PLANT SCIENCE TODAY
Vol 12(sp4): 01-13
https://doi.org/10.14719/pst.10175

eISSN 2348-1900

REVIEW ARTICLE

From genetic markers to gene expression markers: A comprehensive review of methods and applications in associative transcriptomics

Jayabharathi E[1#], Geethanjali Subramaniam [1#*], Thamaraikannan Sivakumar[2], Nithyananth Hemanth Sadhana[1], Raghu Rajasekaran[1], Boominathan Parasuraman[3] & Harish S[4]

[1]Centre for Plant Molecular Biology and Biotechnology, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India
[2]Division of Genomic Resources, Indian Council of Agricultural Research - National Bureau of Plant Genetic Resources, New Delhi 110 012, India
[3]Department of Plant Physiology, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India
[4]Department of Oilseeds, Centre for Plant Breeding and Genetics, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India
#Both authors contributed equally

*Correspondence email - geethanjalitnau@yahoo.com

Received: 22 June 2025; Accepted: 09 August 2025; Available online: Version 1.0: 17 October 2025

Cite this article: Jayabharathi E, Geethanjali S, Thamaraikannan S, Nithyananth HS, Raghu R, Boominathan P, Harish S. From genetic markers to gene expression markers: A comprehensive review of methods and applications in associative transcriptomics. Plant Science Today. 2025;12(sp4):01-13. https:/doi.org/10.14719/pst.10175

Abstract

Associative Transcriptomics (AT) is an advanced integrative approach that synergizes Genome Wide Association Studies (GWAS) with transcriptome profiling to unravel the complex genetic architecture underlying phenotypic traits. Unlike conventional GWAS, which relies solely on DNA marker polymorphism such as Single Nucleotide Polymorphisms (SNPs), AT incorporates both SNPs and Gene Expression Markers (GEMs), thereby enhancing the resolution and sensitivity of trait discovery. This dual-marker strategy is particularly valuable in genetically complex or polyploid species, where linkage disequilibrium patterns and gene redundancy often obscure signals in traditional analyses. Over the past decade, AT has emerged as a powerful tool for identifying trait-associated loci in several agriculturally important crops, including *Brassica species* and *Triticum aestivum.* The increasing availability of high-throughput sequencing platforms, combined with advancements in machine learning and statistical modelling, has accelerated the ability to integrate transcriptomic and genomic data on a large scale. Recent methodological innovations include the use of pan-transcriptomic references, co-expression networks and multi-omics integration to refine trait mapping. Despite its potential, AT faces ongoing challenges, including the management of population structure, transcriptome complexity across tissues and developmental stages and the dynamic influence of environmental factors on gene expression. AT holds considerable promise in supporting precision breeding programs, enabling targeted genetic interventions and advancing the development of climate-resilient crop varieties.

Keywords: associative transcriptomics; gene expression markers; GEM; GWAS; RNA-Seq; SNP

Introduction

The study of genetic variation and trait inheritance has evolved from classical Mendelian genetics to advanced genome wide analyses, shaping our understanding of complex traits. Mendel's pioneering experiments with pea plants in the 19th century provided the foundation for genetic inheritance, which were later expanded in the early 20th century with Sturtevant's development of the first genetic linkage map. By the mid-20th century, population genetics models by Fisher and Wright laid the groundwork for quantitative genetics, leading to the concept of quantitative trait loci (QTLs). In 1923, Karl Sax first demonstrated the concept of QTL mapping by using seed coat colour as a morphological marker, to track the inheritance of seed size in *Phaseolus vulgaris*, suggesting genetic linkage between qualitative and quantitative traits (1). The late 20th century saw the advent of DNA-based molecular markers enabling the first high-resolution QTL mapping and marker-assisted selection (MAS) strategies in crops (2) (Fig. 1).

Since then, conventional QTL mapping has become a widely used approach to identify the genetic loci controlling quantitative traits. QTL mapping involves crossing two parental lines with contrasting phenotypes, followed by genotyping and phenotyping their segregating offspring to map genomic regions associated with trait variation. While this method has been instrumental in understanding the genetic basis of complex traits, it has several constraints. One significant limitation is the reliance on predefined genetic populations such as F_2, backcross and recombinant inbred lines, which may not represent the full spectrum of natural variation in the species, limiting the applicability of findings to diverse environments or broader germplasm (3). The next consideration is population size, as small or intermediate-sized populations often fail to capture sufficient

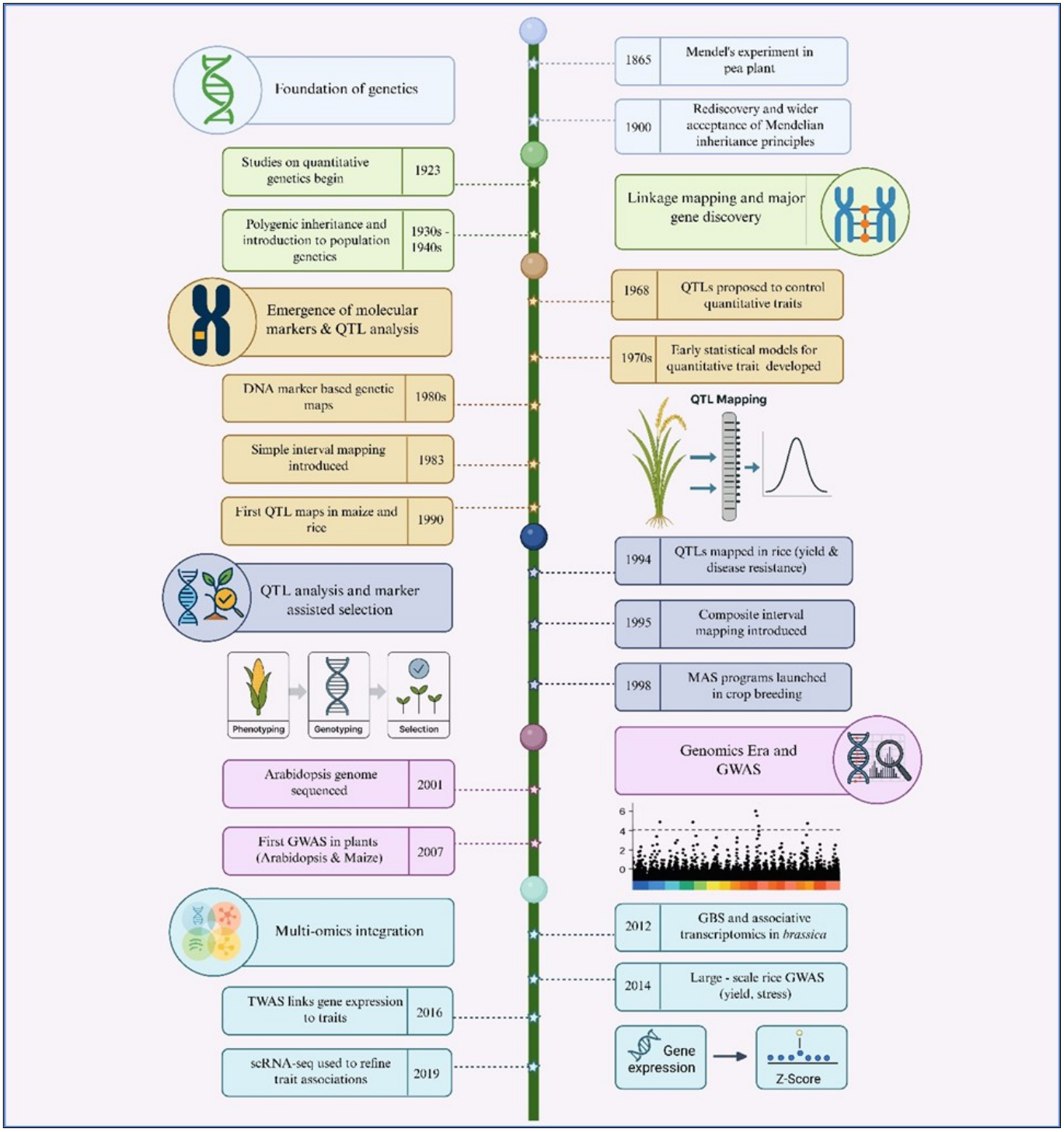

Fig. 1. Chronological timeline of advancements from Mendelian genetics to association mapping in the *Omics* era.

genetic diversity, leading to reduced mapping resolution and statistical power (4). With low-density molecular linkage maps and small population size, genetic recombination between linked markers can complicate the identification of causal QTLs, particularly in regions with low recombination rates (5). Further, many loci identified for quantitative trait variation are small-effect QTLs exhibiting genotype-by-environment (G x E) interactions (4, 6). Conventional QTL mapping often restricted to single-environment studies, fails to account for such G x E interactions, which are critical for understanding the stability and expression of traits across different environments (7). These limitations have driven the development of more advanced approaches, such as GWAS and integrative multi-omics strategies. The completion of whole-genome sequences in model species such as Arabidopsis (8) and humans (9) during the early 2000s facilitated the emergence of GWAS, allowing for the high-throughput identification of genetic loci associated with phenotypic traits. Over the past decade, advancements in next-generation sequencing (NGS) technologies have further refined gene-trait associations to enhance the resolution and power of genetic mapping (10).

