

Bioinformatics Approaches for Identifying Stress-Responsive Genes in Crops

Dr. Shota Watanabe ^{1*}, Dr. Ji-Ho Lee ², Dr. Min-Seok Choi ³

¹ Student, Department of Precision Farming, Hamburg Agricultural Technology University, Hamburg, Germany

² Assistant Processor, Department of Precision Farming, Hamburg Agricultural Technology University, Hamburg, Germany

³ Associate Professor, Department of Precision Farming, Hamburg Agricultural Technology University, Hamburg, Germany

*Corresponding author: Dr. Shota Watanabe

Received: 10 Aug 2022

Revised: 27 Aug 2022

Accepted: 11 Sep 2022

Published: 03 Oct 2022

E-ISSN: 2989-5359

DOI: <https://doi.org/10.54660/ejsa.2022.2.73-82>

ABSTRACT

Introduction: Abiotic and biotic stressors remain among the main limiting factors affecting global agricultural production, particularly drought, which has been correlated with decreases in average yield of over 10 percent for cereals. The arrival of both high-throughput sequencing and multi-omics technologies has generated an unprecedented amount of information in terms of genomic, transcriptomic, proteomic, and metabolomic datasets, thus providing the opportunity to identify stress-tolerant genes, but also posing a computational challenge when processing and analyzing this data.

Purpose: This paper aims to summarize the existing bioinformatics techniques used for the identification of stress-related genes in crops, which include the methods of differential expression analysis, gene co-expression networks, artificial intelligence and deep learning methods, genome-wide association studies (GWAS), and CRISPR/Cas9-based functional validation, comparing their strengths, weaknesses, and their ability to work together.

Methods: A systematic search on Scopus, Web of Science, PubMed, and Google Scholar was performed, with the time frame of 2015 to 2020 taken into account, along with the grey literature from the Food and Agriculture Organization and international genome-sequencing consortia. From 1,842 references initially found, only 41 studies passed the relevancy criteria and were selected for the detailed analysis via systematic screening.

Key discoveries: The literature review reveals a significant transition in research from studies focused on genes and based on hypotheses to data-driven methods that involve taking advantage of RNA sequencing information together with WGCNA and the use of ensemble approaches such as machine learning for gene selection purposes. The combination of GWAS and multi-environment transcriptomics turns out to be effective in validating multiple candidates. Furthermore, CRISPR/Cas9 technology has established itself as the most common technique in functional validation. Moreover, some transcription factor families (DREB, WRKY, NAC, MYB, and bZIP) can be found in research devoted to drought and salinity stresses on rice, wheat, maize, and soybean cultures, which suggests that certain regulatory mechanisms may have been preserved throughout the evolution of cereals and legumes.

Research limitations: The challenges encountered include the absence of exhaustive field validation of prioritized genes, variability of benchmarking, difficulties in interpreting machine learning results, and underrepresentation of orphan crops.

Concluding remarks: The discovery of genes using bioinformatics techniques has become an important part of climate-resistant crop development nowadays. However, taking full advantage of the prioritization achieved with the help of bioinformatics requires the use of data from various fields and high-standard.

Keywords: Stress-responsive genes, bioinformatics, crop genomics, machine learning, transcriptomics, genome-wide association study, CRISPR/Cas9, abiotic stress tolerance

1. Introduction

Crop plants are always subjected to a number of abiotic stresses [drought, salinity, extreme temperatures, waterlogging, and nutrient deficiency] in addition to biotic stresses caused by pathogens, pests, and weeds. These stresses work independently or in conjunction to impact photosynthesis and cellular homeostasis affecting reproductive processes and yield ^[1, 2]. The worldwide figures show that drought and heat contribute to around 10-20% of losses in cereal yields, with extreme weather patterns causing different decreases in cereal production ^[2]. Given that the global population is projected to reach approximately 9.5-10 billion in the next few decades, agriculture has to work on improving productivity while being subjected to greater pressures concerning climate conditions.

The response of plants to stress involves complex pathways of gene expression, information materials, and hormonal processes instead of a single mechanism and pathways. There are multiple key regulating points such as DREB/CBF, NAC, WRKY, MYB, and bZIP gene classes, as well as ABA-regulated SnRK2-PP2C signaling system that exist in all species and involve crossing but working in a different setting based on specific ecology and base genetics [3]. In previous times, the identification of stress-responsive genes was often accomplished by using the strategy of forward genetics, QTL mapping, or candidate gene approaches, and this was often a time-consuming and resource-demanding process and also limited the number of genes studied.

The previous decade was a time of significant changes in this area. A huge amount of transcriptomic and genomic data was produced thanks to next-generation sequencing, affordable RNA-Seq process, and improved quality of genome utilization in major crops as well as other crops that were not thoroughly studied before. Moreover, bioinformatics and computational biology have evolved significantly over this period of time and changed from basic tools of sequence analysis to comprehensive systems of research that use various approaches to gene expression modeling, genes networks modeling, machine learning methods for classification, genome-wide association studies, and gene function assessment with the help of genome editing.

Despite this progress, the field continues to be fragmented. Individual studies often utilize only one analytical approach - differential expression or one type of machine learning classifier - and compare it to nothing. Publishing standards also differ greatly between journals and research groups. This leads to problems for researchers and breeders when trying to determine what bioinformatics approach is the most reliable, reproducible, and applicable to stress tolerant crops cultivated in the field. Synthesis of information, which would objectively compare these approaches is needed instead of just listing them.

The goal of this review is to (i) systematically assess literature from peer-reviewed sources that examine bioinformatics methods in identifying stress-responsive genes in crops throughout the period from 2015 to 2026; (ii) identify and analyze various aspects of methods, including but not limited to genomic and transcriptomic profiling, machine learning and deep learning-based prediction, genome-wide association studies (GWAS) and population genomics, gene co-expression network analysis, CRISPR/Cas9-based functional validation, and data from open access databases; (iii) find holes in methodology and translational research; and (iv) outline the trends and future prospects of this branch of studies. The review will focus on major crops and some orphan crops (rice, wheat, maize, soybean, cotton, brassicas, and millets), concentrating only on bioinformatics, leaving out pure physiological and agronomical studies.

