Review Article

Haoyu Chao, Shilong Zhang, Yueming Hu, Qingyang Ni, Saige Xin, Liang Zhao, Vladimir A. Ivanisenko, Yuriy L. Orlov and Ming Chen*

Integrating omics databases for enhanced crop breeding

https://doi.org/10.1515/jib-2023-0012
Received April 27, 2023; accepted June 12, 2023; published online July 25, 2023

Abstract: Crop plant breeding involves selecting and developing new plant varieties with desirable traits such as increased yield, improved disease resistance, and enhanced nutritional value. With the development of high-throughput technologies, such as genomics, transcriptomics, and metabolomics, crop breeding has entered a new era. However, to effectively use these technologies, integration of multi-omics data from different databases is required. Integration of omics data provides a comprehensive understanding of the biological processes underlying plant traits and their interactions. This review highlights the importance of integrating omics databases in crop plant breeding, discusses available omics data and databases, describes integration challenges, and highlights recent developments and potential benefits. Taken together, the integration of omics databases is a critical step towards enhancing crop plant breeding and improving global food security.

Keywords: crop plant breeding; data integration; databases; omics; plant biology

1 Introduction

Crop plant breeding is a complex and challenging process that requires the identification and selection of desirable traits such as increased yield [1], improved disease resistance [2], and enhanced nutritional value [3]. Over the years, traditional breeding methods have been used to develop new plant varieties by crossing plants with desirable traits to produce offspring with improved traits [4–6]. However, these methods are time-consuming and often limited by the genetic diversity of available plant species. In recent years, the emergence of

Haoyu Chao, Shilong Zhang, Yueming Hu, and Qingyang Ni with equal contribution jointly sharing the first author position.

***Corresponding author: Ming Chen**, Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou 310058, China, E-mail: mchen@zju.edu.cn
Haoyu Chao, **Shilong Zhang**, **Yueming Hu**, **Qingyang Ni**, **Saige Xin and Liang Zhao**, Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou 310058, China
Vladimir A. Ivanisenko, Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia
Yuriy L. Orlov, Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia; Agrarian and Technological Institute, Peoples' Friendship University of Russia, Moscow 117198, Russia; and The Digital Health Institute, I.M. Sechenov First Moscow State Medical University of the Russian Ministry of Health (Sechenov University), Moscow 119991, Russia. https://orcid.org/0000-0003-0587-1609

high-throughput omics technologies has revolutionized crop plant breeding by providing vast amounts of data on the molecular mechanisms underlying plant development [7], and responses to environmental stresses [8]. Genomics is essential in crop breeding, allowing the identification of important genetic traits and accelerating the development of improved varieties. The number of sequenced crop genomes has continued to rapidly grow in recent years (Figure 1A), providing valuable resources for agricultural research. Additionally, epigenomics and transcriptomics have become increasingly important in crop breeding, providing insights into gene regulation and aiding in the identification of desirable traits [9, 10]. The SRA database has seen a continuous increase in epigenomic and transcriptomic data, further emphasizing the significance of these fields for crop breeding (Figure 1B). Proteomics and metabolomics have continued to develop in crop breeding, allowing for a deeper understanding of plant molecular mechanisms [11, 12].

These technologies have enabled the identification of key genes and pathways involved in crop traits, allowing breeders to select and develop new plant varieties with desirable traits more efficiently [4–6]. In recent years, there has been a significant increase in literature focused on the application of omics technologies in crop breeding (Figure 1C), highlighting the growing importance of these approaches in agricultural research. However, the effective use of omics technologies in crop plant breeding requires the integration of diverse datasets from different databases. Integration of omics data is crucial in providing a comprehensive understanding of the biological processes underlying plant traits and their interactions. In recent years, several omics databases have been developed to store and analyze large-scale omics data for different crop species, including rice (Table 1), maize [13], wheat [14], and soybean [15]. These databases provide a wealth of information on the genetic makeup, epigenome regulation, gene expression profiles, protein functions, and metabolic pathways of crops, which can be used to improve breeding programs.

The integration of omics databases can provide several benefits to crop plant breeding. Firstly, it can help to identify novel gene targets that are associated with desirable traits [30]. This can be achieved by integrating genomic, epigenomic, transcriptomic, proteomic, and metabolomic data to identify genes that are differentially expressed or are involved in key metabolic pathways. Secondly, it can help to develop predictive models for crop performance by integrating different omics data and environmental factors. These models can be used to predict the performance of new plant varieties under different environmental conditions and select the best performing varieties for further development [31]. Furthermore, the integration of omics databases can accelerate breeding cycles by providing breeders with a better understanding of the molecular mechanisms underlying crop traits [32]. This can help to reduce the time and cost required to develop new plant varieties with desirable traits. Finally, it can help to improve global food security by providing breeders with the tools and resources needed to develop new crop varieties that are more resilient to environmental stresses and can produce higher yields [33].

In this review, we aim to highlight the importance of integrating omics databases in crop plant breeding and discuss the current state of integration efforts. We will begin by discussing the different types of omics data available for crop plants, including genomic, epigenomic, transcriptomic, proteomic, and metabolomic data. We will then review the different databases that host these omics data and describe their features, strengths, and limitations. Next, we will discuss the challenges associated with integrating omics databases, such as data heterogeneity, scalability, and interoperability. Then, we will highlight some of the recent developments in omics data integration in crop plant breeding and the potential benefits of these efforts. Finally, we will discuss the use of machine learning algorithms and network analysis tools to integrate omics data and identify key genes and pathways associated with desirable traits. Overall, the integration of omics databases is a critical step towards enhancing crop plant breeding and improving global food security. The use of omics technologies and databases can provide breeders with the tools and resources needed to develop new crop varieties with desirable traits more efficiently and sustainably. The integration of omics databases is a rapidly evolving field, and future developments in this area are expected to further enhance our ability to develop crops that are more productive, resilient, and sustainable.

