Liu, X., Nie, Y., Chen, X., Ding, X., Chen, T., Zhao, Y., Zhao, X., Yan, Y., Zhang, M., & Liu, M. (2026). Pangenome-scale analysis of the PYL gene family reveals evolutionary conservation, regulatory diversification and salt-responsive haplotypes in Brassica napus. Frontiers in Plant Science, 17, 1870359. https://doi.org/10.3389/fpls.2026.1870359

Lorenzo, J. L. R., & Cervantes, E. (2026). When genetics meets ecology: Genomics and taxonomy of Vitis species and cultivars. Taxonomy, 6(3), 37. https://doi.org/10.3390/taxonomy6030037

Ma, H., Li, H., Ge, F., Zhao, H., Zhu, B., Zhang, L., Gao, H., Xu, L., Li, J., & Wang, Z. (2024). Improving genomic predictions in multi-breed cattle populations: A comparative analysis of BayesR and GBLUP models. Genes, 15(2), 253. https://doi.org/10.3390/genes15020253

Mahood, E. H., Kruse, L. H., & Moghe, G. D. (2020). Machine learning: A powerful tool for gene function prediction in plants. Applications in Plant Sciences, 8(7), e11376. https://doi.org/10.1002/aps3.11376

Malik, A. A., & Chaudhary, G. (2019). Global food security: A truncated yield of underutilized and orphan crops. In M. Khoobchandani & A. Saxena (Eds.), Biotechnology products in everyday life (pp. 161–171). Springer International Publishing. https://doi.org/10.1007/978-3-319-92399-4_11

McBreen, J., Babar, M. A., Jarquín, D., Ampatzidis, Y., Khan, N., Kunwar, S., Acharya, J. P., Adewale, S., & Brown-Guedira, G. (2025). Enhancing genomic-based forward prediction accuracy in wheat by integrating UAV-derived hyperspectral and environmental data with machine learning under heat-stressed environments. The Plant Genome, 18(1), e20554. https://doi.org/10.1002/tpg2.20554

McGowan, M., Wang, J., Dong, H., Liu, X., Jia, Y., Wang, X., Iwata, H., Li, Y., Lipka, A. E., & Zhang, Z. (2022). Ideas in genomic selection with the potential to transform plant molecular breeding: A review. Plant Breeding Reviews, 45, 273–319. https://doi.org/10.1002/9781119828235.ch7

Meena, V. K., Thribhuvan, R., Dinkar, V., Bhatt, A., Pandey, S., Abhinav, Ahmad, D., Kumar, A., & Singh, A. (2025). Haplotype breeding: Fast-track the crop improvements. Planta, 261(3), 51. https://doi.org/10.1007/s00425-025-04622-3

Mendoza-Revilla, J., Trop, E., Gonzalez, L., Roller, M., Dalla-Torre, H., de Almeida, B. P., Richard, G., Caton, J., Lopez Carranza, N., Skwark, M., Laterre, A., Beguir, K., Pierrot, T., & Lopez, M. (2024). A foundational large language model for edible plant genomes. Communications Biology, 7(1), 835. https://doi.org/10.1038/s42003-024-06465-2

Naithani, S., Deng, C. H., Sahu, S. K., & Jaiswal, P. (2023). Exploring pan-genomes: An overview of resources and tools for unraveling structure, function, and evolution of crop genes and genomes. Biomolecules, 13(9), 1403. https://doi.org/10.3390/biom13091403

Napier, J. D., Heckman, R. W., & Juenger, T. E. (2023). Gene-by-environment interactions in plants: Molecular mechanisms, environmental drivers, and adaptive plasticity. The Plant Cell, 35(1), 109–124. https://doi.org/10.1093/plcell/koac322

Nie, S., Li, F., Li, R., Wang, J., Ma, Y., Chan, C. K. K., Zhao, J., & Hu, H. (2026). Challenges in bringing pangenome research into breeding: A case study in rice. Plant Biotechnology Journal, 24(6), 3679–3690. https://doi.org/10.1111/pbi.70591

