Database, 2023, 1–17
DOI: https://doi.org/10.1093/database/baad088
Review

Genotype and phenotype data standardization, utilization and integration in the big data era for agricultural sciences

Cecilia H. Deng[1,‡,*], Sushma Naithani[2,‡], Sunita Kumari[3,‡], Irene Cobo-Simón[4,5], Elsa H. Quezada-Rodríguez[6,7], Maria Skrabisova[8], Nick Gladman[3,9], Melanie J. Correll[10], Akeem Babatunde Sikiru[11], Olusola O Afuwape[12], Annarita Marrano[13], Ines Rebollo[14], Wentao Zhang[15] and Sook Jung[16,‡] On behalf of the Genotype-Phenotype Working Group, AgBioData

[1]Molecular and Digital Breeding, New Cultivar Innovation, The New Zealand Institute for Plant and Food Research Limited, 120 Mt Albert Road, Auckland 1025, New Zealand
[2]Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331, USA
[3]Cold Spring Harbor Laboratory, 1 Bungtown Rd, Cold Spring Harbor, New York, NY 11724, USA
[4]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA
[5]Institute of Forest Science (ICIFOR-INIA, CSIC), Madrid, Spain
[6]Departamento de Producción Agrícola y Animal, Universidad Autónoma Metropolitana-Xochimilco, Ciudad de México, México
[7]Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Ciudad de México, México
[8]Department of Biochemistry, Faculty of Science, Palacky University, Olomouc, Czech Republic
[9]U.S. Department of Agriculture-Agricultural Research Service, NEA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, NY 14853, USA
[10]Agricultural and Biological Engineering Department, University of Florida, 1741 Museum Rd, Gainesville, FL 32611, USA
[11]Federal University of Agriculture Zuru, PMB 28, Zuru, Kebbi 872101, Nigeria
[12]University of Lagos, Nigeria
[13]Phoenix Bioinformatics, 39899 Balentine Drive, Suite 200, Newark, CA 94560, USA
[14]Universidad de la República, Uruguay
[15]National Research Council Canada, 110 Gymnasium Pl, Saskatoon, Saskatchewan S7N 0W9, Canada
[16]Department of Horticulture, Washington State University, 303c Plant Sciences Building, Pullman, WA 99164-6414, USA

[*]Corresponding author: Tel: +64 21573875; Email: Cecilia.Deng@plantandfood.co.nz
[‡]These authors contributed equally to this work.

Citation details: Deng, C.H., Naithani, S., Kumari, S. *et al.* Genotype and phenotype data standardization, utilization and integration in the big data era for agricultural sciences. *Database* (2023) Vol. 2023: article ID baad088; DOI: https://doi.org/10.1093/database/baad088

Abstract

Large-scale genotype and phenotype data have been increasingly generated to identify genetic markers, understand gene function and evolution and facilitate genomic selection. These datasets hold immense value for both current and future studies, as they are vital for crop breeding, yield improvement and overall agricultural sustainability. However, integrating these datasets from heterogeneous sources presents significant challenges and hinders their effective utilization. We established the Genotype-Phenotype Working Group in November 2021 as a part of the AgBioData Consortium (https://www.agbiodata.org) to review current data types and resources that support archiving, analysis and visualization of genotype and phenotype data to understand the needs and challenges of the plant genomic research community. For 2021–22, we identified different types of datasets and examined metadata annotations related to experimental design/methods/sample collection, etc. Furthermore, we thoroughly reviewed publicly funded repositories for raw and processed data as well as secondary databases and knowledgebases that enable the integration of heterogeneous data in the context of the genome browser, pathway networks and tissue-specific gene expression. Based on our survey, we recommend a need for (i) additional infrastructural support for archiving many new data types, (ii) development of community standards for data annotation and formatting, (iii) resources for biocuration and (iv) analysis and visualization tools to connect genotype data with phenotype data to enhance knowledge synthesis and to foster translational research. Although this paper only covers the data and resources relevant to the plant research community, we expect that similar issues and needs are shared by researchers working on animals.

Database URL: https://www.agbiodata.org.

Introduction

Genotype-to-phenotype (G2P) integration is the process of linking genetic data to measurable qualitative and quantitative phenotypes and traits. Historically, linking genetic markers or genes with desirable traits has led to improved cultivars with higher yields and quality, enhanced

Received 13 April 2023; Revised 17 October 2023; Accepted 28 November 2023
© The Author(s) 2023. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

disease resistance or climate resilience. In the past two decades, the generation of high-throughput omics or 'big data' including plant genomes and pangenomes; genetic variation data including single-nucleotide polymorphisms (SNPs) and structural variations (SVs) including transcriptomes, phenotype, proteomes and metabolomes has changed the scale and scope of data analysis, knowledge synthesis and its application in translational research (1, 2). In addition, researchers and breeders worldwide have collected classic mutant phenotype and trait data, and more recently, there has been a substantial surge in the generation of large-scale phenotype data. Often, specific big datasets are generated by projects and analyzed to address knowledge gaps, but they often remain underutilized for discovering new knowledge. Although it is challenging to survey or predict which scientific data will be reused, by whom and for what purpose (3–5), several studies have shown that data reuse is generally limited (6–8). Researchers tend to primarily employ their own data for hypothesis testing, only occasionally incorporating other's data for baseline purposes (9, 10). Going forward, the different data types produced in various experiments can be reutilized to synthesize new knowledge and to develop data-driven hypotheses for experimental research (11). However, integrating large-scale datasets from diverse sources can be challenging (4, 8) and typically involves infrastructure to support data sharing, data quality check, data re-formatting, curation and re-analysis (12–14). For example, genotype, phenotype and expression data for the same plant accessions were generated from various projects over a decade, each using inconsistent sample identifiers and different plant growth environments. Before utilizing these various datasets to investigate the genetic and environmental factors influencing a particular phenotype, establishing consistent sample names, gene IDs and phenotypes across all datasets will be needed and possibly require modification in the original data format. The fulfillment of the unprecedented potential of big data depends on the data being Findable, Accessible, Interoperable, and Reusable (FAIR) (15–17). To meet the FAIR standards, any dataset should include metadata (18, 19) providing the standard terms and details necessary for data interpretation using controlled vocabularies. Data and metadata standardization can be achieved by developing common community standards of data formats and descriptions so that diverse datasets from different sources can be accessed and interoperated for visualization and knowledge synthesis. A clear, organized and consistent method of capturing and exchanging agricultural data will ensure easier data discovery, comparisons and reuse by various stakeholders.

Making data FAIR requires concerted efforts and communications among all parties involved in data generation and curation. In 2015, the AgBioData Consortium (https://www.agbiodata.org) was formed to identify and promote the means to consolidate and standardize common genomic, genetic and breeding (GGB) database tools and operations, with the goal of increased data interoperability for future research (16). At present, AgBioData comprises over 40 GGB databases and more than 200 scientists, fostering collaborations and open discussions about the common practices, challenges and solutions related to big data generated by agricultural researchers. AgBioData consortium has previously identified challenges facing GGB databases and suggested common guidelines for biocuration, ontologies, metadata, database platforms, programmatic access to data, communication among various partners and stakeholders and sustainability of genomic databases (16). AgBioData aims to (i) identify and address data-related issues by defining community-based standards, (ii) expand the network by involving all the stakeholders of the agricultural research community, (iii) develop educational material to train current and future scientists on database usage and the FAIR principles and (iv) develop a roadmap for a sustainable GGB database ecosystem. As part of this National Science Foundation Research Coordination Networks project, various working groups were established to address significant data-related challenges and requirements. One such group, known as the Genotype-Phenotype working group, was formed in November 2021 with the goal of identifying current challenges in annotating and integrating large-scale genotype and phenotype data. In this study, the Genotype-Phenotype working group summarizes common genotype and phenotype data types, repositories and knowledgebases, and the present state of FAIR practices of genotype and phenotype data (as illustrated in Figure 1). This paper provides (i) a brief introduction of the diverse data types and how they are generated, (ii) primary and secondary data repositories and databases for these data types, (iii) requirements of associated metadata and the minimum standards, (iv) examples of reuse and re-analysis of omics data and (v) limitations of data reuse. Finally, based on our surveys, reviews and community discussions, we list our recommendations that can provide the needed support to the plant genomic communities in making genotype and phenotype data interoperable and reusable for knowledge synthesis and fostering translational research needed for long-term agriculture sustainability.

Genomics and transcriptomics data

Whole-genome and transcriptome sequences

In the past decade, sequencing technology has evolved rapidly from the early days of time-consuming Sanger sequencing to high-throughput massive parallel sequencing that started the era of whole-genome sequencing (WGS) and transcriptome sequencing (20–22). There are three general methods of DNA/cDNA sequencing: (i) Sanger chain termination sequencing and Maxam Gilbert sequencing (23, 24); (ii) short-read sequencing known as next-generation sequencing (NGS) (25–28) including Ion Torrent, Solexa/Illumina and Roche/454 pyrosequencing (29) and (iii) more recent long-read Third Generation Sequencing (3GS) (30–34). Primarily, single-molecule real-time sequencing from Pacific Bioscience (PacBio) and nanopore sequencing from Oxford Nanopore Technologies (ONT) are examples of the 3GS. Illumina is currently the dominant and most popular platform in NGS for both genome and transcriptome sequencing because of its high accuracy, low cost and global distribution of its solutions for sequencing. PacBio and ONT are also gaining popularity and becoming more affordable for high-quality long-read/full-length sequences. Similarly, DNBSeq from MGI Tech, a subsidiary of the Beijing Genomics Institute group, and Ion Torrent Systems are making advances (35, 36). Several file formats are used in WGS, and the most common is the compressed FASTQ format (37) that is used for both NGS and 3GS sequencing. The original file formats for 3GS include legacy h5 format