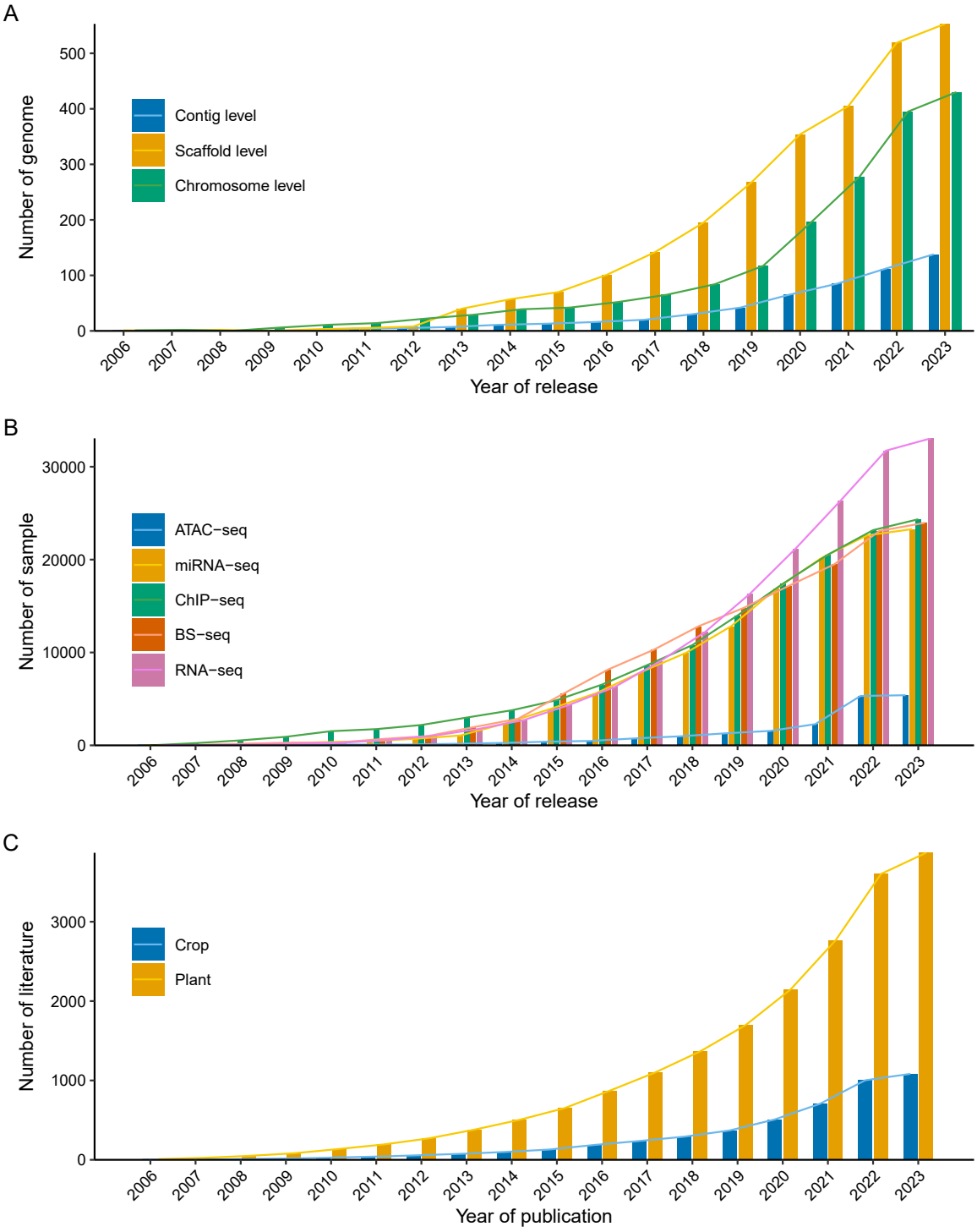

Figure 1: Statistics of land plant omics data and literature. (A) The number of completed genome assemblies for land plants. The data was downloaded from NCBI GENOME REPORTS, which filtered all genomes with the “land plants” tag and a genome size of less than 100 Mb. (B) The amount of epigenomic and transcriptomic data generated from land plants. The number of each omics sample was searched in the NCBI SRA database using a query such as “(((land plants [organism]) AND 2023)) AND RNA-seq [strategy]”. The sample size of “RNA-Seq” is the result after being reduced by 20 times. (C) The number of literatures on land plant omics research. The literature count was searched in the PubMed database by “omics plant” or “omics crop”.

Table 1: Statistics of rice omics database.

Database name	Omics type	Number of accessions	Released year	Link	Ref.
RAP-DB	Genome	1	2006	https://rapdb.dna.affrc.go.jp	[16]
BGI-RIS	Genome	1	2007	http://rice.genomics.org.cn	[17]
RGAP	Genome	1	2005	http://rice.uga.edu	[18]
RIGW	Genome	2	2020	http://rice.hzau.edu.cn/rice_rs3	[19]
RGI	Genome	16	2023	https://riceome.hzau.edu.cn	[20]
RPAN	Genome	3010	2016	https://cgm.sjtu.edu.cn/3kricedb	[21]
RiceENCODE	Epigenome	972	2021	http://glab.hzau.edu.cn/RiceENCODE	[22]
RiceXPro	Transcriptome	194	2011	https://ricexpro.dna.affrc.go.jp	[23]
RED	Transcriptome	284	2017	http://expression.ic4r.org	[24]
PPRD	Transcriptome	11,726	2022	https://plantrnadb.com/ricerna	[25]
RPMD	Proteome	38	2004	http://www.info.chi-biotech.cc	[26]
RKD	Proteome	1429	2007	https://ricephylogenomics.ucdavis.edu	[27]
MCDRP	Proteome	2400	2013	http://www.genomeindia.org/biocuration	[28]
RiceCyc	Metabolome	316	2013	http://pathway.gramene.org/gramene	[29]

2 Omics data and databases for crop plants

Crop plants are complex organisms that have undergone natural selection and human domestication. Omics technologies provide a powerful tool for investigating the genetic and molecular mechanisms underlying plant growth, development, and responses to environmental stresses [30, 31]. With the decreasing cost of high-throughput sequencing, an increasing amount of molecular information on crops is being obtained. This has led to the rapid establishment of large public databases for the sharing of bioinformatics data in various countries, such as the National Genomics Data Center (NGDC) [34], the National Center for Biotechnology Information (NCBI) [35], the DNA Data Bank of Japan (DDBJ) [36], and the European Bioinformatics Institute (EBI) [37] (Figure 2A).