Nobori, T. (2025). Exploring the untapped potential of single-cell and spatial omics in plant biology. New Phytologist, 247(3), 1098–1116. https://doi.org/10.1111/nph.70220

Novielli, P., Romano, D., Pavan, S., Losciale, P., Stellacci, A. M., Diacono, D., Bellotti, R., & Tangaro, S. (2024). Explainable artificial intelligence for genotype-to-phenotype prediction in plant breeding: A case study with a dataset from an almond germplasm collection. Frontiers in Plant Science, 15, 1434229. https://doi.org/10.3389/fpls.2024.1434229

Ohishi, H., & Au Yeung, W. K. (2026). Recent advances in methodologies of epigenomics. Epigenomics, 18(1), 117–135. https://doi.org/10.1080/17501911.2025.2571391

Pelosi, D., Cacciagrano, D., & Piangerelli, M. (2025). Explainability and interpretability in concept and data drift: A systematic literature review. Algorithms, 18(7), 443. https://doi.org/10.3390/a18070443

Phua, D. Y. Z., Li, X., & Nolan, T. M. (2026). Single-cell and spatial transcriptomics in plants: From cell states to inter-tissue coordination. Journal of Experimental Botany. https://doi.org/10.1093/jxb/erag241

Pratap, A., Das, A., Kumar, S., & Gupta, S. (2021). Current perspectives on introgression breeding in food legumes. Frontiers in Plant Science, 11, 589189. https://doi.org/10.3389/fpls.2020.589189

Puglisi, D., Crossa, J., Cuevas, J., Fania, F., Afshari-Behbahanizadeh, S., Vitale, P., & De Vita, P. (2026). Multi-trait and multi-environment genomic prediction enhances yield components improvement in durum wheat. Frontiers in Plant Science, 17, 1759897. https://doi.org/10.3389/fpls.2026.1759897

Raza, A., Bohra, A., Garg, V., & Varshney, R. K. (2023). Back to wild relatives for future breeding through super-pangenome. Molecular Plant, 16(9), 1363–1365. https://doi.org/10.1016/j.molp.2023.08.005

Raza, A., Li, Y., Jan, F., Fernandez, C. G. T., Mir, R. R., Hu, Z., & Varshney, R. K. (2026). From the genome to super-pangenome: A new paradigm for accelerated crop improvement. npj Science of Plants, 2(1), 4. https://doi.org/10.1038/s44383-025-00019-z

Riaz, A., Kanwal, F., Börner, A., Pillen, K., Dai, F., & Alqudah, A. M. (2021). Advances in genomics-based breeding of barley: Molecular tools and genomic databases. Agronomy, 11(5), 894. https://doi.org/10.3390/agronomy11050894

Roy, S., Gadri, H. S., Sharma, V., Chowdhary, M. A., Dwivedi, R., & Bhardwaj, P. (2026). Genome-wide association study bridging genomics–phenomics gap in natural plant populations. Journal of Applied Genetics, 67(1), 1–16. https://doi.org/10.1007/s13353-025-01010-1

SaiVenkat, K. S., Priya, M. S., Sekhar, M. R., Neeraja, C. N., Shanthi, P., & Rao, D. S. (2026). Past, present and future of rice breeding: From domestication to genomics. Plant Molecular Biology Reporter, 44(3), 101. https://doi.org/10.1007/s11105-026-01723-5

Salotti, I., Bove, F., & Rossi, V. (2022). Development and validation of a mechanistic, weather-based model for predicting Puccinia graminis f. sp. tritici infections and stem rust progress in wheat. Frontiers in Plant Science, 13, 897680. https://doi.org/10.3389/fpls.2022.897680

Santangelo, B. E., Apgar, M., Colorado, A. S. B., Martin, C. G., Sterrett, J., Wall, E., Joachimiak, M. P., Hunter, L. E., & Lozupone, C. A. (2024). Integrating biological knowledge for mechanistic inference in the host-associated microbiome. Frontiers in Microbiology, 15, 1351678. https://doi.org/10.3389/fmicb.2024.1351678