Understanding gene regulation and complex traits: Moving from GWAS to associative transcriptomics

Gene regulation is essential for understanding how genetic variation contributes to complex traits such as yield, resistance to biotic and abiotic stresses. GWAS and associative transcriptomics are complementary strategies that together provide a comprehensive understanding of the genetic regulation of complex traits. GWAS identifies genetic variants, particularly SNPs, that are statistically associated with phenotypic traits by

scanning the entire genome across diverse populations including germplasm, thereby reducing the effort and time required compared to QTL mapping. While GWAS is effective, the need for high density DNA-based polymorphic markers and a complete reference genome to localize candidate genes, becomes a limitation, particularly in polyploid crops like *Brassica species*, wheat or sugarcane, where genome complexity due to polyploidy hinders accurate marker placement and trait association (11). Further, GWAS often identifies non-coding or intergenic variants with unclear functional significance and fails to detect regulatory elements that modulate gene expression (12). This restricts its power in uncovering expression QTLs (eQTLs) - regions where gene expression itself is the phenotype of interest. This is where associative transcriptomics and transcriptome wide association studies play a crucial role, as they integrate transcriptome data with genetic variation to establish causal relationships between SNPs, gene expression and phenotypic traits.

Associative transcriptomics *versus* transcriptome wide association studies

Both AT and Transcriptome-Wide Association Studies (TWAS) aim to identify genes linked to traits by integrating genetic and gene expression data, going beyond traditional GWAS. While they share this common goal, their workflows and data requirements differ considerably.

AT is a transcriptome based GWAS approach that correlates trait variation with both quantitative expression of genes and sequence variation of transcripts (13 - 15). AT uniquely captures two types of molecular markers from RNA sequencing data- SNPs, which are derived from expressed transcripts and used to map sequence variation (16) and GEMs, which represent normalized transcript abundance, that are typically measured in reads per kilobase of transcript per million mapped read (RPKM) units. This dual strategy of combining SNPs and GEMs makes it possible to pinpoint candidate genes and regulatory networks that drive complex trait variation (17, 18).

TWAS, by contrast, builds on GWAS by incorporating gene expression information in a two-stage process. First, it uses a reference panel to train models that predict gene expression from genotype. These models are then applied to a larger GWAS cohort with genotypic data only to impute expression levels and test their association with the trait (19). Unlike AT, TWAS does not require direct expression measurements from the study population, making it scalable to larger datasets. However, it relies heavily on the accuracy and tissue relevance of the expression reference panel.

Thus, both AT and TWAS provide complementary frameworks for linking genes to traits through expression, but AT requires matched genotype and transcriptome data from the same individuals, while TWAS uses external expression data to enable large-scale association testing. Together, these approaches enhance our ability to dissect the genetic architecture of complex genetic traits and improve predictive breeding strategies, especially when integrated with emerging multi-omics tools like single-cell RNA-seq and epigenomics (20, 21).

Workflow of AT

AT follows a structured workflow to establish genotype-phenotype associations by combining gene expression profiles with genotypic and phenotypic data. The workflow involves selecting an appropriate mapping panel, phenotyping the panel, collecting plant samples, extracting and sequencing RNA to generate transcriptome profiles. Once the sequencing data is obtained, preprocessing steps are done, including quality control, alignment to a reference genome and normalization to reduce technical biases (22). The core of the protocol is the association of transcriptomic variations with genetic markers or phenotypic traits, typically facilitated by statistical models (11, 13, 15). Differential gene expression is then analysed to identify genes that are significantly correlated with the traits of interest, followed by functional annotation and pathway analysis, which help elucidate the biological relevance of the identified genes (23). By integrating these various steps, AT helps to elucidate the genetic basis of complex traits, though the accuracy of the results relies heavily on the quality of data at each stage (Fig. 2).

Mapping panels

The success of AT studies largely depends on the choice of a well-structured diversity panel, which enhances statistical power and mapping resolution. A genetically diverse panel captures a broad spectrum of natural variations arising due to recombinations and mutations, facilitating the identification of eQTLs and marker-trait associations. Incorporating accessions with distinct transcriptomic profiles under various conditions aids in uncovering gene regulatory mechanisms associated with agronomically valuable traits (24). Avoiding excessive population stratification through careful selection minimizes spurious and false associations, improving the robustness of trait-linked discoveries (25). Moreover, panels that include elite cultivars, ecotypes, breeding lines from diverse geographical backgrounds and wild relatives facilitate the identification of key adaptive traits, accelerating the development of climate-resilient and high-yielding crop varieties through targeted breeding strategies (26). In *Brasscia napus*, AT panel was made up of a subset of 288 accessions derived from the Renewable Industrial Products from Rapeseed (RIPR) diversity population comprising of 56 Modern Winter Oil Seed Rape (OSR), 65 Winter OSR, 6 Winter Fodder, 121 Spring OSR, 26 Swede and 14 Exotic varieties for dissecting the glucosinolate biosynthesis (15). Similarly, different subset of individuals from the same RIPR diversity panel were used for identification of candidate loci governing erucic acid and Vitamin E content through AT (27, 28).

Acquisition of phenotypic data

The reliability of genotype-phenotype associations in AT studies depends on accurate and high-quality phenotypic data, which are essential for ensuring meaningful and reproducible results. Complex phenotypic traits being influenced by both genetic and environmental factors, warrant systematic data collection across multiple locations, seasons and controlled environments to improve repeatability and minimize confounding effects (3, 29). Adequate biological and technical replications is crucial for ensuring the reproducibility of trait measurements and reducing errors arising from environmental interactions (30). Environmental factors such as water, temperature, soil composition and biotic stresses can significantly alter gene expression patterns, making it essential to design experiments that separate genetic effects from environmental noise through multi-season and multi-environment trials (6, 31). High-throughput phenotyping (HTP) technologies, including imaging and remote sensing, enhance the precision and scalability of trait evaluation (32, 33), while advanced statistical models help correct for environmental and population structure effects (34). For example, texture analyser was used to measure and

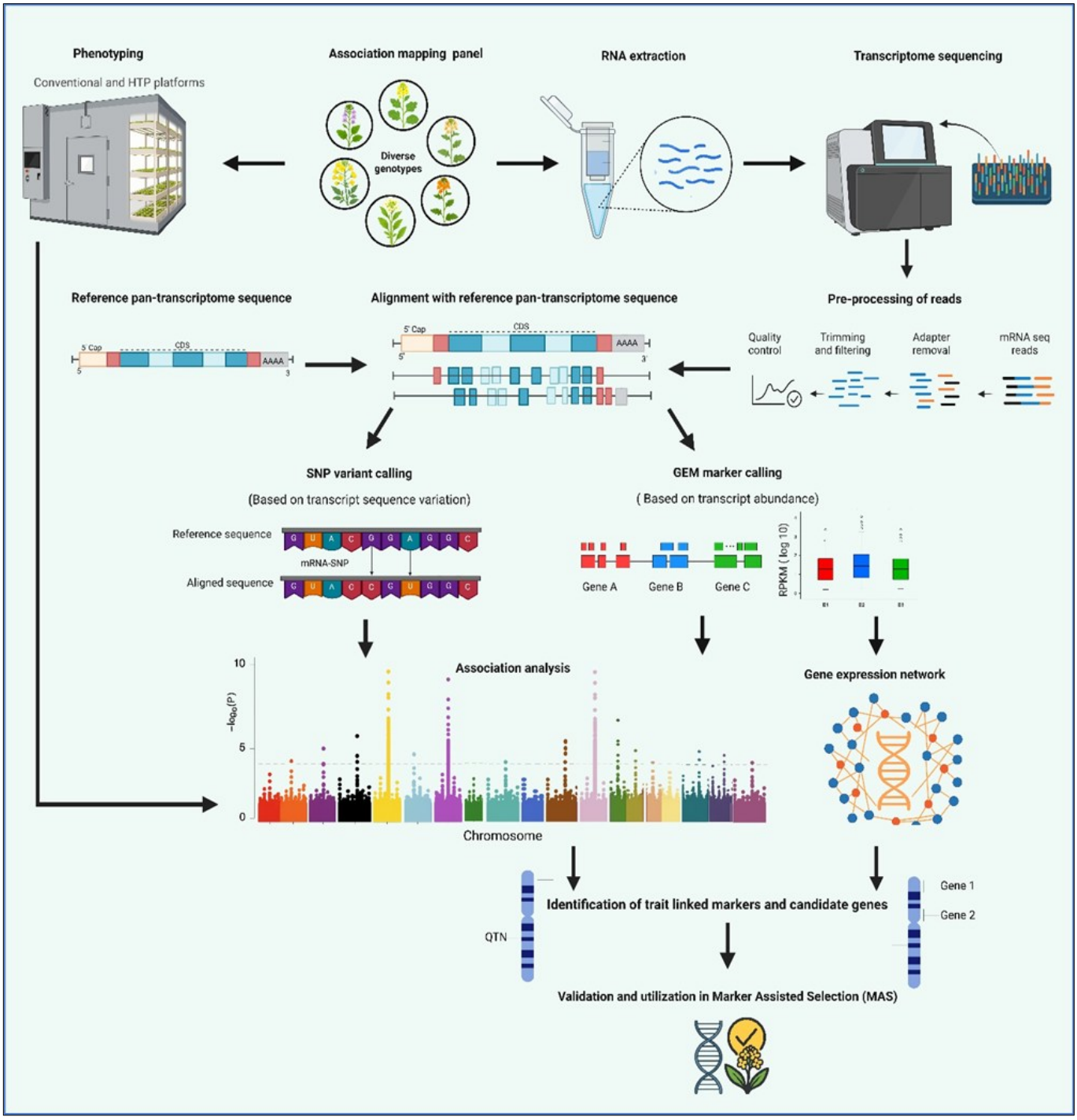

Fig. 2. Workflow of associative transcriptomics.

monitor stem strength in real time to elucidate its genetic basis based on associative transcriptomics in bread wheat (35). In addition to HTP technologies like Fourier transform infra-red screening and OLIgo Mass Profiling, demonstrated the suitability of high density carbohydrate microarrays as a novel cell wall glycomics based phenotyping strategy for association analysis of plant cell wall composition in rape seed (36). Previous studies deployed image processing and advanced machine learning methods to automate stomatal counting as a part of the combined GWAS-TWAS analysis to identify genetic loci associated with physiological traits governing water use efficiency (37). Ensuring well-replicated and seasonally tested phenotypic datasets strengthens the power of AT studies to identify regulatory networks governing complex traits, ultimately advancing crop improvement efforts.

