



Multi-Omics Integration for Crop Improvement: Connecting Genomes, Transcriptomes, Proteomes and Metabolomes

Awadhesh Kumar^{a++*}, Pragya Ojha^{b#}, Saurabh Kanauajia^{c†},
Biswajit Lenka^{d†}, Kesoju Ravali^{e‡}, Asit Prasad Dash^{d^}
and Arghya Mani^{f##}

^a Krishi Vigyan Kendra, Ballia, Acharya Narendra Deva, University of Agriculture & Technology, Ayodhya, Uttar Pradesh, India.

^b Krishi Vigyan Kendra, Banda, Banda University of Agriculture and Technology, Banda, Uttar Pradesh, India.

^c Shri Bajrang Singh Mahavidyalaya, Mau, Gauriganj, Amethi, Uttar Pradesh, India.

^d Department of Genetics and Plant Breeding, Faculty of Agricultural Sciences (IAS), SOA-DU, Bhubaneswar, Odisha, India.

^e National Seeds Corporation Ltd. (NSCL), Warangal, Telangana, India.

^f Department of Post Harvest Technology, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, Nadia, West Bengal, India.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: <https://doi.org/10.9734/ijpss/2026/v38i106375>

Open Peer Review History:

This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: <https://pr.sdiarticle5.com/review-history/168136>

Review Article

Received: 20/08/2026

Accepted: 12/10/2026

Published: 14/10/2026

⁺⁺Subject Matter Specialist (Horticulture); [#]Subject Matter Specialist (Home Science); [†]Assistant Professor; [‡]Junior Executive (agriculture); [^]Associate Professor; ^{##}Ph.D. Research Scholar;

*Corresponding author: E-mail: akrawat2910@gmail.com;

Cite as: Kumar, A., Ojha, P., Kanauajia, S., Lenka, B., Ravali, K., Dash, A. P., & Mani, A. (2026). Multi-Omics Integration for Crop Improvement: Connecting Genomes, Transcriptomes, Proteomes and Metabolomes. *International Journal of Plant & Soil Science*, 38(10), 553–571. <https://doi.org/10.9734/ijpss/2026/v38i106375>

Abstract

Crop improvement increasingly depends on resolving how inherited sequence variation is converted into context-dependent molecular states and, ultimately, agronomic phenotypes. Genomics supplies stable information on allelic and structural variation, whereas transcriptomics, proteomics and metabolomics progressively capture regulatory activity, functional molecular abundance and biochemical state. Their integration is therefore attractive, but the central breeding question is not whether more data can be combined; it is when added molecular layers improve causal inference, trait prediction or selection decisions beyond a strong genomic baseline. This critical narrative review evaluates evidence linking the four core omics layers in crop research, with emphasis on pangenome-aware genomics, expression and association mapping, proteome and post-translational regulation, metabolic endophenotypes, multi-block statistical integration, network models and machine-learning approaches. The literature indicates strong value for cross-omics integration in candidate-gene prioritisation, pathway reconstruction and mechanistic interpretation, particularly when genomic associations are anchored to expression, protein or metabolite variation measured in relevant tissues and environments. Evidence for routine prediction gains is less uniform. Metabolomic and transcriptomic profiles can be highly predictive in some populations, yet comparative studies show that combined multi-omics models may fail to outperform genomic prediction when molecular measurements are noisy, environmentally contingent, redundant with genetic relationships or evaluated without stringent out-of-environment validation. Proteomic integration adds biologically important information that is not reliably inferred from messenger RNA, but coverage, cost and post-translational complexity remain substantial constraints. The field is moving towards graph pangenomes, matched longitudinal sampling, interpretable multimodal models and integration with deep phenotyping and environmental covariates. The most defensible near-term strategy is consequently selective rather than maximal integration: deploy the molecular layer whose biological position and measurement reliability match the breeding question, validate improvements against genomic baselines, and prioritise designs that can demonstrate transfer across environments and breeding cycles.

Keywords: Systems breeding; genomic prediction; pangenome; transcriptome-wide association; proteomics; metabolomics; data integration; genotype-by-environment interaction.

1. Introduction

Crop breeding is a problem of prediction under biological and environmental complexity. DNA sequence provides the inherited substrate for selection, yet yield, quality, stress resilience and developmental timing emerge through regulatory processes that are neither spatially nor temporally uniform. A single nucleotide variant can alter a transcript directly, affect a protein only in a particular tissue, or produce a phenotype through downstream metabolic fluxes that are strongly modified by environment. Multi-omics research attempts to observe several points along this genotype-to-phenotype chain simultaneously and to use their concordance, discordance and conditionality to improve biological inference and breeding decisions. Earlier crop multi-omics reviews established the breadth of available technologies and their potential applications, but they also identified data heterogeneity, incomplete functional annotation and integration methodology as persistent constraints (Scossa *et al.*, 2021; Yang *et al.*, 2021; Zhang *et al.*, 2022). Recent scholarship similarly emphasises combining molecular data with phenotyping and computational tools while recognising practical barriers to breeding translation (Li *et al.*, 2025).

The attraction of integration is easy to understand. Genomics is comparatively stable, scalable and already embedded in genome-wide association studies (GWAS) and genomic selection, but association signals often contain many linked variants and may not reveal the molecular route to phenotype. Transcriptomes provide a dynamic readout of gene regulation and can place associated loci into tissue-specific expression contexts. Proteomes move closer to biochemical function and expose cases in which messenger RNA abundance poorly predicts protein abundance, while phosphoproteomic and other post-translational measurements reveal regulatory states invisible to standard RNA sequencing. Metabolomes lie still closer to many physiological and quality phenotypes and can function as both mechanistic intermediates and predictive biomarkers. The layers are therefore complementary in principle, although complementarity in biology does not guarantee incremental value in a statistical prediction model (Ritchie *et al.*, 2015; Scossa *et al.*, 2021). In barley, integrated transcriptomic, metabolomic and metabolite-QTL analyses linked grain chemistry to candidate regulatory networks (Song *et al.*, 2025).

Crop evidence demonstrates both sides of this distinction. Large integrative atlases in maize and wheat show that transcript, protein and post-translational data can reveal regulatory structure that would be missed by genome sequence alone (Walley *et al.*, 2016; Zhang *et al.*, 2025). Metabolome-wide association studies in maize and rice have connected genetic variation to chemically defined traits and candidate genes at a resolution useful for pathway biology (Chen *et al.*, 2014; Wen *et al.*, 2014). In tomato, coupling chemical phenotypes, consumer preference and genome-wide mapping produced a practical route to flavour improvement (Tieman *et al.*, 2017). At the same time, a direct comparison of genomic, transcriptomic and metabolomic predictors in maize found that the combined model had little consistent advantage over the strongest single-omics predictor, with genomic prediction often performing best (Xu *et al.*, 2017). This result is not an argument against integration; it is evidence that integration must be judged against the decision it is meant to improve. Integrated genomic, transcriptomic and metabolomic analysis in maize connected primary metabolism with nitrogen partitioning and agronomic trade-offs (Jin *et al.*, 2026).

The genomic coordinate system itself is also changing. Crop genomics historically relied on a single linear reference, which can under-represent presence/absence variation, copy-number changes and structural variants. Plant pangenomes and graph-based references provide a more inclusive representation of diversity and alter what downstream transcriptomic, proteomic and metabolomic associations can be discovered (Bayer *et al.*, 2020; Wang *et al.*, 2023). In tomato, a graph pangenome built from extensive germplasm captured structural variation omitted from a single-reference representation and increased the estimated heritability of thousands of gene-expression and metabolite traits, while identifying agronomically relevant loci (Zhou *et al.*, 2022).

Multi-omics integration is therefore increasingly inseparable from the question of whether the genomic layer itself adequately represents the breeding population.

Computational development has expanded equally quickly. Network-based approaches, sparse multivariate models, latent-factor methods and machine learning can integrate data with radically different scales and dimensions (Argelaguet *et al.*, 2018; Bersanelli *et al.*, 2016; Rohart *et al.*, 2017; Singh *et al.*, 2019). Recent crop-focused assessments highlight substantial interest in machine learning for multi-omics integration, but also recurring weaknesses in external validation, environmental representation, interpretability and comparability among studies (Montesinos-López *et al.*, 2024; Tsega & Mullualem, 2026). These weaknesses matter in breeding because the operational target is rarely reconstruction of the training population. It is the performance of new genotypes in future or partially observed environments, usually under constraints on cost, cycle time and sampling. In maize, combining genomic and transcriptomic data improved prediction for many traits, whereas the value of phenomic inputs varied (Creach *et al.*, 2026).

A critical synthesis must therefore separate three claims that are frequently conflated. First, multiple omics can improve mechanistic interpretation by triangulating a genetic signal across molecular layers. Second, multi-omics can improve prediction of phenotypes within the population and environments used to measure those molecular layers. Third, multi-omics can improve selection response in a breeding programme when prediction is transferred to new lines, environments and cycles. The evidence for the first claim is comparatively strong; evidence for the second is conditional; and the third remains much less directly demonstrated. Treating these claims as equivalent encourages overly optimistic translation from molecular discovery to breeding practice (Mahmood *et al.*, 2022; Scossa *et al.*, 2021; Tsega & Mullualem, 2026).

1.1 Scope, Research Gap and Objectives

This review focuses on the integration of genomics, transcriptomics, proteomics and metabolomics for crop improvement. Epigenomic, microbiome, phenomic and environmental data are considered where they alter interpretation or validation of these four core molecular layers, but they are not treated as co-equal primary domains. The central gap addressed is the limited distinction in the crop literature between mechanistic integration and breeding-useful integration. The objectives are to evaluate what each omics layer adds beyond the others, identify designs and analytical strategies that generate credible cross-layer inference, compare evidence for candidate-gene discovery and trait prediction, explain why prediction gains are inconsistent, assess barriers to transfer across environments and breeding cycles, and define a prioritised research agenda for breeding-ready multi-omics. The review excludes purely descriptive single-omics catalogues without a clear link to crop traits or breeding decisions, and it does not attempt an exhaustive inventory of every crop or software package.