Through omics data, crop shape can be improved, including increasing yield, enhancing the root system's nutrient uptake ability, improving plant adaptability to the environment, resistance to adversity, and flavor, among other things (Figure 2B). With the expansion of biological big data, more and more secondary databases have been established to better integrate and analyze multi-omics data, thereby explaining the molecular mechanisms of crops at different levels (Figure 2C). In this article, we will first introduce the five main types of omics data commonly used in crop plant research: genomic, epigenomic, transcriptomic, proteomic, and metabolomic data.

Genomic information has proven to be an invaluable tool for crop improvement [38]. The identification of genes responsible for desirable traits such as resistance to diseases or high yield is facilitated by genomic data. The use of genomic databases such as NCBI Assembly [39], Genome Warehouse [40], EnsemblPlants [41], Phytozome [42], and PlantGDB [43] provides access to genome sequences, gene annotations, and functional annotations for many crop species, including rice, maize, soybean, wheat, and so on (Table 2). This information can help researchers develop molecular markers and breeding programs that produce improved crop varieties with enhanced characteristics. Furthermore, genomics can aid in understanding the evolution and domestication of crops, which can have implications for their conservation and management. Therefore, genomics is the cornerstone of omics research. Overall, genomics has the potential to transform agriculture by improving crop productivity, sustainability, and resilience, contributing to global food security.

Epigenomic data can be integrated with other omics data to gain a more comprehensive understanding of the underlying biological processes. For example, integrating epigenomic data with transcriptomic data can provide insights into how changes in chromatin structure affect gene expression [57]. This can help identify key

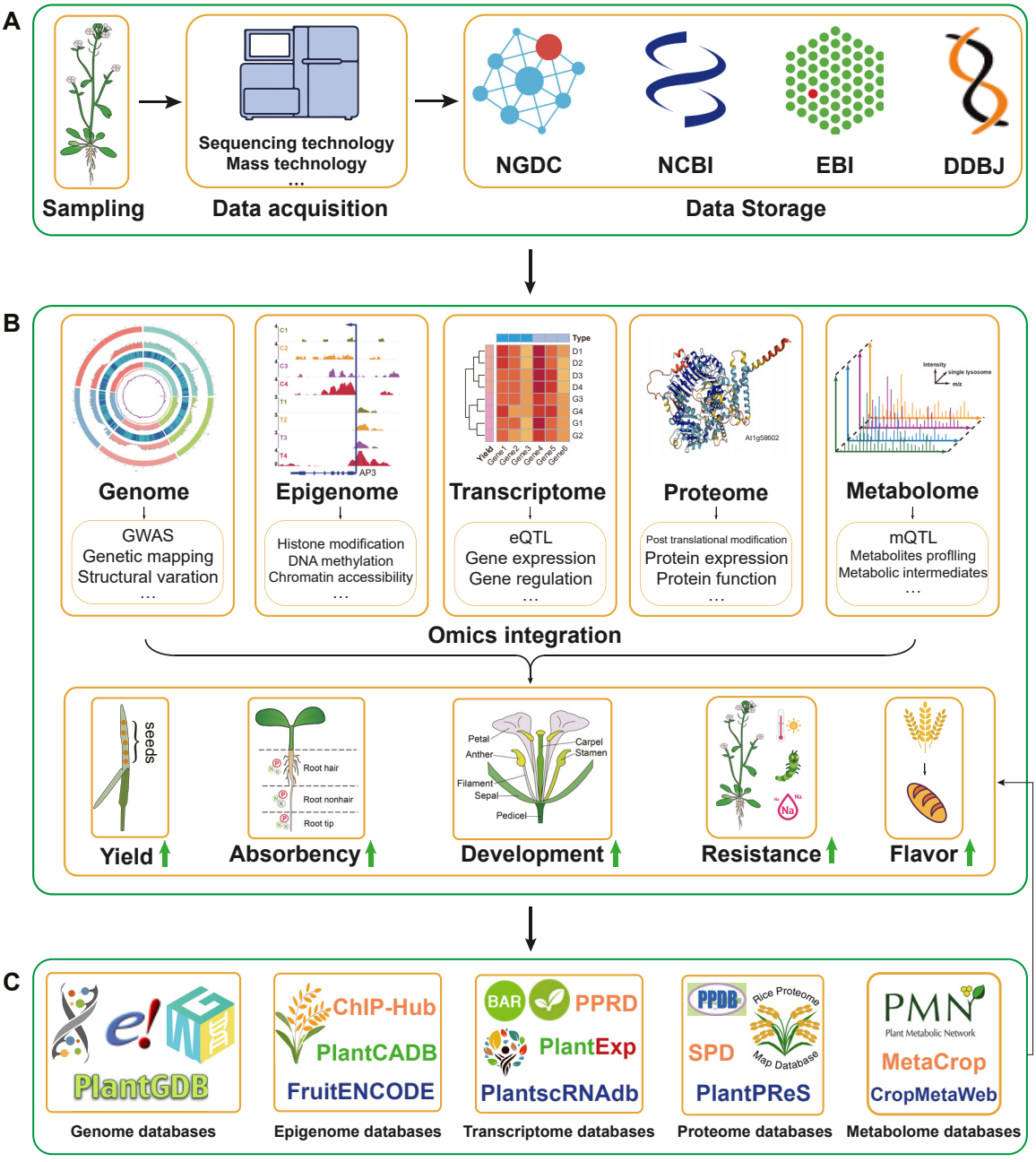

Figure 2: Generation, storage, mining, and integration of crop omics data. (A) Generation and storage of omics data in public databases. (B) Applications of five omics technologies (genomic, epigenomic, transcriptomic, proteomic, and metabolomic data) in crop breeding. (C) Construction of a secondary database based on mining of multi-omics data.

regulatory genes and pathways that can be targeted in crop breeding programs. Despite its potential, integrating epigenomic data poses unique challenges due to the complex nature of epigenetic modifications and the difficulty in accurately measuring them. However, recent advancements in high-throughput epigenomic technologies, such as ATAC-Seq [58], ChIP-seq [59] and BS-Seq [60], have made it possible to generate large amounts of epigenomic data in a cost-effective and efficient manner. Currently, there are Encyclopedia of DNA Elements (ENCODE) projects for human and mouse, but as yet there is no well-defined project for plants. In 2014, the

Table 2: Statistics of comprehensive omics databases about plants.