Sample collection and RNA extraction

Sample collection and preparation including RNA extraction directly impacts the quality and reliability of the resulting data. The selection of appropriate tissues and time points is essential, as the expression of genes varies significantly across different tissues, developmental stages and environmental conditions (38). Selecting tissues relevant to the trait of interest ensures that the transcriptomic data reflects the pertinent biological process. Timing of sample collection can influence gene expression patterns. Temporal variation must be carefully considered when designing experiments, as certain genes may exhibit dynamic expression across different stages of development or under specific environmental stresses (39). For instance, first true leaf was collected close to the petiole at a time point close to the mid photoperiod from 21-day old seedlings of *Brassica napus* for

transcriptome sequencing (13). The base of third leaf and the shoot growth point at the 3-leaf stage during the day were collected for analysing transcript abundance of stomatal patterning genes in a subset of 229 Sorghum accessions (37).

RNA extraction from collected samples requires meticulous handling as it affects the downstream quality of sequencing data. Low-quality or degraded RNA can lead to unreliable results, making sample preparation a critical step in NGS-based studies (40, 41). Common RNA extraction methods include TRIzol reagent-based protocols, column-based methods and combinations of both methods, that offer higher purity and are particularly useful for high-throughput RNA-seq applications (42). The quality of RNA extracted must be rigorously assessed before sequencing to ensure accurate results. Quality control of RNA involves checking both the quantity and integrity of RNA samples. The concentration of RNA is usually determined using spectrophotometric methods (e.g., NanoDrop) or fluorometric assays (e.g., Qubit). Conventionally, RNA integrity is measured using the RNA Integrity Number (RIN), which is determined through electrophoretic methods or automated systems like the Agilent Bioanalyzer (43). Recently, a new quality score based on ratio metric fluorescence-based method has been developed for rapidly assessing RNA degradation and the results are given in terms of RNA Integrity and Quality number (RNA IQ) (40). Both RIN and RNA IQ scales range from 1 to 10, with 1 indicating highly degraded RNA and 10 indicating high quality intact RNA (40). RNA with a RIN or RNA IQ of 7 or above is considered suitable for sequencing and provides reliable data for downstream analyses.

Sequencing platforms for transcriptomics

RNA-sequencing is central to the AT approach, offering a cost-effective method to generate both SNPs and GEMs from a single dataset. Various methods are available for whole transcriptome sequencing, including shotgun sequencing, tag-based sequencing, bulk RNA sequencing, single-cell RNA sequencing and spatial RNA sequencing. These methods differ in resolution, throughput, application scope and their successful implementation depends on the selection of an appropriate sequencing platform. Among the most widely used sequencing platforms for RNA sequencing, Illumina, PacBio and Oxford Nanopore Technologies stand out due to their capabilities in sequencing depth, accuracy, read length and overall performance.

Illumina is the most widely used sequencing platform for transcriptomic studies, valued for its high throughput, cost efficiency and accuracy. It generates high quality short reads (typically 100-150 base pairs) with low error rates, making it suitable for reliable detection of gene expression levels, alternative splicing events and novel transcripts. Its combination of short read lengths and deep coverage ensures precise quantification and robust analysis of differential gene expression across different experimental conditions (44). Most of the transcriptome sequencing done in *Brassica species* for associative transcriptomic analysis have extensively used the Illumina sequencing platforms (11, 13, 45, 46). However, the short-read lengths can be limiting when trying to resolve complex genomic features, such as long genes or intricate splicing patterns. Moreover, Illumina sequencing is prone to GC bias, where regions with high GC content may be underrepresented. Despite these limitations, Illumina remains the gold standard for transcriptomic studies, particularly in well-annotated genomes where high-throughput, cost-effective sequencing is required.

Pacific Biosciences (PacBio) sequencing offers long reads (up to tens of kilobases), which is a significant advantage for transcriptomic studies where full-length transcript sequencing and accurate isoform detection are crucial (47). PacBio's Single Molecule, Real-Time (SMRT) technology provides high-quality, long reads that can span entire exons or even full-length transcripts (48). This capability makes PacBio ideal for studying complex transcriptomic features, such as alternative splicing, gene fusion events and alternative polyadenylation events (49). One of the major advantages of PacBio is its ability to capture complete transcripts without the need for assembly from short reads, reducing the risk of assembly errors.

Oxford Nanopore Technologies offers a unique approach to long-read sequencing using a portable and flexible platform (50). The key advantage of Oxford Nanopore is its ability to sequence directly from RNA, eliminating the need for reverse transcription and cDNA synthesis, thus providing a more accurate snapshot of the native transcriptome. This feature makes Oxford Nanopore ideal for real-time targeted sequencing in dynamic conditions, such as stress responses or developmental processes (51). Nanopore sequencing also has the advantage of being scalable, with both benchtop and handheld devices available for different experimental needs. However, both PacBio and Oxford Nanopore sequencing come with their own set of challenges. These long-read sequencing platforms are more expensive than Illumina sequencing and have higher error rates, particularly with homo-polymeric regions, although recent advancements in base calling algorithms have improved accuracy. Furthermore, the throughput and overall sequencing quality can vary depending on the specific device and run conditions, which may limit the consistency of results.

The choice of sequencing platforms largely depends on the specific goals of the transcriptomic study. Illumina remains the most widely used platform for gene expression quantification, differential expression analysis and detecting known splice variants, due to its high throughput, accuracy and affordability. PacBio and Oxford Nanopore, with their ability to generate long reads, are invaluable for capturing full-length transcripts, isoform diversity and resolving complex splicing patterns, but they come with higher costs and increased error rates. For studies focused on gene expression at the isoform level or investigating novel transcripts in less characterized genomes, long-read platforms like PacBio and Oxford Nanopore may be more appropriate, whereas short-read sequencing with Illumina is ideal for studies that prioritize depth and precision in well-annotated species.

Preprocessing of transcriptome data for SNP identification in AT

In AT, preprocessing mRNA-Seq data from a diversity panel to identify SNPs within unigenes involves a series of steps aimed at ensuring data quality and accuracy in downstream analyses. The first step in this process is performing quality control (QC) on the raw sequencing data, typically using tools such as Trimmomatic (52) or Cutadapt (53) to remove adapter sequences and filter out low-quality reads. A subsequent QC check is conducted with Fast QC (54) to assess the overall quality of the data, ensuring that it meets the necessary standards before further processing.

Once the raw reads are cleaned, they are aligned to a reference transcriptome using alignment tools such as STAR (55), HISAT2 (56), or Bowtie2 (57). The alignment step is followed by post-alignment QC to verify the accuracy and completeness of the alignment, which is commonly done using tools like Samtools (58) to assess alignment statistics and coverage. In an AT study in *Brassica species*, raw reads were aligned against a curated reference unigene set derived from ancestral genomes (*Brassica rapa* for the A genome and *B. oleracea* for the C genome) to distinguish between duplicated genes in polyploid species (59).

SNP and GEM marker calling

The aligned reads are used to generate two key marker data sets viz., SNP matrix and GEM matrix. SNP calling is done using specialized softwares such as GATK (60), FreeBayes (61), Samtools (58) or by using custom scripts. During this step, various filters are applied to retain only high-confidence variants, ensuring the reliability of the detected SNPs. In a study involving *B. napus* (45) filtering for minor allele frequencies and excessive missing data, retained over 144000 high-confidence SNPs for association analysis. Further, in these polyploid species, genome specific SNPs can be distinguished using inter homeolog polymorphisms (IHPs), a method based on detecting allele-specific expression linked to parental genomes. This step is essential for accurate marker positioning in polyploid genomes, where duplicated loci can otherwise confound analysis (62). In *Brassica napus*, where high sequence similarity exists between the A and C sub-genomes RNA-Seq data mapped to a curated unigene set, modified to include both homeologous variants from the two genomes, enabled the accurate distinction of homeologous gene expression and assignment of SNPs to specific sub-genomes (13).

