Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Plant Biology

New perspectives of post-GWAS analyses: From markers to causal genes for more precise crop breeding

Ivana Kaňovská, Jana Biová and Mária Škrabišová

Crop breeding advancement is hindered by the imperfection of methods to reveal genes underlying key traits. Genome-wide Association Study (GWAS) is one such method, identifying genomic regions linked to phenotypes. Post-GWAS analyses predict candidate genes and assist in causative mutation (CM) recognition. Here, we assess post-GWAS approaches, address limitations in omics data integration and stress the importance of evaluating associated variants within a broader context of publicly available datasets. Recent advances in bioinformatics tools and genomic strategies for CM identification and allelic variation exploration are reviewed. We discuss the role of markers and marker panel development for more precise breeding. Finally, we highlight the perspectives and challenges of GWAS-based CM prediction for complex quantitative traits.

Addresses
Department of Biochemistry, Faculty of Science, Palacký University in Olomouc, Šlechtitelů 27, Olomouc 77900, Czech Republic

Corresponding author: Škrabišová, Mária (maria.skrabisova@upol.cz)

Current Opinion in Plant Biology 2024, **82:**102658

This review comes from a themed issue on **Genome studies and molecular genetics 2024**

Edited by **Leena Tripathi** and **Sushma Naithani**

For a complete overview see the Issue and the Editorial

Available online xxx

https://doi.org/10.1016/j.pbi.2024.102658

Keywords
GWAS, Crop breeding, Molecular markers, Causal gene, Causative mutation.

Introduction

As the global population continues to increase, ensuring access to sufficient and nutritious food is becoming increasingly important. To achieve this, it is essential to increase crop production sustainably. While traditional breeding methods have been effective in this endeavor for many years, they are no longer sufficiently powerful to meet the rising demand for increased yield and quality of produce, as these methods are time-consuming and constrained by various factors.

In this context, exploring genomic determinants and identifying causal genes are crucial for accelerating crop breeding. Genome-wide association study (GWAS) is a powerful statistical method that helps uncover the association between genomic variants and phenotypes that represent important agronomical traits. GWAS has been successfully used for well over 15 years in human health [1] and is also used to help with plant breeding improvement. However, GWAS is still limited by numerous factors such as sample size, phenotype quality and distribution in the data set, and genotype quality [2]. Another drawback of GWAS is that it fails to identify rare alleles or multiple independent alleles of one single gene [3*].

Box 1. Post-GWAS Glossary

Candidate gene: A gene hypothesized to be involved in determining
a particular phenotype of a trait but has not yet been proven experimentally (e.g., by genome editing, plant transformation).

Causal gene: A gene that has been proven experimentally (e.g., by genome editing, plant transformation, biochemical characterization of recombinant proteins, etc.) to determine a particular phenotype of a trait.

Causative mutation (CM): A mutation in a causal gene that modifies its encoded protein and thus leads to a phenotype change. CM can directly affect the protein or disrupt its original transcription pattern (e.g., mutations in promoters). Often, there are many polymorphisms in a gene that are inherited as alleles, and currently, the methods that can confirm a polymorphism to be the CM of the causal gene are limiting the progress in crop breeding improvement.

Marker efficiency: A measure of how well a marker can predict a phenotype, an estimate of direct correspondence between a variant position and a phenotype (calculated as accuracy).

Diversity panels: A collection of resequenced samples from data sets aggregated from independent studies that cover every genetic variant present in the worldwide population of an organism, ideally curated for missing data (imputed) and mapped to the same reference genome assembly version.

Omics-based filtering: GWAS associates genomic regions where the CM is often masked by frequent false positive variant positions. Omics-based filtering overlays GWAS data with omics data to reduce false positives (e.g., transcriptomic data are used to search for genes expressed in the associated locus). Another powerful (yet parallel to standard GWAS) way of utilizing omics data is to input it directly into GWAS as phenotypic data (in-GWAS, not covered in this review).

Though GWAS and other association studies identify associated genomic regions that are statistically linked to traits, these regions are not always causative. Regardless of the type of genetic feature underlying a desired phenotype, in GWAS, SNPs are significantly less prone to technical issues of current sequencing methods than Indels. Therefore, often, the highest associated SNP identified in GWAS is not the causal one but is in linkage disequilibrium (LD) with the causative genetic feature. Additionally, regression models that are currently utilized in GWAS are, by their nature, suppressing false positives on the one hand and inflating false negatives on the other hand [4]. Many different approaches can be used following GWAS to narrow down the list of associated variant positions (SNPs and Indels) and thus discover the causal gene. One of the traditional post-GWAS approaches is fine mapping which requires dense phenotyping or sequencing of large populations [5]. However, a more commonly used methodology for gene prioritization for GWAS-based discovery is the integration of omics data. The most recent approach is the implementation of machine learning (ML). In general, these same approaches can also be applied to post-QTL (quantitative trait loci) analysis. However, despite advances in the precision of these methods, the causality of identified mutations can only be confirmed experimentally.

Here, we focus on recent trends and strategies in post-GWAS methods that are described in Figure 1.

Integration of omics data in post-GWAS analysis

Post-GWAS analyses are often coupled with omics data to lower the number of false positives and more accurately identify causal variants. One approach to integrating omics data is to use them directly for association analysis instead of genotype data. For example, a transcriptome-wide association study (TWAS) detects the association between changes in gene expression (expression QTL/eQTL) and a phenotype [5]. Similar methods include metabolome (mGWAS), epigenome (EWAS), and proteome (PWAS) association studies [5]. However, similar to GWAS, the limitation of omic-based associations is that they identify broad regions of associated variants with a likelihood of a high number of false positives [6].

Nevertheless, the omics-based prediction can also inflate false negatives if the analysis is limited to only one type of omics data or a small sample size [7]. Combining multiple types of omics data reduces the likelihood of false negatives and positives [8]. Therefore, integrating multiple omics data types mitigates the limitations of relying on just a single data type.

Another way to integrate GWAS and omics data is by identifying causal variants in both GWAS and eQTL studies [9]. To pinpoint causal genes or mutations from associated loci using eQTL and GWAS integration, first, fine-mapping is carried out to identify candidate variants, and then colocalization methods are used [9] to assess whether the GWAS and eQTL signals overlap. Candidate genes are then prioritized and validated using functional annotations and further confirmed experimentally through methods such as plant transformation and gene editing.