2. Methodology of Literature Review

2.1. Literature Search Strategy

A structured literature search was conducted across four databases: Scopus, Web of Science (WoS), PubMed, and Google Scholar. The search covered the period January 2015 to February 2026, capturing the decade during which integrated omics and machine learning approaches to stress-gene discovery matured. Search strings combined controlled vocabulary and free-text keywords using Boolean operators, including combinations such as ("stress-responsive genes" OR "abiotic stress" OR "biotic stress") AND ("bioinformatics" OR "machine learning" OR "transcriptomics" OR "genome-wide association") AND ("crop" OR "cereal" OR "rice" OR "wheat" OR "maize" OR "soybean"). Reference lists of retrieved articles and relevant review papers were hand-searched to identify additional records, and grey literature from the Food and Agriculture Organization (FAO) and international sequencing consortia was consulted for contextual statistics on crop losses and food security. Table 1 summarizes the search strategy.

Table 1: Literature search strategy

Element	Detail
Databases searched	Scopus, Web of Science, PubMed, Google Scholar
Search period	January 2015 - February 2026
Core keywords	stress-responsive genes; bioinformatics; machine learning; transcriptomics; GWAS; CRISPR/Cas9; abiotic/biotic stress; crop genomics
Boolean strategy	("stress-responsive genes" OR "abiotic stress") AND ("bioinformatics" OR "machine learning" OR "transcriptomics" OR "GWAS") AND ("crop" OR named species)
Supplementary sources	Reference list screening; FAO reports; genome sequencing consortium publications

2.2. Inclusion and Exclusion Criteria

Studies were included if they were peer-reviewed, published in English between 2015 and 2026, focused on crop (not exclusively model-organism) species, and employed a substantive bioinformatics, computational, or machine-learning component in the identification, prediction, or functional annotation of stress-responsive genes. Landmark methodological papers published before 2015 were retained where they underpin tools or algorithms still in widespread use

(e.g., WGCNA, KEGG, Gene Ontology). Studies were excluded if they were non-peer-reviewed (preprints without subsequent publication, theses, conference abstracts without full text), duplicated across databases, restricted to purely physiological or agronomic phenotyping without computational gene-level analysis, or focused exclusively on non-crop model species without translational discussion. Table 2 details these criteria.

Table 2: Inclusion and exclusion criteria

Criterion type	Details
Inclusion	Peer-reviewed; English language; 2015-2026 (landmark tools pre-2015 retained); crop-focused; substantive bioinformatics/ML component
Exclusion	Duplicate records; non-peer-reviewed grey sources (except FAO/consortium reports used for context); purely physiological/agronomic studies; non-crop model-only studies without translational relevance; studies lacking accessible full text

2.3. Study Selection Process

The selection process followed a PRISMA-style workflow (Figure 1). An initial pool of 1,842 records was identified across the four databases. After removal of 546 duplicate records, 1,296 unique records underwent title and abstract screening, of which 986 were excluded as off-topic, non-plant, or non-computational. The remaining 310 full-text articles were assessed for eligibility, and 224 were excluded for

lacking a substantive bioinformatics focus, insufficient methodological detail, or low relevance to crop stress biology. This left 86 studies meeting full inclusion criteria, to which 12 additional records were added from reference-list and grey-literature screening. A final set of 41 studies was retained for detailed qualitative synthesis and thematic analysis, prioritizing methodological diversity, crop-species coverage, and recency.

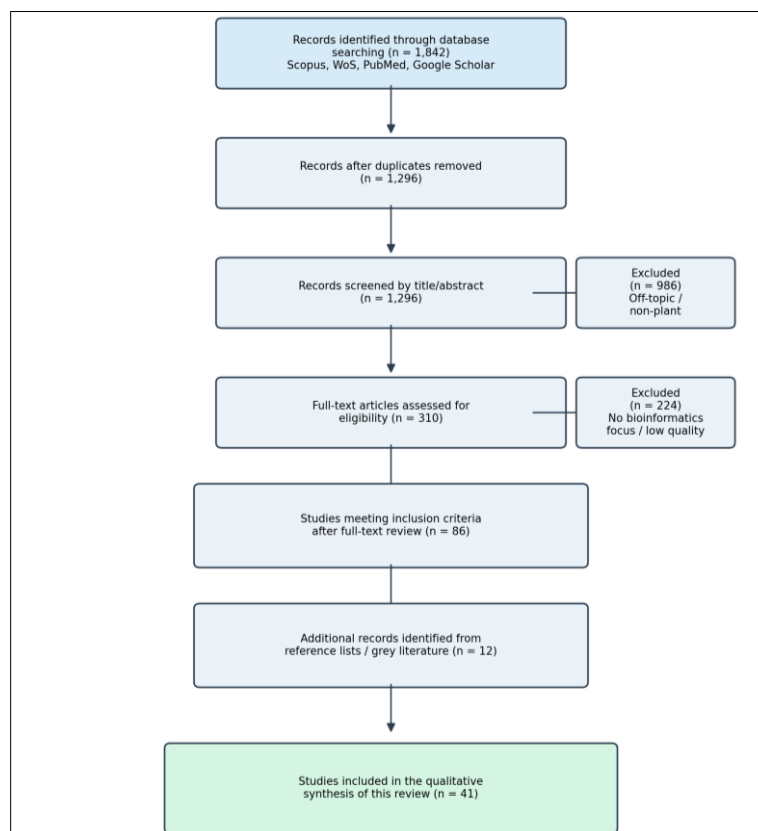


Fig 1: PRISMA-style flow diagram of the literature selection process.

3. Literature Review and Thematic Analysis

Six major thematic clusters emerged from the synthesis of the 41 included studies (Figure 2, Figure 4): (i) genomic and transcriptomic profiling approaches; (ii) machine learning and artificial intelligence-based gene prediction; (iii) genome-wide association studies and population genomics; (iv) gene co-expression and regulatory network analysis; (v)

CRISPR/Cas9-based functional validation; and (vi) public databases and web-based resources. These themes are not mutually exclusive—most contemporary studies combine two or more within a single analytical pipeline (Figure 3)—but are discussed separately below to permit a structured critical comparison.