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

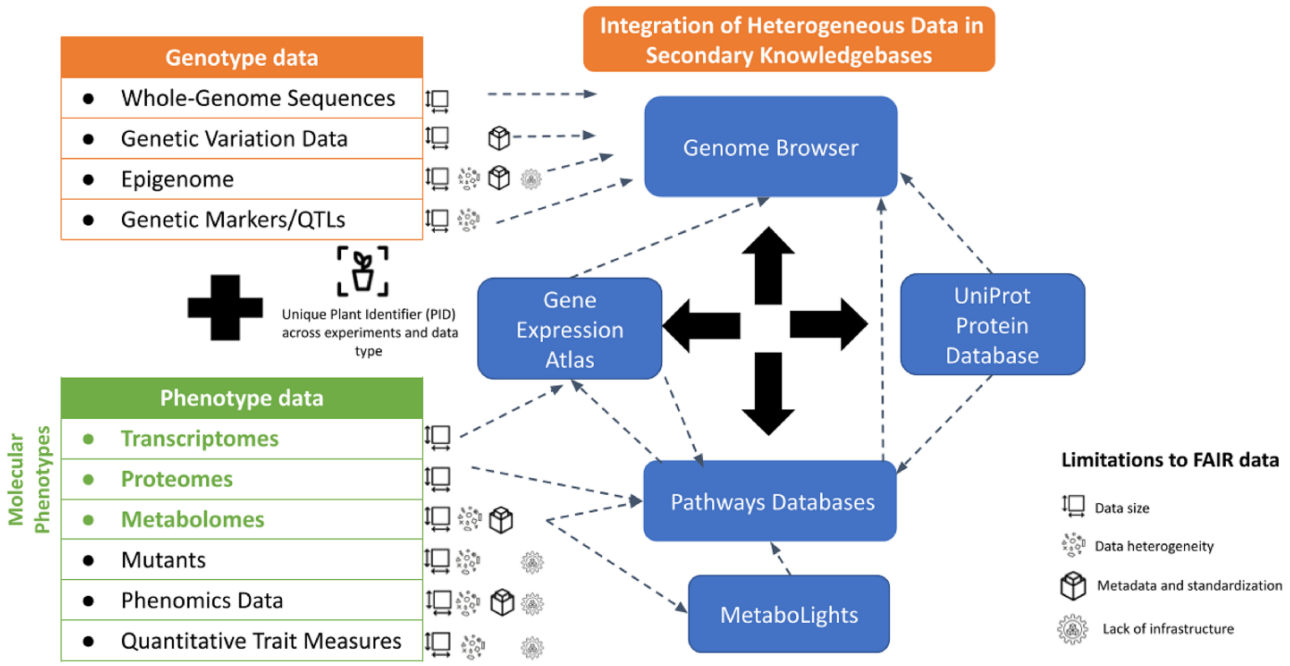

Figure 1. Current status of genotype-to-phenotype data integration. The left side illustrates the diversity of genotype and phenotype data. The right-hand side lists examples of the existing databases and knowledgebases which support the integration of heterogenous data types and their visualization.

for PacBio (http://files.pacb.com/software/instrument/2.0.0/bas.h5%20Reference%20Guide.pdf), the industry-standard Binary Alignment Map (BAM) format (38) and the FAST5 format for ONT (39) that is based on the hierarchical data format (HDF5) (40) used for ONT data storage sequencer. Meanwhile, there are numerous basecallers available for conversion to FASTQ format (35) (https://long-read-tools.org/); in general, we find that sequencing data have achieved standard data formatting. Uniformity among available basecallers is essential because they play a critical role in converting data from various technologies into a standardized format, such as FASTQ. This uniformity facilitates data exchange, analysis and collaboration in the field of genomics and bioinformatics.

Genome sequencing strategies for genotyping

Genotyping technology is a crucial component in linking genotype to phenotype. The process entails the generation of genetic profiles, which comprise DNA fragments, markers or sequences. These profiles serve to differentiate between different accessions, cultivars or siblings within a population in the initial stage. Subsequently, a correlation analysis is conducted to assess the relationship between the genotype profile and the phenotypic traits. The first-generation genotyping marker was Restriction Fragment Length Polymorphisms, which relied upon underlying differences in base pair sequences to create fingerprints after DNA regions were digested with known restriction enzymes (36). Later, the technological advancements led to genotyping by scoring microsatellite markers, simple sequence repeats or short tandem repeats (37). High-throughput low- and high-density SNP arrays (38) provide a cost-effective genotyping solution for studies such as population structures, genomic diversity, gene discovery and molecular breeding (39–42). Furthermore, development of whole-genome arrays made it possible to genotype a large number of samples in a short period of time, and data analysis simpler. However, designing an efficient array with high-quality SNPs for a particular crop usually requires significant investment upfront. As genome sequencing has advanced even further, researchers can now achieve whole-genome profiling through lower- or higher-coverage sequencing strategies such as NGS and 3GS. In essence, researchers now could choose from the various available options depending on their research goals and available budget. For example, sequencing of sub-sampled loci (43) has been widely used in phylogenomics studies for cost-effective large-scale genotyping. Skim sequencing (44) is another low-coverage WGS approach. Target enrichment sequencing investigates specific genomic elements via pre-defined probe sequences (45). Exome sequencing is a common type of target sequencing that focuses on protein-coding regions of genes (46). Amplicon sequencing is a highly targeted approach addressing specific genome loci. Genotyping-by-sequencing (47, 48) and restriction site–associated DNA marker sequencing (49, 50) are two popular, cost-effective sequencing strategies for shearing the genome via restriction enzyme(s).

The advent of high-throughput sequencing has generated immense amounts of data that are being used to capture intraspecies and interspecies genetic diversity and allow exploration of genetic variations. Regarding genotyping data structure, the 1000 Genomes project (https://www.internationalgenome.org/) spearheaded the first Variant Call Format (VCF) for standardizing the SNPs, indels and SV between two or more genomes at a given locus (51). The VCF has become the go-to format for variant data and associated metadata; over time, modifications of the base VCF file have expanded to include experiment-specific modifications, such as genome-wide association study (GWAS)-VCF (52) and genomic variant call format (tinyurl.com/5f8wpmhr), and accommodate variant information of polyploid genomes. In

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

addition to the low-coverage genome sequence, transcriptome sequences are routinely used for genotyping and identification of useful genetic markers. More recently, the integration of single-cell genome sequencing and single-cell transcriptome sequencing tools has facilitated quantifying genetic and expression variability between individual cells (16). Like sequence data, genotyping data has standardized formats.

Public repositories for genomic and transcriptomic data

Regardless of the sequencing platform or strategy used, raw sequencing data in compressed fastq.gz format are submitted to a public data repository such as National Center for Biotechnology Information (NCBI) GenBank, Sequence Read Archive (SRA) (53, 54) (Leinonen *et al.*, 2010; Kodama *et al.*, 2012) and/or Gene Expression Omnibus (55–57) via the NCBI submission portal. NCBI provides BioSample metadata templates based on organism lineage validation. Besides NCBI, the data can be submitted to the DNA DataBank of Japan (DDBJ) (58–60), SRA via the DDBJ submission navigation website or the European Nucleotide Archive (ENA) through the BioStudies portal. DDBJ, ENA and NCBI GenBank (Table 1 and Supplementary Table S1) form the International Nucleotide Sequence Database Collaboration (INSDC) (61, 62) and exchange data daily. Prior to publishing the results, all the life science journals require authors to submit their raw sequence data to the public INSDC repositories—a key component of the data sharing policies in the community of biologists (63). Additional public platforms that host the sequence data include the US Department of Energy Joint Genome Institute (JGI) (64, 65) that makes sequencing data generated by its collaborating projects available immediately to registered users and then follows public release on JGI and NCBI/SRA or GeneBank after a one-year embargo period. JGI also provides Phytozome (66), the Plant Comparative Genomics portal, for genome accessing, comparison and visualization (Table 2). *Nature* and *Scientific Data* request that sample metadata is deposited in one of the INSDC BioSample databases in conjunction with sequence data. It is crucial to use the standardized metadata at both the study and sample level to facilitate the curation and processing of transcriptomics data in a FAIR-compliant way. A few sequence repositories such as Zenodo (https://www.zenodo.org), DRYAD (https://datadryad.org), Figshare (https://figshare.com) and Harvard Dataverse (https://dataverse.harvard.edu) accept data submission in any file format.

Apart from the public databases hosted in the USA and Europe, the Genome Sequence Archive (GSA, https://ngdc.cncb.ac.cn/gsa) in China follows INSDC-compliant data standards (67). The Indian Biological Data Center (IBDC, https://ibdc.rcb.res.in) is also a public repository in India hosting various life science data. For sequencing data, IBDC provides the INSDC-compatible Indian Nucleotide Data Archive (https://inda.rcb.ac.in/home) with data synchronized to NCBI/ENA/DDBJ and the Indian Nucleotide Data Archive-Controlled Access (https://inda.rcb.ac.in/indasecure/home) for private data. In New Zealand, the Aotearoa Genomic Data Repository (AGDR) hosts genomic data, especially for native *taonga* ('treasure' in Maori language) species. The presence of these multiple public databases across different countries and regions could be beneficial for the advancement of research in terms of data availability, collaboration and preservation of unique species. However, challenges related to data harmonization, fragmentation and standardization must be addressed to fully harness these resources' potential for genotype-to-phenotype research.

The recent availability of data submission to cloud storage is also gaining popularity and contributing to the advancement of research focusing on genotype to phenotype. For example, Amazon Web Services (AWS) offers Open Data (https://aws.amazon.com/opendata) source for unregistered users to find and use publicly available datasets, while allowing subscribed customers to search and access even third-party data (https://docs.aws.amazon.com/data-exchange/index.html) for research use. In addition, through Amazon Omics (https://aws.amazon.com/omics/), it provides data on Plant and Animal Genomics (https://aws.amazon.com/solutions/agriculture/plant-animal-genomics/), which is another platform that could be used to facilitate omics data analysis and integration. Similarly, there are other cloud

Table 1. A list of active, maintained and updated public repositories for genomic, genotyping and transcriptome data

Database name	NCBI	DRA	ENA	GSA	IBDC	AGDR[a]	DRYAD[c]	Zenodo[b,c]	FigShare
Genome sequence data	+	+	+	+	+	+	+	+	+
WGS annotations	+	?	?	?	?	?	?	?	+
Genotyping data	+	?	?	?	?	?	?	?	+
Transcriptome sequence data	+	+	+	?	?	?	+	+	+
fq.gz	+	+	+	+	+	+	+	+	+
BAM	+	+	+	+	+	+	+	+	+
SFF	+	+	+	+	+	–	+	+	+
HDF	+	+	+	+	+	–	+	+	+
VCF	+	+	+	?	?	?	+	+	+
INSDC-Source	+	+	+	a	b	c	d	*e*	f

The '+' and '–' symbols indicate the presence and absence of the supported data type and data format, respectively. Databases that support any data type beyond the specified most common types are marked by "¥". Out of the INSDC, source databases were established and maintained by (a) National Genomics Data Centre, China, and China National Center for Bioinformation; (b) The IBDC; (c) New Zealand Ministry for Business Innovation and Employment; (d) University of North Carolina at Chapel Hill, California Digital Library; (e) CERN, the European Organization for Nuclear Research (Conseil européen pour la Recherche nucléaire) and (f) Digital Science. Holtzbricnck Publishing Group, Macmillan Publishers Limited. '?' means that the information is not available. '+/–' means that this data type can be submitted only through a command line or programmatic approach but not by the interactive interface. Detailed information about metadata requirements and database URLs is available in Supplementary Table S1.