Database name	Omics type	Number of accessions	Released year	Link	Ref.
Assembly	Genome	2662	2016	https://www.ncbi.nlm.nih.gov/assembly	[39]
GWH	Genome	1358	2021	https://ngdc.cncb.ac.cn/gwh	[40]
Phytozome	Genome	312	2012	https://phytozome-next.jgi.doe.gov	[42]
PlantGDB	Genome	187	2004	http://plantgdb.org	[43]
EnsemblPlants	Genome	134	2002	https://plants.ensembl.org/index.html	[41]
ChIP-Hub	Epigenome	>10,000	2022	https://biobigdata.nju.edu.cn/ChIPHub	[44]
PlantCADB	Epigenome	649	2022	https://bioinfor.nefu.edu.cn/PlantCADB	[45]
PlantExp	Transcriptome	131,423	2023	https://biotec.njau.edu.cn/plantExp	[46]
PPRD	Transcriptome	∼45,000	2022	http://ipf.sustech.edu.cn/pub/plantrna	[25]
Genevestigator	Transcriptome	>250,000	2006	https://genevestigator.com	[47]
ePlant	Transcriptome	>10,000	2005	http://bar.utoronto.ca	[48]
PlantGenIE	Transcriptome	35,533	2015	https://plantgenie.org	[49]
PsctH	Transcriptome	20	2021	http://jinlab.hzau.edu.cn/PsctH	[50]
PlantscRNAdb	Transcriptome	31	2021	http://ibi.zju.edu.cn/plantscrnadb	[51]
PCMDB	Transcriptome	22	2022	http://www.tobaccodb.org/pcmdb	[52]
PPDB	Proteome	>5000	2004	http://ppdb.tc.cornell.edu	[53]
PlantPReS	Proteome	>20,000	2016	http://www.proteome.ir/	[54]
PMN	Metabolome	9129	2021	https://plantcyc.org	[55]
MetaCrop	Metabolome	392	2008	https://metacrop.ipk-gatersleben.de	[56]

international plant science community launched the Plant ENCODE project [61]. Since then, with the efforts of plant researchers worldwide, several ENCODE databases for various plant species have been established (Table 2), including RiceENCODE, which provides an important platform for studying the epigenome, genetic mechanisms, tissue specificity of rice. Additionally, FruitENCODE [62] has obtained various functional genomic data for 11 fleshy fruits, laying the groundwork for understanding the molecular regulation of fruit ripening. Moreover, comprehensive plant regulome databases called ChIP-Hub [44] and PlantCADB [45] also has been constructed (Table 2). In conclusion, incorporating epigenomic data into omics-based approaches can further enhance crop plant breeding by providing a more comprehensive understanding of the biological processes underlying desirable traits. By integrating diverse omics datasets, researchers can identify key regulatory genes and pathways that can be targeted to develop new plant varieties with improved yield, disease resistance, and nutritional value.

Transcriptomic data is crucial in advancing crop breeding by providing crucial insights into gene expression patterns in different tissues and under varying conditions. In addition to identifying differentially expressed genes, transcriptomic data can help researchers to understand the complex regulatory networks that control gene expression, including the involvement of non-coding RNAs (ncRNAs) such as long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) [63]. These ncRNAs have emerged as important players in gene regulation and can significantly influence crop traits and responses to environmental stimuli. Besides, single-cell transcriptomic analysis is an emerging technology that allows researchers to study gene expression patterns at the level of individual cells [64]. This technology has revolutionized the field of transcriptomics, enabling researchers to identify rare cell types, map developmental trajectories, and uncover novel gene expression patterns that are masked in bulk transcriptomic analyses. By applying single-cell transcriptomics to crop plants, researchers can gain a more comprehensive understanding of gene expression patterns in different cell types and tissues, and the molecular mechanisms that govern crop growth and development. To access bulk transcriptomic data for crop plants, researchers can use established databases such as PlantExp [46], PPRD [25], Genevestigator [47], ePlant [48], and PlantGenIE [49], which provide a variety of transcriptomic data sets for different crop species (Table 2) Besides, with the widespread application of single-cell transcriptomic technology in plants, databases focused on plant single-cell transcriptomics, such as PsctH [50], PlantscRNAdb [51], and PCMDB [52] have been

established successively (Table 2). These databases are critical resources that enable researchers to explore gene expression patterns across different tissues and under varying conditions. The availability of transcriptomic data sets from different crop species and tissues has greatly facilitated the identification of candidate genes and pathways for crop improvement, thereby enabling the development of more productive and resilient crop varieties.

Proteomic data is a valuable tool for understanding the protein content and function of crop plants. By using proteomic data, researchers can identify proteins involved in critical metabolic pathways or associated with specific traits. The Plant Proteome Database (PPDB) [53], Plant stress proteome database (PlantPReS) [54], Rice Proteome Database (RPD) [65], Soybean Proteome Database (SPD) [66], are among the most commonly used proteomic databases for crop plants (Table 2), providing access to proteomic datasets for various crop species, including protein sequences, structures, and functional annotations. Proteomics plays a critical role in crop science research, as it provides researchers with a comprehensive view of the protein content of crop plants. This information can be used to improve crop yield and quality, increase stress tolerance, and develop new crop varieties with improved traits. For example, proteomic data has been used to identify proteins associated with abiotic stress responses, such as drought or salinity [67], and to identify proteins involved in plant-microbe interactions [68], such as those associated with disease resistance. In addition, proteomic data can be used to identify proteins associated with specific crop traits, such as those related to nutritional value or flavor. This information can be used to develop crops with enhanced nutritional content or improved flavor profiles [69], which can increase their value to consumers. Looking forward, proteomics will continue to play an important role in crop science research, as new technologies and methods are developed to analyze and interpret proteomic data. These advances will allow researchers to gain a more detailed understanding of the protein content and function of crop plants, which can be used to develop new crop varieties that are more resilient, productive, and sustainable.