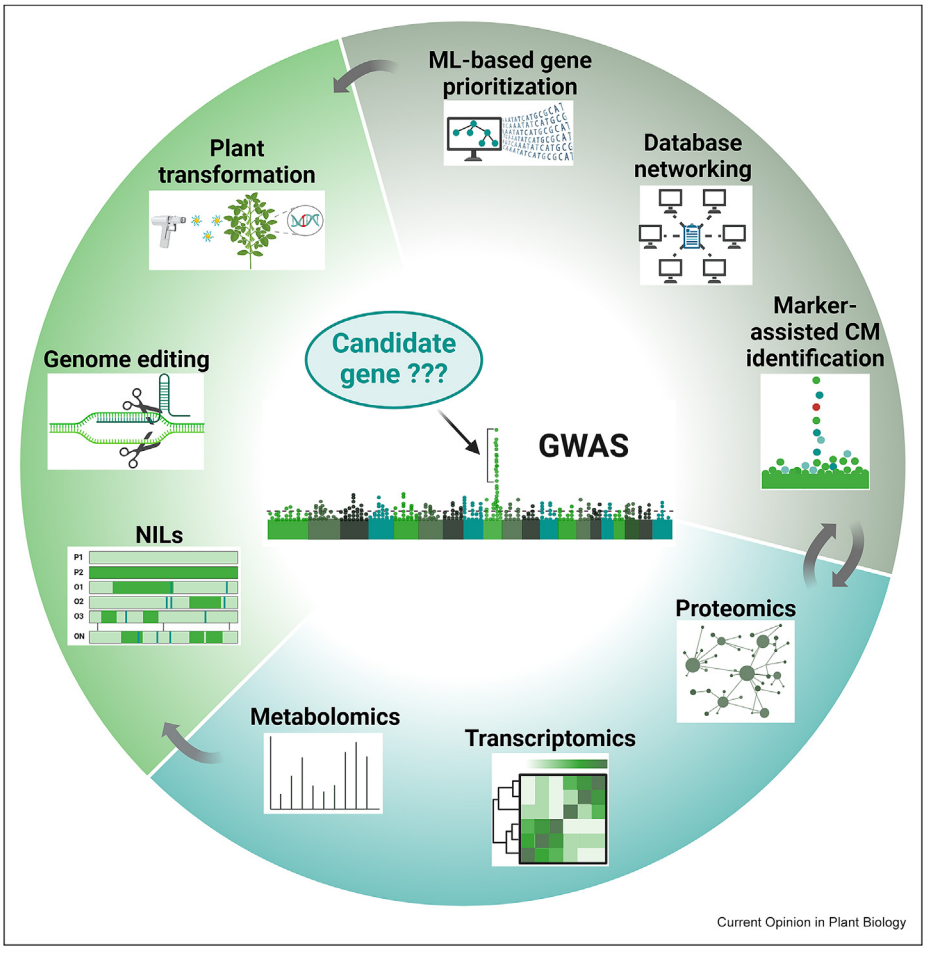
The integration of reference genomes, pan-genomes, and multi-omics data into unified databases is crucial for enhancing data interaction and improving post-GWAS prediction accuracy. Recent efforts in creating vast databases resulted in immense multi-omics resources such as SoyOmics for soybean [10*]; CottonMD database for cotton [11], BnIR for *Brassica* [12] ZEAMAP for maize [13], Gramene [14] and Phytozome [15]. Despite the efforts that were made for these crops, it is crucial to maximize input data quality and sample consistency for every species, as these factors are critical in analyzing multiple datasets from various dimensions.

Although using pan-genome for association studies can be computationally demanding, it allows for the analysis of a broader range of genetic diversity. The same omic-based post-GWAS strategies can then be applied, leading to the identification of new associated loci. For instance, recent eQTL mapping using SNPs from a rice super pan-genome uncovered new candidate genes related to stress tolerance in rice [16]. In addition to the aforementioned databases, many others enable users to analyze multi-omics data in a pan-genome context thoroughly; for example the BnaOmics database for *Brassica napus* [17], grapevine [18], and various other crops [19].

Marker-assisted causative mutation identification at the crossroads between reductionism and pan-genomes

Unlike marker-assisted breeding, where the original purpose is to select individuals with desirable traits using molecular markers without extensive knowledge of causal genes, marker-assisted causative mutation identification aims to identify causative mutations in causal genes with the assistance of molecular markers.

Figure 1



The post-GWAS wheel of methods. The scheme summarizes the most common approaches used in the post-GWAS analysis. The methods are grouped into three areas. The grey area encompasses the methods that utilize the previous knowledge with computational capabilities and algorithms: the common approaches in the group are represented by ML-based gene prioritization [39,44,59,68] utilization of multiple types of data from databases [10,13,17,59,68,69], and marker-assisted CM identification [26,27]. The blue area covers methods that add another layer to GWAS results, the omics data: transcriptomics [16,68,70,71], metabolomics [8,40], proteomics [8,40], or the combination of all of these [7,10,40]. The green area covers methods that are based on the engineering of genetic information: plant transformation [72], genome editing [71,73,74], and the creation of populations of nearly isogenic lines (NILs) [71,74]. The arrows between the areas in the post-GWAS wheel indicate that the approaches can be used subsequently or combined to achieve the best result and compensate for the limits of the individual methods [16,71]. Not all the methods can be used for every crop since the techniques are often limited for certain species; for instance, sufficient and reproducible protocols for transformation in some plant species are still not available [75–77]. This illustration was created using BioRender.com.

Recent advances in sequencing technologies have led to the creation of reference genomes for most crops and have enabled the flourishing of genomics-based predictions [20]. However, despite the decreasing cost of resequencing, it remains a challenge for many crop species. Investing more funds in the extensive resequencing of more individuals across various crop species would improve both the quality and quantity of genotype data, which could then be used to identify the causative mutations (CM) and accelerate crop breeding.

However, the tendency has been quite the opposite. Crop breeders are more inclined to develop marker panels associated with desired traits than to identify the actual CMs underlying the phenotypes of interest. This trend is understandable, as markers are predominantly used in modeling for genomic selection (GS), where the genetic value of individuals is predicted for breeding purposes. Since the genetic value is an individual's genetic makeup (genotype) relative to its observable traits, markers used in GS typically cover complex, multi-genic traits like yield, protein content, and other quantitative traits. In contrast to a simple dominant gene controlling a qualitative trait (e.g., complete loss of pigmentation in soybean flowers [21]), complex traits are controlled cumulatively by independent loci, each

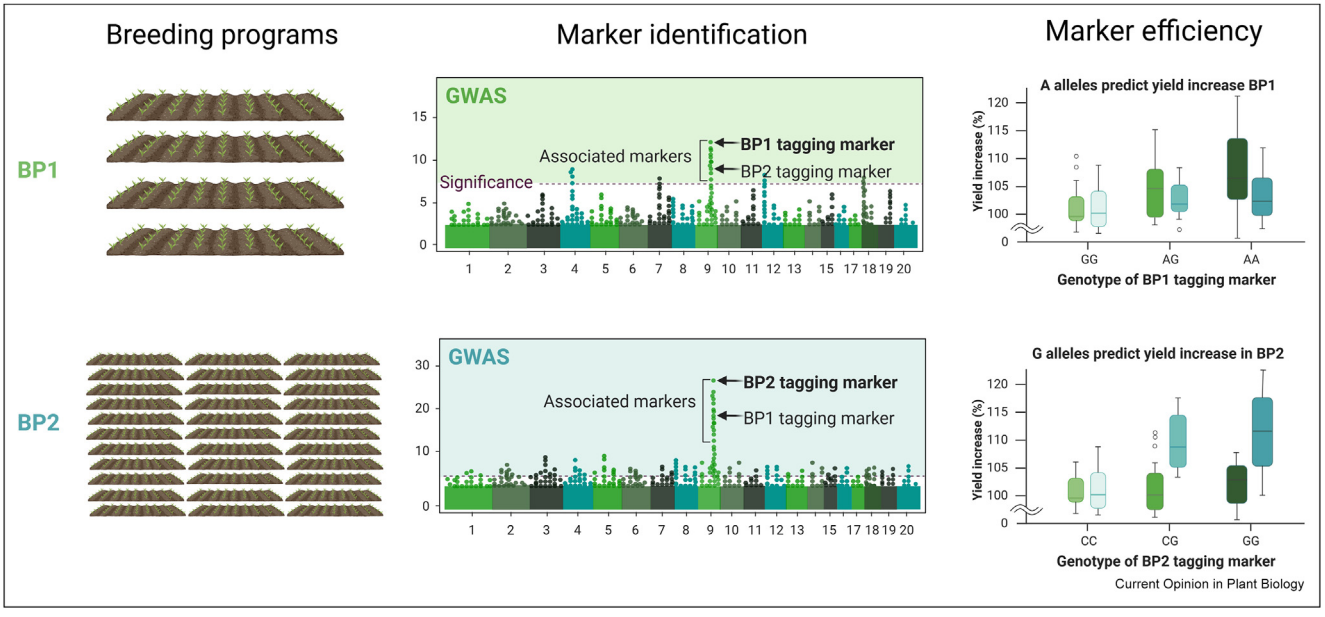
carrying multiple genes with varying allelic effects. Therefore, the markers developed for GS are not optimal in identifying individual small-effect genes of complex traits which is often the aim of markers utilization in GWAS-driven discovery.