[a]data are available upon request.

[b]recommended by FAIRsharing.org.

[c]Databases that support any data type beyond the specified most common types.

Abbreviations: DRA, DDBJ sequence read archive; SFF, Standard Flowgram Format.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

Table 2. List of Crop/clad Community GGB Databases that integrate various data types, including whole-genome data, genotype, phenotype, QTL, GWAS and germplasm data. Refer to Supplementary Table S3 for data types and metadata for each database

Species/Crop	Database	Database URL
Arabidopsis	TAIR	https://www.arabidopsis.org/
Cassava	CassavaBase	https://www.cassavabase.org/
Citrus	Citrus Genome Database	https://www.citrusgenomedb.org/
Citrus/*Diaphorina citri*/Ca. *Liberibacter asiaticus*	Citrus Greening	https://www.citrusgreening.org/
Cotton	CottonGen	https://www.cottongen.org/
Cucurbit	Cucurbit Genomics	http://cucurbitgenomics.org/
Forest trees	TreeGenes	https://treegenesdb.org
	Hardwood Genomics	http://www.hardwoodgenomics.org/
Grains	GrainGenes	https://wheat.pw.usda.gov
	Gramene	https://www.gramene.org/
	SorghumBase	https://www.sorghumbase.org/
	Triticeae toolbox, T3	https://wheat.triticeaetoolbox.org/
	WheatIS	https://wheatis.org
	KitBase	http://kitbase.ucdavis.edu/
Legumes	KnowPulse	https://knowpulse.usask.ca/
	Legume Information System	https://www.legumeinfo.org/
	PeanutBase	https://peanutbase.org
Pulses	Pulse Crop Database	https://www.pulsedb.org/
	Soybase	https://www.soybase.org/
Maize	MaizeGDB	https://maizegdb.org/
Musa	MusaBase	https://www.musabase.org/
Rosaceae	Genome Database for Rosaceae	https://www.rosaceae.org/
Solanaceae	Sol Genomics	https://solgenomics.net/
Sweet Potato	SweetPotatoBase	https://www.sweetpotatobase.org/
Vaccinium	Genome Database for Vaccinium	https://www.vaccinium.org/
Yam	YamBase	https://www.yambase.org/
Comparative genomics database used by multiple communities		
A comparative genomics database for ~300 plant species	Phytozome	https://phytozome-next.jgi.doe.gov/
A comparative genomics database hosting 118 genomes from models, crops, fruits, vegetables, etc.	Gramene	https://www.gramene.org/
Others	AgBase	https://agbase.arizona.edu/
	Bio-Analytic Resource	https://bar.utoronto.ca/

storage options for storing and accessing genetic data for research by users, either corporate and individuals, including Google Cloud Life Sciences (https://cloud.google.com/life-sciences) and Microsoft Genomics (https://azure.microsoft.com/en-in/products/genomics/).

Table 1 contains a list of active, maintained and updated public repositories for genome, genotyping and transcriptome sequence data. Detailed information about metadata file formats related to these repositories is provided in Supplementary Table S1.

The metadata associated with sequence and genotype data promotes a dataset's discoverability and reusability. We note here that many secondary public repositories exist that exclusively host data on promoters, transcription factors, proteomes, various RNA types, epigenomics data and pangenomes (Supplementary Table S2). However, here, we limit our discussion to primary genotype and phenotype data and expect that detailed discussions on other related topics will be provided by the other working groups of the AgBiodata consortium.

Metadata requirements on genomics and transcriptomics datasets

The metadata associated with genome, genotyping and transcriptome sequencing is crucial for data reusability and interoperability. To maximize the implementation of FAIR standards, the metadata should be described with accurate Gene Ontology and Plant Ontology terms with proper evidence codes wherever applicable. Project- and sample-level metadata typically includes taxonomic identifier (for species), tissue type (organism part) from which the sample was taken, disease state, growth or developmental stage of the sample, the biological gender of the sample and collection date. Assay-level metadata is directly related to the preparation of biological materials undergoing the assay, including method details (bulk RNA-seq, scRNA-seq, etc.), library information (single-end or paired-end), replicates (biological or technical), instrument metadata, quality control (QC) and workflow metadata. For example, submission of sequencing data to NCBI GenBank and SRA requires metadata for the submitter (including name, affiliation, and email of the data submitter and other authors), BioProject goals (such as genome sequencing and assembly; raw sequence reads, epigenomics, exome, proteome and variation) and BioSamples information (like organism's name and taxonomic identifier, geographical origin of the sample and tissue type).

We note here that in most repositories, the organism's name is the only required field for biological targets, with optional fields of strain, breed, cultivar, isolate name, label and description. The minimum general information required for a project is the data release date, project title and public description of the study goals. Optional fields include a project's relevance to a field (agricultural, medical, industrial, environmental, evolution, model organism and other), external links to other websites associated with the study, grant

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

information (number, title and grantee), research consortium name and the uniform resource locator (URL), data provider and URL (if different from the submitting organization) and publication information.

Optional but useful metadata for BioSamples includes sample title, BioProject accession, biomaterial provider (laboratory name and address, or a cultural collection identifier), name of the cell line, cell type, collected by and date, culture identifier and source institute (refer to http://www.insdc.org/controlled-vocabulary-culturecollection-qualifier), disease name and stage, observed genotype, growth protocol, height or length measured, the growth environmental, the geographical coordinates of the sample collection, phenotype of sampled organism (compliant with the BioPortal at http://bioportal.bioontology.org), population (filial generation, number of progeny and genetic structure), sample type (cell culture, mixed culture, tissue sample, whole organism, single cell and so on), sex, specimen voucher, temperature of the sample at time of sampling, treatment and sample description (defined in the Phenotype And Trait Ontology of Open Biological and Biomedical Ontology Foundry at http://obofoundry.org/ontology/pato.html).

The mandatory attributes for library construction metadata are BioSample name, library ID, a title, data type and method information (Whole Genome Amplification, Whole Genome Sequencing (WGS), RNA-Seq, Expressed Sequence Tags, ChIP-Seq, and so on), source (GENOMIC, TRANSCRIPTOMIC, GENOMIC SINGLE CELL, METAGENOMIC, etc.), selection (Polymerase Chain Reaction (PCR), RANDOM, Reverse Transcription-PCR, cDNA, DNAse, Restriction Digest, etc.), layout, platform, instrument model, design description, file type and filename(s).

A few other sequence data repositories do not enforce the submission of metadata but encourage data submitters to provide as many details as possible. In this category, AGDR (https://repo.data.nesi.org.nz/DD) requires submitter ID, project ID, project code, project name, program name, database gap accession number, experiment type, number of samples and replicates and data type. In addition, it provides metadata templates for submitting detailed information on samples and methods (sample, aliquot, RNA integrity number-, adapter name and sequence, barcoding, base caller name and version, experiment name, flowcell barcode, fragment sizes, instrument model, lane number, library name and library preparation kits), project, publication, core metadata collection, indigenous governance and indigenous knowledge label templates.

The minimum metadata for a DRYAD submission requires a title describing the data and the study, author(s) information, abstract (dataset structure and concepts, reuse potential, legal or ethical considerations, etc.) and research domain. Optional metadata recommended is funding information, research facility, keywords, technical methods details and publication details. However, the biosample or plant accession metadata is not captured. Figshare recommends metadata submission guidelines similar to INSDC repositories but does not enforce them as a requirement. The storage quota for a free account is 20 GB and up to 100 projects.

Genotyping data submission and metadata requirements

The major repository for submitting non-human VCF files containing genotyping-related data is the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) European Variation Archive (EVA) (68), but a newer repository has also arisen in the Genome Variation Map (69). NCBI hosts the dbSNP and dbVar databases, which are intended for human data. All repositories strive to adhere to FAIR practices, but others have put forth additional recommendations (70). The EVA repository accepts VCF file structures, including hapmap formatted files (71) and SNP genotyping arrays, that are validated using custom EBI VCF Validation Suite software (https://github.com/EBIvariation/vcf-validator) with a minimum number of data fields with accompanying metadata that includes, but is not limited to, project title, sequencing platform information, software, reference organism and genome version and date and data generation location. The data fields for a VCF are the header lines that contain information about the dataset and relevant reference sources (organism, genome version, alignment, mapping method, etc.) followed by the variant site record row data: chromosome number, chromosome position, reference allele, alternate allele, quality, filter tag and additional allele info format (https://gatk.broadinstitute.org). However, the naming structure within some of these fields is not standardized, which can lead to interoperability concerns.