To generate GEMs, RNA-seq reads from multiple genotypes are aligned to a unigene reference using a short-read aligner like MAQ (58). In polyploid species, separate genome versions are included to account for homeologous sequences. Reads mapping equally to multiple locations are randomly picked and evenly assigned to genomic specific unigenes to avoid bias. From the alignment files, gene-level read counts are extracted using custom scripts, followed by quantification and normalized as RPKM (Reads per Kilobase of unigene per Million aligned reads) values. These RPKM values represent GEMs, which are quantitative markers reflecting transcript abundance across genotypes for use in expression-trait association studies.

The aligned SNPs and GEMs are then positioned on pseudochromosomes, which are constructed by ordering unigenes based on homology to existing gene models and physical maps available in that particular crop species. These variants are annotated for their potential functional significance using resources like Ensembl (63) or SnpEff (64), to provide insights into the potential impact of each SNP on gene function and regulation. The bioinformatic tools commonly used for processing transcriptomic data are provided in Table 1.

Population structure and linkage disequilibrium analysis

Population structure and linkage disequilibrium (LD) analysis are essential steps in associative transcriptomics to ensure that the observed associations between genetic variants (SNPs) and gene expression are not confounded by population stratification or unrelated genetic factors. These confounding effects can inflate test statistics, producing spurious associations between molecular markers (SNPs or GEMs) and phenotypic traits resulting in false positives (34, 65). Such confounding effects are particularly prevalent in crops such as *Brassica species*, where population history includes interspecific hybridization and breeding bottlenecks (45).

Population structure refers to the presence of genetic subgroups within a population, which can lead to false associations if not accounted for. In one of the first AT applications, observed that failure to adjust for population structure could result in inflated significance scores, which may overstate the importance of detected loci (13). This issue becomes amplified, particularly in polyploid species, where allelic dosage and gene duplication further complicate marker-trait associations. To assess population structure, tools like PCA (Principal Component Analysis) or STRUCTURE Harvester (66) can be used to detect clusters within the population. PCA, implemented in software like PLINK (67) or R, identifies the major axes of genetic variation, which are visualized in scatter plots to observe clustering of individuals. STRUCTURE (68) and similar software use Bayesian clustering methods to infer population subgroups based on genetic data. If population structure is detected, it is essential to include principal components or population subgroups as covariates in subsequent association analyses to correct for any potential confounding. PSIKO (Population Structure Inference using Kernel Optimization) is a more computationally efficient tool for modelling population structure (69). It uses a probabilistic framework to infer subpopulations from SNP matrices without the computational burden of traditional STRUCTURE runs. PSIKO is particularly advantageous in large datasets where STRUCTURE becomes impractical due to its intensive MCMC sampling. In practical workflows, PSIKO-derived Q matrices can replace STRUCTURE outputs in mixed models, offering similar resolution but with reduced runtime (15). When paired with a kinship matrix, the PSIKO-MLM pipeline effectively balances computational feasibility with statistical rigor in AT analyses.

LD refers to the non-random association of alleles at different loci in a population. High LD between SNPs can lead to spurious associations if not properly accounted for. To assess LD,

Table 1. Computational workflow and bioinformatic resources for implementing associative transcriptomics

S.No	Computational steps in AT	Bioinformatic tools
1	Preprocessing RNA sequence reads	FastQC, Fastp, Cutadadapt, Trimmomatic
2	Read alignment	STAR, HISAT2, Bowtie, TopHat2, MAQ
3	SNP variant calling	GATK, FreeBayes, Samtools
4	GEM calling	edgeR, StringTie
5	Population structure analysis	Structure harvester, Tassel, PLINK, Admixture, FastStructure
6	LD analysis	PLINK, VCFtools, Haploview, TASSEL, PopLDdecay, LD heatmap and SNPRelate in R package
7	Association analysis	TASSEL, Integrated R packages for GWAS (GAPIT, FARMCPU, FASTmrMLM, FASTmrEMMA, ISISEMBLASSO, BLINK)
8	Annotation tools	Ensembl, SnpEff

PLINK (67) or VCFtools (70) can be used to calculate pairwise r^2 values between SNPs across the genome. LD decay is commonly plotted to visualize the distance over which SNPs maintain strong associations. In associative transcriptomics, LD analysis helps to ensure that the SNPs being analyzed for association with gene expression are independent of each other, reducing the risk of confounded results. Tools like Haploview (71) can also be used to visualize LD blocks and identify SNPs that are in high LD with each other, which can guide the selection of independent variants for further analysis. Low levels of LD as observed in winter rapeseed accessions (13, 72) favour precision of association studies when complemented with high density genotyping (35). Proper correction for population structure and LD helps ensure that the associations identified are biologically meaningful and not a result of hidden genetic biases.

Genome wide association analysis using transcriptome data

Once population structure and LD have been accounted for, the SNP and GEM data sets obtained from high-throughput transcriptome sequencing can be subjected to association analysis. In the association models, these genetic variants serve as independent variables, while phenotypic trait values represent the dependent variable, facilitating the identification of regulatory loci influencing transcriptional variation. However, each category demands distinct statistical models tailored to their data structure and biological interpretation.

For SNP based association mapping, statistical models employed to assess marker-trait relationships often involves both single-locus and multi-locus models. Among single-locus models, generalized linear models (GLM) and mixed linear models (MLM) are widely used. In the GLM (Q) model, population structure is accounted for by incorporating the Q-matrix as covariates, whereas in GLM (PCA), the first three principal components (PCs) are used to correct for population stratification and minimize false positives (25). The MLM approach further improves control over population structure and relatedness by integrating both the Q-matrix and a kinship matrix (K), which is derived through identity-by-state analysis (73). Depending on data availability for Q matrix and K matrix, both GLM and MLM models can be implemented in TASSEL, enabling flexible genome-wide association studies in plants (74). In *Brassica napus*, STRUCTURE analysis identified two distinct subpopulations corresponding to different ecotypes, which were used to define the fixed effects in the MLM (13).

In associative transcriptomics, multi-locus models have gained prominence over traditional single-locus methods due to their improved power, accuracy and ability to handle complex traits without the need for stringent multiple testing corrections. The evolution of these models reflects significant methodological and computational advancements. GAPIT (Genome Association and Prediction Integrated Tool) supports large-scale datasets using compressed mixed models for efficient prediction and association analysis (75). FarmCPU (Fixed and Random Model Circulating Probability Unification) iteratively integrates fixed and random effects to control false positives effectively (76). Other key models include mrMLM (multi-locus random-SNP-effect MLM) which bypasses Bonferroni correction and its computationally efficient variant FASTmrMLM (fast multi-locus random-SNP-effect MLM) (77). FASTmrEMMA (fast multi-locus random-SNP-effect EMMA) further reduces computational time while maintaining accuracy (78). ISIS EMBLASSO (Iterative Sure Independence Screening EM-Bayesian LASSO) combines Expectation-Maximization with Bayesian LASSO for robust QTN detection (79). BLINK (Bayesian-information and linkage-disequilibrium iteratively nested keyway) enhances statistical power and speeds up computation compared to Farm CPU (80). Collectively, these models represent a significant advancement in dissecting complex trait genetics, especially in large, diverse populations typical of associative transcriptomics studies. Utilizing all these single- and multi-locus GWAS models, candidate genes associated with flowering, maturity and seed weight in rice bean across two datasets were identified (81). However, no marker was consistently detected across all datasets for the traits studied, underscoring the complexity of the traits and the influence of model choice on detection outcomes.

In contrast to SNPs, GEMs reflecting transcript abundance represent continuous variables and hence their association with traits is evaluated using simple linear regression models. Gene expression marker associations are computed using in-house script available in statistical packages such as R using a fixed effect linear model with each unigene's normalized expression (RPKM) values. Linear regression analysis is performed using RPKM as a predictor value to predict a quantitative outcome of the trait value. In addition to SNP variations in expressed gene sequences, regression analysis of transcript abundance against phenotypic trait values enables the identification of unigenes that exhibit a significant association between their expression levels and the traits of interest. In *B. napus*, this model identified transcription factor unigenes like *BnaC.HAG1a* and *BnaA.HAG1c* whose reduced transcript abundance was significantly correlated with low glucosinolate content, implicating gene deletions as potential causal events (45). However, without correction, linear regression on GEMs can be prone to genomic inflation. Hence, genomic control methods such as Quantile-Quantile (QQ) plots, False Discovery Rate (FDR) or Bonferroni correction are used to mitigate false positives (82). On the other hand, GEM analysis often excludes kinship corrections, as expression data typically have lower linkage than genomic SNPs and may capture more trait-specific regulatory effects. Overly stringent correction may obscure true associations, while lenient thresholds increase the burden of downstream validation. For instance, associations with *FAE1* (fatty acid elongase) and *HAG1* orthologs in *B. napus* remained significant even under stringent thresholds due to their large effect sizes (13, 45). However, relaxing the significance cutoff in conjunction with biological validation strategies, resulted in identifying 17 peak regions that included known and novel candidates. The choice between methods entails a trade-off between stringency and sensitivity. As such, statistical signals are often complemented by co-localization, gene annotations and co -expression data to prioritize biologically meaningful candidates.