A recent trend in GS involves reducing the number of markers in the analysis to avoid the "too much data kills data" effect based on recent experiences caused by overall data availability due to faster data generation than utilization rates [22]. One such example is the genotyping of soybean, where the number of markers has decreased from the initial 200K Axiom® SoyaSNP array [23] and Illumina SoySNP50K DNA bead chip [24] through 6K and 3K to the recent 1K (Agriplex Community Soy 1K SNP) marker panel. These reduced panels of markers have also been used for GWAS, as in the case of the SoySNP6K marker that was used to identify causal genes for soybean cyst nematode resistance [25]. At the same time, the widely used practice is to identify CM based on the predicted annotated function of surrounding genes or the proximity of such genes to the marker. However, lower marker density in the analysis raises the question of how far a CM can be from its associated tagging marker. Recent research has shown that markers obtained from low-density genotypes can be leveraged and used to subsidize missing phenotypes of resequenced data sets from other independent studies, thereby assisting with more precise CM prediction [26**].

Figure 2 illustrates how markers identified as significantly associated (tagging markers) in an idealized GWAS for yield in two independent breeding programs (BP) predict the same phenotype of the other BP with varying efficiency. The scheme rationalizes against marker adoption without accuracy testing [26**] on a simplified example. Although it is a common practice to identify a marker in one BP, validate it in several others, and use it to screen individuals in future breeding populations, similar genetic backgrounds of the testing populations can cause an overestimation of the adopted marker's efficiency to predict a phenotype in other, genetically less related BPs.

The reality of quantitative traits with a polygenic nature in complex genome crops is undoubtedly more vulnerable to the failure of adopted markers. Accuracy-based marker and CM identification have gained support through recent efforts in testing gene-based markers in rice [27**]. Therefore, accuracy testing should be performed routinely to measure the effectiveness of predicting the direct correspondence between markers (SNPs) and variant positions (SNPs, insertions, and

Figure 2



Schematics of an idealized marker identification for two different breeding programs (BP) and their efficiency in predicting yield. BP1 (green) and BP2 (blue) are idealized breeding programs with unspecified genetic backgrounds aiming for yield increase, here, for the sake of simplicity, yield is considered to be a Mendelian trait. Therefore, only one single genetic locus is identified by GWAS for yield in this example (based on the statistical significance of the associated markers). Regardless of the size of the tested populations in BP1 and BP2, the tagging markers identified in GWAS cannot be used interchangeably for yield prediction due to their varying efficiency. Here, the fictive parental lines of the two breeding programs bear different donor alleles of the causal gene underlying high yield. Assuming each marker is in nearly perfect correspondence with its high-yielding alleles, adopting markers from other breeding programs can be counterproductive without accuracy calculation. This illustration was created using BioRender.com.

deletions) or phenotypes. Pinpointing causal features for both simple and complex traits by a marker-assisted causative mutation identification approach relies on accuracy testing, where the variant position with the highest value is identified as causal. This method helps distinguish associated from causal variants.

A fundamentally different approach to marker reductionism involves increasing the number of markers by using pan-genomes as new references. This approach is bolstered by the recognition that linear reference genomes have limitations in genomic-based predictions, as they fail to capture the diversity within a species and can distort our understanding of the genomic basis of traits [28]. However, routine use of pan-genomes is still constrained by factors such as the availability of large-scale data storage and the computational demands of applying them to downstream multi-omics analyses [29].

Solution for rare phenotypes and multiple alleles

GWAS, as a statistical method, relies heavily on sample size, genotype quality, and phenotype distribution. A trending approach to overcome this limitation is to maximize the input data [30]. Since resources to resequence, genotype, or phenotype excess samples are often limited in breeding programs, new pipelines for genomic data aggregation from independent studies have been developed [31**]. This was demonstrated in the Synthetic phenotype association study [26**], where multiple independent resequenced datasets were aggregated into a single diversity panel and where, at the same time, substitution of missing phenotypes by the genotype of a tagging marker resulted in successful post-GWAS CM identification.

Aggregated data can serve as crop-specific diversity panels that allow allele exploration and post-GWAS for rare phenotypes. Recently, strategies for effectively utilizing publicly available genomic datasets without introducing substantial biases have been proposed [32] with a meta-imputation method emerging as a viable solution [33]. Harnessing broad natural diversity has already led to groundbreaking discoveries of treasures that would otherwise stay overlooked, as shown in wheat [34*].

With the integration of the diverse panels in GWAS, the chance of more than one CM in a single gene increases. By its nature, GWAS typically identifies either a variant position that correlates with all the multiple CMs present in the causal gene, or the variant position of the most frequent CM, rather than detecting all CMs simultaneously. Recently, the MADis tool has been developed to enable genomic analysis of natural and artificial selection by multiple alleles prediction [3*]. The MADis tool is available online for a diversity panel of 1066 soybean accessions [31**] and can also be utilized for private datasets or other species in the form of an available Python package. After providing the genotype and phenotype data, the MADis tool tests for a combination of variant positions in a gene that explains most of the phenotype. The MADis tool effectively identifies multiple alleles in a single gene including also rare alleles with only a single occurrence in a dataset.

Machine learning and artificial intelligence in GWAS

Machine learning (ML) offers a wide array of algorithms that can be applied to various types of data analysis. Given its versatility, it is no surprise that ML-based data analysis is also applicable to casual gene discoveries. The choice of the ML algorithm for data analysis depends largely on the data structure and specific results that are required [35–37]. The most commonly used ML algorithms for GWAS-based discoveries are regression algorithms [38,39], dimensional reduction algorithms [38,40–42], algorithms based on decision trees [38,39,43,44], and neural networks [38,41]. One of the great advantages of ML-based data analysis, which also finds application in crop genome research, is its ability to process large, complex datasets, including those that are multi-dimensional or have significant amounts of missing data [35–38]. For these purposes, dimensional reduction algorithms can be used for data simplification [37,38,40,42]. Besides the simplification step, algorithms like random forest [43,44] or neuronal network-based algorithms [38,41,45] can be implemented for data processing of complex data [35–37].