Crop Community GGB Databases

Whole-genome, transcriptome and genotype data can also be submitted to most of the GGB databases such as Genome Database for Rosaceae (72, 73), CottonGen (74, 75), SoyBase (76, 77), Legume Information System (78, 79), Sol Genomics Network (80, 81), MaizeGDB (82, 83), TreeGenes (84, 85), the Arabidopsis Information Resource (TAIR) (86, 87), KnowPulse (88) and InterMine (89, 90) (Table 2). Some of these databases, such as Gramene (91–93), SorghumBase (94) and InterMine (89, 95), do not accept data from authors but obtain from the primary databases. Depending on the GGB databases, different types of data and metadata can be submitted. Typically, these crop GGB databases collect a wide variety of data such as quantitative trait loci (QTL), GWAS, markers, alleles, genetic maps and cultivar/germplasm phenotype data and integrate them with whole-genome, transcriptome and genotyping data. These GGB databases standardize various names that associate the data with various ontologies to integrate data from various sources and of various types. This integration of different kinds of data, not typically done in the primary databases specialized in particular types of data, is one of the key steps in making the data FAIR. Integrating data from diverse sources provides researchers with foundations for subsequent statistical analyses to discover novel associations between different data types, potentially leading to valuable insights and breakthroughs. For example, SNP genotype data and phenotype data from multiple locations of the same germplasm allow further analysis that can reveal how particular genotypes manifest specific phenotypes in distinct environments.

Uses and Applications

WGS data can be reused in genome assembly (96–98); pan-genome construction (99); single-nucleotide variation (100–102), copy number variation (CNV) (103–105) and structure variation (SV) (106–108) discovery; phylogenomics; comparative genomics and other genome research to study genome structures, genome diversity, the evolution of gene families or organisms, crop domestication and improvement (69, 109–111). Genotyping data in VCF format

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

(https://samtools.github.io/hts-specs/VCFv4.2.pdf) can be used for numerous purposes: to store the location of given variants (including GWAS-associated variants), to identify targets of molecular markers for genotyping purposes, to evaluate the effects of given base pair and structural variants on gene function, to perform comparative genomics and evolutionary studies and to assist computational breeding approaches via machine learning and other methods. Data extraction and manipulation of VCF files are easy with the use of existing software toolkits such as VCFtools (51) (https://vcftools.github.io/index.html) and SAMtools (112) and can be utilized in conjunction with existing and ad hoc bioinformatic pipelines due to their command line functionality. By integrating VCF data with RNA-Seq and phenomics data, researchers can use these data sets for quantitative genetic studies, including GWASs (113–116), QTL (115, 117–119) analysis, marker discovery and genome selection (GS) (120–122) to accelerate modern breeding techniques. Integrating transcriptomics data with metabolomics data can help predict biomarkers often associated with biological pathways. This will assist in understanding the mechanism of underlying molecular patterns driving a condition. Integration of genomic, epigenomic and transcriptomic profiles will facilitate the prediction of key genomic variables and biological variation. Integration of gene expression data and CNVs can be used to categorize samples into groups based on their similarity to two datasets.

Phenotype and Phenomics

Data types, Repositories and Knowledge Bases

Plant phenotyping is the key for plant breeding, characterization of biodiversity and genetic and genomic-based approaches for translational research (123). The classical genetic and functional genomics studies in model and crop plants have identified numerous mutants that show distinct morphological and anatomical phenotypes associated with one or more genes, pathways and molecular processes. Table 3 lists databases that host the mutant collections and description of the phenotype of individual mutants and associated genes, including MaizeDIG (83), RIKEN Arabidopsis Genome Encyclopedia (124), Mutant Variety Database (125), Plant Genome Editing Database (126), Tomato Mutants Archive TOMATOMA (127) and Plant Editosome Database (109).

In addition, complex phenotypic traits (i.e. morphological and physiological) related to the fitness and performance of an organism are often quantitative and have multiple genetic determinants (128, 129). Examples of traits determined by multiple genes (known as QTLs) are crop yield, biomass, resistance to pests and pathogens, abiotic stress tolerance, nutritional value and ease of harvest. In addition to crop breeding, trait-based approaches are widespread in ecological research (130), as they provide a general understanding of a wide range of ecological and evolutionary phenomena, such as the impact of climate change and anthropogenic land use on biodiversity (131–133). In Table 3, we provide a list of key databases (or portal of bigger databases) that host information related to traits, QTLs and associated data including the Gramene QTL database (134), QTL database for wheat (135), Global Plant Trait Network Database (GLOPNET) (136), TRY Plant Trait Database (137), a database of Ecological Flora of the Britain and Ireland (138), BIOPOP Database of Plant Traits (139), GRIN (140), the United States Department of Agriculture (USDA) PLANTS Database, BiolFlor (141), LEDA Traitbase (142), BROT database of plant traits for Mediterranean basin species (143) and AusTraits (144). Trait and QTL data are also integrated with other types of data in various crop community databases listed in Table 2.

Phenomics is the systematic analysis for the refinement and characterization of phenotypes on a genome-wide scale. With the advent of high-throughput platforms, it became possible to collect phenomics data at a single-cell, organismal and/or population-wide scale (145). Phenomics can be used for species recognition and biodiversity characterization (146), stress quantification (146–148) and crop yield prediction (149, 150). Thus, phenomics datasets are very large and have different formats (e.g. JavaScript Object Notation files). Some of the databases that host phenomics data include Genoplante Information System (GnpIS) (151, 152), Plant Genomics & Phenomics Research Data Repository (PGP) (153), Cartograplant (154), AgData commons (https://data.nal.usda.gov/ (155), PathoPlant (156, 157), PncStress (158) and Ozone Stress Responsive Gene Database (159).

Despite its analogy to genomes, it is not possible to fully characterize phenomes due to heterogeneity and multifaceted nature of phenotype data with added layers reflecting complexities at the cell, tissue and whole plant levels that have further variations according to the development stages and growth environment (145, 160). Thus, phenomics approaches may focus on specific factors of phenotypic data. For example, an intensive phenomics study may focus on high-throughput digital imaging across different stages and tissues of an organism in different growth stages or growth environments and may include quantitative data about plant height, biomass, flowering time, yield and photosynthesis efficiency. Another study may employ orthomosaic or time series Red, Green, Blue images and remote sensing to monitor the algal blooms in the ocean (161) or lesions in maize leaves (162). As phenomics data can be highly variable, necessary metadata includes information about plant species, tissue, developmental stage, environmental conditions, experimental design and data collection, processing, and analysis.

In addition to traditional phenotypes, molecular phenotypes include changes in the chromatin organization, transcripts, proteins, metabolites and ions (163–165). The quantitative changes in the gene expression, proteins and metabolite profiles in plants have far-reaching consequences for (i) the nutritional values of cereals, legumes, fruits and vegetables; (ii) the quality of bioproducts such as wine, beverages, vinegar, oil and fuel; (iii) the ability of plants to adapt in response to various abiotic stress conditions and (iv) the innate ability to defend against pests, pathogens and herbivores (166–170).

Proteome and metabolome datasets allow a deeper understanding of an organism's metabolic processes at the level of organ, tissue and cell, as well as how these processes change in response to intrinsic developmental programs and environmental factors. Proteome datasets further confirm the subcellular localization, their comparative abundance between different tissues and cells, protein–protein interactions and post-translational modifications (171). Once the original proteomic datasets and associated metadata/manuscript have been submitted to public data repositories such as Proteomics Identifications database (171–173), MassIVE (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp), Japan Proteome Standard Database (174, 175), Integrated Proteome

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

Table 3. List of public repositories, databases and secondary knowledgebases hosting or integrating various types of phenotypes, phenomics and molecular phenotype data

Category	Databases	URLs
Species-specific mutant collections	Database of images and genome (MaizeDIG)	https://maizedig.maizegdb.org/
	Mutant Variety Database	https://nucleus.iaea.org/sites/mvd/SitePages/Home.aspx
	Plant Genome Editing Database	http://plantcrispr.org/cgi-bin/crispr/index.cgi
	RIKEN Arabidopsis Genome Encyclopedia	http://rarge-v2.psc.riken.jp/line
	TOMATOMA	https://tomatoma.nbrp.jp/index.jsp
	Plant Editosome	https://ngdc.cncb.ac.cn/ped/
Traits and QTLs	Gramene QTL	https://archive.gramene.org/qtl/
	Wheatqtl	http://www.wheatqtldb.net/
	GLOPNET	http://bio.mq.edu.au/~iwright/glopian.htm
	TRY database	https://www.try-db.org/TryWeb/Home.php
	Ecological Flora of the Britain and Ireland	http://ecoflora.org.uk/
	BIOPOP	http://www.landeco.uni-oldenburg.de/Projects/biopop/main.htm
	FloraWeb	https://www.floraweb.de/
	USDA GRIN	https://www.ars-grin.gov/
	BiolFlor	https://wiki.ufz.de/biolflor/index.jsp
	LEDA	https://uol.de/en/landeco/research/leda
	USDA PLANTS	https://plants.usda.gov/home
	BROT	https://www.uv.es/jgpausas/brot.htm
	AusTraits	https://austraits.org/
	Community Databases in Table 2 and Supplementary Table S3	
Phenomics	GnpIS	https://urgi.versailles.inra.fr/gnpis
	PGP Repository	https://edal-pgp.ipk-gatersleben.de/
	Cartograplant	https://cartograplant.org/
	AgData commons Plants & Crops	https://data.nal.usda.gov/ag-data-commons-hierarchy/plants-crops
	PathoPlant	http://www.pathoplant.de/
	PncStress	http://bis.zju.edu.cn/pncstress/
	Indian Crop Phenome DB	https://ibdc.rcb.res.in/icpd/
Gene Expression	Ozone Stress Responsive Gene Database	https://www.osrgd.com
	EBI-Plant Expression Atlas	https://www.ebi.ac.uk/gxa/plant/experiments
	CoNeKT	https://conekt.sbs.ntu.edu.sg/
Protein, peptides and proteomes	Expath	http://expath.itps.ncku.edu.tw/
	Proteome Xchange	https://wwwz.proteomexchange.org
	Plant Proteome Database	http://ppdb.tc.cornell.edu/
	PlantMWpIDB	https://plantmwpidb.com/
	Heat Shock Proteins database	http://hsfdb.bio2db.com/
	WallProtDB	https://www.polebio.lrsv.ups-tlse.fr/WallProtDB/
	Aramemnon	http://aramemnon.botanik.uni-koeln.de/
	PhosPhAt	https://phosphat.uni-hohenheim.de/db.html
	Database of Phospho-sites in Plants	http://dbppt.biocuckoo.org/browse.php
	Plant Protein Phosphorylation Database	https://www.p3db.org/home
	qPTMplants	http://qptmplants.omicsbio.info/
	Plant PTM viewer	https://www.psb.ugent.be/webtools/ptm-viewer/
	PlaPPISite	http://zzdlab.com/plappisite/index.ph
	M. truncatula Small Secreted Peptide Database	https://mtsspdb.zhaolab.org/database
	PlantPepDB	http://14.139.61.8/PlantPepDB/index.php
	Arabidopsis PeptideAtlas	http://www.peptideatlas.org/builds/arabidopsis/
	Indian Structural Data Archive	https://isda.rcb.ac.in/
Metabolites, biochemical and small chemical entities	Antimicrobial plant peptides (PhytAMP)	http://phytamp.pfba-lab-tun.org/main.php
	PubChem	https://pubchem.ncbi.nlm.nih.gov
	ChEBI	https://www.ebi.ac.uk/chebi
	Metabolomics Workbench	https://www.metabolomicsworkbench.org
Secondary Knowledgebase	MetaboLights	https://www.ebi.ac.uk/metabolights/index
	PoDP	https://pairedomicsdata.bioinformatics.nl/
	Plant Reactome pathway knowledgebase	https://plantreactome.gramene.org
	MetaCyc	https://metacyc.org
	PMN	https://plantcyc.org/data
	KEGG pathways	https://www.genome.jp/kegg/pathway.html
	PlantPathMarks (PPMdb)	http://ppmdb.easyomics.org/
	The Bio-Analytic Resource	https://bar.utoronto.ca
	The protein–protein interaction database for Maize	https://mai.fudan.edu.cn/ppim/