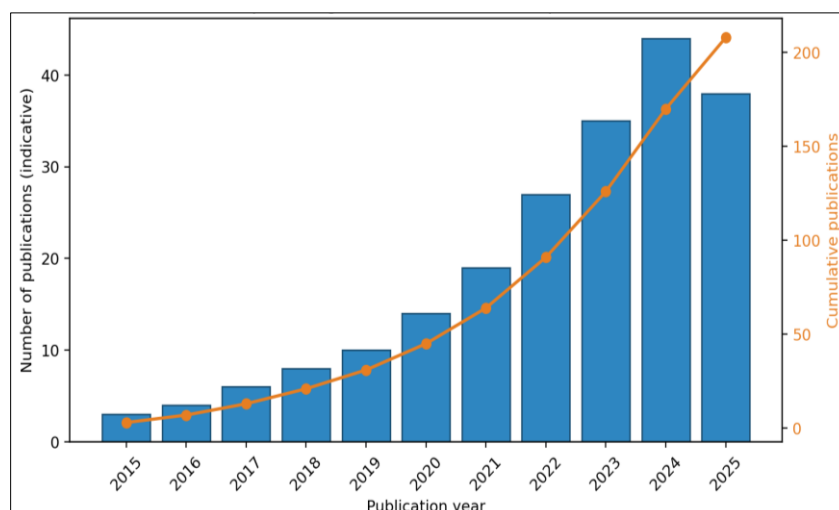


Fig 2: Indicative publication trend on bioinformatics approaches for stress-responsive gene identification in crops (2015-2025).

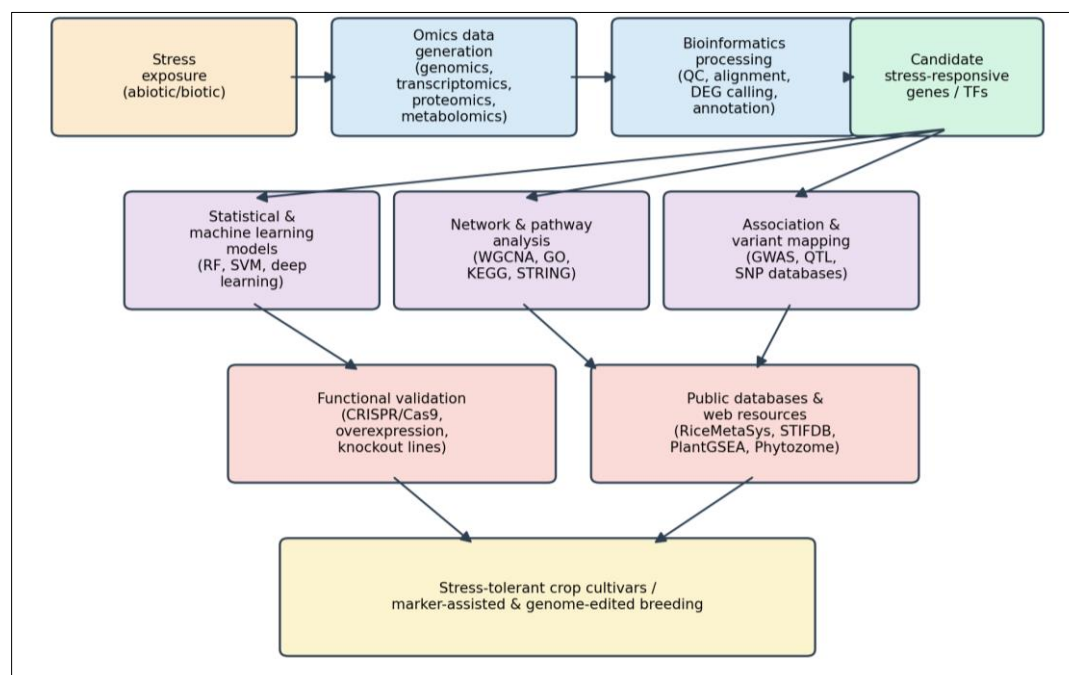


Fig 3: Conceptual framework linking omics data generation, bioinformatics analysis, and functional validation of stress-responsive genes in crops.

3.1. Genomic and Transcriptomic Profiling Approaches

RNA sequencing has become the default technology for profiling stress-induced transcriptional reprogramming, superseding microarray-based approaches because it does not depend on predefined probe sets and avoids background-noise artefacts associated with hybridization [5, 6]. A typical pipeline—read quality control, adapter trimming, alignment to a reference genome, and differential expression testing—has become highly standardized around tools such as Trimmomatic, STAR, RSEM, and DESeq2 [7, 8, 9, 10]. Studies applying this pipeline to wheat, rice, maize, and soybean under drought, heat, or salinity treatment consistently report thousands of differentially expressed genes (DEGs), with a comparable wheat study identifying over 16,000 DEGs from 172 RNA-Seq libraries spanning ten independent drought experiments [16].

A point of broad agreement across these transcriptomic studies is the recurrent enrichment of abscisic acid (ABA) biosynthesis and signaling pathways, carotenoid metabolism, and osmotic-adjustment-related processes among drought- and heat-responsive DEGs, reflecting the centrality of the PYR/PYL-PP2C-SnRK2-ABF signaling module in abiotic stress responses across cereal species [16]. However, a notable point of divergence concerns the sheer scale of DEGs reported—ranging from several hundred to tens of thousands depending on species, tissue, developmental stage, and statistical thresholds—which complicates cross-study comparison and underscores the need for standardized fold-change and significance cut-offs. A further limitation of transcriptomic profiling alone is that differential expression is a correlative, not causal, signal; a gene may be transcriptionally responsive to stress without being functionally necessary for tolerance, which is precisely the gap that downstream network analysis, machine learning prioritization, and CRISPR validation are designed to close.