ML plays a significant role in GWAS-based discoveries, as it can be applied at various stages of the process [37]. ML algorithms can be used prior to the GWAS to acquire and pre-process the input data that covers both phenotype [38,40,41,43] and genotype data [40,42]. This step includes dimensional reduction, which can be performed on both phenotype (omic, e-traits, and other complicated data used as phenotype) and genotype data [40,42]. For phenotyping purposes, neural networks are particularly useful for image analysis, enabling the acquisition of phenotype data for large sample sets [38,41].

During the GWAS itself, an ML-based model can be implemented [39,46]. Finally, the ML algorithms assist in gene prioritization in the post-GWAS phase, where the decision tree-based algorithm random forest is often used for crops and serves for choosing the causal gene candidates from the associated loci [39,44]. Last but not least, phenotype can be predicted from genotype data by utilizing the Genotype-to-Phenotype (G2P) strategy, where all regression, decision tree-based, and neural network algorithms can be implemented [45,47]. To summarize, ML can improve causal gene identification in various aspects, depending on the diverse types of

data used for identification and reflecting the natural diversity and complexity of genomes.

Current computational tools for post-GWAS analysis

This review aims to serve as an inspiring resource for the current post-GWAS strategies that can be adopted for more precise and accelerated pre-breeding of various crops and plant species. Therefore, we have compiled available tools and platforms for post-GWAS analysis of crop and model species data, focusing on those that are actively supported, maintained, and updated (Table 1). It is important to note that none of the platforms offer an extensive resource that would integrate all the omics data categories: genomics, phenomics, transcriptomics, proteomics, metabolomics, and epigenomics. The most comprehensive omics data platform to date is the maize

Table 1

Current Computational Tools for post-GWAS Analysis. The table summarizes the computational tools for crop causal gene identification, including the species, processed data type, source, and short description.

Tool/Platform name	Species	Omics data types	Integration objective and description	Access/Code source	Reference
QTG-Finder	Arabidopsis, rice, sorghum	Genomics	Gene prioritization, ML-based algorithm for causal gene identification in associated loci.	https://github.com/carnegie/QTG_Finder	[44]
MODAS	Maize	Genomics, phenomics, transcriptomics, proteomics, metabolomics	ML-based, gene prioritization, dimensional reduction, eGWAS/metabGWAS, Mendelian randomization algorithms, and gene annotation integration.	https://modas-bio.github.io/	[40]
MaizeNetome	Maize	Genomics, phenomics, transcriptomics	Data analysis by networking.	http://minteractome.ncpgr.cn/qtgfinder.php	[48]
Milletdb	Millets	Genomics, phenomics, transcriptomics, epigenomics	Data integration, analysis and visualization of results.	http://milletdb.novogene.com	[50]
BnIR, BnaOmics	Rapeseed	Genomics, phenomics, transcriptomics, epigenomics	Database with tools for data visualization.	https://yanglab.hzau.edu.cn/BnIR, https://bnaomics.ocri-genomics.net/	[12,17]
RicePilaf	Rice	Genomics, phenomics, transcriptomics	Gene prioritization, integrates publically available data.	https://github.com/bioinfodlsu/rice-pilaf	[59]
MBKbase	Rice	Genomics, phenomics	Data analysis and visualization.	https://mbkbase.org/rice	[49]
SoyHUB	Soybean	Genomics, phenomics	Exploration of natural and artificial soybean diversity, gene prioritization. A suite of tools for soybean applied genomics with a curated diversity panel of soybean accessions.	https://soykb.org/soyhub.php/	[3,26,31,51–54]
SoyOmics	Soybean	Genomics, phenomics, transcriptomics	Database with tools for data visualization and analysis. Include GWAS and QTL, omic, genotype, and phenotype data.	https://ngdc.cncb.ac.cn/soyomics	[10]
KBCommons	Soybean, Arabidopsis, maize, other crops	Genomics, phenomics, transcriptomics	Crop genomic data integration, allele discovery and analysis, gene prioritization.	https://kbcommons.org/	[31,53]

MODAS [40], which, along with QTL-Finder [44], is one of the few tools with ML-based analysis capabilities. In addition to ML-based tools, MaizeNetome offers data analysis by networking [48]. While platforms like MBKbase [49] and Soyomics [10*] are limited in the type of omics data they cover, they both support the visualization of analysis outputs. Milletdb [50], BnIR [12], and BnaOmics [17] stand out by integrating epigenomics data, a feature uncommon even in continuously and actively maintained and data-rich databases for crops such as maize, rice, or soybean.

A suite of tools for post-GWAS analysis of soybean data [3*,26**,31**,51-54] is available at SoyHUB hosted at SoyKB [55,56]. KBCommons [57,58], an extension of SoyKB, supports tools and applied genomics strategies for other crops and plant species such as rice [53], maize and *Arabidopsis* [31**]. Another platform, that integrates publicly available data, similar to SoyHUB and KBCommons, is the RicePilaf platform for rice [59*].

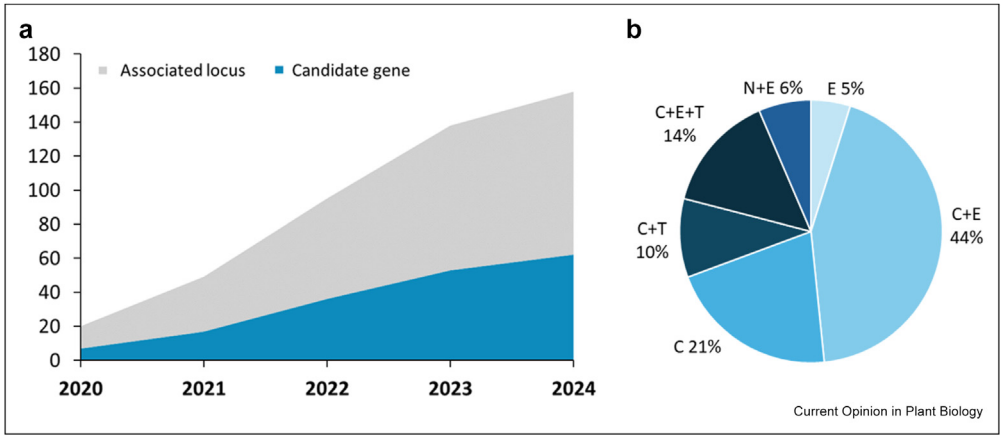
Perspectives and challenges of GWAS-based CM prediction for complex quantitative traits

Improving the predictive power of GWAS for complex traits remains challenging due to multidimensional colinearity and polygenicity, as recently reviewed in the context of neuropsychiatric disorders [60]. The challenge has been tackled by various approaches, spanning fitting model improvement and additional data implementation. An epigenetics-based association study (EWAS) can identify genes via confounding or reverse causation, offering an alternative to GWAS [61*]. Another approach aims for computationally more efficient ML implementation of a whole-genome regression fitting model [62]. Another promising approach in crop breeding is focused on genome-to-phenome predictions utilizing ML [47,63]. The complexity of polygenic traits is complicated by the limited methodology that would facilitate the prediction of the contribution of the existing alleles to the mosaic of effects of the multiple genes underlying a quantitative phenotype.