Abbreviations: PTM, Post-translational modifications; PMN, Plant metabolic network; KEGG, Kyoto encyclopedia of genes and genomes.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

resources (176, 177), Panorama (178) and Peptide Atlas (179, 180), they are made available for re-analysis and further exploration by other researchers. Metabolomics provides a comprehensive overview of the metabolite profile of an organism, tissues, cells or subcellular component at a specific time point and is used to identify nutritional, medicinal, flavor and disease resistance compounds as well as chemical interactions between plants and other biological systems. A recent comprehensive review of the methodologies to explore the highly complex and diverse metabolites of plants and associated methodologies can be found in (181). The types of data collected for metabolomics depend on the method of chemical fingerprinting. As an example, in mass spectrometry (MS), a typical dataset would consist of a matrix containing information on the retention time and index, the mass-to-charge ratio (*m/z*) and peak characteristics such as the number and width. These data go through pre-processing, which converts raw instrument data into organized formats using background subtraction, noise reduction, curve resolution, peak picking, peak thresholding and spectral deconvolution. There are various software tools for analyzing metabolite data, each of which may be specific to a particular method of detection or instrument used in the analysis. The most popular software packages are MZmine, XCMS, MSdial, metaMS, Progenesis QI and MetAlign. For annotation of unknown metabolites, popular software tools include MS-FINDER, MetDNA, MetFamily and GNPS, among others. Raw file formats generated by the machines included raw, idb, cdf, wiff, scan, dat, cmp, cdf.cmp, lcd, abf, jpf, xps and mgf. Derived file formats are mzml, nmrml, mzxml, xml, mzdata, cef, cnx, peakml, xy, smp and scan. Several initiatives were undertaken due to the complexity of metabolomic data. The Chemical Analyses Working group started the 'Metabolomics Standard Initiative' to develop metabolomic standards (182, 183) with revisions suggested by (184). Community-driven Metabolomics Society has a Data Standards Task Group focusing on metabolomics data standardization and sharing. This was followed by the 'Coordination of Standards in Metabolomics' (185), and MetaboLights (186), for developing tools to ease the submission of metabolomic data (187). ProteomeCentral and Omics DI serve as central repositories for these datasets, which are then reused in protein knowledge bases (UniProt and NeXtProt), genome browsers (Ensembl and University of California Santa Cruz), proteomics resources and other bioinformatics resources (e.g. OpenProt and LNCipedia). The ProteomeXchange (PX) datasets are re-analyzed by different proteomics resources of the PX consortium, making data more reliable. The Paired Omics Data Platform (PoDP) (188) links the metabolomics data submitted to MassIVE or MetaboLights to genomes stored in NCBI or JGI. In Table 3, we list the two major repositories available for submission of raw and processed metabolome data, the National Institutes of Health Common Fund's National Metabolomics Data Repository portal and the Metabolomics Workbench and MetaboLights.

Some gene expression and metabolic phenotypes often culminate in visible phenotypes, which can be described using the Plant Ontology terms (189–191). More recently, Plant Ontology terms have been extended to large-scale phenomics data from a single species (192) to support the comparative phenomics in plants (193) and describe trait phenotypes expressed under a specific developmental stage or specific environment and stress (194). To cover the genotype–phenotype gap, we need to integrate multiple types of data, including genotypic, large-scale phenome, gene expression, proteome and metabolome data, described using defined and standardized ontologies.

Finally, making phenotype data FAIR requires developing additional public repositories and community guidelines for standardization and formatting phenotype data with well-described metadata. Furthermore, new tools and features are in development for the visualization of phenotype data on genome browser (195). The phenotype data and derived information can also be integrated into plant metabolic networks (196, 197), system-level plant pathways (198–200), expression Atlas, metabolic models and so on. The integration of genotype and phenotype information in the secondary knowledge bases is of primary importance to plant researchers for formulating data-driven hypotheses as well as for analyzing the high-throughput omics data (196). Here, we provide a list of public repositories and knowledgebases currently hosting various types of phenotype data in Table 3.

Phenotype data formats, standards and metadata

The structure and characteristics of data types and any additional metadata are crucial for enabling future data reuse and re-analysis by other researchers. The most relevant metadata shared across the various data types (generated by a diverse set of methods and platforms) includes taxonomic identification of the plant, the individual or cultivar name or accession ID, georeferences or growth conditions, field sampling or experimental design, cell, tissue, organ information (e.g. whole plant, leaf, root, flower, shoot, single cell, etc.), plant maturity and health status, measurement date (season, time of the day) and the type of phenotype measured (quantitative or qualitative) (137, 201). These metadata can be entered as a simple text format during the submission of the raw data to any primary repository and are easily exported from one database to another as text files.

Furthermore, plant phenotypes can be classified as categorical (qualitative and ordinal) or quantitative (continuous) traits (15). Some phenotypes are rather stable within species (mostly categorical traits), and some of these can be systematically compiled from species checklists and floras (202). Thus, not all phenotypes can be mapped from one species to another. It is also important to note here that often, a phenotype is a cumulative outcome of the genotype, the environment and their interaction. Many important agronomic traits, such as seed or fruit quality, yield, abiotic stress tolerance and pathogen resistance, have a quantitative genetic architecture involving minor and major genes or QTLs. Thus, the research question and the method become important to set the scope and goals of the study and require specific metadata and standards. For instance, most traits relevant to ecology and earth system sciences are characterized by intraspecific variability and trait–environment relationships (mostly quantitative traits). These traits must be measured on individual plants in their particular environmental context. Each such trait measurement has high information content as it captures the specific response of a given genome to the prevailing environmental conditions (137). Thus, the collection of these quantitative phenotype data and their essential environmental covariates is of vital importance. While trait measurements themselves

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

may be relatively simple, the selection of the adequate entity (e.g. a representative plant in a community or a representative leaf on a tree) and obtaining the relevant ancillary data (taxonomic identification, soil and climate properties, disturbance history, etc.) may require sophisticated instruments and a high degree of expertise and experience. Besides, these data are most often individual measurements with a low degree of automation. This limits the number of measurements and causes a high risk of errors, which need to be corrected *a posteriori*, requiring substantial human work. Hence, the integration of these data from different sources into a consistent dataset requires a carefully designed workflow with sufficient data quality assurance. These measurements of quantitative traits are single sampling events for particular individuals at certain locations and times, which preserve relevant information on intraspecific variation and provide the necessary detail to address questions at the level of populations or communities (201). Hence, an accurate and careful collection of data, including their associated metadata and ancillary data, is key to correctly preserving this valuable information and performing a suitable data integration across studies, species and data types.

GWAS and QTL mapping

GWAS (203–206) and QTL (207–209) mapping are statistical methods used to identify marker-trait associations and candidate genes (causative mutations) controlling traits of interest (210). Both approaches rely on the linkage disequilibrium between the tested markers and the functional polymorphisms at the causative genes (211, 212). However, they differ on the type of genetic populations used for the study (204, 213): GWAS relies on a diversity panel (*e.g.* germplasm collections) of, ideally, unrelated individuals; on the contrary, QTL mapping investigates the co-segregation of genetic markers with desired phenotypes in progeny purposely generated (e.g. F_2 population or recombinant inbred lines) (210). Although both these analytical methods, their results can be used as data inputs for other types of analysis (e.g. meta-analysis, estimation of polygenic scores) (214). The genomic and genetic positions of trait-associated markers from GWAS and QTL studies can also be integrated with other types of data, enabling data transfer among related species. Thus, their outputs can be considered a data type, and consequently, they require metadata collection and the use of standards in order to make them FAIR. Therefore, the FAIRness of the association mapping outputs is vital in linking genotype and phenotype in the multi-omics era.

The primary output of a GWAS analysis is a list of variant positions, SNP ID or indel positions, allele, strand information, effect size and associated standard error, *P*-value and corrected *p*-value, test statistics, minor allele frequency and sample size (215). One critical metadata for GWAS/QTL data is the statistical method used to calculate and correct the *P*-values (GWAS/QTL). Regarding the SNPs, the most important metadata includes the model species and the version of the reference genome against which these SNPs are mapped (refer to the Genotype data section). The metadata required to make the traits interoperable and reusable is explained in the laboratory/field traits section. In the case of QTL analysis, a linkage map and pedigree information of the individuals, as well as the heritability of each SNP, is also important to be collected (115).