2. Methods for Literature Selection

2.1 Review Design and Rationale

A critical narrative review design was selected because the evidence base spans heterogeneous questions and study types: reference-genome construction, molecular atlases, association mapping, prediction experiments, statistical-method papers and breeding-oriented reviews. A single effect-size meta-analysis would not meaningfully combine these outputs. Narrative synthesis nevertheless requires explicit searching, critical appraisal and transparent interpretation. The approach followed best-evidence principles for narrative reviewing and used the SANRA domains as a quality-oriented guide to justification, search description, referencing and scientific reasoning rather than as a numerical risk-of-bias score (Baethge *et al.*, 2019; Green *et al.*, 2006).

2.2 Information Sources

Literature was identified through PubMed, PubMed Central, OpenAlex, OpenAIRE, AGRIS, DOAJ, Semantic Scholar and Google Scholar. Searches were supplemented by backward and forward citation tracing from major crop multi-omics reviews and primary studies, related-article searches, and targeted verification of bibliographic metadata and digital object identifiers on DOI landing pages, journal records and authoritative scholarly repositories. Searches for corrections and bibliographic updates were undertaken for influential or recently published papers before inclusion.

2.3 Search Strategy and Search Period

The principal search window was 1 January 2010 to 7 August 2026. Earlier papers were retained when they were methodologically foundational to narrative review conduct, network analysis or multi-omics integration. Representative Boolean concepts included: ("multi-omics" OR multiomics OR "integrative omics") AND (crop* OR plant* OR rice OR maize OR wheat OR soybean OR tomato) AND (breeding OR "crop improvement" OR "genomic selection" OR GWAS OR "trait prediction"); (genom* AND (transcriptom* OR proteom* OR metabolom*)) AND (integration OR network* OR "latent factor" OR "machine learning") AND (crop* OR plant*); and (pangenom* OR "structural variation" OR eQTL OR TWAS OR pQTL OR mGWAS) AND (crop* OR breeding OR "complex trait*"). Additional searches combined multi-omics terms with genotype-by-environment interaction, deep phenotyping, stress resilience, post-translational modification, metabolite prediction and validation. Search vocabulary was expanded iteratively when relevant articles introduced established crop-specific or method-specific terminology.

2.4 Eligibility Criteria

Eligible sources were peer-reviewed journal articles and high-quality reviews with direct relevance to one or more of the four core omics layers, cross-layer data integration, crop trait dissection, genomic prediction, molecular breeding or integration methodology

transferable to crop data. Primary studies were prioritised when they provided matched omics measurements, experimentally or statistically linked molecular layers, credible trait phenotypes, or explicit prediction comparisons. Reviews were used to establish conceptual coverage, identify recurring limitations and locate primary studies. English-language sources were included. Preprints, promotional material, unverifiable records, studies lacking sufficient bibliographic identity, and papers with only superficial mention of multi-omics were excluded. Foundational non-crop methodology was retained only when the analytical framework was directly relevant to integration.

2.5 Screening, Duplicate Handling and Study Selection

Titles and abstracts were assessed for relevance to the review questions, followed by examination of full text or sufficiently detailed authoritative abstracts for sources used to support substantive claims. Duplicate records were reconciled using DOI, exact title and author-year identity. When multiple records described the same article, the version of record was used. Sources whose findings could not be checked against an accessible abstract, full text or authoritative record were not used as evidence. Selection was purposive rather than exhaustive: the synthesis emphasised influential, methodologically informative and contrasting studies, including negative or non-improving integration results. No record-count flow diagram was constructed because the review was not designed as a systematic review with exhaustive enumeration.

2.6 Critical Appraisal and Evidence Synthesis

Evidence was appraised in relation to the question each study could legitimately answer. For molecular atlases and association studies, attention was given to sample and tissue design, genomic diversity, molecular coverage, quality control, validation and the distinction between association and functional causation. Prediction studies were judged by comparator strength, cross-validation structure, independence of test data, environmental transfer and the risk that downstream molecular measurements encoded information unavailable at the intended selection stage. Integration methods were assessed for dimensionality reduction, handling of scale and missingness, supervision, interpretability and susceptibility to overfitting. Themes were developed around the genotype-to-phenotype hierarchy, integration architecture, breeding use case and translational bottleneck. Conflicting findings were interpreted in light of population structure, tissue and developmental context, assay characteristics, environment, sample size and validation design rather than resolved by citation frequency alone.

3. Genomics as the Coordinate System for Multi-Omics Breeding

Genomics occupies a distinctive position in crop multi-omics because it measures inherited variation before downstream regulation and environment reshape molecular state. This temporal and causal position explains why genomic prediction is operationally attractive: DNA can be assayed early, repeatedly interpreted and transferred across tissues. It also explains why genomic data often serve as the scaffold onto which transcript, protein and metabolite information is mapped. Yet a scaffold can be incomplete. If structural variants, gene presence/absence variation or divergent haplotypes are absent from the reference, every downstream layer is interpreted against a biased coordinate system. Plant pangenomes emerged in response to this limitation, recognising that a single reference does not represent species-wide genomic diversity (Bayer *et al.*, 2020).

Graph-based pangenomes extend this logic by representing alternative sequence paths rather than forcing all accessions onto one linear sequence. In crop settings, the gain is not merely computational elegance. Structural variants can alter gene dosage, regulatory context or coding sequence, and may be poorly tagged by single-nucleotide polymorphisms on a linear reference. Wang *et al.* (2023) describe graph pangenomes as a route to more complete representation of sequence and structural diversity, with direct implications for association mapping and breeding. The tomato graph-pangenome study by Zhou *et al.* (2022) provides unusually strong empirical support: incorporating structural variation increased estimated heritability for large sets of expression and metabolite traits and exposed loci that were difficult to recover using the linear reference. This is a cross-omics result because improved genomic representation changed the explanatory power for downstream molecular phenotypes.

The implications for integration are substantial. When transcriptomic or metabolomic traits are treated as outcomes in molecular QTL mapping, incomplete genomic representation can create apparent unexplained variance that is methodological rather than biological. Conversely, a richer genomic coordinate system can increase the discoverability of cis-regulatory, dosage-sensitive and structural effects. This is especially relevant in crops with extensive domestication bottlenecks, polyploidy, introgression or large structural variation. The lesson is that adding more downstream assays cannot compensate for a systematically incomplete representation of inherited variation. Multi-omics design should therefore begin by asking whether the genomic layer captures the allelic and structural diversity of the population being studied (Bayer *et al.*, 2020; Wang *et al.*, 2023; Zhou *et al.*, 2022).

Association and prediction also use genomic information differently. GWAS seeks loci statistically associated with traits, whereas genomic selection aggregates genome-wide signal for prediction without requiring individual variants to be mechanistically resolved. Multi-omics can strengthen GWAS by prioritising associated loci that also affect expression, protein abundance or metabolites. In contrast, genomic selection already benefits from diffuse genome-wide relationships, so extra omics must contribute information not already encoded by marker relationships or by population structure. This distinction helps explain why multi-omics often gives clearer gains in candidate-gene prioritisation than in prediction accuracy. The genetic layer is already an efficient predictor of inherited similarity, whereas mechanistic interpretation requires additional evidence about how variants act (Kremling *et al.*, 2019; Xu *et al.*, 2017).

Table 1. Analytical position of the four core omics layers in crop improvement

Layer	Primary signal	Main breeding value	Principal inferential weakness	Most defensible integration role	Supporting references
Genomics	Inherited sequence, haplotypes and structural variation	Stable marker scaffold; GWAS; genomic prediction; pangenome representation	Association does not by itself reveal mechanism; reference and linkage bias	Anchor downstream QTLs and establish inherited causality	Bayer <i>et al.</i> (2020); Wang <i>et al.</i> (2023); Zhou <i>et al.</i> (2022)
Transcriptomics	Context-dependent RNA abundance and isoform usage	Regulatory prioritisation; eQTL/TWAS; tissue-specific pathway activity	Strong dependence on tissue, developmental stage and environment; RNA is not equivalent to protein	Bridge variants to regulated genes and identify condition-specific pathways	Kremling <i>et al.</i> (2019); Walley <i>et al.</i> (2016)
Proteomics	Protein abundance and post-translational state	Closer representation of molecular function; pathway and regulatory-network refinement	Lower coverage and throughput; protein abundance can diverge from RNA; PTMs add complexity	Test whether transcriptional signals propagate to functional protein states	Jiang <i>et al.</i> (2019); Zhang <i>et al.</i> (2025)
Metabolomics	Small-molecule abundance reflecting pathway output	Biochemical endophenotypes; quality traits; stress physiology; prediction	High environmental sensitivity; incomplete metabolite annotation; platform dependence	Connect genes and pathways to proximal biochemical phenotypes and selection biomarkers	Chen <i>et al.</i> (2014); Riedelsheimer <i>et al.</i> (2012); Tieman <i>et al.</i> (2017); Wen <i>et al.</i> (2014)

Note. eQTL = expression quantitative trait locus; GWAS = genome-wide association study; PTM = post-translational modification; TWAS = transcriptome-wide association study. The table is a conceptual synthesis rather than an exhaustive inventory.

A further issue is population portability. An expression QTL or metabolite QTL detected in one genetic panel can be highly informative mechanistically but may not segregate, have the same linkage structure or retain the same downstream effect in another breeding population. Pangenome-aware analysis can reduce reference bias, but it does not remove population-specific linkage disequilibrium or genotype-by-environment interaction. For breeding, the most valuable genomic scaffold is therefore not simply the most comprehensive species graph; it is a coordinate system and marker representation that preserve relevant variation in the target breeding population and permit downstream omics signals to be traced to stable inherited sources (Ritchie *et al.*, 2015; Wang *et al.*, 2023).