The SNP and GEM-based GWAS pinpoint QTLs governing trait variation and are visualised using Manhattan plots. The identification of eQTLs through GEM-based GWAS can reveal both cis-eQTLs and trans-eQTLs. Cis-eQTLs are genetic variants located near the gene they regulate, often within the gene's promoter or intronic regions, whereas trans-eQTLs are genetic variants located on different chromosomes or regions distant from the gene but still influencing its expression (83). When SNP and GEM peaks co-localize, they provide compelling evidence for candidate gene involvement, especially in complex traits influenced by both

structural and regulatory variation.

Weighted Gene Co-Expression Network Analysis (WGCNA)

The relationship between gene expression and phenotypic traits can be further explored using WGCNA (84). It starts with transcript quantification data, which represents the gene expression levels across a panel of lines or samples. WGCNA identifies clusters, or modules of genes that are co-expressed, meaning they have similar expression patterns across all the samples. These modules are then linked to traits of interest such as yield, nutritional quality and stress tolerance, by calculating how the overall expression of each module correlates with the trait. This helps in identifying genes or gene networks that may play a role in regulating these traits. By analyzing the gene modules, researchers can uncover the biological processes or pathways underlying complex traits, which is particularly useful for improving breeding strategies or understanding the genetic basis of diseases. AT coupled with WCGNA analysis in European ash tree (*Fraxinus excelsior*) revealed a single module comprising of 56 genes inclusive of two GEM markers, that were highly correlated with canopy damage caused due to crown dieback symptom by the fungal pathogen *Hymenoscyphus fraxineus*. A MADS box gene SVP harboured in this module was involved in inducing age related resistance besides flowering pathway (85).

Applications of AT in crop plants

AT was initially adapted from a microarray-based study that investigated the use of expression markers to predict hybrid vigour in plants. This method was validated as a proof of concept in rapeseed (13). Illumina based transcriptome sequencing in a panel of 84 *B. napus* accessions, detected 101644 SNPs within 11743 unigenes. In addition, transcriptome quantification, allowed for the identification of GEMs. Together, these two marker types were successfully utilised to identify genes controlling seed erucic acid and glucosinolate content in a subset of 53 lines. SNP based association mapping identified markers associated with two QTLs (*eru 1* and *eru2*) gover ning seed erucic acid content, that are orthologs of *Arabidopis thaliana FAE1* gene. GEM based GWAS identified genetic loci governing seed glucosinolate content on linkage groups A9 and C2, that were orthologs of *Arabidopsis thaliana* HAG1 (*MYB28* transcription factor) gene, controlling the biosynthesis of aliphatic glucosinolates. Further, WCGNA analysis also correlated to a core module containing several unigenes involved in glucosinolate biosynthesis (13). By using an extended panel of 101 accessions, high density SNP and GEM based association analysis pinpointed to two additional candidate genes, *BnaA.GTR2a* located on chromosome A2 and *BnaC.HAG3b* on C9, that explained for 25.8 % of the total phenotypic variation in seed glucosinolate synthesis. Based on gene co-expresssion analysis, the unigene *BnaC.BAT5* was found to be primarily involved in aliphatic GS biosynthesis by serving as targets for these aliphatic GS regulators HAG1 and HAG3. The deletion polymorphism in *BnaC.HAG3b* has been converted into diagnostic PCR based markers for marker assisted selection of low seed glucosinolate lines. Similarly, using 201 SNP and 147 GEMs, 22 candidate genes were identified and a tropinone reductase encoding gene (*BnTR1*), was confirmed to be closely linked to transpiration rate and enhanced low temperature tolerance under freezing conditions in *Brassica napus* (80). AT in a diverse panel of 69 accessions identified microRNA *miR172D* and a flowering time gene *BoFLC.C2* as candidate regulators of the vernalization response in *Brassica oleracea* (46).

While AT has been extensively utilised in the polyploid *Brassica species*, the strategy is now being extended to other crops as well (Table 2). Utilising this approach in a diverse panel comprising of 100 accessions, several candidate genes governing key agronomic traits such as flowering, maturity and seed weight in rice bean (*Vigna umbellata*) have been identified (81). The candidate genes HSC80, P-II PsbX, phospholipid-transporting ATPase-9, pectin-acetylesterase-8 and E3 ubiquitin-protein ligase RHG1A were linked to flowering, WRKY1 and DEAD-box RH27 were associated with seed weight, PIF3 and pentatricopeptide repeat-containing genes were associated with both maturity and seed weight (81). The candidates for plant height, stem strength and lodging resistance in bread wheat (28, 35) and dieback tolerance in *Fraxia* species, have also been identified through associative transcriptomics approach (85).

Even in species lacking genomic resources, a systematic approach has been suggested to facilitate genetic analysis (13). This involves de novo transcriptome assembly to generate a unigene reference from mRNA-Seq data, construction of high-density linkage maps from transcriptome-derived SNPs obtained across mapping populations, utilising these linkage maps to infer the putative genomic order of unigenes by aligning them with the genome sequence of a closely related species. Finally, GWAS can be conducted using both SNP markers and GEMs derived from mRNA-Seq data across diverse genetic collections, enabling the identification of loci associated with phenotypic variation.

Prospects of AT

Unlike single-gene traits, polygenic traits are influenced by multiple loci, making them challenging to study using conventional methods. A holistic understanding of complex traits demands integration across molecular layers. AT enables the identification of key genes and regulatory networks governing these traits by correlating expression data with phenotypic variation across genetically diverse populations. Combining AT with metabolomics, proteomics and phenomics offers a systems biology perspective. The integration of AT with pangenomics can provide insights into structural and transcriptional variation across diverse accessions. The construction of a pan-transcriptome enabled high-resolution detection of SNPs and GEMs, improving trait mapping efficiency in *Brassica juncea* (86). Such integration allows researchers to account for presence-absence variation and gene copy number variation across populations (87). In parallel, the inclusion of epigenomic layers-such as DNA methylation and histone modifications, could help in interpreting unexplained expression variation and gene dosage compensation (88). DNA methylation variations in the promoters of MADS-box genes such as *SOC1*, *SVP* -like and *JOINTLESS* have been linked to the differential responses of tolerant and susceptible trees to dieback disease caused by the fungus *Hymenoscyphus fraxineus* (89). In *B. napus*, genomic deletions correlated with loss of gene expression and reduced glucosinolate content, suggesting the need to integrate epigenomic data for deeper understanding (13).

A notable strength of AT is its independence from fully annotated genomes, making it highly suitable for underutilized and non-model crops. In *Vigna umbellata* (rice bean), AT enabled the identification of SNPs and GEMs associated with agronomically desirable traits such as seed weight and flowering

Table 2. An overview of associative transcriptomic studies in crops summarizing mapping panels, marker resources and candidate genes linked to phenotypic traits