Unlike the human and animal GWAS and QTL data, open access resources such as the National Human Genome Research Institute-EBI GWAS Catalog, GWAS Atlas (216), OpenGWAS, Animal QTL database, and Animal Genome Informatics resources (USDA national infrastructure National Research Support Project: A National Animal Genome Research Program), QTL and GWAS data for plant species and major crops are mostly stored in crop community database (Table 2). The databases typically integrate the QTL and GWAS data with other types of data, playing a crucial role in improving the findability and accessibility of plant GWAS data that would have otherwise been buried in publications. AraGWAS Catalog (216) contains recomputed GWAS results using a standardized GWAS pipeline on all publicly available phenotypes from AraPheno (217).

Meta-analysis is a widely used analysis for integrating the summary statistics from multiple GWAS/QTL studies (218, 219). It is a set of methods that allows the quantitative combination of data from numerous studies and the evaluation of the consistency, inconsistency or heterogeneity of the results across multiple datasets (218). Meta-analysis of GWAS/QTL datasets can improve the power to detect association signals by increasing sample size and examining more variants throughout the genome than each dataset alone (220). However, in order to integrate datasets coming from different studies in meta-analysis, standardized data and metadata collection across the studies are needed (16). In addition, the genotype and phenotype data from the GWAS/QTL studies can be reused for further knowledge discovery, especially for QTL by environment interaction, predicting plant response in new environments and linking genomes to complex phenotypes across species (221, 222).

Data reusability limitations and challenges

Accessing, reusing and integrating analytic data from various data types remain difficult (223). Despite the significant progress made in agricultural research due to advances in genotyping and phenotyping technologies, most of the data used and generated in research studies are not shared. Here, we discuss limitations to data reuse in genotype-to-phenotype studies in three aspects: challenges with data, resources and funding and implementation of FAIR data policy.

Challenges with data

Data diversity and data format heterogeneity

Agriculture and horticulture research involves a wide range of genotypic, phenotypic and environmental data, often from different experimental protocols, data generation technologies and data processing workflows. As a result, data formats can be highly heterogeneous, making it challenging to integrate data from different sources and reuse them in future studies (137). This issue is even more significant for phenotypic data, especially with emerging high-throughput phenotyping technologies. Digital imaging and remote sensing allow researchers to explore new levels of trait variability that were previously inaccessible using traditional and manual phenotyping methods. However, the large diversity of data and metadata generated by these technologies can be highly variable in terms of file size, format and content. The heterogeneity of the data analysis pipeline

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

also contributes to the complexity of standardization in phenomics.

Data size, quality and versioning

Most genomics, transcriptomics, epigenetics and phenomics data are extremely large in file size and computationally intensive. For example, whole-genome sequencing data used for variant calling or VCF files that collate multi-individual genome-wide variants can be computationally challenging to handle, limiting their sharing in FAIR public repositories and making data manipulation difficult. Also, data quality and integrity may be compromised before or during the submission process, preventing reuse.

Object identification

Data submitted to a public domain often lack a unique data object identifier (DOI) and any plant or accession identifier (PID), which makes it challenging to trace and integrate different types of data generated from the same individual plant across experiments and research laboratories. Having a universal DOI associated with its PID would be desirable to improve data findability and reuse. However, most data used and generated in research studies are not shared, inaccessible or reusable because of missing fundamental metadata or improper data format.

Metadata and data standardization

Metadata is any type of data descriptor that can facilitate data interpretation and reuse. It is very common that when data are submitted to public domains they are accompanied by incomplete, inconsistent or missing metadata. Developing and promoting standard data formats and metadata can improve data discovery and reuse, facilitate data integration and interoperability and allow data from different sources. Some data standards for genomics and phenomics data have been developed, such as the Minimum Information About a Genome Sequence from the Genomic Data Standards Consortium, the Plant Phenotype Ontology and the Minimum Information About a Plant Phenotyping Experiment. For GWAS data, GWAS-VCF format has been proposed. However, promoting and consistently applying these standards across different research groups and databases remain challenging. For instance, if there are standards for collecting and describing trait measurements, they are organism-specific (e.g. International Organization of Vine and Wine; www.oiv.int) or based on model species.

The metabolomic research community faces similar challenges. The USA Plant, Algae, and Microbial Metabolomics Research Coordination Network (224) coordinated an initiative to identify the grand challenges of metabolomic research. As noted, the data obtained from metabolomic analyses can often result in different chemical feature values even in the same biological treatments due to the variability associated with biological systems, equipment differences and protocol and reagents. Therefore, identifying metabolites with confidence and the limited metabolome depth of coverage are the key challenges in metabolomic research (225). A recent review of liquid chromatography-mass spectrometry (LC-MS) literature found a lack of details reported on the methodology and level of confidence for metabolites in most of the reviewed research articles (226). To address these challenges, multidimensional analysis methods, the use of standard libraries for metabolite characterization and tools that simplify the submission of metadata and data are being developed (187, 227).

Resources and funding

The submission of different data types (i.e. genomics, transcriptomics, proteomics, metabolomics and phenomics data) to separate and specialized primary repositories is a common practice, resulting in a heterogeneity of data repositories and multiple PIDs, limiting data interoperability. It is challenging to locate phenotypic datasets for a particular set of plants that have been characterized at the genomic or transcriptomic level due to the absence of common standards among data repositories.

Incompatible software or hardware among different data platforms also makes interoperability challenging. Bulk data download or data movement across repositories is another issue due to data size and a lack of standardization. Software development and maintenance are required for fast data search and retrieval and sufficient user support. For some types of plant data, such as QTL and GWAS, there are no primary repositories where researchers can submit their data. Community GGB databases (Table 2) address this issue by collecting, curating and integrating various data types from different sources and related species, which is pivotal in data integration. However, not all plant GWAS data are timely stored in databases due to either a lack of crop community databases or funding for curation. In addition, community GGB databases often have limited computational and personnel resources for curation and inclusion of all types of omics data due to limited funding and lack of understanding of the importance of curation by funders. Additionally, there is a lack of appropriate infrastructure for the raw data deposition in community databases.

Implementation of FAIR data policy

FAIR data policy refers to the list of 15 guidelines elaborated to facilitate data search, access and reuse by human-driven and machine-driven activities (228). These principles apply to every type of scholarly digital object archived in a repository, and their implementation has started in many different research fields (224, 229, 230). In summary, these principles recommend that, when data are submitted, they are very well described using detailed metadata and are assigned a globally unique and persistent identifier that allows everybody to find them in a searchable resource. Data should be formatted according to community-based standards if available or in a way that human and computer systems can easily interpret and exchange. Controlled vocabularies and ontologies are strongly encouraged to facilitate data interoperability across database resources. FAIR data, however, do not mean open access or free but refer to clarity and transparency about the conditions governing access and reuse (e.g. credential system to access and download data (231). These principles aim to increase data transparency and improve data reuse for new research purposes, enhancing data value over time.

The implementation of the FAIR data policy, however, can be challenging due to several reasons. First, making data FAIR requires additional efforts and time commitment from researchers for which adequate training and funding may not exist. Second, many scientists are unaware of the FAIR principles, community-based standards, ontologies and

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

public databases. Third, there is also a need for long-term sustainability of databases and bioinformatic training/outreach, which requires ongoing funding and infrastructure support. Many databases struggle to secure funding and may face difficulties in maintaining FAIR data and providing training to its users. Other barriers for FAIR data include considerations of data privacy and confidentiality, legal and ethical issues, concerns of ownership and lack of incentive if not credited for sharing data.

Recommendations

Big data generated by recent high-throughput sequencing and phenotyping technologies allow researchers to use datasets to explore how an organism's genetic code influences its physical traits. To accelerate G2P research, we propose the following recommendations for ensuring data interoperability and reuse for discovering new knowledge and promoting translational research.

(i) Standardization of data collection protocols: Standardizing data collection protocols and using common data formats are recommended to be developed by each crop community to ensure that data are collected consistently and comparably. Using metadata standards will make sharing and comparing data across different studies easier.

(ii) Centralized or interoperable data sharing platform: It is beneficial to have a centralized data sharing platform, but it is also recognized that multiple database resources can be built with different strengths. It is recommended for these resources to use standardized data models and exchange formats and the deployment of existing and emerging software components to facilitate the sharing of genotypic and phenotypic data. It includes the use of online databases and repositories that are specifically designed for the storage and sharing of plant genetic and phenotypic data.

(iii) Consistent data annotation: It is recommended that researchers consistently annotate data with relevant information such as the genotype, phenotype and experimental treatments to make the data more easily searchable and usable by other researchers.

(iv) Data QC: It is recommended that researchers use more automated management of data flows and implement data QC such as data curation and validation to ensure that the data are accurate, are reliable and can be used to make valid conclusions.

(v) Data integration: It is recommended that databases adopt new database technologies and develop robust data standards that can facilitate the global integration of G2P data in the future. Data integration from different resources such as genomics, transcriptomics, proteomics and metabolomics can help to better understand the complex relationship between the genotype and the phenotype.

(vi) Community-driven efforts: It is recommended that researchers and funders make more community-driven efforts such as open-source projects, workshops and collaborations that can help to promote the sharing and use of data among researchers, which in turn will lead to a better understanding of the G2P relationship. There should be encouragement for integrated science training plans that enable biologists to think quantitatively and facilitate collaboration with experts in physical, computational and engineering sciences. It can help scientists get familiar with the development of computational pipelines and workflows that will be essential for researchers to acquire, analyze and critically interpret G2P data.

(vii) Data storage infrastructure, data management software and data curation tools: Funders are recommended to recognize that these tools are necessary to handle large volumes of data in diverse formats and have researchers to have a separate funding for this type of work, ideally in collaboration with the existing community databases instead of reinventing the wheel.

(viii) A concerted effort to make multi-omics datasets interoperable by biocuration with controlled ontology terms will help address this issue. Community databases address some of these issues by collecting, curating and integrating various data of different types from different sources and from different but related species. However, community databases need to have sustainable funding.

(ix) Data security, backup and recovery must be considered and implemented for sustainability.

(x) Data compliance with data sharing policies, privacy regulations and laws should be enforced.

Supplementary Material

Supplementary material is available at *Database* online.

Acknowledgements

We acknowledge the funding to AgBioData Consortium through the National Science Foundation (16) for the Research Coordination Network (RCN) project (award abstract #2126334), USDA National Institute of Food and Agriculture Specialty Crop Research Initiative project [2022-51181-38449], and USDA National Institute of Food and Agriculture National Research Support Project 10.