Metabolomics is a powerful tool for investigating the genetic basis of metabolic variation, providing insights into the complex biochemical cascades that connect the genome, transcriptome, and proteome to phenotype [70]. By analyzing a wide range of sample types, including primary cells, tissues, biofluids, and entire organisms, metabolomics can determine the relative and absolute amounts of various metabolites, such as sugars, lipids, amino acids, and nucleotides. In crop science research, metabolomics is also an important tool that offers a comprehensive view of the metabolite content and function of crop plants. Using metabolomics data, researchers can identify metabolites that are involved in critical metabolic pathways or associated with specific traits in crops [71]. Several commonly used metabolomic databases for crop plants include Plant Metabolic Network (PMN) [55], and MetaCrop [56], a detailed database of crop plant metabolism (Table 2). Metabolomics not only aids in identifying individual metabolites but also contributes to the development of crops that have superior stress tolerance or nutritional content. One application of metabolomics is the detection of biomarkers linked to abiotic stress responses, which can then be utilized to cultivate crops that are more resistant to these conditions. Additionally, metabolomics can determine metabolic pathways that influence specific crop traits, such as those affecting nutritional value or flavor, and can be utilized to create crops with superior nutritional content or flavor profiles.

Different from the micro-level molecular omics, macro-level crop phenomics is the focal point of breeders. Therefore, in recent years, phenomics has also emerged as a field of study. Phenomics refers to a comprehensive and systematic approach to studying and describing the phenotypes of organisms, combining the terms “phenotype” and “omics”. Phenomics methods often utilize high-throughput techniques and large-scale data analysis to collect and analyze phenotype data. These techniques may include image analysis, genomics, transcriptomics, metabolomics, proteomics, and others, in order to obtain comprehensive information about an individual’s phenotype. By integrating phenotype data with genomic and environmental data, researchers can identify genes or environmental factors associated with specific phenotypic features, revealing the genetic basis and regulatory mechanisms underlying phenotypes.

In conclusion, the application of omics technologies and databases offers a powerful tool for investigating the intricate genetic and molecular mechanisms that underlie the growth, development, and responses of plants to environmental stresses. The diverse types of omics data complement each other in providing a comprehensive

view of the molecular processes involved in crop plants. However, it is important to consider the distinct features, strengths, and limitations of the different databases that host these omics data when selecting an appropriate database for a particular research question. A judicious selection of a database can facilitate the integration of omics data, and provide valuable insights into the biology of crop plants, leading to the development of new crop varieties with desirable traits at an accelerated pace.

3 Challenges of integrating omics from databases

Integrating omics data from different databases presents several challenges due to the heterogeneity of the data, differences in data formats, and varying levels of data quality. One of the primary challenges is the integration of data from different omics technologies, which often use different data formats and produce data with different levels of complexity (Figure 3). For example, genomic data typically consists of large, complex data sets, while proteomic data may contain information on thousands of individual proteins. Overcoming these challenges requires the development of standardized data formats and integration tools that can handle diverse data types.

Another challenge associated with integrating omics databases is scalability. As the number of omics data generated increases, it becomes increasingly difficult to store, process, and analyze the data. For example, a single genome sequence for a crop plant may require hundreds of gigabytes of storage, while a large-scale proteomic study may generate terabytes of data. Scalability can be addressed through the use of cloud-based storage and computing resources, as well as the development of efficient algorithms and data compression techniques.

Interoperability is another challenge that arises when integrating omics databases. Different databases may use different ontologies and vocabularies to describe the same biological concepts, making it difficult to integrate data from different sources. Furthermore, data may be stored in different formats or with different levels of annotation, making it difficult to compare and analyze the data. This is especially common in single-cell transcriptome studies. Interoperability can be improved through the use of common data standards and the development of ontology-based integration tools that can map data from different sources onto a common framework.

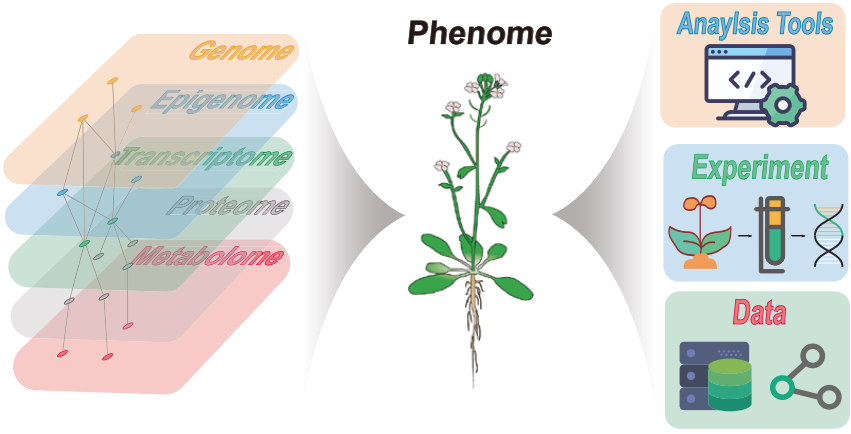

Figure 3: Challenges of integrating omics data. The left side represents the challenges of integrating omics data stored in various databases, where the dots and lines indicate the potential mutual regulation of different omics levels. The right side represents the challenges of data processing at different levels, including challenges in using bioinformatics analysis tools for data processing, challenges in obtaining experimental data, and challenges in data storage and sharing. The middle section represents the growth phenotype of plants.

In conclusion, integrating omics data from different databases presents several challenges that must be addressed in order to fully exploit the potential of these data for crop plant research. Overcoming these challenges requires the development of standardized data formats, scalable storage and computing resources, and ontology-based integration tools. With the development of these tools and resources, the integration of omics databases can provide valuable insights into the biology of crop plants and help to accelerate the development of new crop varieties with desirable traits.

4 Recent developments in omics data integration for crop plant breeding

Recent developments in omics data integration for crop plant breeding have shown promise in accelerating the development of new crop varieties with desirable traits [72, 73]. One approach is the use of machine learning algorithms to integrate data from different omics technologies and predict the performance of different crop varieties under different environmental conditions. Today, in the field of computer science, machine learning has produced numerous excellent algorithms and frameworks. Currently, machine learning methods applied in biological research can be roughly categorized into unsupervised, supervised, and reinforcement learning.

Unsupervised learning aims to extract latent data features or structures from unlabeled biological data. For example, unsupervised dimensionality reduction techniques such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) can be applied to crop sequencing and quantitative samples, along with clustering methods like K-means clustering and hierarchical clustering. These approaches can help us better understand the characteristics of the omics data. In contrast, supervised learning requires the use of labeled training data, where the input and output are known, to build models. These models are then used to predict and classify new inputs, using algorithms such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forests, Decision Trees, Naive Bayes, etc. Supervised learning can be applied, for example, to differentiate between good and bad genotypes based on molecular data. Finally, unlike unsupervised and supervised learning, reinforcement learning focuses more on iterative experimentation (trial and error) and delayed rewards. It continuously optimizes the correspondence between states and actions based on feedback (rewards) provided by the environment. One application of reinforcement learning is in protein structure prediction [74].