S.No	Crop	Mapping panel	No. of SNP markers	GEM markers/ unigenes	Candidate gene/QTL/marker	Trait	Reference
1	*Brassica napus*	53 accessions	62980	57343-A genome 59559-C genome	*eru1* and *eru2* (orthologs of *FAE 1* gene) *HAG1*	Erucic acid content in seed oil Seed Glucosinolate (GSL) content	(13)
2	*B. napus*	101 accessions	144131	49,599- A genome 50,935- C genome	*BnaA.GTR2a* *BnaC.HAG3b*	Seed Glucosinolate content	(45)
3	*B. napus*	84 accessions	62980	189116	Nitrate homeostasis - *CLC A and 3(2),5-bisphosphate nucleotidase, SAL2* Phosphate homeostasis -*H^+-ATPASE1* Sulfate homeostasis - *Cys synthase C*	Nutrient homeostais	(93)
4	*B. napus*	383 accessions	256397	116,098	Ca and Mg transporter genes - *ACA8 and MGT7* Flowering time regulator genes viz., *Flowering Locus C (FLC) and Suppressor Of Overexpression Of CO1 (SOC1)*	Calcium and Magnesium accumulation	(94)
5	*B. napus*	331 accessions	256397	116098	*Irregular Xylem 14 (IRX14)* *Glucuronic Acid Substitution Of Xylan 1 (GUX1)*	Xylan synthesis and branching	(36)
6	*B. napus*	79 accessions	144131	189,116	*Pectin Methyl Esterase Inhibitor (PMEI)* *Microtubule Organisation 1 (MOR1) BRI1 (BRASSINOSTEROID-INSENSITIVE 1) SUPPRESSOR 1*	Stem strength Plant height	(28)
7	*B. napus*	383 accessions (RIPR panel)	355536	53889 25834-A genome 28055- C genome	*Bo2g050970.1, an orthologue of a γ-tocopherol methyl transferase gene VTE4* *Cab035983.1 and Bo3g168810.1, (orthologues of FAE1)* *Cab033920.1 (ortholog of fatty acid hydroxylase)*	vitamin E content Seed erucic acid content	(27)
8	*B. napus*	288 accessions from RIPR panel	256397	53,889	*Bna.HAG1.A9 and Bna.HAG1.C2* *Bna.HAG3.A3*	Aliphatic GSL content in seeds and leaves Aromatic GSL content in root	(15)
9	*B. napus*	245 accessions from RIPR panel	256397	53,889	*BnaA03g45000D, TIR-NBS-LRR class genes, pentacyclic triterpene synthase 1* *cytokinin response factor 2*	Clubroot resistance	(95)
10	*B. napus*	123 accessions from RIPR panel	256397	53,889	*Tropinone reductase gene (BnTR1)*	Freezing stress tolerance	(96)
11	*B. napus*	126 accessions			*ADP glucose pyrophosphorylase large subunit 1 (BnaC2APL1), PS II Oxygen-Evolving Complex 1 (BnaC9PSBO1)*	Starch biosynthesis and photosynthesis	(97)
12	*B. napus*	Subsets from 193 accessions, included in RIPR panel	219454	53,883	*KIN7.2, EDS5H, BBX18, XK1, FBA1*	Quantitative disease resistance to fungal pathogens	(98)
13	*B. napus*	189 accessions of RIPR panel	355536	53,889	*BR-Signalling Kinase1 (BSK1)*	Responsiveness to Necrosis and Ethylene-inducing peptide 1-like proteins (NLPs)	(99)
14	*B. oleraceae*	69 accessions from Diversity Fixed Foundation Set	36631	65017	*miR172D, BoFLC.C2*	Vernalization response	(46)
15	*B. juncea*	204 inbred accessions (CGAT panel)	171196	48975 25698 -A genome 23 277- B genome	*BjA.TTL* *Cab016066.1:252:G* *Cab038799.1*	Seed weight Seed colour Vitamin E content	(11)
16	*Fraxinus excelsior*	182 trees (Danish ash panel)	174346	32,441	*Gene_22343_Predicted_mRNA_scaffold3139; Gene_19216_Predicted_mRNA_scaffold2427 and Gene_23247_Predicted_mRNA_scaffold3380*	Dieback disease tolerance	(85)
17	*Pinus massoniana*	204 wild accessions	94194	-	*MADS-box genes- SOC-1, SVP-like proteins and Dormancy MADS-box JOINTLESS* *CYP720B, cytochrome P450, AP2/ERFB*	Oleoresin yield	(100)
18	*Triticum aestivum*	100 accessions	12456	94,060	*Acetyl xylan esterase Orthologue of COP9 Signalosome Subunit5B (CSN5B)* *Small Auxin Up RNA (SAUR) genes*	Stem strength Plant height	(28)
19	*Vigna umbellata*	100 accessions	49271	87	HSC80, P-II PsbX, phospholipid-transporting-ATPase-9, pectin-acetylesterase-8 and E3-ubiquitin-protein-ligase-RHG1A WRKY1 and DEAD-box-RH27 PIF3 and pentatricopeptide-repeat-containing-gene Aldo-keto-reductase	Flowering Seed weight Maturity, seed weight Flowering, maturity	(81)

time, despite the lack of a complete reference genome (81). The future of AT will be shaped by its integration with machine learning (ML) approaches. ML algorithms can exploit high-dimensional SNP and GEM data for trait prediction. In European ash tree, expression-based models were able to predict dieback tolerant and susceptible phenotypes, suggesting that ML can enhance predictive power in breeding pipelines demonstrating its applicability even in long-lived perennials. These examples highlight how AT can accelerate functional genomics in neglected species important for food security and biodiversity (85). Incorporating environmental data and developmental time points into AT workflows will further enable the construction of dynamic predictive models for complex traits.

Challenges in AT

AT relies on natural genetic variation, necessitating the use of large, well-characterized populations for generating robust and reproducible associations. However, the availability of such populations is often limited, particularly in the case of underutilized and orphan crops. One of the major limitations in AT is the dynamic nature of gene expression, which is significantly influenced by developmental stages, environmental fluctuations and epigenetic modifications. Standardizing transcriptomic studies to account for these variables remains a critical challenge. Accurate genome assemblies serve as the foundation for effective AT applications. Unfortunately, many economically important crop species still lack high quality reference genomes, which compromises the precision of transcriptomic association studies. In polyploid crops, the high sequence similarity between homeologs further complicates read mapping and quantification (58). Moreover, trans-acting eQTLs often exert weak regulatory signals, making their detection more challenging. Another bottleneck in AT lies in the integration of high-dimensional pan genomic and transcriptomic datasets, that requires sophisticated computational tools and extensive bioinformatics resources (90). To minimize spurious associations and enhance predictive power, ML algorithms and advanced statistical models are essential. Factors such as LD decay, marker density and distribution, as well as the stringency and sensitivity of association models, critically influence the reliability of marker trait associations. Additionally, artificial intelligence (AI) and deep learning models can facilitate the analysis of complex datasets by identifying hidden patterns and improving phenotype predictions (91). The functional validation of candidate genes identified through AT necessitates precise techniques like CRISPR-Cas9 and RNA interference (RNAi), which, despite their potential, are time-intensive and subject to stringent regulatory frameworks (92).

Conclusion

AT offers a powerful framework for dissecting complex gene trait relationships and accelerating genetic improvement in crops, especially polyploid species. In crop species with limited genomic resources, AT enables the identification of functional markers without requiring fully assembled genomes. The information derived from AT can substantially benefit breeding programs through development of trait-linked markers useful for marker assisted selection. Applying AT to orphan crops and wild relatives of cultivated species can reveal untapped genetic resources, crucial for developing climate-resilient and nutritionally enhanced cultivars. Despite existing challenges, continuous advancements in sequencing platforms, bioinformatics tools and functional genomics techniques are set to expand the utility of AT. Future research must prioritize the integration of AT with multi-omics strategies and AI-driven data analytics to fully realize its potential in precision breeding and sustainable crop development.

Authors' contributions

Compilation of information was done by JE. Conceptualization, writing of original draft, review and editing was performed by GS. Image designing and workflow was done by TS and NHS. RR, BP and HS reviewed of the manuscript. All authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest: The authors do not have any conflict of interest to declare.

Ethical issues: None

References

1. Sax K. The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics. 1923;8(6):552. https://doi.org/10.1093/genetics/8.6.552
2. Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. 1989;121 (1):185-99. https://doi.org/10.1093/genetics/121.1.185
3. Mackay I, Piepho HP, Garcia AAF. Statistical methods for plant breeding. Handbook of Statistical Genomics: Two Volume Set. 2019:501-20. https://doi.org/10.1002/9781119487845.ch17
4. Mackay TF, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. Nature Reviews Genetics. 2009;10(8):565-77. https://doi.org/10.1038/nrg2612
5. Tanksley SD. Mapping polygenes. Annual Review of Genetics. 1993;27 (1):205-33. https://doi.org/10.1146/annurev.ge.27.120193.001225
6. Des Marais DL, Hernandez KM, Juenger TE. Genotype-by-environment interaction and plasticity: exploring genomic responses of plants to the abiotic environment. Annual Review of Ecology, Evolution and Systematics. 2013;44(1):5-29. https://doi.org/10.1146/annurev-ecolsys-110512-135806
7. Bernardo R. Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Science. 2008;48 (5):1649-64. https://doi.org/10.2135/cropsci2008.03.0131
8. Bevan M, Mayer K, White O, Eisen JA, Preuss D, Bureau T, et al. Sequence and analysis of the *Arabidopsis* genome. Current Opinion in Plant Biology. 2001;4(2):105-10. https://doi.org/10.1016/S1369-5266(00)00144-8
9. Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. Science. 2003;300(5617):286-90. https://doi.org/10.1126/science.1084564
10. Mahmood U, Li X, Fan Y, Chang W, Niu Y, Li J, et al. Multi-omics revolution to promote plant breeding efficiency. Frontiers in Plant Science. 2022;13:1062952. https://doi.org/10.3389/fpls.2022.1062952
11. Harper AL, He Z, Langer S, Havlickova L, Wang L, Fellgett A, et al. Validation of an associative transcriptomics platform in the polyploid crop species *Brassica juncea* by dissection of the genetic architecture of agronomic and quality traits. The Plant Journal. 2020;103(5):1885-93. https://doi.org/10.1111/tpj.14876
12. Zhong W, Liu W, Chen J, Sun Q, Hu M, Li Y. Understanding the

function of regulatory DNA interactions in the interpretation of non-coding GWAS variants. Frontiers in Cell and Developmental Biology. 2022;10:957292. https://doi.org/10.3389/fcell.2022.957292

13. Harper AL, Trick M, Higgins J, Fraser F, Clissold L, Wells R, et al. Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. Nature Biotechnology. 2012;30(8):798-802. https://doi.org/10.1038/nbt.2302

14. Stower H. Associative transcriptomics. Nature Reviews Genetics. 2012;13(9):597. https://doi.org/10.1038/nrg3318

15. Kittipol V, He Z, Wang L, Doheny-Adams T, Langer S, Bancroft I. Genetic architecture of glucosinolate variation in *Brassica napus*. Journal of Plant Physiology. 2019;240:152988. https://doi.org/10.1016/j.jplph.2019.06.001

16. Duitama J, Srivastava PK, Măndoiu II. Towards accurate detection and genotyping of expressed variants from whole transcriptome sequencing data. BMC Genomics. 2012;13:1-10. https://doi.org/10.1186/1471-2164-13-S2-S6

17. Degtyareva AO, Antontseva EV, Merkulova TI. Regulatory SNPs: altered transcription factor binding sites implicated in complex traits and diseases. International Journal of Molecular Sciences. 2021;22(12):6454. https://doi.org/10.3390/ijms22126454