Data availability

No new data were generated or analysed in support of this research.

Conflict of interest

None declared.

References

1. Scossa,F., Alseekh,S. and Fernie,A.R. (2021) Integrating multi-omics data for crop improvement. *J. Plant Physiol.*, **257**, 153352.
2. Yang,W., Feng,H., Zhang,X. *et al.* (2020) Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol. Plant*, **13**, 187–214.
3. Borgman,C.L. (2015) *Big Data, Little Data, No Data: Scholarship in the Networked World. The MIT Press, Cambridge, MA*.
4. Mosconi,G., Li,Q., Randall,D. *et al.* (2019) Three gaps in opening science. *Comput. Support Coop. Work (CSCW)*, **28**, 749–789.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

5. Federer,L.M. (2019) Who, what, when, where, and why? Quantifying and understanding biomedical data reuse. University of Maryland.
6. Wallis,J.C., Rolando,E. and Borgman,C.L. (2013) If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS One*, **8**, e67332.
7. Pasquetto,I.V., Randles,B.M. and Borgman,C.L. (2017) On the reuse of scientific data. *Data Sci. J.*, **16**, 1–9.
8. Culina,A., Crowther,T.W., Ramakers,J.J.C. *et al.* (2018) How to do meta-analysis of open datasets. *Nat. Ecol. Evol.*, **2**, 1053–1056.
9. He,L. and Nahar,V. (2016) Reuse of scientific data in academic publications: an investigation of Dryad digital repository. *J. Inf. Manag.*, **65**, 478–494.
10. Pasquetto,I.V., Borgman,C.L. and Wofford,M.F. (2019) Uses and reuses of scientific data: the data creators' advantage. *Harv. Data Sci. Rev.*, **1**.
11. Rung,J. and Brazma,A. (2013) Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.*, **14**, 89–99.
12. Karasti,H. and Blomberg,J. (2018) Studying infrastructuring ethnographically. *Comput. Support. Coop. Work (CSCW)*, **27**, 233–265.
13. Hanson,B., Sugden,A. and Alberts,B. (2011) Making data maximally available. *Science*, **331**, 649.
14. Leonelli,S. (2013) Integrating data to acquire new knowledge: three modes of integration in plant science. *Stud. Hist. Philos. Sci. Part C*, **44**, 503–514.
15. Kattge,J., Bonisch,G., Diaz,S. *et al.* (2020) TRY plant trait database – enhanced coverage and open access. *Glob. Chang. Biol.*, **26**, 119–188.
16. Harper,L., Campbell,J., Cannon,E.K.S. *et al.* (2018) AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. *Database*, **2018**, bay088.
17. Adam-Blondon,A.F., Alaux,M., Pommier,C. *et al.* (2016) Towards an open grapevine information system. *Hortic. Res.*, **3**, 16056.
18. Dempsey,L. and Heery,R. (1998) Metadata: a current view of practice and issues. *J. Doc.*, **54**, 145–172.
19. Mayernik,M.S. and Acker,A. (2018) Tracing the traces: the critical role of metadata within networked communications. *J. Assoc. Inf. Sci. Technol.*, **69**, 177–180.
20. Edwards,D. (2016) The impact of genomics technology on adapting plants to climate change. In: Edwards D, Batley J (eds) *Plant Genomics and Climate Change*. Springer, New York, NY, pp. 173–178.
21. Hu,T., Chitnis,N., Monos,D. *et al.* (2021) Next-generation sequencing technologies: an overview. *Hum. Immunol.*, **82**, 01–811.
22. Smith,L.M., Fung,S., Hunkapiller,M.W. *et al.* (1985) The synthesis of oligonucleotides containing an aliphatic amino group at the 5′ terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res.*, **13**, 2399–2412.
23. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 5463–5467.
24. Crossley,B.M., Bai,J., Glaser,A. *et al.* (2020) Guidelines for Sanger sequencing and molecular assay monitoring. *J. Vet. Diagn. Invest.*, **32**, 767–775.
25. Mardis,E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.*, **24**, 133–141.
26. van Dijk,E.L., Auger,H., Jaszczyszyn,Y. *et al.* (2014) Ten years of next-generation sequencing technology. *Trends Genet.*, **30**, 418–426.
27. Buermans,H.P. and den Dunnen,J.T. (2014) Next generation sequencing technology: advances and applications. *Biochim. Biophys. Acta*, **1842**, 1932–1941.
28. Slatko,B.E., Gardner,A.F. and Ausubel,F.M. (2018) Overview of next-generation sequencing technologies. *Curr. Protoc. Mol. Biol.*, **122**, e59.
29. Ekblom,R. and Wolf,J.B. (2014) A field guide to whole-genome sequencing, assembly and annotation. *Evol. Appl.*, **7**, 1026–1042.
30. English,A.C., Richards,S., Han,Y. *et al.* (2012) Mind the gap: upgrading genomes with pacific biosciences RS long-read sequencing technology. *PLoS One*, **7**, e47768.
31. Huddleston,J., Ranade,S., Malig,M. *et al.* (2014) Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.*, **24**, 688–696.
32. Wang,Y., Zhao,Y., Bollas,A. *et al.* (2021) Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.*, **39**, 1348–1365.
33. Marx,V. (2023) Method of the year: long-read sequencing. *Nat. Methods*, **20**, 6–11.
34. Chen,P., Sun,Z., Wang,J. *et al.* (2023) Portable nanopore-sequencing technology: trends in development and applications. *Front Microbiol.*, **14**, 1043967.
35. Wick,R.R., Judd,L.M. and Holt,K.E. (2019) Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.*, **20**, 129.
36. Grodzicker,T., Williams,J., Sharp,P. *et al.* (1975) Physical mapping of temperature-sensitive mutations of adenoviruses. *Cold Spring Harb. Symp. Quant. Biol.*, **39**, 439–446.
37. Yang,W., Kang,X., Yang,Q. *et al.* (2013) Review on the development of genotyping methods for assessing farm animal diversity. *J. Anim. Sci. Biotechnol.*, **4**, 2.
38. Carvalho,B., Bengtsson,H., Speed,T.P. *et al.* (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, **8**, 485–499.
39. Chagne,D., Crowhurst,R.N., Troggio,M. *et al.* (2012) Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS One*, **7**, e31745.
40. Bayer,M.M., Rapazote-Flores,P., Ganal,M. *et al.* (2017) Development and evaluation of a barley 50k iSelect SNP Array. *Front. Plant Sci.*, **8**, 1792.
41. Verde,I., Jenkins,J., Dondini,L. *et al.* (2017) The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genomics*, **18**, 225.
42. Ganal,M.W., Polley,A., Graner,E.M. *et al.* (2012) Large SNP arrays for genotyping in crop plants. *J. Biosci.*, **37**, 821–828.
43. McKain,M.R., Johnson,M.G., Uribe-Convers,S. *et al.* (2018) Practical considerations for plant phylogenomics. *Appl. Plant Sci.*, **6**, e1038.
44. Kumar,P., Choudhary,M., Jat,B.S. *et al.* (2021) Skim sequencing: an advanced NGS technology for crop improvement. *J. Genet.*, **100**, 1–10.
45. Schmickl,R., Liston,A., Zeisek,V. *et al.* (2016) Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African Oxalis (Oxalidaceae). *Mol. Ecol. Resour.*, **16**, 1124–1135.
46. Head,S.R., Komori,H.K., LaMere,S.A. *et al.* (2014) Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, **56**, 61–64, 66, 68, passim.
47. Deschamps,S., Llaca,V. and May,G.D. (2012) Genotyping-by-Sequencing in Plants. *Biology*, **1**, 460–483.
48. Elshire,R.J., Glaubitz,J.C., Sun,Q. *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.
49. Andrews,K.R., Good,J.M., Miller,M.R. *et al.* (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.*, **17**, 81–92.
50. Miller,M.R., Dunham,J.P., Amores,A. *et al.* (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.*, **17**, 240–248.
51. Danecek,P., Auton,A., Abecasis,G. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