Data storage and sharing

Data science is a critical component of genomics, and advancements in science and technology have enabled data to provide unprecedented levels of information. This has led to the proposal of the FAIR principles [64,65]. However, the practical implementation of these principles can be hindered by several complicating factors. One key issue is the 'half-life' of data, which is significantly influenced by the stability of the data storage repository. For optimal implementation of FAIR principles, a stable, long-term repository is essential. Another challenge lies in the diversity of data types and formats, which require comprehensive metadata for effective FAIR applications [64,66*]. Another major issue in data sharing arises from the identification of samples. Without a unique identifier for plant materials that is consistently used across all databases, the ability to connect and efficiently reuse data is compromised, diminishing its overall value [64,66*]. Similarly, complications can occur with the naming of genes or gene products if not standardized correctly. Ensuring unique and consistent identifiers for both plant materials and

Figure 3



Progress in causal gene identification on an example of soybean. (a), The plot summarizes the open-access studies listed on the Web of Science as published between January 2000 and June 2024, where GWAS was performed on soybean. The two categories in the plot group are the publications based on the GWAS-associated locus identification (grey) and the publications where a subsequent methodology was applied to identify a candidate gene (blue). **(b)**, This plot further divides the "candidate gene" group from A into the following categories based on the used methodology: C – computational, *in-silico* prediction, E – expression-based or other omics data type-based prediction, T – transgenic approach (plant transformation or genome editing), N – NILs population-based prediction.

gene products is essential to maintain the integrity and utility of shared data [64].

Confirmation of predicted causal genes lags behind the identification of candidate genes

Based on our review of the soybean GWAS studies (Figure 3a), in less than 40 % of these studies, the associated loci were analyzed by subsequent post-GWAS analysis. Although our data visualization might suggest a steady trend, our detailed follow-up investigation revealed that only 24 % of the candidate genes were confirmed by plant transformation or genome editing (Figure 3b). We deliberately limited our search to *in vivo* confirmations, thus excluding biochemical characterization of recombinant proteins encoded by the candidate genes or other such *in vitro* methods. This insight into the candidate gene confirmation is based on soybean, however, the conclusion that biological confirmation lags behind the identification of candidate genes can be generalized to other crops with an even wider gap between the identified versus confirmed genes since, in contrast to other crops, soybean transformation has been successfully used for years [67].

Conclusion

Post-GWAS approaches play a critical role in reducing the number of false positives and facilitating causal gene identification. By integrating multi-omics data, utilizing existing data for CM identification, incorporating global genetic diversity, including pan-genomes, and employing ML and artificial intelligence, crop breeding can become more precise and efficient. The limitations of current approaches in biologically confirming candidate genes highlight the need for more reliable prediction methods. Reducing uncertainty in candidate gene lists would undoubtedly encourage transgenic experiments. Post-GWAS methods address this need and at the same time, increased data availability enhances predictive power. Here we advocate for the generation of high-quality genomes and reference genomes, adherence to FAIR principles, adoption of the latest genomic strategies, and enhancement of data storage and computing infrastructure (see Box 1).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Open access funding for this publication was provided by the Department of Biochemistry, Faculty of Science, Palacký University in Olomouc, Olomouc, Czechia. This work was supported by the Legume Generation (Boosting innovation in breeding for the next generation of legume crops for Europe) project that has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No. 101081329. It also receives support from the governments of the United Kingdom, Switzerland and New Zealand. Further, this work was also supported by the project ToWArds Next GENeration Crops, reg. No. CZ.02.01.01/00/22_008/0004581 of the ERDF Programme Johannes Amos Comenius by the Ministry of Education, Youth and Sport of the Czech Republic.

Data availability

No data was used for the research described in the article.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Abdellaoui A, Yengo L, Verweij KJH, Visscher PM: **15 years of GWAS discovery: realizing the promise**. *Am J Hum Genet* 2023, **110**:179–194, https://doi.org/10.1016/j.ajhg.2022.12.011.

2. Tibbs Cortes L, Zhang Z, Yu J: **Status and prospects of genome-wide association studies in plants**. *Plant Genome* 2021, **14**, e20077, https://doi.org/10.1002/tpg2.20077.

3. * Biová J, Kaňovská I, Chan YO, Immadi MS, Joshi T, Bilyeu K, Škrabišová M: **Natural and artificial selection of multiple alleles revealed through genomic analyses**. *Front Genet* 2024, **14**, https://doi.org/10.3389/fgene.2023.1320652.
A Multiple Alleles discovery (MADis) tool identifies multiple CMs that arose independently in a single causal gene.

4. Sesia M, Bates S, Candès E, Marchini J, Sabatti C: **False discovery rate control in genome-wide association studies with population structure**. *Proc Natl Acad Sci USA* 2021, **118**, e2105841118, https://doi.org/10.1073/pnas.2105841118.

5. Gupta PK, Kulwal PL, Jaiswal V: **Chapter Two - association mapping in plants in the post-GWAS genomics era**. In *Advances in genetics*. Edited by Kumar D, Academic Press; 2019: 75–154, https://doi.org/10.1016/bs.adgen.2018.12.001. vol. 104.

6. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K, *et al.*: **Opportunities and challenges for transcriptome-wide association studies**. *Nat Genet* 2019, **51**:592–599, https://doi.org/10.1038/s41588-019-0385-z.

7. Baranger DAA, Hatoum AS, Polimanti R, Gelernter J, Edenberg HJ, Bogdan R, Agrawal A: **Multi-omics cannot replace sample size in genome-wide association studies**. *Gene Brain Behav* 2023, **22**, e12846, https://doi.org/10.1111/gbb.12846.

8. Yang Y, Saand MA, Huang L, Abdelaal WB, Zhang J, Wu Y, Li J, Sirohi MH, Wang F: **Applications of multi-omics technologies for crop improvement**. *Front Plant Sci* 2021, **12**, 563953, https://doi.org/10.3389/fpls.2021.563953.

9. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, *et al.*: **Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets**. *Nat Genet* 2016, **48**: 481–487, https://doi.org/10.1038/ng.3538.

10. * Liu Y, Zhang Y, Liu X, Shen Y, Tian D, Yang X, Liu S, Ni L, Zhang Z, Song S, *et al.*: **SoyOmics: a deeply integrated database on soybean multi-omics**. *Mol Plant* 2023, **16**:794–797, https://doi.org/10.1016/j.molp.2023.03.011.
This work describes the creation, compilation, and utilization of multi-omics data as a one-stop solution for big data mining for soybean with a friendly and interactive interface.

11. Yang Z, Wang J, Huang Y, Wang S, Wei L, Liu D, Weng Y, Xiang J, Zhu Q, Yang Z, *et al.*: **CottonMD: a multi-omics database for cotton biological study**. *Nucleic Acids Res* 2023, **51**: D1446–D1456, https://doi.org/10.1093/nar/gkac863.

12. Yang Z, Wang S, Wei L, Huang Y, Liu D, Jia Y, Luo C, Lin Y, Liang C, Hu Y, *et al.*: **BnIR: a multi-omics database with various tools for Brassica napus research and breeding**. *Mol Plant* 2023, **16**:775–789, https://doi.org/10.1016/j.molp.2023.03.007.