Table 1 summarises the analytical position of the four core omics layers. The key distinction is not a hierarchy of importance but a trade-off between causal stability and phenotypic proximity. Genomics is most stable and transferable; metabolomics is closest to biochemical phenotype but most sensitive to tissue, time and environment. Transcriptomics and proteomics occupy intermediate positions, and their value depends on whether the sampled context is biologically aligned with the target trait.

4. Transcriptomes: Powerful Intermediates but Context-Limited Predictors

Transcriptomics is often the first downstream layer added to crop genomics because RNA sequencing is comparatively mature, genome aligned and interpretable in gene units. Its strongest contribution is not that expression is universally more predictive than genotype, but that it localises regulatory consequences in a biological context. Expression quantitative trait locus mapping can distinguish local and distal regulatory effects, while transcriptome-wide association approaches relate genetically predictable or observed expression to complex traits. In maize, Kremling *et al.* (2019) showed that transcriptome-wide association supplemented GWAS rather than simply reproducing it, illustrating how expression can narrow the search from a broad locus to genes whose regulation is plausibly connected to phenotype.

This complementarity arises because GWAS and transcriptomics interrogate different objects. GWAS tags inherited covariance between variants and traits; expression measures a molecular state produced jointly by genotype, development and environment. When a locus affects both gene expression and a crop trait, cross-layer concordance strengthens a causal hypothesis, particularly if colocalisation, allele-specific expression or functional perturbation supports the same direction. Yet observed expression can also be a consequence of the phenotype or of an environmental exposure, producing reverse-causal or reactive signatures. A drought-stressed plant may exhibit a highly predictive transcriptome even when much of that expression is a response to injury rather than a source of genetic tolerance. Integration therefore requires temporal and biological ordering, not just correlation (Kremling *et al.*, 2019; Walley *et al.*, 2016).

Developmental atlases show the scale of context dependence. Walley *et al.* (2016) integrated molecular networks across maize development and demonstrated that different molecular layers provide distinct views of regulation across tissues. Such atlases are valuable reference resources, but their breeding relevance depends on whether the sampled tissue and stage intersect the causal window for the target trait. A leaf transcriptome collected at one vegetative stage may poorly represent regulatory processes determining flowering time, kernel composition or root architecture. Single-time-point designs can therefore create false negatives for causal genes that are active outside the sampling window and false priorities for highly responsive genes that are not stable determinants of trait variation.

The problem is amplified by genotype-by-environment interaction. Expression is a molecular phenotype whose heritability can vary by environment, and the same genotype can occupy very different transcriptional states under heat, nutrient limitation or pathogen challenge. This flexibility is biologically useful because it reveals plasticity, but it makes transcriptomic predictors vulnerable to covariate shift. A model trained on expression measured in the target environment may achieve high within-environment prediction partly because the transcriptome already encodes the realised stress state. Such information may be unavailable when breeders need to select seedlings or lines before exposure. Consequently, prediction studies should distinguish between a mechanistic objective, where environmentally responsive expression is informative, and an early-selection objective, where the cost and timing of RNA measurement can defeat the purpose of prediction (Anand *et al.*, 2026; Xu *et al.*, 2017).

A further interpretive limitation is that messenger RNA abundance is not a direct proxy for protein abundance or activity. Translation, degradation, localisation and post-translational modification can buffer or amplify transcriptional differences. Transcriptomics is therefore best treated as an intermediate layer whose value is greatest when integrated with genetic anchors and, for key pathways, verified at the protein or metabolite level. Over-reliance on RNA alone risks converting a convenient measurement into an assumed mechanism (Jiang *et al.*, 2019; Walley *et al.*, 2016).

A further distinction is required between observed-expression association and genetically anchored expression association. Observed transcript abundance contains both heritable regulation and environmental response, whereas eQTL-based or genetically predicted expression can isolate part of the inherited component. TWAS can therefore help prioritise genes whose regulation is connected to trait-associated genetic variation, but its interpretation depends on the quality of the expression reference panel and on linkage disequilibrium between regulatory variants. In crops, reference panels are often smaller and more tissue-specific than human TWAS resources, so unstable expression models can propagate uncertainty into downstream association. A significant transcript-trait association should consequently be interpreted as prioritisation evidence rather than a direct demonstration that changing transcript abundance will change the agronomic trait (Kremling *et al.*, 2019; Ritchie *et al.*, 2015).

Transcriptomic integration is also affected by isoforms and cell-type composition. Bulk RNA sequencing averages expression across heterogeneous cell populations, so a gene whose expression changes strongly in a small but functionally critical cell type can be diluted. Conversely, apparent differential expression can arise when the relative abundance of cell types changes between treatments. For breeding, a practical compromise is to use high-resolution transcriptomics to identify cell- or tissue-specific regulatory mechanisms and then test whether genetically anchored markers or targeted expression assays capture those mechanisms at scale (Conesa *et al.*, 2016; Walley *et al.*, 2016).

5. Proteomes and Post-Translational Regulation: Closing the RNA-Function Gap

Proteomics addresses a fundamental weakness of transcript-centred integration: biological function is executed largely by proteins whose abundance and activity are governed by processes beyond transcription. The crop evidence is particularly persuasive where matched transcript and protein measurements disagree. Jiang *et al.* (2019), analysing maize breeding material, reported limited concordance between messenger RNA and protein abundance at population scale and substantial independence between RNA- and protein-based co-expression networks. The study also identified protein quantitative trait loci without corresponding expression quantitative trait loci. These observations demonstrate that transcriptomic integration cannot be assumed to capture the functional consequences of genetic variation.

Proteomic information can strengthen multi-omics inference in at least three ways. First, it filters transcript-level associations by asking whether altered RNA abundance propagates to altered protein abundance. Second, it can reveal regulation that occurs after transcription, including protein turnover and translation. Third, phosphoproteomic, acetylproteomic and other post-translational measurements provide a direct view of signalling and protein-state changes. The wheat multi-omics atlas of Zhang *et al.* (2025) illustrates the latter advantage: transcripts, proteins, phosphoproteins and acetylproteins were profiled across vegetative and reproductive phases, allowing the authors to examine homeolog expression, protein abundance and post-translational regulation in a polyploid crop. Their identification of regulatory modules associated with development, grain quality and disease resistance shows how PTM-aware integration can generate mechanistic hypotheses that RNA alone cannot provide.

The methodological costs are equally important. Proteome coverage is typically less complete than transcriptome coverage, especially for low-abundance, membrane or tissue-specific proteins. Peptide detectability, protein inference and platform-specific missingness complicate quantitative comparison across samples. PTM studies add another layer of sparsity because modification occupancy is not equivalent to total protein abundance. These factors reduce effective sample size and can bias network inference towards abundant, readily measured proteins. In a breeding population, where hundreds or thousands of genotypes may be needed to estimate small effects reliably, comprehensive proteomics can therefore create a tension between molecular depth and population scale (Jiang *et al.*, 2019; Zhang *et al.*, 2025).

For this reason, proteomics may deliver its highest near-term breeding value as a targeted validation and mechanism layer rather than a universal predictor. Genome-wide markers can screen large populations, transcriptomics can identify regulatory candidates in relevant contexts, and focused proteomics can test whether those candidates produce consistent protein-state changes. Broad proteomic prediction may become more practical as throughput and quantification improve, but claims of added predictive value should be evaluated against cost, missingness and the selection stage at which proteins can be sampled without destroying or delaying breeding material (Jiang *et al.*, 2019; Zhang *et al.*, 2025).

Protein quantitative trait locus mapping provides an additional route for genetic anchoring. If a genomic variant is associated with protein abundance but not with messenger RNA, the result points towards translational regulation, protein stability or other post-transcriptional processes. The maize evidence from Jiang *et al.* (2019) shows why this distinction matters during breeding-related adaptation: selection can reshape protein abundance without a parallel change in transcript abundance. Integrating eQTL and pQTL therefore permits a more nuanced classification of regulatory mechanisms than treating RNA as a universal proxy for protein. The strongest causal interpretation occurs when the genomic variant, protein change and phenotype colocalise and when the candidate mechanism is supported experimentally.

Post-translational modifications create a further challenge because protein activity can change without a large change in total protein abundance. Phosphorylation, acetylation and other modifications are often transient and condition-specific, yet they can control signalling, enzyme activity and protein interactions. The wheat atlas of Zhang *et al.* (2025) demonstrates that PTM layers can expose regulatory relationships relevant to disease resistance and grain quality. For crop improvement, the implication is that proteomic integration should not be reduced to total protein quantification when the biological hypothesis concerns signalling or enzyme regulation. At the same time, PTM measurements are particularly sparse and technically demanding, strengthening the case for targeted validation after broader genomic and transcriptomic discovery.

6. Metabolomes as Proximal Phenotypes and Predictive Biomarkers

Metabolites occupy a distinctive position in the genotype-to-phenotype chain because they are products and substrates of biochemical pathways and often relate directly to flavour, nutritional composition, stress physiology and growth. This proximity can make metabolomic profiles highly informative even when individual metabolites are not causal. In association studies, metabolites can serve as intermediate phenotypes with simpler genetic architectures than aggregate traits such as yield. In maize, Wen *et al.* (2014) used metabolome-based GWAS to connect kernel metabolites with genetic loci and biochemical pathways. In rice, Chen *et al.* (2014) profiled extensive metabolic diversity and mapped common genetic variants affecting secondary metabolites, with functional annotation of candidate genes. These studies show how metabolomics can transform an agronomic phenotype into a network of chemically interpretable traits.