A recent study used machine learning algorithms to integrate genomic and phenotypic data and predict the performance of different varieties under drought conditions [75, 76]. The results showed that the algorithm was able to accurately predict the performance of different varieties, and identified several new candidate genes that may be involved in drought tolerance.

Another approach is the use of multi-omics data integration to identify key regulatory networks and pathways that underlie specific traits or responses to environmental stresses [77]. For example, a recent study used multi-omics data integration to identify key regulatory networks involved in salt stress tolerance in tomato [78]. The study integrated transcriptomic, proteomic, and metabolomic data and identified several key regulatory pathways involved in salt stress tolerance, including the production of osmoprotectants and the regulation of ion transport. Overall, The potential benefits of omics data integration for crop plant breeding are numerous [79]. By integrating data from different omics technologies, researchers can gain a more comprehensive understanding of the molecular processes underlying crop growth, development, and responses to environmental stresses [80]. The functional annotation of the plant gene and gene-phenotype association could be found by new text mining tools [81, 82]. This knowledge can be used to develop new crop varieties with improved yields, disease resistance, and stress tolerance, as well as to identify new targets for crop improvement. In addition, the integration of omics data can help to accelerate the breeding process by reducing the time and resources required for traditional breeding methods.

5 Discussion

The integration of omics data in crop plant breeding has brought about tremendous advances in the development of new crop varieties that possess desirable traits such as increased yield, improved disease resistance, and enhanced nutritional value [1–3]. The integration of diverse omics datasets from different databases provides a comprehensive understanding of the underlying biological processes and interactions that influence plant traits. Machine learning algorithms and multi-omics data integration have emerged as powerful tools for analyzing and interpreting omics data, enabling the development of predictive models that can accelerate the breeding process and reduce the time and resources required for traditional breeding methods. These models can identify genetic factors underlying desirable traits and predict the performance of different varieties under specific environmental conditions, allowing for more efficient and effective selection and crossing of plants.

Moreover, the integration of omics data can be used to develop new crop varieties with improved yields, disease resistance, and stress tolerance, as well as to identify new targets for crop improvement. However, the integration of omics data poses several challenges, including the sheer volume of data generated by different omics technologies, the complexity of integrating multiple omics data sets, and the need for advanced computational tools and expertise to analyze and interpret these data. Additionally, the ethical and social considerations associated with the use of omics data in crop breeding cannot be ignored. The development of new crop varieties with desirable traits can have significant impacts on the environment, local communities, and the wider agricultural system [83]. Therefore, transparent and inclusive decision-making processes that involve all stakeholders, including farmers, consumers, and policymakers, are essential.

In conclusion, the integration of omics data in crop plant breeding is a critical step towards enhancing crop productivity and improving global food security. Addressing the challenges associated with the integration of omics data will require collaboration and coordination among researchers, policymakers, and stakeholders across the agricultural sector. The benefits of omics data integration in crop breeding are enormous, and it is essential that efforts are made to leverage these technologies for the betterment of humanity.

Acknowledgment: The authors are grateful to the members of Ming Chen's laboratory for helpful discussions and valuable comments.
Author contributions: M. C. and Y. O. conceived the idea for this review. H.C., S. Z., Y. H. and Q. N. performed the data analyses and manuscript writing. S.X., L.Z., and V.I. supervised the work and provided scientific advice.
Research funding: This work was supported by the National Natural Science Foundation of China [32070677; 32270709; 32261133526]; The 151 Talent Project, and S&T Innovation Leader of Zhejiang Province; Jiangsu Collaborative Innovation Center for Modern Crop Production and Collaborative Innovation Center for Modern Crop Production co-sponsored by province and ministry. Russian team was supported by RSF-NSFC Cooperation project [23-44-00030].
Conflict of interest statement: The authors declare no competing interests.

References

1. Zhu XG, Long SP, Ort DR. Improving photosynthetic efficiency for greater yield. Annu Rev Plant Biol 2010;61:235—61.
2. Nelson R, Wiesner-Hanks T, Wisser R, Balint-Kurti P. Navigating complexity to breed disease-resistant crops. Nat Rev Genet 2018;19:21—33.
3. Goicoechea N, Antolín MC. Increased nutritional value in food crops. Microb Biotechnol 2017;10:1004—7.
4. Hu B, Wang W, Ou S, Tang J, Li H, Che R, et al. Variation in NRT1.1B contributes to nitrate-use divergence between rice subspecies. Nat Genet 2015;47:834—8.
5. Zhang H, Yu F, Xie P, Sun S, Qiao X, Tang S, et al. A Gγ protein regulates alkaline sensitivity in crops. Science 2023;379:eade8416.
6. Zhai K, Liang D, Li H, Jiao F, Yan B, Liu J, et al. NLRs guard metabolism to coordinate pattern- and effector-triggered immunity. Nature 2022;601:245—51.