Schreiber, M., Jayakodi, M., Stein, N., & Mascher, M. (2024). Plant pangenomes for crop improvement, biodiversity and evolution. Nature Reviews Genetics, 25(8), 563–577. https://doi.org/10.1038/s41576-024-00691-4

Scott, M. F., Fradgley, N., Bentley, A. R., Brabbs, T., Corke, F., Gardner, K. A., Horsnell, R., Howell, P., Ladejobi, O., Mackay, I. J., Mott, R., & Cockram, J. (2021). Limited haplotype diversity underlies polygenic trait architecture across 70 years of wheat breeding. Genome Biology, 22(1), 137. https://doi.org/10.1186/s13059-021-02354-7

Sheikh, M., Iqra, F., Ambreen, H., Pravin, K. A., Ikra, M., & Chung, Y. S. (2024). Integrating artificial intelligence and high-throughput phenotyping for crop improvement. Journal of Integrative Agriculture, 23(6), 1787–1802. https://doi.org/10.1016/j.jia.2023.10.019

Shen, P., Yang, Q., Shi, Z., Liu, Z., Feng, Y., Lamlom, S. F., Ren, H., Liu, F., Ma, J., Zhang, Y., & Yan, L. (2026). Genome-wide association studies reveal environment-dependent genetic associations for plant architecture in regional soybean germplasm. The Plant Genome, 19(2), e70251. https://doi.org/10.1002/tpg2.70251

Simeunović, M., Ratković, K., Kovač, N., Racković, T., & Fernandes, A. (2025). A knowledge-driven framework for a decision support platform in sustainable viticulture: Integrating climate data and supporting stakeholder collaboration. Sustainability, 17(4), 1387. https://doi.org/10.3390/su17041387

Singh, P., Sundaram, K. T., Vinukonda, V. P., Venkateshwarlu, C., Paul, P. J., Pahi, B., Gurjar, A., Singh, U. M., Kalia, S., Kumar, A., Singh, V. K., & Sinha, P. (2024). Superior haplotypes of key drought-responsive genes reveal opportunities for the development of climate-resilient rice varieties. Communications Biology, 7(1), 89. https://doi.org/10.1038/s42003-024-05769-7

Sinha, P., Singh, V. K., Saxena, R. K., Khan, A. W., Abbai, R., Chitikineni, A., Desai, A., Molla, J., Upadhyaya, H. D., Kumar, A., & Varshney, R. K. (2020). Superior haplotypes for haplotype-based breeding for drought tolerance in pigeonpea (*Cajanus cajan* L.). Plant Biotechnology Journal, 18(12), 2482–2490. https://doi.org/10.1111/pbi.13422

Sun, H., Zhang, J., Liao, S., Guo, S., Zhou, Z., Zhao, X., Wu, S., Zhao, J., Gong, G., Wang, J., Li, M., Yu, Y., Ren, Y., Tian, S., Li, S., Zhang, H., Hammar, S. A., McGregor, C., Jarret, R., ... Xu, Y. (2026). Population-level super-pangenome reveals genome evolution and empowers precision breeding in watermelon. Nature Genetics, 58(5), 1176–1185. https://doi.org/10.1038/s41588-026-02598-8

Sun, L., Lai, M., Ghouri, F., Nawaz, M. A., Ali, F., Baloch, F. S., Nadeem, M. A., Aasim, M., & Shahid, M. Q. (2024). Modern plant breeding techniques in crop improvement and genetic diversity: From molecular markers and gene editing to artificial intelligence—A critical review. Plants, 13(19), 2676. https://doi.org/10.3390/plants13192676

Tanin, M. J., Saini, D. K., Alagappan, P., & Kaur, J. P. (2026). A review of AI-driven phenomics, genomics, and predictive breeding in wheat. Discover Plants, 3(1), 210. https://doi.org/10.1007/s44372-026-00696-9

Tay Fernandez, C. G., Nestor, B. J., Danilevicz, M. F., Marsh, J. I., Petereit, J., Bayer, P. E., Batley, J., & Edwards, D. (2022). Expanding gene-editing potential in crop improvement with pangenomes. International Journal of Molecular Sciences, 23(4), 2276. https://doi.org/10.3390/ijms23042276