13. Gui S, Yang L, Li J, Luo J, Xu X, Yuan J, Chen L, Li W, Yang X, Wu S, *et al.*: **ZEAMAP, a comprehensive database adapted to the maize multi-omics era**. *iScience* 2020, **23**, 101241, https://doi.org/10.1016/j.isci.2020.101241.

14. Tello-Ruiz MK, Naithani S, Gupta P, Olson A, Wei S, Preece J, Jiao Y, Wang B, Chougule K, Garg P, *et al.*: **Gramene 2021: harnessing the power of comparative genomics and pathways for plant research**. *Nucleic Acids Res* 2021, **49**: D1452–d1463, https://doi.org/10.1093/nar/gkaa979.

15. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, *et al.*: **Phytozome: a comparative platform for green plant genomics**. *Nucleic Acids Res* 2011, **40**:D1178–D1186, https://doi.org/10.1093/nar/gkr944.

16. Wei H, Wang X, Zhang Z, Yang L, Zhang Q, Li Y, He H, Chen D, Zhang B, Zheng C, *et al.*: **Uncovering key salt-tolerant regulators through a combined eQTL and GWAS analysis using the super pan-genome in rice**. *Natl Sci Rev* 2024, **11**, nwae043, https://doi.org/10.1093/nsr/nwae043.

17. Cui X, Hu M, Yao S, Zhang Y, Tang M, Liu L, Cheng X, Tong C, Liu S: **BnaOmics: a comprehensive platform combining pan-genome and multi-omics data from Brassica napus**. *Plant Commun* 2023, **4**, 100609, https://doi.org/10.1016/j.xplc.2023.100609.

18. Chougule K, Tello-Ruiz MK, Wei S, Olson A, Lu Z, Kumari S, Kumar V, Contreras-Moreira B, Naamati G, Dyer S, *et al.*: *Pan genome resources for grapevine*. Leuven, Belgium: International Society for Horticultural Science (ISHS); 2024:257–266, https://doi.org/10.17660/ActaHortic.2024.1390.31.

19. Naithani S, Deng CH, Sahu SK, Jaiswal P: **Exploring pan-genomes: an overview of resources and tools for unraveling structure, function, and evolution of crop genes and genomes**. In *Biomolecules*; 2023, https://doi.org/10.3390/biom13091403. vol. 13.

20. Varshney RK, Nayak SN, May GD, Jackson SA: **Next-generation sequencing technologies and their implications for crop genetics and breeding**. *Trends Biotechnol* 2009, **27**:522–530, https://doi.org/10.1016/j.tibtech.2009.05.006.

21. Zabala G, Vodkin LO: **A rearrangement resulting in small tandem repeats in the F3′5′H gene of white flower genotypes is associated with the soybean W1 locus**. *Crop Sci* 2007, **47**: 113–124, https://doi.org/10.2135/cropsci2006.12.0838tpg.

22. Feltus FA, Breen 3rd JR, Deng J, Izard RS, Konger CA, Ligon 3rd WB, Preuss D, Wang KC: **The widening Gulf between genomics data generation and consumption: a practical guide to big data transfer technology**. *Bioinf Biol Insights* 2015, **9**:9–19, https://doi.org/10.4137/BBI.S28988.

23. Lee YG, Jeong N, Kim JH, Lee K, Kim KH, Pirani A, Ha BK, Kang ST, Park BS, Moon JK, *et al.*: **Development, validation and genetic analysis of a large soybean SNP genotyping array**. *Plant J* 2015, **81**:625–636, https://doi.org/10.1111/tpj.12755.

24. Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB: **Development and evaluation of SoySNP50K, a high-density genotyping array for soybean**. *PLoS One* 2013, **8**, e54985, https://doi.org/10.1371/journal.pone.0054985.

25. Mahmood A, Bilyeu KD, Škrabišová M, Biová J, De Meyer EJ, Meinhardt CG, Usovsky M, Song Q, Lorenz AJ, Mitchum MG, *et al.*: **Cataloging SCN resistance loci in North American public soybean breeding programs**. *Front Plant Sci* 2023, **14**, https://doi.org/10.3389/fpls.2023.1270546.

26. ** Skrabisova M, Dietz N, Zeng S, Chan YO, Wang J, Liu Y, Biova J, Joshi T, Bilyeu KD: **A novel Synthetic phenotype association study approach reveals the landscape of association for genomic variants and phenotypes**. *J Adv Res* 2022, **42**: 117–133, https://doi.org/10.1016/j.jare.2022.04.004.
This work introduces three novel concepts for the improvement of GWAS-based discoveries. The authors demonstrate, that Accuracy testing of variant positions in GWAS genotype deflates false positives. Thus, Accuracy leads to the selection of more accurate markers and serves as a post-GWAS evaluation criterion in CMs identification.

27. ** Anilkumar C, Muhammed Azharudheen TP, Sah RP, Sunitha NC, Devanna BN, Marndi BC, Patra BC: **Gene based markers improve precision of genome-wide association studies and accuracy of genomic predictions in rice breeding**. *Heredity* 2023, **130**:335–345, https://doi.org/10.1038/s41437-023-00599-5.
This article evaluates the effectiveness of GWAS significant and gene-based markers in GS of rice. The study compares regression and machine learning-based models in GWAS and demonstrates that gene-based markers possess improved prediction accuracy and concludes that candidate gene-based markers are more effective in GS.

28. Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D: **Plant pan-genomes are the new reference**. *Nat Plants* 2020, **6**: 914–920, https://doi.org/10.1038/s41477-020-0733-0.

29. Li W, Liu J, Zhang H, Liu Z, Wang Y, Xing L, He Q, Du H: **Plant pan-genomics: recent advances, new challenges, and roads ahead**. *J Genet Genomics* 2022, **49**:833–846, https://doi.org/10.1016/j.jgg.2022.06.004.

30. Chen Z, Boehnke M, Fuchsberger C: **Combining sequence data from multiple studies: impact of analysis strategies on rare variant calling and association results**. *Genet Epidemiol* 2020, **44**:41–51, https://doi.org/10.1002/gepi.22261.

31. ** Chan YO, Dietz N, Zeng S, Wang J, Flint-Garcia S, Salazar-Vidal MN, Skrabisova M, Bilyeu K, Joshi T: **The Allele Catalog Tool: a web-based interactive tool for allele discovery and analysis**. *BMC Genom* 2023, **24**:107, https://doi.org/10.1186/s12864-023-09161-3.
This work delineates a pipeline for the aggregation of raw sequenced reads from independent studies into a single diversity panel, the variant calling pipeline (SnakyVC) that is species independent. Further, it describes the utilization of the diversity panel in allele discovery and analysis by a newly developed tool, the Allele Catalog Tool.

32. Kim J, Karyadi DM, Hartley SW, Zhu B, Wang M, Wu D, Song L, Armstrong GT, Bhatia S, Robison LL, *et al.*: **Inflated expectations: rare-variant association analysis using public controls**. *PLoS One* 2023, **18**, e0280951, https://doi.org/10.1371/journal.pone.0280951.