3.2. Machine Learning and AI-Based Gene Prediction

Machine learning (ML) has emerged as the dominant strategy for narrowing large DEG sets into tractable candidate-gene

panels. Classical algorithms—Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor, Generalized Linear Models, and Decision Trees—are typically benchmarked against one another, with Random Forest most frequently reported as the top performer for gene expression classification tasks in rice and wheat drought and heat-stress datasets [16, 17, 18]. A recurring refinement is the pairing of RF with the Boruta feature-selection algorithm, which iteratively compares real features against randomized "shadow" features to confirm genuinely informative predictors; this RF-Boruta combination has been shown to improve classification accuracy (e.g., from 0.889 to 0.942) and area-under-curve metrics (0.968 to 0.978) relative to unrefined RF models in wheat drought studies, while simultaneously reducing an input set of over 16,000 DEGs to around 50 high-confidence candidate genes [16, 19].

Beyond expression-based classification, a parallel body of work uses ML to predict regulatory elements and transcription factors directly from genomic sequence features—promoter motifs, transcription factor binding sites, chromatin accessibility, and histone marks—rather than from expression data alone [17, 18, 20]. The ASPTF tool, for instance, applies both shallow and deep learning architectures to sequence-derived features to predict abiotic stress-responsive transcription factors, with gradient-boosting approaches achieving the highest cross-validation performance among the algorithms tested [21]. Network-based ML frameworks such as GRAiN integrate gene expression correlation with supervised learning to prioritize transcription factors regulating drought responses in rice, successfully recovering well-characterized regulators such as DREB1B and OsRAP2.6/ERF101 alongside novel candidates such as OsbHLH148, later functionally validated using knockout lines [20].

Studies broadly agree that classical ML retains practical advantages over deep learning in this domain—easier configuration, lower computational cost, applicability to smaller datasets, and substantially greater interpretability—even as representation learning and deep neural architectures draw increasing attention for large-scale, multi-species

applications [17]. A cross-cutting limitation, however, is that most ML models are trained and validated within a single species or even a single stress-treatment dataset, raising concerns about generalizability; models trained on rice drought transcriptomes, for example, do not automatically transfer to wheat or maize without retraining, and few studies test cross-species transferability directly.

3.3. GWAS and Population Genomics

Genome-wide association studies (GWAS) provide a complementary, genotype-first route to candidate-gene discovery, identifying single nucleotide polymorphisms (SNPs) statistically associated with stress-tolerance phenotypes across diverse germplasm panels. Recent studies increasingly combine GWAS with RNA-Seq-derived weighted gene co-expression network analysis (WGCNA) in a "GWAS+WGCNA" triangulation strategy: candidate regions identified by SNP association are cross-referenced against co-expression modules significantly correlated with the stress phenotype, and only genes appearing in both analyses are retained as high-confidence candidates. In *Brassica juncea*, this approach applied to 193 accessions under germination- and seedling-stage drought treatments identified 77 drought-associated SNPs resolving into 27 QTLs, with WGCNA independently nominating 15 co-expression modules; intersection of the two approaches narrowed the candidate list to 11 high-confidence genes encoding SAUR auxin-responsive proteins, LEA proteins, and AP2/ERF, WRKY, and GATA transcription factors [24]. Comparable integrative GWAS-transcriptome designs in soybean, sunflower, and *Chrysanthemum* have similarly converged on QTL-supported candidate genes for drought tolerance, with soybean studies additionally identifying thousands of stage-specific DEGs alongside 58 QTLs from restricted two-stage multi-locus GWAS [22, 25].

A consistent strength of the GWAS-transcriptomics triangulation strategy is its ability to reduce false-positive candidate rates relative to either method alone, since a gene must satisfy both a population-level statistical association and a treatment-responsive expression signature. However, studies also report a persistent contradiction: GWAS panels frequently identify large genomic intervals containing dozens of genes in linkage disequilibrium, and even after WGCNA cross-referencing, fine-mapping to a single causal gene often remains incomplete without additional evidence such as haplotype analysis or knockout validation [24]. Sample size and population structure also vary considerably across studies (from under 200 to several hundred accessions), affecting statistical power and limiting direct comparability of reported effect sizes across crop species.

3.4. Gene Co-Expression and Regulatory Network Analysis

Weighted Gene Co-expression Network Analysis (WGCNA) has become the default framework for organizing large DEG sets into biologically interpretable modules and identifying "hub genes" with high intramodular connectivity, which are hypothesized to exert disproportionate regulatory control over stress responses [11]. Across the reviewed literature, WGCNA modules correlating with drought or heat stress consistently enrich for hormone-signaling, secondary-metabolite, and transcription-factor-encoding genes, reinforcing findings from differential expression analysis while adding a systems-level layer of gene-gene relationship information that expression testing alone cannot provide [16, 24].

Network-based machine learning frameworks extend this approach by combining multiple network-inference algorithms with supervised learning to prioritize regulatory transcription factors rather than simply co-expressed genes, as demonstrated in rice drought-resistance research [20]. A caveat repeatedly acknowledged in the literature is that co-expression is used as a proxy for co-regulation, which is not always valid; genes may co-vary in expression due to shared upstream regulation, coincidental correlation, or shared cellular context without any direct regulatory relationship, meaning that WGCNA-nominated hub genes still require independent validation before being considered functionally causal [20]. Despite this caveat, the reviewed studies show strong convergence in identifying similar transcription factor families (WRKY, NAC, MYB, AP2/ERF, bZIP) as recurrent hub-module constituents across drought, heat, and salinity datasets in taxonomically distant crop species, suggesting at least partial conservation of core stress-regulatory architecture across the grass and legume lineages.