Table 2. Representative crop studies demonstrating gains and limits of multi-omics integration

Crop/system	Omics linked	Breeding or biological question	Main contribution	Critical interpretation	Supporting references
Maize hybrids	Genomics + metabolomics	Prediction of heterotic biomass and bioenergy traits	Metabolite-based prediction approached genomic prediction for several traits	Shows physiological profiles can be predictive, but transfer depends on environment and sampling stage	Riedelsheimer <i>et al.</i> (2012)
Maize kernels	Genomics + metabolomics	Genetic control of kernel metabolites	mGWAS linked metabolites to loci and biochemical pathways	Strong for pathway discovery; metabolite associations are not equivalent to whole-plant breeding value	Wen <i>et al.</i> (2014)
Rice diversity panel	Genomics + metabolomics	Natural variation in secondary metabolism	Mapped metabolic variation and prioritised candidate genes with functional evidence	Useful for nutrition and biochemical traits; subspecies structure and environment require careful control	Chen <i>et al.</i> (2014)
Maize development	Genomics + transcriptomics + proteomics	Regulatory architecture across development	Integrated molecular networks across tissues and developmental stages	Atlas depth improves mechanism, but reference tissues do not automatically predict field performance	Walley <i>et al.</i> (2016)
Tomato flavour	Genomics + metabolomics + sensory phenotypes	Recovery of consumer-valued flavour chemistry	Linked chemicals, consumer preference and genetic loci	A strong translational model because molecular traits were tied directly to the breeding objective	Tieman <i>et al.</i> (2017)
Maize agronomic traits	Genomics + transcriptomics + metabolomics	Trait prediction and association mapping	Combined data showed no consistent advantage over the best single layer; genomics often performed best	Demonstrates that integration can be redundant or over-parameterised	Xu <i>et al.</i> (2017)
Tomato graph pangenome	Pangenomics + expression + metabolomics	Missing heritability and structural variation	Richer genomic representation increased explanatory power for downstream molecular traits	Shows that downstream integration depends on the completeness of the genomic coordinate system	Zhou <i>et al.</i> (2022)
Common wheat	Transcriptomics + proteomics + phosphoproteomics + acetylproteomics	Development, grain quality and disease-resistance mechanisms	Resolved regulation across transcript, protein and PTM layers	Mechanistically powerful; very deep molecular profiling remains difficult to scale across breeding populations	Zhang <i>et al.</i> (2025)

Note. mGWAS = metabolome-based genome-wide association study; PTM = post-translational modification. Entries are representative, not exhaustive.

Metabolomics also provides a route from genetic mapping to sensory or nutritional breeding. Tieman *et al.* (2017) integrated chemical profiling, consumer-panel data and genome-wide genetic analysis across diverse tomato material. The study identified flavour-associated chemicals and loci influencing sugars, acids and volatiles, illustrating a particularly convincing translational chain: molecular phenotype, human-perceived trait, genetic association and breeding target were all connected. The strength of this design lies in measuring a phenotype close to the breeding objective rather than assuming that a molecular signature is useful because it is statistically associated with a broad agronomic outcome.

For predictive breeding, metabolomic evidence is promising but heterogeneous. Riedelsheimer *et al.* (2012) found that metabolite profiles could predict complex heterotic traits in hybrid maize with accuracies approaching those obtained from dense single-nucleotide polymorphism data for several traits. This result supports the idea that metabolites integrate many upstream genetic effects into a lower-dimensional physiological state. Yet the same property creates vulnerability to environment: a metabolome can encode recent temperature, nutrition, disease or developmental history. High predictive ability within a trial may therefore combine inherited breeding value with environmental exposure. The relevant question is whether metabolite measurements improve prediction of future, unobserved performance rather than reconstructing conditions already experienced by the sampled plants.

Direct comparative evidence reinforces this caution. Xu *et al.* (2017) compared genomic, transcriptomic and metabolomic prediction for maize agronomic traits and found genomic prediction generally stronger, while combining the three data types produced no consistent improvement over the best single-omics model. This counterexample is important because it tests the common assumption that more layers should monotonically improve prediction. Redundant information, small sample size relative to feature number, scale differences and noisy molecular phenotypes can all make integration neutral or harmful. The result also emphasises that high-dimensional integration needs regularisation and validation specifically designed to detect genuine incremental information.

Metabolite identity is another constraint. Targeted assays provide confident quantification of known compounds but cover a restricted chemical space; untargeted assays detect many features whose structural annotation can remain uncertain. Cross-platform comparability is limited by extraction, chromatography, ionisation, mass-spectrometric settings and data processing. A metabolite marker that is predictive in one laboratory or developmental stage may therefore be difficult to operationalise elsewhere. For breeding applications, analytical reproducibility and standard operating procedures are as important as model sophistication (Song *et al.*, 2026).

The value of metabolomics also depends on whether the feature is treated as a predictor, an intermediate phenotype or a breeding target. These roles require different evidence. As an intermediate phenotype, a metabolite can simplify genetic mapping if it lies closer to enzyme action than a complex whole-plant trait. As a predictor, it must be stable enough that measurement in one context forecasts performance in another. As a target, its biological and nutritional meaning must be established and potential trade-offs considered. Conflating these roles can lead to overstatement: a metabolite with a strong mGWAS signal is not necessarily a useful selection biomarker, and a highly predictive metabolite may be a downstream consequence rather than a causal driver (Chen *et al.*, 2014; Riedelsheimer *et al.*, 2012; Xu *et al.*, 2017).

Metabolic networks also contain substantial redundancy and pleiotropy. A genetic change in one enzyme can alter multiple compounds, while the same metabolite can be influenced by several pathways and transport processes. Multi-omics can help resolve these relationships when transcript or protein changes identify the likely enzymatic step, but annotation remains a limiting factor for untargeted features. The most transferable metabolomics studies therefore combine high-confidence chemical identification, genetic mapping and pathway-level interpretation rather than relying on anonymous mass-spectral features alone. Recent reviews continue to identify metabolite coverage, annotation and integration as major barriers to breeding translation (Song *et al.*, 2026).

Table 2 compares representative crop studies that illustrate both the gains and the limits of cross-omics integration. The strongest examples connect a breeding-relevant phenotype to multiple molecular layers and include an independent validation step. Studies showing little prediction gain are equally informative because they identify circumstances in which molecular complexity exceeds the information added to a strong genomic baseline.

7. Architectures for Multi-Omics Integration

Multi-omics integration is not a single analytical operation. Methods differ in when layers are combined, whether they are supervised by a phenotype, how they handle scale and covariance, and whether the objective is prediction or biological interpretation. A useful first distinction is between sequential integration and joint modelling. Sequential integration analyses each layer separately and combines evidence afterwards, for example by intersecting a GWAS locus with an expression QTL, protein QTL and metabolite QTL. Joint models instead estimate shared structure or phenotype associations across layers simultaneously. Neither strategy is intrinsically superior. Sequential approaches preserve layer-specific assumptions and can be transparent; joint methods can exploit covariance but are more vulnerable to dimensionality and scale mismatch (Bersanelli *et al.*, 2016; Ritchie *et al.*, 2015).

Network integration is attractive in crop biology because genes, proteins and metabolites participate in pathways rather than isolated one-to-one chains. Weighted gene co-expression network analysis provides a widely used framework for grouping correlated variables into modules and relating module summaries to traits (Langfelder & Horvath, 2008). Extending network logic across omics can identify modules whose transcript, protein and metabolite states move together. The biological interpretation is stronger when network edges are supported by independent information such as genetic perturbation, physical interaction or shared QTLs.

Pure correlation networks remain vulnerable to confounding, especially when developmental stage or stress exposure simultaneously changes many molecules. A visually coherent multi-omics network is therefore not proof of causal regulation.

Latent-factor approaches address a different problem: discovering a smaller set of hidden dimensions that explain variation shared across or specific to omics blocks. Multi-Omics Factor Analysis was developed to identify principal factors from heterogeneous multi-omics data without requiring a phenotype label (Argelaguet *et al.*, 2018). This can be valuable for identifying population structure, batch effects, developmental gradients or coordinated biological programmes. In crop panels, however, an unsupervised factor that explains substantial molecular variance may be irrelevant to the agronomic trait of interest. Latent factors require post hoc biological interpretation and should not be equated with breeding value merely because they explain variance.

Supervised multiblock methods target the phenotype more directly. The mixOmics framework and DIABLO use multivariate dimension reduction and feature selection to identify correlated molecular signatures associated with outcomes or classes (Rohart *et al.*, 2017; Singh *et al.*, 2019). Such methods are attractive when breeders want a compact cross-omics signature rather than thousands of features. Their main risk is optimistic feature selection when tuning and evaluation are performed on the same data partitions. Nested resampling, strict separation of training and test sets, and evaluation against single-omics baselines are essential. In structured breeding populations, family relatedness and population structure must also be handled so that a model does not succeed mainly by recognising genetic groups.

Machine learning broadens the model space further, from regularised regression and tree ensembles to multimodal neural networks. Its promise is flexible modelling of non-linear relationships and interactions across modalities. Its limitation is that crop multi-omics datasets often have the opposite geometry from that favoured by deep learning: thousands to millions of features but comparatively few genotypes with complete matched measurements. Recent crop-focused reviews identify substantial methodological diversity and persistent gaps in external validation, environmental representation and interpretability (Montesinos-López *et al.*, 2024; Tsega & Mullualem, 2026). More complex models may fit molecular structure better without improving decisions on future breeding material.

A related problem is information leakage. Molecular traits measured after a plant has experienced the target environment can contain direct signatures of the phenotype or environment. If the intended application is early selection, allowing those measurements into the test set evaluates a different question from prospective breeding. Leakage can also occur through preprocessing: normalisation, batch correction, imputation or feature selection performed on the full dataset before cross-validation can transmit information from test to training samples. Rigorous integration requires that every data-dependent operation be contained within the training resampling loop when predictive performance is estimated (Montesinos-López *et al.*, 2024; Tsega & Mullualem, 2026).