33. Yu K, Das S, LeFaive J, Kwong A, Pleiness J, Forer L, Schonherr S, Fuchsberger C, Smith AV, Abecasis GR: **Meta-imputation: an efficient method to combine genotype data after imputation with multiple reference panels**. *Am J Hum Genet* 2022, **109**:1007–1015, https://doi.org/10.1016/j.ajhg.2022.04.002.

34. * Cheng S, Feng C, Wingen LU, Cheng H, Riche AB, Jiang M, Leverington-Waite M, Huang Z, Collier S, Orford S, *et al.*: **Harnessing landrace diversity empowers wheat breeding**. *Nature* 2024, https://doi.org/10.1038/s41586-024-07682-9.
Describes systematic utilization of genetic diversity in crop improvement for the introduction of new beneficial alleles into breeding programs.

35. Mahood EH, Kruse LH, Moghe GD: **Machine learning: a powerful tool for gene function prediction in plants**.

Applications in Plant Sciences 2020, **8**, e11376, https://doi.org/10.1002/aps3.11376.

36. Greener JG, Kandathil SM, Moffat L, Jones DT: **A guide to machine learning for biologists**. *Nat Rev Mol Cell Biol* 2022, **23**: 40–55, https://doi.org/10.1038/s41580-021-00407-0.

37. Yan J, Wang X: **Machine learning bridges omics sciences and plant breeding**. *Trends Plant Sci* 2023, **28**:199–210, https://doi.org/10.1016/j.tplants.2022.08.018.

38. Ferguson JN, Fernandes SB, Monier B, Miller ND, Allen D, Dmitrieva A, Schmuker P, Lozano R, Valluru R, Buckler ES, *et al.*: **Machine learning-enabled phenotyping for GWAS and TWAS of WUE traits in 869 field-grown sorghum accessions**. *Plant Physiol* 2021, **187**:1481–1500, https://doi.org/10.1093/plphys/kiab346.

39. Sun J, Wu Q, Shen D, Wen Y, Liu F, Gao Y, Ding J, Zhang J: **TSLRF: two-stage algorithm based on least angle regression and random forest in genome-wide association studies**. *Sci Rep* 2019, **9**, 18034, https://doi.org/10.1038/s41598-019-54519-x.

40. Liu S, Xu F, Xu Y, Wang Q, Yan J, Wang J, Wang X, Wang X: **MODAS: exploring maize germplasm with multi-omics data association studies**. *Sci Bull* 2022, **67**:903–906, https://doi.org/10.1016/j.scib.2022.01.021.

41. Falk KG, Jubery TZ, Mirnezami SV, Parmley KA, Sarkar S, Singh A, Ganapathysubramanian B, Singh AK: **Computer vision and machine learning enabled soybean root phenotyping pipeline**. *Plant Methods* 2020, **16**:5, https://doi.org/10.1186/s13007-019-0550-5.

42. Yano K, Morinaka Y, Wang F, Huang P, Takehara S, Hirai T, Ito A, Koketsu E, Kawamura M, Kotake K, *et al.*: **GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture**. *Proc Natl Acad Sci U S A* 2019, **116**: 21262–21267, https://doi.org/10.1073/pnas.1904964116.

43. Wang W, Guo W, Le L, Yu J, Wu Y, Li D, Wang Y, Wang H, Lu X, Qiao H, *et al.*: **Integration of high-throughput phenotyping, GWAS, and predictive models reveals the genetic architecture of plant height in maize**. *Mol Plant* 2023, **16**:354–373, https://doi.org/10.1016/j.molp.2022.11.016.

44. Lin F, Fan J, Rhee SY: **QTG-finder: a machine-learning based algorithm to prioritize causal genes of quantitative trait loci in Arabidopsis and rice**. *G3 (Bethesda)* 2019, **9**:3129–3138, https://doi.org/10.1534/g3.119.400319.

45. Jeong S, Kim JY, Kim N: **GMStool: GWAS-based marker selection tool for genomic prediction from genomic data**. *Sci Rep* 2020, **10**, 19653, https://doi.org/10.1038/s41598-020-76759-y.

46. Zhao Y, Zhu H, Lu Z, Knickmeyer RC, Zou F: **Structured genome-wide association studies with Bayesian hierarchical variable selection**. *Genetics* 2019, **212**:397–415, https://doi.org/10.1534/genetics.119.301906.

47. Shen Z, Shen E, Yang K, Fan Z, Zhu QH, Fan L, Ye CY: **BreedingAIDB: a database integrating crop genome-to-phenotype paired data with machine learning tools applicable to breeding**. *Plant Commun* 2024, 100894, https://doi.org/10.1016/j.xplc.2024.100894.

48. Feng J-W, Han L, Liu H, Xie W-Z, Liu H, Li L, Chen L-L: **MaizeNetome: a multi-omics network database for functional genomics in maize**. *Mol Plant* 2023, **16**:1229–1231, https://doi.org/10.1016/j.molp.2023.08.002.

49. Peng H, Wang K, Chen Z, Cao Y, Gao Q, Li Y, Li X, Lu H, Du H, Lu M, *et al.*: **MBKbase for rice: an integrated omics knowledgebase for molecular breeding in rice**. *Nucleic Acids Res* 2020, **48**:D1085–D1092, https://doi.org/10.1093/nar/gkz921.

50. Sun M, Yan H, Zhang A, Jin Y, Lin C, Luo L, Wu B, Fan Y, Tian S, Cao X, *et al.*: **Milletdb: a multi-omics database to accelerate the research of functional genomics and molecular breeding of millets**. *Plant Biotechnol J* 2023, **21**:2348–2357, https://doi.org/10.1111/pbi.14136.

51. Zeng S, Škrabišová M, Lyu Z, Chan YO, Bilyeu KD, Joshi T: **SNPViz v2.0: a web-based tool for enhanced haplotype analysis using large scale resequencing datasets and discovery of phenotypes causative gene using allelic variations**. In *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)*; 2020:1408–1415, https://doi.org/10.1109/BIBM49941.2020.9313539.

52. Zeng S, Škrabišová M, Lyu Z, Chan YO, Dietz N, Bilyeu K, Joshi T: **Application of SNPViz v2.0 using next-generation sequencing data sets in the discovery of potential causative mutations in candidate genes associated with phenotypes**. *Int J Data Min Bioinf* 2021, **25**:65, https://doi.org/10.1504/IJDMB.2021.116886. -65.

53. Chan YO, Biova J, Mahmood A, Dietz N, Bilyeu K, Skrabisova M, Joshi T: **Genomic Variations Explorer (GenVarX): a toolset for annotating promoter and CNV regions using genotypic and phenotypic differences**. *Front Genet* 2023, **14**, 1251382, https://doi.org/10.3389/fgene.2023.1251382.

54. Biová J, Dietz N, Chan YO, Joshi T, Bilyeu K, Škrabišová M: **AccuCalc: a Python package for accuracy calculation in GWAS**. *Genes* 2023, **14**, https://doi.org/10.3390/genes14010123.

55. Joshi T, Patil K, Fitzpatrick MR, Franklin LD, Yao Q, Cook JR, Wang Z, Libault M, Brechenmacher L, Valliyodan B, *et al.*: **Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics**. *BMC Genom* 2012, **13**:S15, https://doi.org/10.1186/1471-2164-13-s1-s15.