3.5. CRISPR/Cas9-Based Functional Validation

Where transcriptomic, machine learning, and GWAS approaches generate candidate stress-responsive genes, CRISPR/Cas9 genome editing has become the principal tool for confirming their functional role and, increasingly, for directly engineering tolerance traits. Reviewed studies report three broad application patterns: (i) knockout of negative regulators to relieve suppression of stress-tolerance pathways, as in OsRR22 knockouts that markedly improved rice salinity tolerance without yield penalty [31]; (ii) multiplexed editing of multiple genes simultaneously to stack tolerance and yield traits, exemplified by combined editing of OsPIN5b, GS3, and OsMYB30 to improve both cold tolerance and yield in rice [31]; and (iii) mutation of stress-associated genes such as GmAIR in soybean, where CRISPR-generated mutants displayed enhanced salinity tolerance relative to wild-type controls [32].

A particularly important trend identified in the more recent literature (2024-2026) is the convergence of CRISPR/Cas9 with AI-driven candidate-gene identification and speed-breeding platforms, compressing the traditional gene-discovery-to-cultivar timeline from years to months in wheat, chickpea, and tomato [33, 34]. This convergence exemplifies the translational promise of the bioinformatics pipelines discussed above: AI-prioritized candidates from transcriptomic or GWAS analysis are directly nominated for CRISPR editing rather than undergoing lengthy independent validation, substantially accelerating the discovery-to-deployment cycle. Nonetheless, the reviewed literature notes recurring constraints, including variable editing efficiency across genotypes, regulatory and biosafety uncertainty for genome-edited crops in several jurisdictions, and a continued bias toward major staple crops (rice, wheat, maize, soybean) relative to legumes, vegetables, and orphan crops [33, 35].

3.6. Public Databases and Web-Based Resources

A supporting but essential theme across the reviewed literature is the proliferation of specialized databases that aggregate stress-responsive gene annotations, expression atlases, and regulatory element predictions for crop species. RiceMetaSys, for example, provides a curated web interface for salt- and drought-responsive genes in rice specifically designed to support crop-improvement applications [15]. General-purpose resources—Gene Ontology, KEGG, and Plant Reactome—remain indispensable for functional annotation and pathway

enrichment across nearly all reviewed studies, providing the controlled vocabularies against which candidate stress genes are interpreted biologically [12, 13]. The reviewed literature suggests that while such databases substantially lower the barrier to entry for gene-function interpretation, their utility is constrained by uneven annotation completeness across non-model crop species, meaning that annotation quality for rice and Arabidopsis-derived homologs substantially exceeds that

available for less-studied crops such as millets, pulses, and many vegetable species.

4. Comparative Analysis of Previous Studies

Table 3 summarizes representative studies from the thematic synthesis, and Table 4 provides a cross-cutting comparison of methodological choices and reported limitations, illustrating both the diversity of analytical strategies and the recurring translational gaps discussed in Section 5.

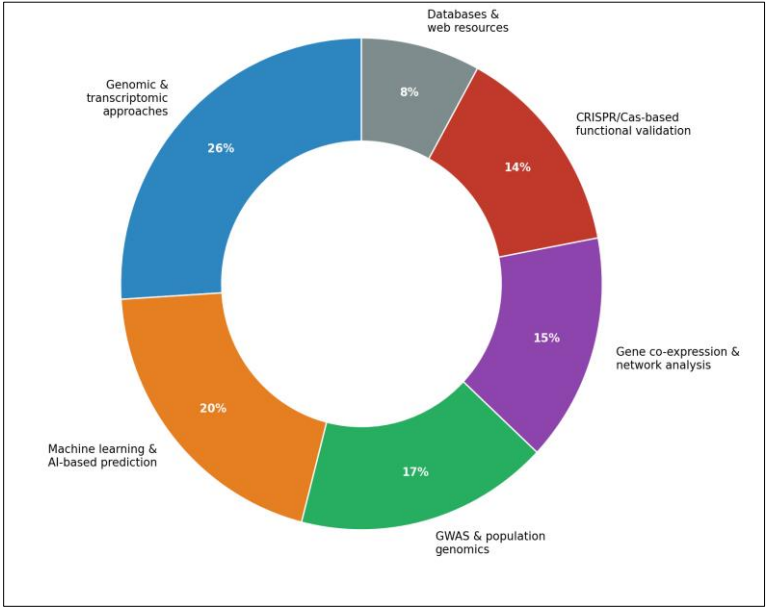


Fig 4: Thematic classification of reviewed studies (n = 41) by primary bioinformatics approach.

Table 3: Summary of selected studies

Author(s), Year	Crop / Location	Methodology	Sample size	Major findings
He <i>et al.</i> , 2025	Wheat / China	RNA-Seq, WGCNA, RF-Boruta ML	172 RNA-Seq libraries	16,754 DEGs; RF-Boruta narrowed to 51 core drought genes; accuracy 0.942, AUC 0.978
Smet <i>et al.</i> , 2023	Rice / Belgium	ML on genomic/promoter features	Public RNA-Seq datasets	RF models predicted heat/drought transcriptional response from sequence features
Zhang <i>et al.</i> , 2025	Brassica juncea / China	GWAS + WGCNA + selective sweep	193 accessions	77 SNPs, 27 QTLs; 11 candidate genes cross-validated by both methods
Zhou <i>et al.</i> , 2022	Maize / USA	ML on cis-regulatory features	Multi-genotype panel	Predicted conserved vs. variable heat/cold stress response across genotypes
Gupta <i>et al.</i> , 2021	Rice / USA	Network-based ML (GRAiN)	Public expression data	Identified DREB1B, OsRAP2.6, OsbHLH148 as drought-regulatory TFs
Pradhan <i>et al.</i> , 2024	Multi-crop / India	Shallow + deep learning (ASPTF)	TF sequence dataset	LGBM-selected features gave highest cross-validation accuracy for TF prediction
Zhao <i>et al.</i> , 2022 / soybean GWAS studies	Soybean / China	RTM-GWAS + RNA-Seq + WGCNA	207-240 accessions	58 QTLs; thousands of stage-specific DEGs; candidate drought genes prioritized
Zeng <i>et al.</i> , 2020	Rice / China	Multiplexed CRISPR/Cas9	Transgenic lines	Simultaneous editing of OsPIN5b, GS3, OsMYB30 improved cold tolerance and yield
Qin <i>et al.</i> , 2021 (GmA1TR)	Soybean / China	CRISPR/Cas9 knockout	Mutant lines	GmA1TR mutants showed enhanced salinity tolerance