Tessema, A. M., Abebe, F. M., Kidane, Y. G., Hailegiorgis, D., Tehulie, N. S., Hassen, A., Abebe, S., & Ibrahim, S. (2026). Crop biofortification for enhanced nutritional density: Breeding, agronomical, and biotechnological strategies. Crop Science, 66, e70246. https://doi.org/10.1002/csc2.70246

Tsega, A., & Mullualem, D. (2026). Machine learning for multi-omics data integration in crop improvement: A systematic review. BMC Bioinformatics, 27(1), 81. https://doi.org/10.1186/s12859-026-06438-8

Ugwu, O. P.-C., Ogenyi, F. C., Alum, E. U., Eze, V. H. U., Basajja, M., Ugwu, J. N., Ugwu, C. N., Ejemot-Nwadiaro, R. I., Okon, M. B., Egba, S. I., & Ejim, U. D. (2025). Implementing artificial intelligence and machine learning algorithms for optimized crop management: A systematic review on data-driven approach to enhancing resource use and agricultural sustainability. Cogent Food & Agriculture, 11(1), 2569982. https://doi.org/10.1080/23311932.2025.2569982

Varshney, R. K., Bohra, A., Yu, J., Graner, A., Zhang, Q., & Sorrells, M. E. (2021). Designing future crops: Genomics-assisted breeding comes of age. Trends in Plant Science, 26(6), 631–649. https://doi.org/10.1016/j.tplants.2021.03.010

Vetriventhan, M., Azevedo, V. C. R., Upadhyaya, H. D., Nirmalakumari, A., Kane-Potaka, J., Anitha, S., Ceasar, S. A., Muthamilarasan, M., Bhat, B. V., Hariprasanna, K., Bellundagi, A., Cheruku, D., Backiyalakshmi, C., Santra, D., Vanniarajan, C., & Tonapi, V. A. (2020). Genetic and genomic resources, and breeding for accelerating improvement of small millets: Current status and future interventions. The Nucleus, 63(3), 217–239. https://doi.org/10.1007/s13237-020-00322-3

Vikas, V. K., Pradhan, A. K., Budhlakoti, N., Mishra, D. C., Chandra, T., Bhardwaj, S. C., Kumar, S., Sivasamy, M., Jayaprakash, P., Nisha, R., Shajitha, P., Peter, J., Geetha, M., Mir, R. R., Singh, K., & Kumar, S. (2022). Multi-locus genome-wide association studies (ML-GWAS) reveal novel genomic regions associated with seedling and adult plant stage leaf rust resistance in bread wheat (*Triticum aestivum* L.). Heredity, 128(6), 434–449. https://doi.org/10.1038/s41437-022-00525-1

Vorbrugg, S., Bezrukov, I., Bao, Z., Xian, W., & Weigel, D. (2024). Gfa2bin enables graph-based GWAS by converting genome graphs to pan-genomic genotypes. bioRxiv, 2024.12.05.626966. https://doi.org/10.1101/2024.12.05.626966

Wambugu, P. W., & Henry, R. (2022). Supporting in situ conservation of the genetic diversity of crop wild relatives using genomic technologies. Molecular Ecology, 31(8), 2207–2222. https://doi.org/10.1111/mec.16402

Wang, B., Jia, P., Gao, S., Zhao, H., Zheng, G., Xu, L., & Ye, K. (2025). Long and accurate: How HiFi sequencing is transforming genomics. Genomics, Proteomics & Bioinformatics, 23(1), qzaf003. https://doi.org/10.1093/gpbjnl/qzaf003

Wang, H., Yan, S., Wang, W., Chen, Y., Hong, J., He, Q., Diao, X., Lin, Y., Chen, Y., Cao, Y., Guo, W., & Fang, W. (2025). Cropformer: An interpretable deep learning framework for crop genomic prediction. Plant Communications, 6(3), 101223. https://doi.org/10.1016/j.xplc.2024.101223