56. Joshi T, Wang J, Zhang H, Chen S, Zeng S, Xu B, Xu D: **The evolution of soybean knowledge base (SoyKB)**. In *Plant genomics databases: methods and protocols*. Edited by Dijk ADJv, New York: Springer; 2017:149–159, https://doi.org/10.1007/978-1-4939-6658-5_7.

57. Zeng S, Lyu Z, Narisetti SRK, Xu D, Joshi T: **Knowledge Base Commons (KBCommons) v1.1: a universal framework for multi-omics data integration and biological discoveries**. *BMC Genom* 2019, **20**:947, https://doi.org/10.1186/s12864-019-6287-8.

58. Zeng S, Lyu Z, Narisetti SRK, Xu D, Joshi T: **Knowledge Base Commons (KBCommons) v1.0: a multi OMICS' web-based data integration framework for biological discoveries**. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*; 2018:589–594, https://doi.org/10.1109/BIBM.2018.8621369. 3-6 Dec. 2018.

59. * Shrestha AMS, Gonzales MEM, Ong PCL, Larmande P, Lee HS, Jeung JU, Kohli A, Chebotarov D, Mauleon RP, Lee JS, *et al.*: **RicePilaf: a post-GWAS/QTL dashboard to integrate pangenomic, coexpression, regulatory, epigenomic, ontology, pathway, and text-mining information to provide functional insights into rice QTLs and GWAS loci**. *GigaScience* 2024, **13**, giae013, https://doi.org/10.1093/gigascience/giae013.
This work describes a web app for post-GWAS/QTL analysis that integrates multiple types of data (pangenome, omics, epigenomics, regulatory) for rice from publicly available databases into the results.

60. Uffelmann E, Posthuma D: **Emerging methods and resources for biological interrogation of neuropsychiatric polygenic signal**. *Biol Psychiatr* 2021, **89**:41–53, https://doi.org/10.1016/j.biopsych.2020.05.022.

61. * Battram T, Gaunt TR, Relton CL, Timpson NJ, Hemani G: **A comparison of the genes and genesets identified by GWAS and EWAS of fifteen complex traits**. *Nat Commun* 2022, **13**: 7816, https://doi.org/10.1038/s41467-022-35037-3.
Epigenetics data-based genome-wide association study (EWAS) captures different aspects of the biology of complex traits

62. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, Benner C, O'Dushlaine C, Barber M, Boutkov B, *et al.*: **Computationally efficient whole-genome regression for quantitative and binary traits**. *Nat Genet* 2021, **53**:1097–1103, https://doi.org/10.1038/s41588-021-00870-7.

63. Zeng S, Mao Z, Ren Y, Wang D, Xu D, Joshi T: **G2PDeep: a web-based deep-learning framework for quantitative phenotype prediction and discovery of genomic markers**. *Nucleic Acids Res* 2021, **49**:W228–W236, https://doi.org/10.1093/NAR/GKAB407.

64. Reiser L, Harper L, Freeling M, Han B, Luan S: **FAIR: a call to make published data more findable, accessible, interoperable, and reusable**. *Mol Plant* 2018, **11**:1105–1108, https://doi.org/10.1016/j.molp.2018.07.005.

65. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship**. *Sci Data* 2016, **3**, 160018, https://doi.org/10.1038/sdata.2016.18.

66. * Deng CH, Naithani S, Kumari S, Cobo-Simon I, Quezada-Rodriguez EH, Skrabisova M, Gladman N, Correll MJ, Sikiru AB, Afuwape OO, *et al.*: **Genotype and phenotype data standardization, utilization and integration in the big data era for agricultural sciences**. *Database* 2023, **2023**, baad088, https://doi.org/10.1093/database/baad088.
This review identifies the current limitations of FAIR principles in the big data in agriculture.

67. Xu H, Guo Y, Qiu L, Ran Y: **Progress in soybean genetic transformation over the last decade**. *Front Plant Sci* 2022, **13**, 900318, https://doi.org/10.3389/fpls.2022.900318.

68. Sheoran S, Jaiswal S, Raghav N, Sharma R, Sabhyata Gaur A, Jaisri J, Tandon G, Singh S, Sharma P, *et al.*: **Genome-wide association study and post-genome-wide association study analysis for spike fertility and yield related traits in bread wheat**. *Front Plant Sci* 2021, **12**, 820761, https://doi.org/10.3389/fpls.2021.820761.

69. Shen Y, Zhou G, Liang C, Tian Z: **Omics-based interdisciplinarity is accelerating plant breeding**. *Curr Opin Plant Biol* 2022, **66**, 102167, https://doi.org/10.1016/j.pbi.2021.102167.

70. Liang T, Hu Y, Xi N, Zhang M, Zou C, Ge F, Yuan G, Gao S, Zhang S, Pan G, *et al.*: **GWAS across multiple environments and WGCNA suggest the involvement of ZmARF23 in embryonic callus induction from immature maize embryos**. *Theor Appl Genet* 2023, **136**:93, https://doi.org/10.1007/s00122-023-04341-x.

71. Qin C, Li YH, Li D, Zhang X, Kong L, Zhou Y, Lyu X, Ji R, Wei X, Cheng Q, *et al.*: **PH13 improves soybean shade traits and enhances yield for high-density planting at high latitudes**. *Nat Commun* 2023, **14**:6813, https://doi.org/10.1038/s41467-023-42608-5.

72. Xia EH, Tong W, Wu Q, Wei S, Zhao J, Zhang ZZ, Wei CL, Wan XC: **Tea plant genomics: achievements, challenges and perspectives**. *Hortic Res* 2020, **7**:7, https://doi.org/10.1038/s41438-019-0225-4.

73. Schneider HM, Lor VS, Zhang X, Saengwilai P, Hanlon MT, Klein SP, Davis JL, Borkar AN, Depew CL, Bennett MJ, *et al.*: **Transcription factor bHLH121 regulates root cortical aerenchyma formation in maize**. *Proc Natl Acad Sci U S A* 2023, **120**, e2219668120, https://doi.org/10.1073/pnas.2219668120.

74. Yin P, Fu X, Feng H, Yang Y, Xu J, Zhang X, Wang M, Ji S, Zhao B, Fang H, *et al.*: **Linkage and association mapping in multi-parental populations reveal the genetic basis of carotenoid variation in maize kernels**. *Plant Biotechnol J* 2024, https://doi.org/10.1111/pbi.14346. n/a.

75. Budeguer F, Enrique R, Perera MF, Racedo J, Castagnaro AP, Noguera AS, Welin B: **Genetic transformation of sugarcane, current status and future prospects**. *Front Plant Sci* 2021, **12**, 768609, https://doi.org/10.3389/fpls.2021.768609.

76. Cardi T, Murovec J, Bakhsh A, Boniecka J, Bruegmann T, Bull SE, Eeckhaut T, Fladung M, Galovic V, Linkiewicz A, *et al.*: **CRISPR/Cas-mediated plant genome editing: outstanding challenges a decade after implementation**. *Trends Plant Sci* 2023, **28**:1144–1165, https://doi.org/10.1016/j.tplants.2023.05.012.

77. Choudhury A, Rajam MV: **Genetic transformation of legumes: an update**. *Plant Cell Rep* 2021, **40**:1813–1830, https://doi.org/10.1007/s00299-021-02749-7.