18. Mogil LS, Andaleon A, Badalamenti A, Dickinson SP, Guo X, Rotter JI, et al. Genetic architecture of gene expression traits across diverse populations. PLoS Genetics. 2018;14(8):e1007586. https://doi.org/10.1371/journal.pgen.1007586

19. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nature Genetics. 2016;48(3):245-52. https://doi.org/10.1038/ng.3506

20. Liu S, Trapnell C. Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Research. 2016;5:F1000 Faculty Rev-182. https://doi.org/10.12688/f1000research.7223.1

21. Rhaman MS, Ali M, Ye W, Li B. Opportunities and challenges in advancing plant research with single-cell omics. Genomics, Proteomics & Bioinformatics. 2024;22(2):qzae026. https://doi.org/10.1093/gpbjnl/qzae026

22. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Computational Biology. 2017;13(5):e1005457. https://doi.org/10.1371/journal.pcbi.1005457

23. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols. 2012;7(3):562-78. https://doi.org/10.1038/nprot.2012.016

24. Xu C, Jiao C, Sun H, Cai X, Wang X, Ge C, et al. Draft genome of spinach and transcriptome diversity of 120 *Spinacia* accessions. Nature Communications. 2017;8(1):15275. https://doi.org/10.1038/ncomms15275

25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006;38(8):904-9. https://doi.org/10.1038/ng1847

26. Varshney RK, Thudi M, Roorkiwal M, He W, Upadhyaya HD, Yang W, et al. Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. Nature genetics. 2019;51(5):857-64. https://doi.org/10.1038/s41588-019-0401-3

27. Havlickova L, He Z, Wang L, Langer S, Harper AL, Kaur H, et al. Validation of an updated Associative Transcriptomics platform for the polyploid crop species *Brassica napus* by dissection of the genetic architecture of erucic acid and tocopherol isoform variation in seeds. The Plant Journal. 2018;93(1):181-92. https://doi.org/10.1111/tpj.13767

28. Miller CN, Harper AL, Trick M, Wellner N, Werner P, Waldron KW, Bancroft I. Dissecting the complex regulation of lodging resistance in *Brassica napus*. Molecular Breeding. 2018;38:1-18. https://doi.org/10.1007/s11032-018-0781-6

29. Xu Y, Li P, Yang Z, Xu C. Genetic mapping of quantitative trait loci in crops. The Crop Journal. 2017;5(2):175-84. https://doi.org/10.1016/j.cj.2016.06.003

30. Schielzeth H, Rios Villamil A, Burri R. Success and failure in replication of genotype-phenotype associations: How does replication help in understanding the genetic basis of phenotypic variation in outbred populations? Molecular Ecology Resources. 2018;18(4):739-54. https://doi.org/10.1111/1755-0998.12780

31. Li P, Fan Y, Yin S, Wang Y, Wang H, Xu Y, et al. Multi-environment QTL mapping of crown root traits in a maize RIL population. The Crop Journal. 2020;8(4):645-54. https://doi.org/10.1016/j.cj.2019.12.006

32. Fiorani F, Schurr U. Future scenarios for plant phenotyping. Annual review of plant biology. 2013;64(1):267-91. https://doi.org/10.1146/annurev-arplant-050312-120137

33. He S, Li X, Chen M, Xu X, Tang F, Gong T, et al. Crop HTP technologies: applications and prospects. Agriculture. 2024;14(5):723. https://doi.org/10.3390/agriculture14050723

34. Sul JH, Martin LS, Eskin E. Population structure in genetic studies: Confounding factors and mixed models. PLoS genetics. 2018;14(12):e1007309. https://doi.org/10.1371/journal.pgen.1007309

35. Miller CN, Harper AL, Trick M, Werner P, Waldron K, Bancroft I. Elucidation of the genetic basis of variation for stem strength characteristics in bread wheat by Associative Transcriptomics. BMC genomics. 2016;17:1-11. https://doi.org/10.1186/s12864-016-2775-2

36. Wood IP, Pearson BM, Garcia-Gutierrez E, Havlickova L, He Z, Harper AL, et al. Carbohydrate microarrays and their use for the identification of molecular markers for plant cell wall composition. Proceedings of the National Academy of Sciences. 2017;114(26):6860-5. https://doi.org/10.1073/pnas.1619033114

37. Ferguson JN, Fernandes SB, Monier B, Miller ND, Allen D, Dmitrieva A, et al. Machine learning-enabled phenotyping for GWAS and TWAS of WUE traits in 869 field-grown sorghum accessions. Plant Physiology. 2021;187(3):1481-500. https://doi.org/10.1093/plphys/kiab346

38. Alvarez M, Schrey AW, Richards CL. Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution? Mol Ecol. 2015;24(4):710-25. https://doi.org/10.1111/mec.13055

39. Zhang H, Hu Z, Yang Y, Liu X, Lv H, Song B-H, et al. Transcriptome profiling reveals the spatial-temporal dynamics of gene expression essential for soybean seed development. BMC Genomics. 2021;22(1):453. https://doi.org/10.1186/s12864-021-07783-z

40. Li S, Liu J, Zhao M, Su Y, Cong B, Wang Z. RNA quality score evaluation: A preliminary study of RNA integrity number (RIN) and RNA integrity and quality number (RNA IQ). Forensic Science International. 2024;357:111976. https://doi.org/10.1016/j.forsciint.2024.111976

41. Lu W, Zhou Q, Chen Y. Impact of RNA degradation on next-generation sequencing transcriptome data. Genomics. 2022;114(4):110429. https://doi.org/10.1016/j.ygeno.2022.110429

42. Brown RA, Epis MR, Horsham JL, Kabir TD, Richardson KL, Leedman PJ. Total RNA extraction from tissues for microRNA and target gene expression analysis: not all kits are created equal. BMC Biotechnology. 2018;18:1-11. https://doi.org/10.1186/s12896-018-0421-6

43. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Molecular Biology. 2006;7:1-14. https://doi.org/10.1186/1471-2199-7-3

44. Modi A, Vai S, Caramelli D, Lari M. The Illumina sequencing protocol and the NovaSeq 6000 system. Bacterial pangenomics: methods and protocols. Springer; 2021. p. 15-42. https://doi.org/10.1007/978-1-0716-1099-2_2

45. Lu G, Harper AL, Trick M, Morgan C, Fraser F, O'Neill C, Bancroft I. Associative transcriptomics study dissects the genetic architecture of seed glucosinolate content in *Brassica napus*. DNA Research. 2014;21(6):613-25. https://doi.org/10.1093/dnares/dsu024

46. Woodhouse S, He Z, Woolfenden H, Steuernagel B, Haerty W, Bancroft I, et al. Validation of a novel associative transcriptomics pipeline in *Brassica oleracea*: identifying candidates for vernalisation response. BMC Genomics. 2021;22:1-13. https://doi.org/10.1186/s12864-021-07805-w

47. Byrne A, Cole C, Volden R, Vollmers C. Realizing the potential of full-length transcriptome sequencing. Philosophical Transactions of the Royal Society B. 2019;374(1786):20190097. https://doi.org/10.1098/rstb.2019.0097

48. Rhoads A, Au KF. PacBio sequencing and its applications. Genomics, Proteomics & Bioinformatics. 2015;13(5):278-89. https://doi.org/10.1016/j.gpb.2015.08.002

49. Zhao L, Zhang H, Kohnen MV, Prasad KV, Gu L, Reddy AS. Analysis of transcriptome and epitranscriptome in plants using PacBio Iso-Seq and nanopore-based direct RNA sequencing. Frontiers in Genetics. 2019;10:253. https://doi.org/10.3389/fgene.2019.00253

50. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION sequencing and genome assembly. Genomics, Proteomics & Bioinformatics. 2016;14(5):265-79. https://doi.org/10.1016/j.gpb.2016.05.004

51. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology. 2016;17:1-11. https://doi.org/10.1186/s13059-016-1103-0

52. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. https://doi.org/10.1093/bioinformatics/btu170

53. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet Journal. 2011;17(1):10-2. https://doi.org/10.14806/ej.17.1.200

54. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010.

55. 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. https://doi.org/10.1093/bioinformatics/bts635

56. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature Methods. 2015;12(4):357-60. https://doi.org/10.1038/nmeth.3317

57. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9(4):357-9. https://doi.org/10.1038/nmeth.1923

58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. https://doi.org/10.1093/bioinformatics/btp352

59. Higgins J, Magusin A, Trick M, Fraser F, Bancroft I. Use of mRNA-seq to discriminate contributions to the transcriptome from the constituent genomes of the polyploid crop species *Brassica napus*. BMC Genomics. 2012;13:1-14. https://doi.org/10.1186/1471-2164-13-247

60. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20(9):1297-303. https://doi.org/10.1101/gr.107524.110

61. Richter F, Morton SU, Qi H, Kitaygorodsky A, Wang J, Homsy J, et al. Whole genome de novo variant identification with FreeBayes and neural network approaches. bioRxiv. 2020:2020.03.24.994160. https://doi.org/10.1101/2020.03.24.994160

62. Bancroft I, Morgan C, Fraser F, Higgins J, Wells R, Clissold L, et al. Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. Nature Biotechnology. 2011;29(8):762-6. https://doi.org/10.1038/nbt.1926

63. Harrison PW, Amode MR, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, et al. Ensembl 2024. Nucleic Acids Research. 2024;52(D1):D891-9.