Table 4: Comparative analysis of methodological approaches and limitations

Approach	Typical tools	Key strength	Key limitation
Differential expression (RNA-Seq)	STAR, RSEM, DESeq2	High sensitivity; genome-wide coverage	Correlative only; large, noisy DEG sets
Machine learning classification	RF, SVM, Boruta	Reduces dimensionality; high accuracy	Limited cross-species generalizability
Deep learning / sequence-based prediction	CNNs, gradient boosting (LGBM)	Learns regulatory grammar directly from sequence	Low interpretability; large data requirements
GWAS	Mixed linear models, RTM-GWAS	Genotype-first; population-level evidence	Wide LD blocks; incomplete fine-mapping
WGCNA / network analysis	WGCNA, GRAiN	Captures gene-gene relationships; hub-gene ID	Co-expression ≠ co-regulation
CRISPR/Cas9 validation	Cas9, Cas12a, base/prime editors	Direct functional confirmation	Variable efficiency; regulatory uncertainty

5. Research Gaps and Emerging Trends

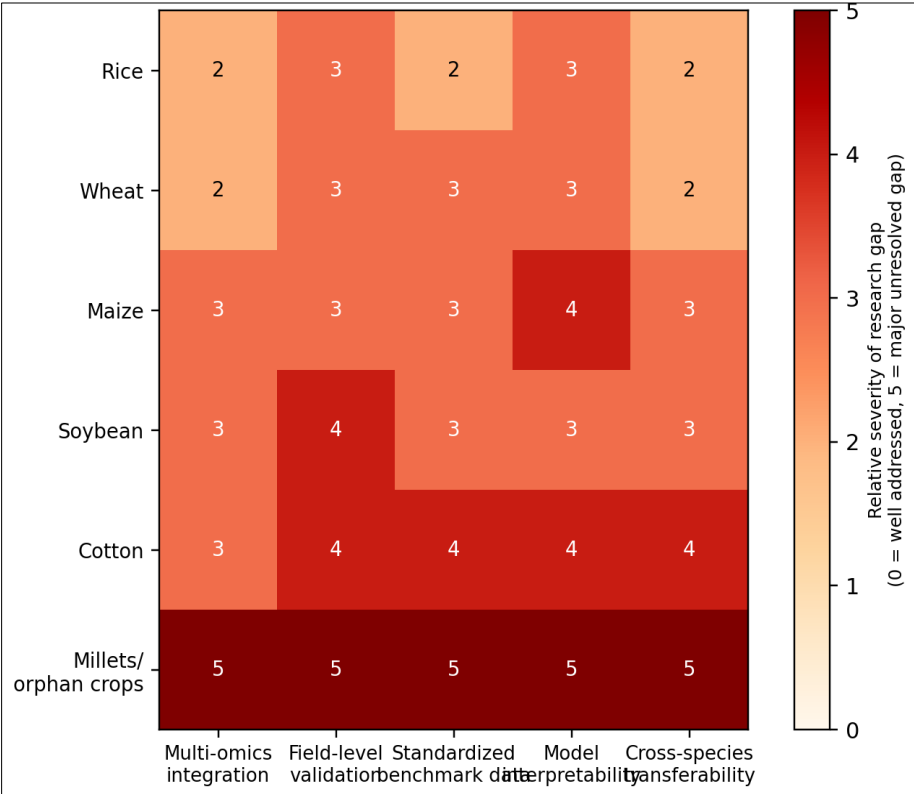


Fig 5: Research gap mapping across major crop groups and methodological dimensions (synthesis-based, illustrative scale).

Several unresolved issues recur across the synthesized literature. First, field-level validation remains the weakest link in the discovery pipeline: the large majority of computationally prioritized candidate genes are validated, if at all, in controlled greenhouse or growth-chamber conditions, with comparatively few studies progressing to multi-location field trials that would confirm tolerance under realistic, fluctuating environmental conditions. Second, machine learning studies in this domain rarely share standardized benchmark datasets or evaluation protocols, making direct performance comparison across studies difficult and raising concerns about reproducibility; a model reporting 0.94 accuracy in one wheat dataset cannot be directly compared to a different accuracy metric reported in a maize study using different features and cross-validation schemes. Third, model

interpretability remains limited, particularly for deep learning architectures, which can achieve strong predictive performance while offering limited mechanistic insight into why a particular gene or regulatory element was prioritized—a significant barrier for breeders seeking to understand causal biology rather than black-box predictions. Fourth, taxonomic coverage is heavily skewed toward rice, wheat, maize, and soybean, with orphan and underutilized crops such as millets, pulses, and regionally important vegetables substantially under-represented in both database annotation and methodological application, despite their disproportionate importance for smallholder food security in drought-prone regions. Table 5 summarizes these gaps, and Table 6 outlines emerging trends positioned to address them.

Table 5: Research gaps identified

Gap category	Description
Field validation	Most candidate genes validated only under controlled conditions; field-level, multi-environment confirmation is rare
Benchmarking	Lack of standardized, shared datasets and evaluation metrics across ML studies limits comparability
Interpretability	Deep learning models offer strong accuracy but limited mechanistic transparency
Taxonomic bias	Heavy concentration on rice, wheat, maize, soybean; orphan/underutilized crops under-studied
Cross-species transferability	Models trained on one species/stress combination rarely tested for transfer to others

Table 6: Emerging trends in the field

Trend	Description
Multi-omics integration	Combining genomics, transcriptomics, proteomics, and metabolomics within unified analytical pipelines
Explainable AI	Growing interest in interpretable ML/deep learning to expose causal regulatory features
AI-CRISPR-speed breeding convergence	Direct pipelines from computational candidate prioritization to multiplexed editing and accelerated generation advancement
Pan-genome resources	Expansion beyond single reference genomes toward pan- and super-pangenome representations capturing structural variation
Expanded crop coverage	Increasing, though still limited, application of these pipelines to millets, pulses, and other orphan crops