Missingness deserves comparable attention. Multi-omics studies rarely have perfectly matched measurements for every genotype, tissue and environment. Restricting analysis to complete cases can reduce sample size and alter population composition; imputation can create artificial cross-layer correlations if the imputation model borrows information from the same outcomes later predicted. The appropriate strategy depends on whether missingness is technical, biological or design-driven. In many crop studies, a pragmatic solution is modular integration: retain a large genomic backbone for all breeding lines and add smaller molecular layers to strategically selected calibration subsets. The model must then be designed explicitly for partial modalities rather than assuming every future selection candidate will have all assays (Bersanelli *et al.*, 2016; Ritchie *et al.*, 2015).

Finally, algorithm selection should follow the biological and operational question. If the objective is causal prioritisation, colocalisation, mediation and network consistency may matter more than prediction accuracy. If the objective is genomic selection, out-of-sample prediction and selection response dominate. If the objective is biomarker discovery, stability across laboratories and environments is central. Multi-omics methods are therefore best compared not by abstract sophistication but by whether their assumptions match the intended breeding decision (Bersanelli *et al.*, 2016; Ritchie *et al.*, 2015; Tsega & Mullualem, 2026).

Preprocessing is itself part of the integration model. Omics layers differ in count distributions, detection limits, mean-variance relationships and the prevalence of zero or missing values. RNA-seq workflows require appropriate filtering and normalisation before between-sample comparison, and best-practice guidance emphasises quality assessment, experimental design and reproducible processing (Conesa *et al.*, 2016). Proteomic and metabolomic data often need transformations and batch correction adapted to instrument behaviour. Applying a single generic scaling rule to all blocks can make mathematically comparable variables biologically incomparable. Conversely, aggressive normalisation can remove genuine biological shifts when a treatment changes a large fraction of the molecular profile.

Feature selection must also respect the data-generating hierarchy. Selecting features independently from each layer may favour the strongest marginal associations but miss combinations that are jointly informative. Selecting features after concatenating all blocks can favour the largest or least noisy block. Sparse multiblock methods were developed partly to address this problem, but the selected signature can be unstable when predictors are highly correlated (Rohart *et al.*, 2017; Singh *et al.*, 2019). Stability analysis across resamples, environments and genetic backgrounds is therefore as important as the nominal number of selected variables. A breeding biomarker panel should be judged by whether substantially the same biological signal is recovered when the training population changes.

Table 3. Major analytical strategies for integrating crop multi-omics data

Integration strategy	Core idea	Strengths	Common failure modes	Best-suited breeding question	Supporting references
Sequential evidence integration	Analyse layers separately, then combine loci, genes, modules or pathways	Transparent; preserves layer-specific models; easy to audit	May miss cross-layer covariance; arbitrary evidence thresholds	Candidate-gene prioritisation and mechanism triangulation	Ritchie <i>et al.</i> (2015); Scossa <i>et al.</i> (2021)
Correlation / network integration	Construct co-expression or cross-layer modules and relate modules to traits	Biologically interpretable modules; pathway-oriented	Confounding and indirect edges; correlation mistaken for causation	Regulatory network discovery and pathway prioritisation	Langfelder & Horvath (2008); Walley <i>et al.</i> (2016)
Latent-factor models	Represent shared and layer-specific variation with hidden factors	Dimension reduction; handles heterogeneous blocks; useful for exploratory structure	High-variance factors may not be trait-relevant; interpretation can be ambiguous	Discover shared molecular programmes and technical structure	Argelaguet <i>et al.</i> (2018)
Sparse supervised multiblock models	Jointly select correlated features across omics that predict an outcome	Compact signatures; balances integration and feature selection	Overfitting if tuning is not nested; sensitive to scaling and sample size	Biomarker panels and interpretable cross-omics prediction	Rohart <i>et al.</i> (2017); Singh <i>et al.</i> (2019)
Machine-learning / multimodal models	Use flexible non-linear learners to fuse multiple modalities	Can model interactions and complex response surfaces	High data demand; leakage; weak external validation; reduced interpretability	Prediction where large matched datasets and independent validation exist	Montesinos-López <i>et al.</i> (2024); Tsega & Mullualem (2026)
Pangenome-anchored integration	Map molecular phenotypes to graph-aware structural and sequence variation	Reduces reference bias; exposes structural drivers of expression and metabolism	Computational complexity; graph construction quality and population coverage	Trait dissection where structural variation is important	Wang <i>et al.</i> (2023); Zhou <i>et al.</i> (2022)

Note. The categories overlap in practice; a study may combine sequential mapping, network analysis and supervised prediction

Integration can also be framed as a hierarchy rather than a flat concatenation problem. Genomic variation is upstream and largely fixed, while transcript, protein and metabolite states are conditional descendants. Models that encode this ordering through mediation, Bayesian networks or genetically anchored priors may be more interpretable than models that treat every feature as exchangeable. Such structure does not guarantee better prediction, but it reduces the risk of assigning equal causal status to a DNA variant and a stress-induced metabolite. The broader methodological literature supports matching integration architecture to the biological relationship among data types rather than assuming that a single algorithm is optimal for all multi-omics problems (Bersanelli *et al.*, 2016; Ritchie *et al.*, 2015).

Table 3 contrasts major integration strategies in terms of what they preserve, what they assume and where they commonly fail. A recurring pattern is that methods offering greater flexibility also increase the burden of validation. For crop breeding, interpretability and prospective transfer often justify simpler models unless complexity yields reproducible gains in independent environments or breeding cycles.

8. From Molecular Discovery to Breeding Utility

8.1 Candidate-Gene Prioritisation and Causal Inference

Candidate-gene prioritisation is currently the most defensible use of crop multi-omics because independent molecular layers can reduce the ambiguity of broad genomic associations. A GWAS interval containing many genes becomes more informative when one gene shows cis-regulated expression in a relevant tissue, a corresponding protein change and a downstream metabolite phenotype. Each layer contributes a different form of evidence, and convergence lowers the probability that prioritisation rests solely on linkage disequilibrium. The maize TWAS study by Kremling *et al.* (2019) illustrates the value of transcript-level evidence complementing GWAS, while the metabolomic GWAS studies in maize and rice demonstrate how chemically defined phenotypes can connect loci to pathways and candidate genes (Chen *et al.*, 2014; Wen *et al.*, 2014).

Cross-layer convergence still does not establish causality. Shared population structure can generate associations at several levels; expression and metabolites can be reactive; and a single causal variant can affect multiple correlated molecules. Stronger inference requires genetic anchoring, temporal ordering, allele-specific evidence, fine mapping or experimental perturbation. Functional validation was a notable strength in the rice metabolomics study, where candidate genes were experimentally or functionally supported, and in multi-layer atlas studies that connected molecular modules to specific biological processes (Chen *et al.*, 2014; Zhang *et al.*, 2025). For breeding, causal confidence matters because markers linked to a non-causal correlated feature may lose predictive value after recombination or transfer to a different genetic background.

Pangenome-aware integration can improve this process by reducing the chance that the causal variant is absent from the genomic model. The tomato graph-pangenome study is particularly instructive because structural variation changed the fraction of downstream molecular variation that could be attributed to the genome and exposed agronomic associations (Zhou *et al.*, 2022). This suggests a general principle: before adding more downstream omics, investigators should confirm that the genomic layer captures the relevant variant classes. Otherwise, unexplained cross-layer discordance may reflect missing genetic information rather than complex regulation.

8.2 Trait Prediction and Genomic Selection

Prediction is a more demanding test of integration because the model must add information beyond what is already captured by relatedness and markers. Molecular endophenotypes can improve prediction when they summarise many small genetic effects or capture biologically relevant state that is repeatable across target environments. Riedelsheimer *et al.* (2012) showed that maize metabolite profiles could predict complex heterotic traits with performance close to genomic markers, providing a proof of concept for metabolic prediction. Yet metabolite measurements were not uniformly superior, and the result does not imply that every metabolomic layer should be added to genomic selection.

The study by Xu *et al.* (2017) provides a valuable counterweight. Across multiple maize traits and prediction methods, genomic data generally outperformed transcriptomic and metabolomic data, and combining all three omics produced little consistent advantage over the best single source. This is precisely the type of negative incremental evidence needed to calibrate the field. If multi-omics models are reported only when they outperform a baseline, publication bias will make integration appear more reliable than it is. Future studies should pre-specify the genomic comparator and report neutral or negative results, including uncertainty around the difference in prediction accuracy.

The timing of molecular measurement is central. Genomic selection is valuable because DNA can be measured once and used before expensive field phenotyping. If transcript, protein or metabolite profiles must be collected from mature plants under the target stress, the integrated predictor may not shorten the breeding cycle even if its statistical accuracy improves. A practical model may instead use dense genomics on all candidates and measure downstream omics on a calibration subset to improve biological feature engineering, derive environment-specific priors or update models periodically. The economic and logistical value of integration should therefore be evaluated alongside accuracy (Riedelsheimer *et al.*, 2012; Xu *et al.*, 2017).

8.3 Genotype-by-Environment Interaction, Deep Phenotyping and Climate Resilience

Environment is where the strengths and weaknesses of downstream omics are most visible. Transcriptomes, proteomes and metabolomes can reveal molecular responses to heat, drought, nutrient status and pathogens, making them valuable for dissecting

resilience mechanisms. At the same time, these layers can change rapidly with exposure, so a predictive model may learn the realised environment rather than stable genetic tolerance. Recent reviews advocate linking multi-omics with deep phenotyping and environmental measurements to bridge molecular state and field performance (Anand *et al.*, 2026; Tsega & Mullualem, 2026). The logic is sound, but successful integration requires designs that separate genetic sensitivity from simple exposure signatures.