64. Cingolani P. Variant annotation and functional prediction: SnpEff. Variant calling: methods and protocols. Springer; 2012. p. 289-314. https://doi.org/10.1007/978-1-0716-2293-3_19

65. Zhao S, Crouse W, Qian S, Luo K, Stephens M, He X. Adjusting for genetic confounders in transcriptome-wide association studies improves discovery of risk genes of complex traits. Nature Genetics. 2024;56(2):336-47. https://doi.org/10.1038/s41588-023-01648-9

66. Earl DA, VonHoldt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources. 2012;4:359-61. https://doi.org/10.1007/s12686-011-9548-7

67. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007;81(3):559-75. https://doi.org/10.1086/519795

68. Pritchard JK, Wen X, Falush D. Documentation for structure software: Version 2.3. University of Chicago, Chicago, IL. 2010;1:37.

69. Popescu A-A, Huber KT. PSIKO2: a fast and versatile tool to infer population stratification on various levels in GWAS. Bioinformatics. 2015;31(21):3552-4. https://doi.org/10.1093/bioinformatics/btv396

70. Cook DE, Andersen EC. VCF-kit: assorted utilities for the variant call format. Bioinformatics. 2017;33(10):1581-2. https://doi.org/10.1093/bioinformatics/btx011

71. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21(2):263-5. https://doi.org/10.1093/bioinformatics/bth457

72. Ecke W, Clemens R, Honsdorf N, Becker HC. Extent and structure of linkage disequilibrium in canola quality winter rapeseed (*Brassica napus* L.). Theoretical and Applied Genetics. 2010;120:921-31. https://doi.org/10.1007/s00122-009-1221-0

73. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics. 2006;38(2):203-8. https://doi.org/10.1038/ng1702

74. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23(19):2633-5. https://doi.org/10.1093/bioinformatics/btm308

75. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, et al. GAPIT: genome association and prediction integrated tool. Bioinformatics. 2012;28(18):2397-9. https://doi.org/10.1093/bioinformatics/bts444

76. Kusmec A, Schnable PS. FarmCPUpp: efficient large-scale genomewide association studies. Plant Direct. 2018;2(4):e00053. https://doi.org/10.1002/pld3.53

77. Zhang Y-W, Tamba CL, Wen Y-J, Li P, Ren W-L, Ni Y-L, et al. mrMLM v4.0.2: an R platform for multi-locus genome-wide association studies. Genomics Proteomics Bioinformatics. 2020;18(4):481-7. https://doi.org/10.1016/j.gpb.2020.06.006

78. Wen Y, Zhang Y, Zhang J, Feng J, Zhang Y. The improved FASTmrEMMA and GCIM algorithms for genome-wide association and linkage studies in large mapping populations. The Crop Journal. 2020;8(5):723-32. https://doi.org/10.1016/j.cj.2020.04.008

79. Tamba CL, Ni Y-L, Zhang Y-M. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. PLoS Computational Biology. 2017;13

(1):e1005357. https://doi.org/10.1371/journal.pcbi.1005357

80. Huang M, Liu X, Zhou Y, Summers RM, Zhang Z. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience. 2019;8 (2):giy154. https://doi.org/10.1093/gigascience/giy154

81. Sahu TK, Verma SK, Gayacharan, Singh NP, Joshi DC, Wankhede D, et al. Transcriptome-wide association mapping provides insights into the genetic basis and candidate genes governing flowering, maturity and seed weight in rice bean (*Vigna umbellata*). BMC Plant Biology. 2024;24(1):379. https://doi.org/10.1186/s12870-024-04976-y

82. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences. 2003;100(16):9440-5. https://doi.org/10.1073/pnas.1530509100

83. Wade AR, Duruflé H, Sanchez L, Segura V. eQTLs are key players in the integration of genomic and transcriptomic data for phenotype prediction. BMC Genomics. 2022;23(1):476. https://doi.org/10.1186/s12864-022-08690-7

84. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:1-13. https://doi.org/10.1186/1471-2105-9-559

85. Harper AL, McKinney LV, Nielsen LR, Havlickova L, Li Y, Trick M, et al. Molecular markers for tolerance of European ash (*Fraxinus excelsior*) to dieback disease identified using Associative Transcriptomics. Scientific Reports. 2016;6(1):19335. https://doi.org/10.1038/srep19335

86. Paritosh K, Yadava SK, Singh P, Bhayana L, Mukhopadhyay A, Gupta V, et al. A chromosome-scale assembly of allotetraploid *Brassica juncea* (AABB) elucidates comparative architecture of the A and B genomes. Plant Biotechnology Journal. 2021;19(3):602-14. https://doi.org/10.1111/pbi.13492

87. Jayakodi M, Schreiber M, Stein N, Mascher M. Building pan-genome infrastructures for crop plants and their use in association genetics. DNA Research. 2021;28(1):dsaa030. https://doi.org/10.1093/dnares/dsaa030

88. Sollars ES, Buggs RJ. Genome-wide epigenetic variation among ash trees differing in susceptibility to a fungal disease. BMC Genomics. 2018;19:1-15. https://doi.org/10.1186/s12864-018-4874-8

89. Franco Ortega S, Bedford JA, James SR, Newling K, Ashton PD, Boshier DH, et al. *Fraxinus excelsior* updated long-read genome reveals the importance of MADS-box genes in tolerance mechanisms against ash dieback. bioRxiv. 2024:2024.12.20.629733. https://doi.org/10.1101/2024.12.20.629733

90. Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. Nature Plants. 2020;6(8):914-20. https://doi.org/10.1038/s41477-020-0733-0

91. Vadapalli S, Abdelhalim H, Zeeshan S, Ahmed Z. Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. Briefings in Bioinformatics. 2022;23(5):bbac191. https://doi.org/10.1093/bib/bbac191

92. Ahmad S, Shahzad R, Jamil S, Tabassum J, Chaudhary MAM, Atif RM, et al. Regulatory aspects, risk assessment and toxicity associated with RNAi and CRISPR methods. CRISPR and RNAi systems. Elsevier; 2021. p. 687-721. https://doi.org/10.1016/B978-0-12-821910-2.00013-8

93. Koprivova A, Harper AL, Trick M, Bancroft I, Kopriva S. Dissection of the control of anion homeostasis by associative transcriptomics in *Brassica napus*. Plant Physiology. 2014;166 (1):442-50. https://doi.org/10.1104/pp.114.239947

94. Alcock TD, Havlickova L, He Z, Bancroft I, White PJ, Broadley MR, et al. Identification of candidate genes for calcium and magnesium accumulation in *Brassica napus* L. by association genetics. Frontiers in Plant Science. 2017;8:1968. https://doi.org/10.3389/fpls.2017.01968

95. Hejna O, Havlickova L, He Z, Bancroft I, Curn V. Analysing the genetic architecture of clubroot resistance variation in *Brassica napus* by associative transcriptomics. Molecular Breeding. 2019;39:1-13. https://doi.org/10.1007/s11032-019-1021-4

96. Huang Y, Hussain MA, Luo D, Xu H, Zeng C, Havlickova L, et al. A *Brassica napus* reductase gene dissected by associative transcriptomics enhances plant adaption to freezing stress. Frontiers in Plant Science. 2020;11:971. https://doi.org/10.3389/fpls.2020.00971

97. Xu J, Zhan H, Xie Y, Tian G, Xie L, Xu B, et al. Associative transcriptomics study dissects the genetic architecture of seedling biomass-related traits in rapeseed (*Brassica napus* L.). Plant Breeding. 2021;140(2):285-93. https://doi.org/10.1111/pbr.12898

98. Jacott CN, Schoonbeek HJ, Sidhu GS, Steuernagel B, Kirby R, Zheng X, et al. Pathogen lifestyle determines host genetic signature of quantitative disease resistance loci in oilseed rape (*Brassica napus*). Theoretical and Applied Genetics. 2024;137 (3):65. https://doi.org/10.1007/s00122-024-04569-1

99. Yalcin HA, Jacott CN, Ramirez-Gonzalez RH, Steuernagel B, Sidhu GS, Kirby R, et al. A complex receptor locus confers responsiveness to necrosis and ethylene-inducing like peptides in *Brassica napus*. The Plant Journal. 2024;119(1):266-82. https://doi.org/10.1111/tpj.16760

100. Liu Q, Xie Y, Liu B, Yin H, Zhou Z, Feng Z, et al. A transcriptomic variation map provides insights into the genetic basis of *Pinus massoniana* Lamb. evolution and the association with oleoresin yield. BMC Plant Biology. 2020;20:1-14. https://doi.org/10.1186/s12870-020-02577-z

Additional information

Peer review: Publisher thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

Reprints & permissions information is available at https://horizonepublishing.com/journals/index.php/PST/open_access_policy

Publisher's Note: Horizon e-Publishing Group remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Indexing: Plant Science Today, published by Horizon e-Publishing Group, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, NAAS, UGC Care, etc
See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting

Publisher information: Plant Science Today is published by HORIZON e-Publishing Group with support from Empirion Publishers Private Limited, Thiruvananthapuram, India.