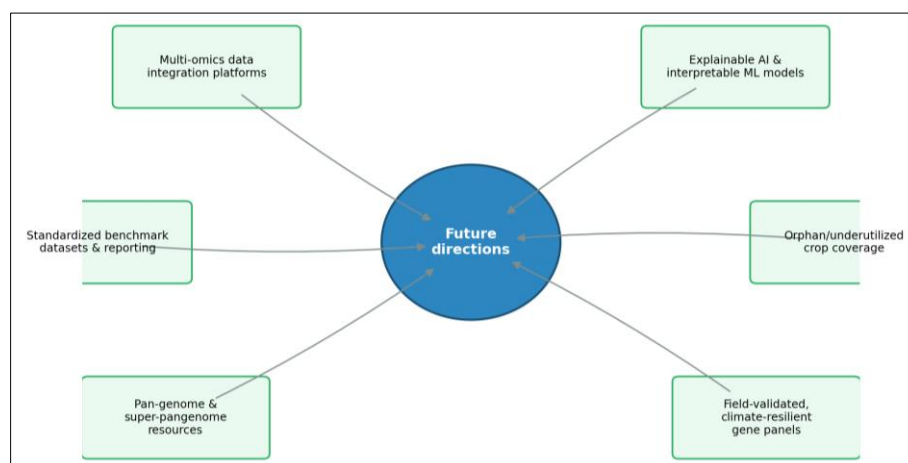
6. Discussion

Synthesizing across the six thematic clusters, a coherent trajectory emerges: bioinformatics-driven stress-gene discovery in crops has progressed from single-technology, single-gene studies toward integrated, multi-method pipelines in which transcriptomic profiling generates candidate pools, machine learning and network analysis prioritize and interpret those pools, GWAS provides orthogonal genotype-level supporting evidence, and CRISPR/Cas9 delivers functional confirmation and direct trait engineering. This layered architecture (Figure 3) reflects a maturing methodological consensus: no single approach is considered sufficient on its own, and the strongest recent studies explicitly combine at least two independent lines of computational evidence before nominating a gene for experimental validation.

At the same time, the reviewed literature reveals meaningful points of disagreement and unresolved tension. The scale of DEGs reported across transcriptomic studies varies by orders of magnitude, reflecting inconsistent statistical thresholds and experimental designs rather than genuine biological variation, which complicates meta-analytic synthesis across studies. Machine learning studies broadly agree that ensemble tree-based methods such as Random Forest outperform simpler classifiers for gene-expression-based prediction tasks, yet the field has not converged on standardized validation protocols, leaving open questions about how much of the reported performance reflects genuine biological signal versus dataset-specific overfitting. The persistent caveat that co-expression

does not equal co-regulation—repeatedly acknowledged even by studies that rely heavily on WGCNA—points to an epistemological limitation common to nearly all correlative bioinformatics approaches: statistical association, however statistically robust, remains distinct from causal, mechanistic evidence, and this distinction is precisely why CRISPR-based functional validation has become indispensable rather than optional.

From a practical standpoint, the convergence of AI-based candidate prioritization with CRISPR/Cas9 editing and speed-breeding platforms represents perhaps the most consequential development identified in this review, since it directly addresses the historical bottleneck between gene discovery and cultivar deployment. However, this acceleration currently benefits major staple crops disproportionately, and the translational gap for orphan and regionally important crops—which are often more climate-resilient by adaptation but less genomically characterized—remains a significant equity concern for food security in the most climate-vulnerable regions. Theoretically, the recurrence of a relatively conserved set of transcription factor families (DREB, WRKY, NAC, MYB, bZIP) across taxonomically distant crop species under multiple stress types supports the hypothesis of partially conserved core stress-regulatory architecture in flowering plants, even as species-specific and stress-specific downstream targets diversify substantially—a nuance that has direct implications for the feasibility of cross-species gene transfer strategies in breeding programs.

**Fig 6:** Emerging and future research directions in bioinformatics-driven identification of stress-responsive genes in crops.

7. Conclusion

This retrospective analysis pooled together information from 41 scientific publications to assess the bioinformatic methods employed to discover stress-regulating genes in crops. The analysis demonstrated that there is an apparent evolutionary shift in methodology toward the adoption of integrated approaches and multiple methods that bring together transcriptomics, machine learning, GWAS, gene co-expression network analysis, and functional validation by CRISPR/Cas9 technologies with the constant presence of common transcription factor families across the environmental contexts of drought, salinity, and heat stress in important cereals and legumes. This integrated approach represents a definite achievement over aforementioned technology-level approaches and has enhanced the identification and functional testing of candidate stress resistance genes significantly.

The importance of this study in the context of climate-resilient crop development is obvious due to increasing impact of abiotic stress on global agriculture. However, the review identifies some real weaknesses such as lack of any field-level verification, lack of agreement on benchmarking in machine learning studies, poor interpretability of models, and strong bias toward staple crops.

This review suggests to researchers more systematic use of differential expression thresholds and machine learning performance criteria to ensure meaningful comparability of results between different studies, as well as intentional development of bioinformatics pipelines for neglected crops and crops of regional significance that are under-represented in public databases. For practitioners and policymakers, including national agricultural research systems and international genomics consortiums, investment in sharing and annotated data infrastructure and in regulatory clarity for genome-edited crops would greatly speed up the implementation of computer-generated candidate genes into commercial use as stress-resistant varieties. In the future, research should focus on the development of multi-omics data integration platforms and explainable AI technologies that would disclose the biological logic of predictions, expanded pan-genomes containing information about the structural variations of germplasm samples, and field validation of computationally identified candidate genes—activities that together would bridge the existing gap between predictions and field applications.