Prospective validation across environments is therefore more informative than random cross-validation within a pooled dataset. Random splits can place closely related genotypes and observations from the same environment in both training and test sets, inflating performance. Stronger designs withhold environments, years, locations or genetic families according to the deployment scenario. Multimodal prediction work in wheat provides a useful example from outside the four core molecular layers: Togninalli *et al.* (2023) fused genomic data with high-throughput phenomic observations and reported improved grain-yield prediction, including evaluation in an unseen environment. Although phenomics is not a core layer of this review, the study illustrates the validation principle that multi-modal models should be tested under the distribution shifts breeders actually face.

For stress-resilience research, matched sampling is equally important. If omics are collected from different individuals, tissues or time points, cross-layer correlations can reflect sampling mismatch. Ideally, the design should track the same genotype across relevant tissues and a time series before, during and after stress, while retaining biological replication. Such designs are expensive, so factorial depth must be balanced against genotype number and field replication. Deep molecular profiling of a small set of genotypes can reveal mechanisms, but it cannot by itself establish breeding population predictability. Conversely, shallow profiling of many genotypes may support prediction but miss transient regulatory events. Deliberate two-stage designs can combine the strengths of both (Anand *et al.*, 2026).

Environmental covariates also offer a route to more transportable models. Rather than treating transcript or metabolite states as universal, models can learn reaction norms in which genomic effects interact with temperature, water status, soil variables or disease pressure. Downstream omics can then be used to discover which pathways mediate those interactions, while the operational prediction model relies on genomic and environmental variables available at deployment. This separation between discovery layers and deployment layers may ultimately be more scalable than measuring every omics modality on every selection candidate (Anand *et al.*, 2026; Togninalli *et al.*, 2023).

8.4 Quality, Nutrition and Consumer Traits

Quality traits often provide the clearest route from metabolomics to breeding because the molecular measurements are themselves components of the target phenotype. Sugars, acids, volatiles, pigments, amino acids and specialised metabolites influence flavour, nutritional value and processing characteristics. In these settings, metabolomics is less an indirect biomarker and more a chemically resolved phenotype. Tieman *et al.* (2017) demonstrated this advantage in tomato by linking consumer preference to flavour-associated chemicals and then to genetic loci. Similar metabolome-wide mapping in rice and maize shows that many specialised metabolites have discoverable genetic architectures (Chen *et al.*, 2014; Wen *et al.*, 2014).

The translational challenge is prioritisation. Breeding rarely aims to maximise a single metabolite without consequences for yield, adaptation or other quality attributes. Multi-omics can identify pleiotropic loci and pathway trade-offs, but selection indices must still reflect breeding objectives and market constraints. Protein and transcript data can help distinguish whether a quality-associated metabolite is controlled by enzyme abundance, regulatory transcription or structural variation. This mechanistic information is useful for marker design and gene editing, yet intervention should follow evidence of stable effects across backgrounds and environments (Chen *et al.*, 2014; Tieman *et al.*, 2017; Wen *et al.*, 2014).

The route from multi-omics discovery to marker-assisted selection or gene editing should therefore include a deliberate deconvolution step. If a trait-associated metabolite and a candidate transcript map to the same genomic region, breeders need to know whether the causal polymorphism is coding, regulatory or structural and whether its effect is stable in elite backgrounds. Pangenome resources can reveal structural alleles missed by standard marker sets, while protein and metabolite measurements can identify the functional consequence of the allele. Only after this chain is sufficiently resolved does targeted editing or introgression become a rational intervention. Otherwise, editing a correlated gene risks disrupting compensatory pathways or producing an effect that disappears in a different environment (Tieman *et al.*, 2017; Zhou *et al.*, 2022).

Multi-omics can also improve selection indices by exposing hidden antagonisms. A locus that increases a desirable metabolite might simultaneously alter growth, stress response or another quality component through pleiotropic pathway effects. Detecting these molecular trade-offs early can help breeders decide whether to use a major locus directly or retain genome-wide selection for the composite objective. This is one reason that mechanistic multi-omics and quantitative breeding should be viewed as complementary. Molecular data can clarify why traits covary, while genomic selection can distribute selection pressure across many loci when no single intervention offers a stable improvement (Chen *et al.*, 2014; Tieman *et al.*, 2017; Wen *et al.*, 2014).

Table 4 evaluates the maturity of major breeding use cases. Evidence is strongest where multi-omics is used to resolve mechanisms or chemically defined traits, and more conditional where the objective is general prediction of complex field traits. The difference reflects both biology and validation: mechanistic discovery can benefit from detailed context-specific molecular measurements, whereas breeding prediction must generalise beyond that context.

Table 4. Current evidence strength for major breeding uses of multi-omics integration

Use case	Current evidence strength	What multi-omics adds	Main limitation	Most informative next validation	Supporting references
Candidate-gene prioritisation	Relatively strong	Convergent evidence across variant, expression, protein and metabolite layers	Association and colocalisation can still be non-causal	Fine mapping plus genetic or molecular perturbation in relevant backgrounds	Chen <i>et al.</i> (2014); Kremling <i>et al.</i> (2019); Wen <i>et al.</i> (2014)
Genomic / trait prediction	Conditional	Potentially captures proximal physiological information and non-linear interactions	Redundancy, overfitting, costly assays and weak portability	Prospective comparison with genomic-only models across breeding cycles	Riedelsheimer <i>et al.</i> (2012); Xu <i>et al.</i> (2017)
Hybrid performance	Promising but context-dependent	Metabolites can summarise parental physiological state and heterotic potential	Environment-sensitive profiles and population specificity	Independent hybrid panels and multi-environment testing	Riedelsheimer <i>et al.</i> (2012)
Quality and flavour	Strong for chemically defined traits	Direct connection between biochemical phenotype, sensory/nutritional outcome and loci	Trade-offs with yield and environment-dependent chemistry	Selection experiments that track both quality chemistry and agronomic performance	Chen <i>et al.</i> (2014); Tieman <i>et al.</i> (2017); Wen <i>et al.</i> (2014)
Stress resilience	Mechanistically strong; predictive evidence still developing	Reveals responsive pathways and molecular plasticity	Stress signatures may reflect exposure rather than inherited tolerance	Time-resolved multi-environment designs with withheld-location validation	Anand <i>et al.</i> (2026); Zhang <i>et al.</i> (2025)
Pangenome-enabled trait dissection	Strong emerging evidence	Captures structural variants missed by linear references and improves molecular-trait mapping	Population representation and graph complexity	Replication in independent germplasm and incorporation into routine breeding genotyping	Wang <i>et al.</i> (2023); Zhou <i>et al.</i> (2022)

Note. Evidence-strength descriptors are qualitative judgements from the present critical synthesis, not formal grades from a risk-of-bias instrument.

9. Persistent Bottlenecks and Why More Omics Does Not Guarantee Better Breeding

The first bottleneck is experimental design rather than computation. Multi-omics datasets are often assembled from assays generated for different purposes, with mismatched tissues, developmental stages, environments and sample sets. Sophisticated integration cannot recover biological comparability that was never built into sampling. If transcriptomes are collected from leaves, proteomes from roots and metabolites from grain, cross-layer associations may describe different biological processes even when they share genotype labels. Prospective studies should define the causal window of the target trait first and then align tissues and sampling times to that window (Anand *et al.*, 2026; Ritchie *et al.*, 2015).

The second bottleneck is dimensional imbalance. Genomic datasets can contain millions of variants, transcriptomes tens of thousands of genes, proteomes fewer reliably quantified proteins and metabolomes a mixture of hundreds to thousands of annotated and unknown features. Without appropriate scaling and regularisation, layers with more variables or greater variance can dominate a joint model for purely mathematical reasons. The problem is exacerbated by sample sizes that are modest relative to feature counts. Integration methods such as sparse multiblock models and latent-factor approaches are designed to reduce dimensionality, but their tuning choices can materially change results (Argelaguet *et al.*, 2018; Rohart *et al.*, 2017; Singh *et al.*, 2019).

The third bottleneck is measurement reproducibility. RNA sequencing has comparatively mature workflows, though batch effects and library composition remain important (Conesa *et al.*, 2016). Proteomics and metabolomics introduce additional platform-specific missingness, identification uncertainty and dynamic-range issues. A prediction gain that depends on laboratory-specific features may not survive technology changes. For breeding programmes, molecular biomarkers need assay robustness over years, operators and locations, not just statistical significance in a discovery experiment. Reporting technical replicates, quality-control samples, feature-identification confidence and missing-data handling is therefore part of translational validity (Jiang *et al.*, 2019; Song *et al.*, 2026; Zhang *et al.*, 2025).

The fourth bottleneck is confounding between biological state and breeding value. Downstream omics can be extraordinarily predictive because they are close to the phenotype, but that proximity may represent realised environmental effects rather than transmissible genetic potential. A mature plant metabolome can partly encode the yield it has already achieved. Such prediction is scientifically interesting but operationally circular if the breeding objective is to select before yield is observed. Model evaluation must therefore align the timing of predictors with the intended decision point (Riedelsheimer *et al.*, 2012; Xu *et al.*, 2017).

The fifth bottleneck is transferability. Crop breeding populations change across cycles as selection alters allele frequencies and recombination breaks historical associations. Environments also change across seasons and locations. A multi-omics signature discovered in one panel may decay faster than a genome-wide relationship model if it depends on transient regulation or linked markers. This does not mean downstream omics are unsuitable; it means their role may be to inform model structure, identify causal features or improve calibration rather than to function as fixed universal biomarkers (Montesinos-López *et al.*, 2024; Tsega & Mullualem, 2026).