Wang, Q., Sun, X., Sun, Y., Cheng, Z., Cheng, Z., Han, S., Feng, Y., Yang, W., Li, H., Zhu, M., Wu, X., Zhang, J., Tang, J., Li, H., Cao, Y., Duan, C., & Shi, Y. (2026). Integrated single-cell and spatial transcriptomics reveal cell-type-specific immune regulatory networks in maize responding to southern corn rust. Advanced Science, 13(27), e12295. https://doi.org/10.1002/advs.202512295

Wang, T., Tong, R., Yuan, Z., & Li, Y. (2026). Beyond QTL and GWAS: How deep learning, graph models, and multi-omics are reshaping plant genomic prediction analysis. Frontiers in Genetics, 17, 1783939. https://doi.org/10.3389/fgene.2026.1783939

Wang, X., Han, L., Li, J., Shang, X., Liu, Q., Li, L., & Zhang, H. (2023). Next-generation bulked segregant analysis for Breeding 4.0. Cell Reports, 42(9), 113039. https://doi.org/10.1016/j.celrep.2023.113039

Wang, Y., Yao, D., Song, J., Zhang, X., Zhu, Y., Luo, Z., Lai, J., Song, W., & Yang, T. (2025). Identification of yield-related QTLs and their applications using DH population in maize. BMC Genomics, 26(1), 719. https://doi.org/10.1186/s12864-025-11920-3

Wei, Y., Zhang, H., Fan, J., Cai, Q., Zhang, Z., Wang, J., Zhang, M., Yan, F., Jiang, J., Xie, H., Luo, X., Wei, L., Lin, Y., He, W., Qu, M., Zhang, X., Zhu, Y., Xie, H., & Zhang, J. (2024). Multiplex-genome-editing based rapid directional improvement of complex traits in rice. Plant Biotechnology Journal, 22(9), 2624–2628. https://doi.org/10.1111/pbi.14375

Winters, G., Beer, S., Willette, D. A., Viana, I. G., Chiquillo, K. L., Beca-Carretero, P., Villamayor, B., Azcárate-García, T., Shem-Tov, R., Mwabvu, B., Migliore, L., Rotini, A., Oscar, M. A., Belmaker, J., Gamliel, I., Alexandre, A., Engelen, A. H., Procaccini, G., & Rilov, G. (2020). The tropical seagrass Halophila stipulacea: Reviewing what we know from its native and invasive habitats, alongside identifying knowledge gaps. Frontiers in Marine Science, 7, 300. https://doi.org/10.3389/fmars.2020.00300

Wu, L., Guo, J., Zhang, L., Cheng, H., Ye, X., & Sun, Y. (2026). Postharvest abiotic stress reshaping flavor profiles in fresh fruits and vegetables: Challenges beyond shelf life. Food Reviews International, 1–27. https://doi.org/10.1080/87559129.2026.2686832

Wu, X., Heffelfinger, C., Zhao, H., & Dellaporta, S. L. (2019). Benchmarking variant identification tools for plant diversity discovery. BMC Genomics, 20(1), 701. https://doi.org/10.1186/s12864-019-6057-7

Yan, J., & Wang, X. (2026). Rethinking genomic selection under environmental uncertainty: Toward learnable and dynamic environmental representations. Frontiers in Plant Science, 16, 1754446. https://doi.org/10.3389/fpls.2025.1754446

Yang, R., & Zhang, H. (2025). Application of artificial intelligence technology in plant microRNA research: Progress, challenges, and prospects. International Journal of Molecular Sciences, 26(24), 11854. https://doi.org/10.3390/ijms262411854

Yao, Z., You, F. M., N'Diaye, A., Knox, R. E., McCartney, C., Hiebert, C. W., Pozniak, C., & Xu, W. (2020). Evaluation of variant calling tools for large plant genome re-sequencing. BMC Bioinformatics, 21(1), 360. https://doi.org/10.1186/s12859-020-03704-1