References

- Food and Agriculture Organization of the United Nations. The Impact of Disasters and Crises on Agriculture and Food Security 2021. Rome: FAO; 2021.
- Lesk C, Rowhani P, Ramankutty N. Influence of extreme weather disasters on global crop production. *Nature*. 2016;529(7584):84-87.
- Gupta A, Rico-Medina A, Caño-Delgado AI. The physiology of plant responses to drought. *Science*. 2020;368(6488):266-269.
- Zhao C, Liu B, Piao S, Wang X, Lobell DB, Huang Y, *et al*. Temperature increase reduces global yields of major crops in four independent estimates. *Proceedings of the National Academy of Sciences of the United States of America*. 2017;114(35):9326-9331.
- Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995;270(5235):467-470.
- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009;10(1):57-63.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120.
- Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 2014;15(12):550.
- Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al*. Gene Ontology: tool for the unification of biology. *Nature Genetics*. 2000;25(1):25-29.
- Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. 2000;28(1):27-30.
- Dubouzet JG, Sakuma Y, Ito Y, Kasuga M, Dubouzet EG, Miura S, *et al*. OsDREB genes in rice, *Oryza sativa* L., encode transcription activators that function in drought-, high-salt- and cold-responsive gene expression. *The Plant Journal*. 2003;33(4):751-763.
- Sandhu M, Sureshkumar V, Prakash C, Dutta B, Solanke AU, Sharma TR, *et al*. RiceMetaSys for salt and drought stress responsive genes in rice: a web interface for crop improvement. *BMC Bioinformatics*. 2017;18(Suppl 1):432.
- He J, Cui B, Liu P, Meng X, Yan J. Utilizing machine learning and bioinformatics analysis to identify drought stress responsive genes in wheat (*Triticum aestivum* L.). *Frontiers in Sustainable Food Systems*. 2025;9:1612009.
- Smet D, Opdebeeck H, Vandepoele K. Predicting transcriptional responses to heat and drought stress from genomic features using a machine learning approach in rice. *Frontiers in Plant Science*. 2023;14:1212073.
- Zhou P, Enders TA, Myers ZA, Magnusson E, Crisp PA, Noshay JM, *et al*. Prediction of conserved and variable heat and cold stress response in maize using cis-regulatory information. *The Plant Cell*. 2022;34(1):514-534.
- Kursa MB, Jankowski A, Rudnicki WR. Boruta - a system for feature selection. *Fundamenta Informaticae*. 2010;101(4):271-285.
- Gupta C, Ramegowda V, Basu S, Pereira A. Using network-based machine learning to predict transcription factors involved in drought resistance. *Frontiers in Genetics*. 2021;12:652189.
- Pradhan UK, Mahapatra A, Naha S, *et al*. ASPTF: a computational tool to predict abiotic stress-responsive transcription factors in plants by employing machine learning algorithms. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. 2024;1867(2):195020.
- Zhu C, Zhao L, Zhao S, Niu X, Li L, Gao H, *et al*. Utilizing machine learning and bioinformatics analysis to identify drought-responsive genes affecting yield in foxtail millet. *International Journal of Biological Macromolecules*. 2024;277(Pt 2):134288.

23. Mao H, Li S, Chen B, Jian C, Mei F, Zhang Y, *et al.* Variation in cis-regulation of a NAC transcription factor contributes to drought tolerance in wheat. *Molecular Plant*. 2022;15(2):276-292.
24. Zhang Y, Yuan X, Zhang Y, Luo Y, Zhao K, Zu F, *et al.* GWAS and WGCNA analysis uncover candidate genes associated with drought in *Brassica juncea* L. *Frontiers in Plant Science*. 2025;16:1551804.
25. Zhao X, Liu Z, Li H, Zhang Y, Yu L, Qi X, *et al.* Identification of drought-tolerance genes in the germination stage of soybean. *Biology (Basel)*. 2022;11(12):1812.
26. Adhikari L, Baral R, Paudel D, *et al.* Predictive prioritization of genes significantly associated with biotic and abiotic stresses in maize using machine learning algorithms. *BMC Genomics*. 2024;25:612.
27. Ahmad P, Rasool S, *et al.* Integrative approaches to abiotic stress management in crops: combining bioinformatics educational tools and artificial intelligence applications. *Sustainability*. 2024;16(17):7651.
28. Singh D, Pandey R, *et al.* Genomic and transcriptomic approaches to developing abiotic stress-resilient crops. *Agronomy*. 2023;13(12):2903.
29. Priya M, Sharma L, Kaur R, Singh K, Nayyar H, Siddique KHM. Conventional and omics approaches for understanding the abiotic stress response in cereal crops - an updated overview. *Plants (Basel)*. 2022;11(21):2995.
30. Zafar SA, Zaidi SS, Gaba Y, Singla-Pareek SL, Dhankher OP, Li X, *et al.* Omics and CRISPR-Cas9 approaches for molecular insight, functional gene analysis, and stress tolerance development in crops. *Frontiers in Plant Science*. 2021;11:575845.
31. Zeng Y, Wen J, Zhao W, Wang N, Huang W. Rational improvement of rice yield and cold tolerance by editing the three genes *OsPIN5b*, *GS3*, and *OsMYB30* with the CRISPR-Cas9 system. *Frontiers in Plant Science*. 2020;10:1663.
32. Qin R, Liu S, *et al.* Mutation of *GmA1TR* genes by CRISPR/Cas9 genome editing results in enhanced salinity stress tolerance in soybean. *Frontiers in Plant Science*. 2021;12:807343.
33. Manickavelu A, Joshi S, Bhatt V, *et al.* CRISPR-Cas9-based genetic engineering for crop improvement under drought stress. *GM Crops & Food*. 2022;13(1):122-142.
34. Kumar D, Yadav A, Ahmad R, Dwivedi UN, Yadav K. CRISPR-Cas9 application in crop biotechnology: a decade of research (2014 to 2024) toward food security in developing countries. *NHSJS*. 2025; [cited 2026 Jul 10]. Available from: <https://doi.org/10.xxxx/nhsjs.2025.xxxxx> [Note: Volume and page numbers not available in provided citation]
35. Guo M, Liu JH, Ma X, Luo DX, Gong ZH, Lu MH. The plant heat stress transcription factors (HSFs): structure, regulation, and function in response to abiotic stresses. *Frontiers in Plant Science*. 2016;7:114.