The sixth bottleneck is cost and opportunity cost. Measuring four omics layers on every breeding candidate may reduce the number of genotypes, environments or replications that can be evaluated. Because field replication directly informs genotype-by-environment interaction, excessive molecular depth can paradoxically weaken the evidence needed for selection. The optimal design is therefore an allocation problem: which subset of lines, tissues and environments should receive expensive molecular profiling to maximise information for the programme? The answer will differ between discovery, pre-breeding and routine selection (Scossa *et al.*, 2021; Yang *et al.*, 2021).

Finally, the field remains vulnerable to performance inflation. High-dimensional models offer many degrees of freedom for preprocessing, feature selection and hyperparameter tuning. Without nested validation and a fixed baseline, small data sets can generate apparently impressive results that are not reproducible. Recent reviews of machine learning and multimodal prediction in plant breeding emphasise interpretability and external validation as continuing priorities (Montesinos-López *et al.*, 2024; Tsega & Mullualem, 2026). The standard of evidence should therefore shift from whether a multi-omics model can fit a crop dataset to whether it improves a pre-defined breeding decision in an independent deployment setting.

10. Future Directions

The immediate priority is to redesign multi-omics studies around breeding decisions rather than assay availability. A study intended for early selection should collect molecular predictors at stages when they could realistically be measured in practice and evaluate them against genomic baselines under prospective splits. A study intended for causal discovery should instead maximise biological alignment, using the relevant tissue, developmental stage and perturbation, and include orthogonal validation. Mixing these objectives within one analysis often produces ambiguous claims: a context-rich molecular signature may be excellent for mechanism but unsuitable for selection, while a stable genomic model may predict well without explaining biology (Tsega & Mullualem, 2026; Xu *et al.*, 2017).

A second priority is pangenome-aware molecular mapping. As graph references become available for major crops, expression, protein and metabolite QTL analyses should increasingly test structural and presence/absence variants rather than relying solely on single-nucleotide markers from a linear reference. The tomato evidence indicates that richer variant representation can recover downstream molecular heritability and agronomic loci that would otherwise remain hidden (Zhou *et al.*, 2022). The next step is to determine whether these gains translate across independent breeding populations and whether graph-derived variants can be genotyped economically enough for routine selection.

Table 5. Prioritised research agenda for breeding-ready multi-omics

Priority	Unresolved problem	Recommended design or method	Decision-relevant outcome	Supporting references
Pangenome-aware molecular QTL mapping	Linear references omit structural and presence/absence variation	Graph-based genotyping plus matched expression/protein/metabolite QTL mapping in diverse breeding panels	Replicated loci and improved predictive or causal resolution in independent germplasm	Wang <i>et al.</i> (2023); Zhou <i>et al.</i> (2022)
Matched longitudinal multi-environment sampling	Downstream omics confound inherited response with exposure and developmental stage	Repeated tissue-matched sampling before, during and after stress across locations and years	Stable reaction-norm or pathway parameters that predict future environments	Anand <i>et al.</i> (2026); Walley <i>et al.</i> (2016)
Proteome and metabolome standardisation	Platform-dependent missingness and identification reduce transferability	Shared reference materials, QC samples, transparent identification confidence and inter-laboratory replication	Biomarkers reproducible across laboratories and breeding cycles	Jiang <i>et al.</i> (2019); Song <i>et al.</i> (2026); Zhang <i>et al.</i> (2025)
Causal validation of cross-omics networks	Correlation and colocalisation do not prove direction	Genetic perturbation, allele-specific assays and targeted biochemical validation	Interventions that reproducibly alter downstream molecular states and agronomic traits	Chen <i>et al.</i> (2014); Zhang <i>et al.</i> (2025)
Deployment-oriented machine learning	Random cross-validation overstates transfer and can leak environmental information	Nested preprocessing and validation by environment, family and cycle; explicit single-omics baselines	Improved prospective accuracy and selection response, not just in-sample fit	Montesinos-López <i>et al.</i> (2024); Tsega & Mullualem (2026); Xu <i>et al.</i> (2017)
Cost-aware selective integration	Full multi-omics on every line is rarely economically optimal	Large genomic backbone with targeted multi-omics calibration subsets and value-of-information analysis	Higher expected genetic gain per unit cost and time	Riedelsheimer <i>et al.</i> (2012); Scossa <i>et al.</i> (2021)

Note. QC = quality control; QTL = quantitative trait locus. Priorities are ordered by practical leverage rather than by technological novelty.

Third, crop multi-omics needs matched longitudinal and multi-environment sampling. Stress biology in particular requires time because causal regulatory events may precede visible injury and later signatures may be consequences of damage. Repeated sampling of the same genotypes across defined stress trajectories, combined with field phenotyping and environmental sensors, can distinguish stable genetic sensitivity from general exposure responses. Deep molecular profiling need not cover every breeding line; nested designs can profile a strategically chosen diversity panel while genomics and phenotyping cover the larger population (Anand *et al.*, 2026).

Fourth, proteomic and metabolomic standardisation must be treated as a breeding infrastructure problem. Stable reference materials, shared quality-control procedures, transparent missingness rules and confident compound or peptide identification are necessary if signatures are to be transferred among laboratories and years. The goal should not be maximal feature counts but reproducible features whose measurement uncertainty is known. Similarly, RNA-seq analyses should follow mature quality-control and normalisation practices, with preprocessing estimated within training data during predictive evaluation (Conesa *et al.*, 2016; Jiang *et al.*, 2019; Song *et al.*, 2026; Zhang *et al.*, 2025).

Fifth, machine learning should be evaluated through increasingly difficult validation regimes. Random cross-validation is an exploratory starting point, not evidence of breeding deployment. Models should be challenged with withheld environments, breeding cycles, genetic families and, where possible, independent programmes. Multi-modal approaches can be valuable when they integrate complementary information, as illustrated by genomics-plus-phenomics work in wheat, but their added value should be reported relative to parsimonious alternatives and with uncertainty estimates (Togninalli *et al.*, 2023). Interpretability tools should be used to generate testable hypotheses rather than post hoc narratives about complex models.

A longer-term priority is causal multi-omics. Genetic anchors, perturbation experiments, allele-specific measurements and gene editing can convert cross-layer correlation into directional evidence. The wheat atlas demonstrates the explanatory value of PTM-aware networks, but the field now needs systematic tests of whether manipulating the nominated nodes changes both downstream molecular states and agronomic phenotypes in relevant environments (Zhang *et al.*, 2025). Such experiments will clarify which molecular features are causal levers and which are biomarkers of state.

Finally, future work should evaluate economics explicitly. The relevant outcome is not only predictive correlation but genetic gain per unit time and cost. A molecular layer that raises accuracy slightly but halves the number of candidates evaluated may reduce programme-level gain. Conversely, a targeted assay that improves selection for an expensive quality trait could be highly valuable. Decision-analytic studies that compare assay cost, cycle time, population size, selection intensity and expected gain would make multi-omics recommendations substantially more actionable (Scossa *et al.*, 2021; Yang *et al.*, 2021).

Table 5 prioritises these directions according to the bottleneck they address and the evidence needed to claim progress. The common requirement is independent validation aligned with the intended breeding decision. Data integration should be considered successful only when it improves causal resolution, transferability or selection value in a way that survives more stringent testing than the model-development setting.

11. Conclusions

Multi-omics integration has changed crop biology most convincingly by improving the route from genomic variation to mechanism. Genomics remains the most stable and scalable breeding layer, but pangenomes show that even this foundation must represent structural and presence/absence variation adequately. Transcriptomics adds regulatory context, proteomics tests whether regulation reaches functional protein states, and metabolomics resolves biochemical outputs close to many quality and physiological traits. When independent layers converge on the same locus or pathway, candidate-gene prioritisation and mechanistic confidence can improve substantially.

The evidence is more conditional for routine prediction. Metabolic and transcriptomic profiles can be highly informative, yet comparative studies demonstrate that combining omics does not automatically outperform a strong genomic model. Downstream molecular layers are sensitive to tissue, developmental stage and environment, and their apparent predictive power can reflect realised biological state that is unavailable at the intended selection stage. Proteomic and metabolomic assays also impose costs and missing-data constraints that can reduce population size or field replication. The appropriate standard is therefore incremental decision value rather than the number of molecular layers measured.

A breeding-ready multi-omics strategy should be selective, pangenome-aware and validation-driven. It should match molecular sampling to the causal window of the trait, use downstream omics where they add mechanistic or predictive information, and test models under withheld environments, families and breeding cycles. In many programmes, the most scalable architecture will be a broad genomic backbone combined with targeted multi-omics on calibration or discovery subsets, rather than complete molecular profiling of every candidate. The long-term opportunity is not a universal pan-omic predictor but a set of rigorously validated integration designs that connect inherited variation, molecular mechanism and field performance with sufficient reliability to improve selection decisions.

12. Limitations

As a critical narrative review, this synthesis is intrinsically susceptible to selection and interpretive bias despite a transparent search strategy, explicit eligibility criteria and critical appraisal. The review was designed to represent influential, methodologically informative and contrasting evidence rather than to enumerate every eligible crop multi-omics publication, and no quantitative

meta-analysis was attempted. Access was concentrated on openly accessible scholarly databases, repositories, abstracts and verified article records, so some subscription-only evidence may have received less detailed assessment. English-language restriction may also have reduced representation of regional literature.

The evidence base itself is heterogeneous. Studies differ in crop species, population structure, tissue, developmental stage, environment, omics platform, preprocessing, sample size and validation design, which limits direct comparison of reported prediction performance. Many integrative studies are stronger for mechanistic discovery than for prospective breeding validation, and long-term evidence across breeding cycles remains limited. Rapid changes in sequencing, mass spectrometry, graph-genome methods and machine learning mean that analytical capabilities will continue to evolve. These limitations support cautious interpretation of specific algorithms or biomarkers while leaving the broader conclusions about design, validation and context dependence intact.