Yu, H., Lin, T., Meng, X., Du, H., Zhang, J., Liu, G., Chen, M., Jing, Y., Kou, L., Li, X., Gao, Q., Liang, Y., Liu, X., Fan, Z., Liang, Y., Cheng, Z., Chen, M., Tian, Z., Wang, Y., ... Li, J. (2021). A route to de novo domestication of wild allotetraploid rice. Cell, 184(5), 1156–1170.e14. https://doi.org/10.1016/j.cell.2021.01.013

Zafar, T., Mehra, A., Das, P., Shaik, B., & Malik, A. A. (2024). Demystifying the advanced interventions of genetics and modern breeding approaches for nutritional security and sustainability of neglected and underused crop species (NUCS). Genetic Resources and Crop Evolution, 71(2), 559–577. https://doi.org/10.1007/s10722-023-01823-1

Zanini, S. F., Bayer, P. E., Wells, R., Snowdon, R. J., Batley, J., Varshney, R. K., Nguyen, H. T., Edwards, D., & Golicz, A. A. (2022). Pangenomics in crop improvement—from coding structural variations to finding regulatory variants with pangenome graphs. The Plant Genome, 15(1), e20177. https://doi.org/10.1002/tpg2.20177

Zeibig, F., Kilian, B., Özkan, H., Pantha, S., & Frei, M. (2024). Phenotyping and identification of target traits for de novo domestication of wheat wild relatives. Food and Energy Security, 13(1), e497. https://doi.org/10.1002/fes3.497

Zhai, J., Gokaslan, A., Schiff, Y., Berthel, A., Liu, Z.-Y., Lai, W.-Y., Miller, Z. R., Scheben, A., Stitzer, M. C., Romay, M. C., Buckler, E. S., & Kuleshov, V. (2025). Cross-species modeling of plant genomes at single-nucleotide resolution using a pretrained DNA language model. Proceedings of the National Academy of Sciences, 122(24), e2421738122. https://doi.org/10.1073/pnas.2421738122

Zhang, Y., Zhao, M., Tan, J., Huang, M., Chu, X., Li, Y., Han, X., Fang, T., Tian, Y., Jarret, R., Lu, D., Chen, Y., Xue, L., Li, X., Qin, G., Li, B., Sun, Y., Deng, X. W., Deng, Y., ... He, H. (2024). Telomere-to-telomere Citrullus super-pangenome provides direction for watermelon breeding. Nature Genetics, 56(8), 1750–1761. https://doi.org/10.1038/s41588-024-01823-6

Zhang, Z., Zhang, J., Kang, L., Qiu, X., Xu, S., Xu, J., Guo, Y., Niu, Z., Niu, B., Bi, A., Zhao, X., Xu, D., Wang, J., Yin, C., & Lu, F. (2024). Structural variation discovery in wheat using PacBio high-fidelity sequencing. The Plant Journal, 120(2), 687–698. https://doi.org/10.1111/tpj.17011

Zhao, D., Lin, M., Taniguti, C. H., Sandercock, A. M., Chen, S., Babiker, E., Bassil, N. V., Brummer, E. C., Camacho, J. R., Chatwin, W., Chen, S.-Y., Clare, S. J., da Silva Pereira, G., David, M., Fraher, S. P., Hardigan, M., Hilton, A., Hislop, L. M., Irish, B. M., ... Sheehan, M. J. (2026). Standardized microhaplotype databases and frameworks for assessing and mining crop genetic diversity. Theoretical and Applied Genetics, 139(9), 244. https://doi.org/10.1007/s00122-026-05340-4

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., Zhang, J., Lyu, H., Lin, T., Gao, Q., Saha, S., Mueller, L., Fei, Z., Städler, T., Xu, S., ... Huang, S. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. Nature, 606(7914), 527–534. https://doi.org/10.1038/s41586-022-04808-9

Zhu, Z., Jhingan, S., Chen, E., & Stein, N. (2026). Functional cereal pan-genomics: Harnessing structural and regulatory variation for precision crop design. Current Opinion in Biotechnology, 97, 103418. https://doi.org/10.1016/j.copbio.2025.103418