7. Wang P, Clark NM, Nolan TM, Song G, Bartz PM, Liao CY, et al. Integrated omics reveal novel functions and underlying mechanisms of the receptor kinase FERONIA in Arabidopsis thaliana. Plant Cell 2022;34:2594—614.
8. Zander M, Lewsey MG, Clark NM, Yin L, Bartlett A, Guzmán JPS, et al. Integrated multi-omics framework of the plant response to jasmonic acid. Nat Plants 2020;6:290—302.
9. Yang L, Zhang P, Wang Y, Hu G, Guo W, Gu X, et al. Plant synthetic epigenomic engineering for crop improvement. Sci China Life Sci 2022;65:2191—204.
10. Luo C, Fernie AR, Yan J. Single-cell genomics and epigenomics: technologies and applications in plants. Trends Plant Sci 2020;25:1030—40.
11. Pechanova O, Takáč T, Samaj J, Pechan T. Maize proteomics: an insight into the biology of an important cereal crop. Proteomics 2013;13:637—62.
12. Fernie AR, Schauer N. Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet 2009;25:39—48.
13. Lawrence CJ. MaizeGDB. Methods Mol Biol 2007;406:331—45.
14. Ma S, Wang M, Wu J, Guo W, Chen Y, Li G, et al. WheatOmics: a platform combining multiple omics data to accelerate functional genomics studies in wheat. Mol Plant 2021;14:1965—8.
15. Liu Y, Zhang Y, Liu X, Shen Y, Tian D, Yang X, et al. SoyOmics: a deeply integrated database on soybean multi-omics. Mol Plant 2023;16:794—7.
16. Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y, et al. Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol 2013;54:e6.
17. He X, Wang J. BGI-RIS V2. Methods Mol Biol 2007;406:275—99.
18. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR rice genome annotation resource: improvements and new features. Nucleic Acids Res 2007;35:D883—7.
19. Song JM, Lei Y, Shu CC, Ding Y, Xing F, Liu H, et al. Rice information GateWay: a comprehensive bioinformatics platform for indica rice genomes. Mol Plant 2018;11:505—7.
20. Yu Z, Chen Y, Zhou Y, Zhang Y, Li M, Ouyang Y, et al. Rice gene index: a comprehensive pan-genome database for comparative and functional genomics of Asian rice. Mol Plant 2023;16:798—801.
21. Sun C, Hu Z, Zheng T, Lu K, Zhao Y, Wang W, et al. RPAN: rice pan-genome browser for ∼3000 rice genomes. Nucleic Acids Res 2017;45:597—605.
22. Xie L, Liu M, Zhao L, Cao K, Wang P, Xu W, et al. RiceENCODE: a comprehensive epigenomic database as a rice Encyclopedia of DNA Elements. Mol Plant 2021;14:1604—6.
23. Sato Y, Antonio BA, Namiki N, Takehisa H, Minami H, Kamatsuki K, et al. RiceXPro: a platform for monitoring gene expression in japonica rice grown under natural field conditions. Nucleic Acids Res 2011;39:D1141—8.
24. Xia L, Zou D, Sang J, Xu X, Yin H, Li M, et al. Rice Expression database (RED): an integrated RNA-Seq-derived gene expression database for rice. J Genet Genom 2017;44:235—41.
25. Yu Y, Zhang H, Long Y, Shu Y, Zhai J. Plant public RNA-seq database: a comprehensive online database for expression analysis of ∼45 000 plant public RNA-Seq libraries. Plant Biotechnol J 2022;20:806—8.
26. Komatsu S, Kojima K, Suzuki K, Ozaki K, Higo K. Rice proteome database based on two-dimensional polyacrylamide gel electrophoresis: its status in 2003. Nucleic Acids Res 2004;32:D388—92.
27. Dardick C, Chen J, Richter T, Ouyang S, Ronald P. The rice kinase database. A phylogenomic database for the rice kinome. Plant Physiol 2007;143:579—86.
28. Gour P, Garg P, Jain R, Joseph SV, Tyagi AK, Raghuvanshi S. Manually curated database of rice proteins. Nucleic Acids Res 2014;42:D1214—21.
29. Dharmawardhana P, Ren L, Amarasinghe V, Monaco M, Thomason J, Ravenscroft D, et al. A genome scale metabolic network for rice and accompanying analysis of tryptophan, auxin and serotonin biosynthesis regulation under biotic stress. Rice 2013;6:15.
30. Zenda T, Liu S, Dong A, Li J, Wang Y, Liu X, et al. Omics-facilitated crop improvement for climate resilience and superior nutritive value. Front Plant Sci 2021;12:774994.
31. Peng B, Guan K, Tang J, Ainsworth EA, Asseng S, Bernacchi CJ, et al. Towards a multiscale crop modelling framework for climate change adaptation assessment. Nat Plants 2020;6:338—48.
32. Watson A, Ghosh S, Williams MJ, Cuddy WS, Simmonds J, Rey MD, et al. Speed breeding is a powerful tool to accelerate crop research and breeding. Nat Plants 2018;4:23—9.
33. Gao C. Genome engineering for crop improvement and future agriculture. Cell 2021;184:1621—35.
34. Xue Y, Bao Y, Zhang Z, Zhao W, Xiao J, He S, et al. Database resources of the national genomics data center, China national center for bioinformation in 2022. Nucleic Acids Res 2022;50:D27—38.
35. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res 2022;50:D20—6.
36. Okido T, Kodama Y, Mashima J, Kosuge T, Fujisawa T, Ogasawara O. DNA data bank of Japan (DDBJ) update report 2021. Nucleic Acids Res 2022;50:D102—5.
37. Thakur M, Bateman A, Brooksbank C, Freeberg M, Harrison M, Hartley M, et al. EMBL's European bioinformatics Institute (EMBL-EBI) in 2022. Nucleic Acids Res 2023;51:D9—17.

38. Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, Mayer KF. Plant genome sequencing - applications for crop improvement. Curr Opin Biotechnol 2014;26:31—7.
39. Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, et al. Assembly: a resource for assembled genomes at NCBI. Nucleic Acids Res 2016;44:D73—80.
40. Chen M, Ma Y, Wu S, Zheng X, Kang H, Sang J, et al. Genome warehouse: a public repository housing genome-scale data. Dev Reprod Biol 2021;19:584—9.
41. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, et al. The ensembl genome database project. Nucleic Acids Res 2002;30:38—41.
42. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 2012;40:D1178—86.
43. Dong Q, Schlueter SD, Brendel V. PlantGDB, plant genome database and analysis tools. Nucleic Acids Res 2004;32:D354—9.
44. Fu LY, Zhu T, Zhou X, Yu R, He Z, Zhang P, et al. ChIP-Hub provides an integrative platform for exploring plant regulome. Nat Commun 2022;13:3413.
45. Ding K, Sun S, Luo Y, Long C, Zhai J, Zhai Y, et al. PlantCADB: a comprehensive plant chromatin accessibility database. Dev Reprod Biol 2022;S1672-0229(22)00133-4. https://doi.org/10.1016/j.gpb.2022.10.005.
46. Liu J, Zhang Y, Zheng Y, Zhu Y, Shi Y, Guan Z, et al. PlantExp: a platform for exploration of gene expression and alternative splicing based on public plant RNA-seq samples. Nucleic Acids Res 2023;51:D1483—91.
47. Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, et al. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinform 2008;2008:420747.
48. Toufighi K, Brady SM, Austin R, Ly E, Provart NJ. The botany array resource: e-northerns, expression angling, and promoter analyses. Plant J 2005;43:153—63.
49. Sundell D, Mannapperuma C, Netotea S, Delhomme N, Lin YC, Sjödin A, et al. The plant genome integrative explorer resource: PlantGenIE.org. New Phytol 2015;208:1149—56.
50. Xu Z, Wang Q, Zhu X, Wang G, Qin Y, Ding F, et al. Plant single cell transcriptome hub (PsctH): an integrated online tool to explore the plant single-cell transcriptome landscape. Plant Biotechnol J 2022;20:10—2.
51. Chen H, Yin X, Guo L, Yao J, Ding Y, Xu X, et al. PlantscRNAdb: a database for plant single-cell RNA analysis. Mol Plant 2021;14:855—7.
52. Jin J, Lu P, Xu Y, Tao J, Li Z, Wang S, et al. PCMDB: a curated and comprehensive resource of plant cell markers. Nucleic Acids Res 2022;50:D1448—55.
53. Sun Q, Zybailov B, Majeran W, Friso G, Olinares PD, van Wijk KJ. PPDB, the plant proteomics database at cornell. Nucleic Acids Res 2009;37:D969—74.
54. Mousavi SA, Pouya FM, Ghaffari MR, Mirzaei M, Ghaffari A, Alikhani M, et al. PlantPReS: a database for plant proteome response to stress. J Proteonomics 2016;143:69—72.
55. Charles H, Daniel G, Kangmei Z, William D, Bo X, Angela X, et al. Plant Metabolic Network 15: A resource of genome-wide metabolism databases for 126 plants and algae. J Integr Plant Biol 2021;63:1888—905.
56. Grafahrend-Belau E, Weise S, Koschützki D, Scholz U, Junker BH, Schreiber F. MetaCrop: a detailed database of crop plant metabolism. Nucleic Acids Res 2008;36:D954—8.
57. Alway SE, MacDougall JD, Sale DG, Sutton JR, McComas AJ. Functional and structural adaptations in skeletal muscle of trained athletes. J Appl Physiol 1988;64:1114—20.
58. Grandi FC, Modi H, Kampman L, Corces MR. Chromatin accessibility profiling by ATAC-seq. Nat Protoc 2022;17:1518—52.
59. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science 2007;316:1497—502.
60. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc 2011;6:468—81.
61. Lane AK, Niederhuth CE, Ji L, Schmitz RJ. pENCODE: a plant encyclopedia of DNA elements. Annu Rev Genet 2014;48:49—70.
62. Lü P, Yu S, Zhu N, Chen YR, Zhou B, Pan Y, et al. Genome encode analyses reveal the basis of convergent evolution of fleshy fruit ripening. Nat Plants 2018;4:784—91.
63. Chao H, Hu Y, Zhao L, Xin S, Ni Q, Zhang P, et al. Biogenesis, functions, interactions, and resources of non-coding RNAs in plants. Int J Mol Sci 2022;23:3695.
64. Ryu KH, Zhu Y, Schiefelbein J. Plant cell identity in the era of single-cell transcriptomics. Annu Rev Genet 2021;55:479—96.
65. Khan MM, Komatsu S. Rice proteomics: recent developments and analysis of nuclear proteins. Phytochemistry 2004;65:1671—81.
66. Ohyanagi H, Sakata K, Komatsu S. Soybean proteome database 2012: update on the comprehensive data repository for soybean proteomics. Front Plant Sci 2012;3:110.
67. Wu X, Gong F, Cao D, Hu X, Wang W. Advances in crop proteomics: PTMs of proteins under abiotic stress. Proteomics 2016;16:847—65.
68. Fang X, Chen J, Dai L, Ma H, Zhang H, Yang J, et al. Proteomic dissection of plant responses to various pathogens. Proteomics 2015;15:1525—43.
69. Shi J, Wang J, Lv H, Peng Q, Schreiner M, Baldermann S, et al. Integrated proteomic and metabolomic analyses reveal the importance of aroma precursor accumulation and storage in methyl jasmonate-primed tea leaves. Hortic Res 2021;8:95.

70. Zhu G, Wang S, Huang Z, Zhang S, Liao Q, Zhang C, et al. Rewiring of the fruit metabolome in tomato breeding. Cell 2018;172:249—61.
71. Sharma V, Gupta P, Priscilla K, SharanKumar, Hangargi B, Veershetty A, et al. Metabolomics intervention towards better understanding of plant traits. Cells 2021;10:346.
72. Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Barrón-López JA, Martini JWR, Fajardo-Flores SB, et al. A review of deep learning applications for genomic selection. BMC Genom 2021;22:19.
73. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet 2015;16:321—32.
74. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583—9.
75. Elvanidi A, Katsoulas N. Machine learning-based crop stress detection in greenhouses. Plants 2022;12:52—71.
76. Pham HT, Awange J, Kuhn M, Nguyen BV, Bui LK. Enhancing crop yield prediction utilizing machine learning on satellite-based vegetation health indices. Sensors 2022;22:719.
77. Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J 2021;19:3735—46.
78. Saand MA, Xu YP, Li W, Wang JP, Cai XZ. Cyclic nucleotide gated channel gene family in tomato: genome-wide identification and functional analyses in disease resistance. Front Plant Sci 2015;6:303.
79. Dergilev AI, Orlova NG, Dobrovolskaya OB, Orlov YL. Statistical estimates of multiple transcription factors binding in the model plant genomes based on ChIP-seq data. J Integr Bioinform 2022;19:20200036.
80. Orlov YL, Ivanisenko VA, Dobrovolskaya OB, Chen M. Plant biology and biotechnology: focus on genomics and bioinformatics. Int J Mol Sci 2022;23:6759.
81. Ivanisenko TV, Saik OV, Demenkov PS, Ivanisenko NV, Savostianov AN, Ivanisenko VA. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. BMC Bioinf 2020;21:228.
82. Ivanisenko TV, Demenkov PS, Kolchanov NA, Ivanisenko VA. The new version of the ANDDigest tool with improved AI-based short names recognition. Int J Mol Sci 2022;23:14934.
83. Salgotra RK, Chauhan BS. Genetic diversity, conservation, and utilization of plant genetic resources. Genes 2023;14:174.