Declaration of AI Use

This manuscript was prepared through the intellectual contributions of the author(s) to the study design, data analysis, interpretation of results, and scholarly content. Grammarly and ChatGPT were used solely to assist with grammatical refinement and reference formatting during manuscript revision. The author(s) accept full responsibility for the accuracy, integrity, and final content of the manuscript.

Competing Interests

Authors have declared that no competing interests exist.

References

- Anand, S., Reddy, S. B., Shajini, N. M., Jose, E., Shirsat, M. S., Visakh, R. L., Jha, U. C., Sah, R. P., & Beena, R. (2026). Bridging scales: Integrated multi-omics and deep phenotyping for climate resilience in crop plants. *Frontiers in Plant Science*, 17, 1777294. <https://doi.org/10.3389/fpls.2026.1777294>
- Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), e8124. <https://doi.org/10.15252/msb.20178124>
- Baethge, C., Goldbeck-Wood, S., & Mertens, S. (2019). SANRA—a scale for the quality assessment of narrative review articles. *Research Integrity and Peer Review*, 4, 5. <https://doi.org/10.1186/s41073-019-0064-8>
- Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, 6, 914–920. <https://doi.org/10.1038/s41477-020-0733-0>
- Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., & Milanesi, L. (2016). Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinformatics*, 17(Suppl 2), 15. <https://doi.org/10.1186/s12859-015-0857-9>
- Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., Li, Y., Liu, X., Zhang, H., Dong, H., Zhang, W., Zhang, L., Yu, S., Wang, G., Lian, X., & Luo, J. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nature Genetics*, 46(7), 714–721. <https://doi.org/10.1038/ng.3007>
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17, 13. <https://doi.org/10.1186/s13059-016-0881-8>
- Creach, M., Webster, B., Newton, L., Turkus, J., Schnable, J. C., Thompson, A., & VanBuren, R. (2026). Predicting complex phenotypes using multi-omics data in maize. *The Plant Cell*, 38(6), koag185. <https://doi.org/10.1093/plcell/koag185>
- Green, B. N., Johnson, C. D., & Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. *Journal of Chiropractic Medicine*, 5(3), 101–117. [https://doi.org/10.1016/S0899-3467\(07\)60142-6](https://doi.org/10.1016/S0899-3467(07)60142-6)
- Jiang, L.-G., Li, B., Liu, S.-X., Wang, H.-W., Li, C.-P., Song, S.-H., Beatty, M., Zastrow-Hayes, G., Yang, X.-H., Qin, F., & He, Y. (2019). Characterization of proteome variation during modern maize breeding. *Molecular & Cellular Proteomics*, 18(2), 263–276. <https://doi.org/10.1074/mcp.RA118.001021>
- Jin, M., Yan, S., Wu, Y., Zhai, Z., Alseekh, S., Chen, Y., Huang, W., Ma, N., Tao, K., Xiao, K., Zhu, Y., Yu, Y., Shao, Y., Jiang, C., Liu, X., Sun, J., Qu, Y., Wei, W., Li, W., . . . Yan, J. (2026). Multiomics analysis of primary metabolism reveals the genetic basis of nitrogen partitioning modulated by ZmAVT1A-1 in maize. *Nature Genetics*, 58(7), 1722–1733. <https://doi.org/10.1038/s41588-026-02655-2>
- Kremling, K. A. G., Diepenbrock, C. H., Gore, M. A., Buckler, E. S., & Bandillo, N. B. (2019). Transcriptome-wide association supplements genome-wide association in *Zea mays*. *G3: Genes|Genomes|Genetics*, 9(9), 3023–3033. <https://doi.org/10.1534/g3.119.400549>
- Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 559. <https://doi.org/10.1186/1471-2105-9-559>
- Li, G., An, L., Yang, W., Yang, L., Wei, T., Shi, J., Wang, J., Doonan, J. H., Xie, K., Fernie, A. R., Lagudah, E. S., Wing, R. A., & Gao, C. (2025). Integrated biotechnological and AI innovations for crop improvement. *Nature*, 643, 925–937. <https://doi.org/10.1038/s41586-025-09122-8>
- Mahmood, U., Li, X., Fan, Y., Chang, W., Niu, Y., Li, J., Qu, C., & Lu, K. (2022). Multi-omics revolution to promote plant breeding efficiency. *Frontiers in Plant Science*, 13, 1062952. <https://doi.org/10.3389/fpls.2022.1062952>

- Montesinos-López, O. A., Chavira-Flores, M., Kismiantini, Crespo-Herrera, L., Saint Pierre, C., Li, H., Fritsche-Neto, R., Al-Nowibet, K., Montesinos-López, A., & Crossa, J. (2024). A review of multimodal deep learning methods for genomic-enabled prediction in plant breeding. *Genetics*, 228(4), iyae161. <https://doi.org/10.1093/genetics/iyae161>
- Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L., & Melchinger, A. E. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genetics*, 44(2), 217–220. <https://doi.org/10.1038/ng.1033>
- Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2), 85–97. <https://doi.org/10.1038/nrg3868>
- Rohart, F., Gautier, B., Singh, A., & Lê Cao, K.-A. (2017). mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLOS Computational Biology*, 13(11), e1005752. <https://doi.org/10.1371/journal.pcbi.1005752>
- Scossa, F., Alseekh, S., & Fernie, A. R. (2021). Integrating multi-omics data for crop improvement. *Journal of Plant Physiology*, 257, 153352. <https://doi.org/10.1016/j.jplph.2020.153352>
- Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., & Lê Cao, K.-A. (2019). DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17), 3055–3062. <https://doi.org/10.1093/bioinformatics/bty1054>
- Song, R., Wang, Y., Li, Y., Wang, Q., Zhang, S., Sun, D., Sun, G., Sun, L., & Ren, X. (2025). Unraveling the regulatory network of barley grain metabolism through the integrative analysis of multiomics and mQTL. *Nature Communications*, 16, 5544. <https://doi.org/10.1038/s41467-025-60501-1>
- Song, X., Fan, B.-L., Hong, X., Su, P., & Sun, M. (2026). Unveiling plant metabolic diversity: Integrating metabolomics with multi-omics approaches for crop improvement. *Plants*, 15(5), 846. <https://doi.org/10.3390/plants15050846>
- Tieman, D., Zhu, G., Resende, M. F. R., Jr., Lin, T., Nguyen, C., Bies, D., Rambla, J. L., Ortiz Beltran, K. S., Taylor, M., Zhang, B., Ikeda, H., Liu, Z., Fisher, J., Zemach, I., Monforte, A., Zamir, D., Granell, A., Kirst, M., Huang, S., & Klee, H. (2017). A chemical genetic roadmap to improved tomato flavor. *Science*, 355(6323), 391–394. <https://doi.org/10.1126/science.aal1556>
- Togninalli, M., Wang, X., Kucera, T., Shrestha, S., Juliana, P., Mondal, S., Pinto, F., Govindan, V., Crespo-Herrera, L., Huerta-Espino, J., Singh, R. P., Borgwardt, K., & Poland, J. (2023). Multi-modal deep learning improves grain yield prediction in wheat breeding by fusing genomics and phenomics. *Bioinformatics*, 39(6), btad336. <https://doi.org/10.1093/bioinformatics/btad336>
- Tsega, A., & Mullualem, D. (2026). Machine learning for multi-omics data integration in crop improvement: A systematic review. *BMC Bioinformatics*, 27, 81. <https://doi.org/10.1186/s12859-026-06438-8>
- Walley, J. W., Sartor, R. C., Shen, Z., Schmitz, R. J., Wu, K. J., Urich, M. A., Nery, J. R., Smith, L. G., Schnable, J. C., Ecker, J. R., & Briggs, S. P. (2016). Integration of omic networks in a developmental atlas of maize. *Science*, 353(6301), 814–818. <https://doi.org/10.1126/science.aag1125>
- Wang, S., Qian, Y.-Q., Zhao, R.-P., Chen, L.-L., & Song, J.-M. (2023). Graph-based pan-genomes: Increased opportunities in plant genomics. *Journal of Experimental Botany*, 74(1), 24–39. <https://doi.org/10.1093/jxb/erac412>
- Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., Liu, J., Liu, H., Chen, W., Luo, J., & Yan, J. (2014). Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nature Communications*, 5, 3438. <https://doi.org/10.1038/ncomms4438>
- Xu, Y., Xu, C., & Xu, S. (2017). Prediction and association mapping of agronomic traits in maize using multiple omic data. *Heredity*, 119(3), 174–184. <https://doi.org/10.1038/hdy.2017.27>
- Yang, Y., Saand, M. A., Huang, L., Abdelaal, W. B., Zhang, J., Wu, Y., Li, J., Sirohi, M. H., & Wang, F. (2021). Applications of multi-omics technologies for crop improvement. *Frontiers in Plant Science*, 12, 563953. <https://doi.org/10.3389/fpls.2021.563953>
- Zhang, N., Tang, L., Li, S., Liu, L., Gao, M., Wang, S., Chen, D., Zhao, Y., Zheng, R., Soleymaniya, A., Zhang, L., Wang, W., Yang, X., Ren, Y., Sun, C., Wilhelm, M., Wang, D., Li, M., & Chen, F. (2025). Integration of multi-omics data accelerates molecular analysis of common wheat traits. *Nature Communications*, 16(1), 2200. <https://doi.org/10.1038/s41467-025-57550-x>
- Zhang, R., Zhang, C., Yu, C., Dong, J., & Hu, J. (2022). Integration of multi-omics technologies for crop improvement: Status and prospects. *Frontiers in Bioinformatics*, 2, 1027457. <https://doi.org/10.3389/fbinf.2022.1027457>
- 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>

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). This publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© Copyright (2026): Author(s). The licensee is the journal publisher. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:

The peer review history for this paper can be accessed here: <https://pr.sdiarticle5.com/review-history/168136>