52. Lyon,M.S., Andrews,S.J., Elsworth,B. *et al.* (2021) The variant call format provides efficient and robust storage of GWAS summary statistics. *Genome Biol.*, **22**, 32.
53. Leinonen,R., Sugawara,H., Shumway,M. *et al.* (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–21.
54. Kodama,Y., Shumway,M., Leinonen,R. *et al.* (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–56.
55. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
56. Barrett,T. and Edgar,R. (2006) Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol.*, **411**, 352–369.
57. Clough,E. and Barrett,T. (2016) The gene expression omnibus database. In: *Statistical Genomics: Methods and Protocols*, pp. 93–110.
58. Tateno,Y., Imanishi,T., Miyazaki,S. *et al.* (2002) DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res.*, **30**, 27–30.
59. Miyazaki,S., Sugawara,H., Ikeo,K. *et al.* (2004) DDBJ in the stream of various biological data. *Nucleic Acids Res.*, **32**, D31–34.
60. Ogasawara,O., Kodama,Y., Mashima,J. *et al.* (2020) DDBJ Database updates and computational infrastructure enhancement. *Nucleic Acids Res.*, **48**, D45–D50.
61. Cochrane,G., Karsch-Mizrachi,I., Nakamura,Y. *et al.* (2011) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **39**, D15–18.
62. Cochrane,G., Karsch-Mizrachi,I., Takagi,T. *et al.* (2016) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **44**, D48–50.
63. (2020) Promoting best practice in nucleotide sequence data sharing. *Sci. Data*, **7**, 152.
64. Nordberg,H., Cantor,M., Dusheyko,S. *et al.* (2014) The genome portal of the department of energy joint genome institute: 2014 updates. *Nucleic Acids Res.*, **42**, D26–31.
65. Sreedasyam,A., Plott,C., Hossain,M.S. *et al.* (2023) JGI Plant Gene Atlas: an updateable transcriptome resource to improve functional gene descriptions across the plant kingdom. *Nucleic Acids Res.*, **51**, 8383–8401.
66. Goodstein,D.M., Shu,S., Howson,R. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–1186.
67. Members,C.-N. and Partners (2021) Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2021. *Nucleic Acids Res.*, **49**, D18–D28.
68. Cezard,T., Cunningham,F., Hunt,S.E. *et al.* (2022) The European Variation Archive: a FAIR resource of genomic variation for all species. *Nucleic Acids Res.*, **50**, D1216–D1220.
69. Song,S., Tian,D., Li,C. *et al.* (2018) Genome Variation Map: a data repository of genome variations in BIG Data Center. *Nucleic Acids Res.*, **46**, D944–D949.
70. Chang,Y., Song,X., Zhang,Q. *et al.* (2022) Robust CRISPR/Cas9 mediated gene editing of JrWOX11 manipulated adventitious rooting and vegetative growth in a nut tree species of walnut. *Sci. Hortic.*, **303**, 111199.
71. International Hapmap,C. (2003) The International HapMap Project. *Nature*, **426**, 789–796.
72. Jung,S., Jesudurai,C., Staton,M. *et al.* (2004) GDR (Genome Database for Rosaceae): integrated web resources for Rosaceae genomics and genetics research. *BMC Bioinf.*, **5**, 130.
73. Jung,S., Lee,T., Cheng,C.H. *et al.* (2019) 15 years of GDR: new data and functionality in the Genome Database for Rosaceae. *Nucleic Acids Res.*, **47**, D1137–D1145.
74. Yu,J., Jung,S., Cheng,C.H. *et al.* (2014) CottonGen: a genomics, genetics and breeding database for cotton research. *Nucleic Acids Res.*, **42**, D1229–1236.
75. Yu,J., Jung,S., Cheng,C.H. *et al.* (2021) CottonGen: the community database for cotton genomics, genetics, and breeding research. *Plants*, **10**, 2805.
76. Grant,D., Nelson,R.T., Cannon,S.B. *et al.* (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.*, **38**, D843–846.
77. Brown,A.V., Conners,S.I., Huang,W. *et al.* (2021) A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.*, **49**, D1496–D1501.
78. Gonzales,M.D., Archuleta,E., Farmer,A. *et al.* (2005) The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic Acids Res.*, **33**, D660–665.
79. Dash,S., Campbell,J.D., Cannon,E.K. *et al.* (2016) Legume information system (LegumeInfo.org): a key component of a set of federated data resources for the legume family. *Nucleic Acids Res.*, **44**, D1181–1188.
80. Fernandez-Pozo,N., Menda,N., Edwards,J.D. *et al.* (2015) The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. *Nucleic Acids Res.*, **43**, D1036–1041.
81. Foerster,H., Bombarely,A., Battey,J.N.D. *et al.* (2018) SolCyc: a database hub at the Sol Genomics Network (SGN) for the manual curation of metabolic networks in Solanum and Nicotiana specific databases. *Database (Oxford)*, **2018**, bay035.
82. Lawrence,C.J. (2007) MaizeGDB. *Methods Mol. Biol.*, **406**, 331–345.
83. Portwood,J.L., 2nd, Woodhouse,M.R., Cannon,E.K. *et al.* (2019) MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Res.*, **47**, D1146–D1154.
84. Wegrzyn,J.L., Lee,J.M., Tearse,B.R. *et al.* (2008) TreeGenes: a forest tree genome database. *Int. J. Plant Genomics*, **2008**, 412875.
85. Falk,T., Herndon,N., Grau,E. *et al.* (2019) Growing and cultivating the forest genomics database, TreeGenes. *Database*, **2019**, bay084.
86. Garcia-Hernandez,M., Berardini,T.Z., Chen,G. *et al.* (2002) TAIR: a resource for integrated Arabidopsis data. *Funct. Integr. Genomics*, **2**, 239–253.
87. Poole,R.L. (2007) The TAIR database. *Methods Mol. Biol.*, **406**, 179–212.
88. Sanderson,L.A., Caron,C.T., Tan,R. *et al.* (2019) KnowPulse: A web-resource focused on diversity data for pulse crop improvement. *Front. Plant Sci.*, **10**, 965.
89. Smith,R.N., Aleksic,J., Butano,D. *et al.* (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, **28**, 3163–3165.
90. Kalderimis,A., Lyne,R., Butano,D. *et al.* (2014) InterMine: extensive web services for modern biology. *Nucleic Acids Res.*, **42**, W468–472.
91. Tello-Ruiz,M.K., Jaiswal,P. and Ware,D. (2022) Gramene: a resource for comparative analysis of plants genomes and pathways. *Methods Mol. Biol.*, **2443**, 101–131.
92. Ware,D. (2007) Gramene. *Methods Mol. Biol.*, **406**, 315–329.
93. Ware,D.H., Jaiswal,P., Ni,J. *et al.* (2002) Gramene, a tool for grass genomics. *Plant Physiol.*, **130**, 1606–1613.
94. Gladman,N., Olson,A., Wei,S. *et al.* (2022) SorghumBase: a web-based portal for sorghum genetic information and community advancement. *Planta*, **255**, 35.
95. Lyne,R., Sullivan,J., Butano,D. *et al.* (2015) Cross-organism analysis using InterMine. *Genesis*, **53**, 547–560.
96. Paajanen,P., Kettleborough,G., Lopez-Girona,E. *et al.* (2019) A critical comparison of technologies for a plant genome sequencing project. *Gigascience*, **8**, giy163.
97. Sun,Y., Shang,L., Zhu,Q.H. *et al.* (2022) Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.*, **27**, 391–401.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

98. Pucker,B., Irisarri,I., de Vries,J. *et al.* (2022) Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. *Quant. Plant Biol.*, **3**, e5.
99. Shi,J., Tian,Z., Lai,J. *et al.* (2023) Plant pan-genomics and its applications. *Mol. Plant*, **16**, 168–186.
100. Ho,S.S., Urban,A.E. and Mills,R.E. (2020) Structural variation in the sequencing era. *Nat. Rev. Genet.*, **21**, 171–189.
101. Quan,C., Lu,H., Lu,Y. *et al.* (2022) Population-scale genotyping of structural variation in the era of long-read sequencing. *Comput. Struct. Biotechnol. J.*, **20**, 2639–2647.
102. Sun,S., Wang,X., Wang,K. *et al.* (2020) Dissection of complex traits of tomato in the post-genome era. *Theor. Appl. Genet.*, **133**, 1763–1776.
103. Lye,Z.N. and Purugganan,M.D. (2019) Copy number variation in domestication. *Trends Plant Sci.*, **24**, 352–365.
104. Hovhannisyan,G., Harutyunyan,T., Aroutiounian,R. *et al.* (2019) DNA copy number variations as markers of mutagenic impact. *Int. J. Mol. Sci.*, **20**, 4723.
105. Dolatabadian,A., Patel,D.A., Edwards,D. *et al.* (2017) Copy number variation and disease resistance in plants. *Theor. Appl. Genet.*, **130**, 2479–2490.
106. Yuan,Y., Bayer,P.E., Batley,J. *et al.* (2021) Current status of structural variation studies in plants. *Plant Biotechnol. J.*, **19**, 2153–2163.
107. Alonge,M., Wang,X., Benoit,M. *et al.* (2020) Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, **182**, 145–161 e123.
108. Chawla,H.S., Lee,H., Gabur,I. *et al.* (2021) Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant. *Plant Biotechnol. J.*, **19**, 240–250.
109. Li,M., Xia,L., Zhang,Y. *et al.* (2019) Plant editosome database: a curated database of RNA editosome in plants. *Nucleic Acids Res.*, **47**, D170–D174.
110. Thao,N.P. and Tran,L.S. (2016) Enhancement of plant productivity in the post-genomics era. *Curr. Genomics*, **17**, 295–296.
111. Pan,Q., Wei,J., Guo,F. *et al.* (2019) Trait ontology analysis based on association mapping studies bridges the gap between crop genomics and Phenomics. *BMC Genomics*, **20**, 443.
112. Danecek,P., Bonfield,J.K., Liddle,J. *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**, giab008.
113. Brachi,B., Morris,G.P. and Borevitz,J.O. (2011) Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol.*, **12**, 232.
114. Gali,K.K., Sackville,A., Tafesse,E.G. *et al.* (2019) Genome-wide association mapping for agronomic and seed quality traits of field pea (Pisum sativum L.). *Front. Plant Sci.*, **10**, 1538.
115. Khan,S.U., Saeed,S., Khan,M.H.U. *et al.* (2021) Advances and challenges for QTL analysis and GWAS in the plant-breeding of high-yielding: a focus on rapeseed. *Biomolecules*, **11**, 1516.
116. Tibbs Cortes,L., Zhang,Z. and Yu,J. (2021) Status and prospects of genome-wide association studies in plants. *Plant Genome.*, **14**, e20077.
117. Liu,J., Hua,W., Hu,Z. *et al.* (2015) Natural variation in ARF18 gene simultaneously affects seed weight and silique length in polyploid rapeseed. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E5123–5132.
118. Christeller,J.T., McGhie,T.K., Johnston,J.W. *et al.* (2019) Quantitative trait loci influencing pentacyclic triterpene composition in apple fruit peel. *Sci. Rep.*, **9**, 18501.
119. Chagné,D., Ryan,J., Saeed,M. *et al.* (2019) A high density linkage map and quantitative trait loci for tree growth for New Zealand mānuka (Leptospermum scoparium). *N. Z. J. Crop Hortic. Sci.*, **47**, 261–272.
120. Budhlakoti,N., Kushwaha,A.K., Rai,A. *et al.* (2022) Genomic selection: a tool for accelerating the efficiency of molecular breeding for development of climate-resilient crops. *Front. Genet.*, **13**, 832153.
121. Bhat,J.A., Ali,S., Salgotra,R.K. *et al.* (2016) Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Front. Genet.*, **7**, 221.
122. Crossa,J., Perez-Rodriguez,P., Cuevas,J. *et al.* (2017) Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.*, **22**, 961–975.
123. Fasoula,D.A., Ioannides,I.M. and Omirou,M. (2019) Phenotyping and plant breeding: overcoming the barriers. *Front. Plant Sci.*, **10**, 1713.
124. Akiyama,K., Kurotani,A., Iida,K. *et al.* (2014) RARGE II: an integrated phenotype database of Arabidopsis mutant traits using a controlled vocabulary. *Plant Cell Physiol.*, **55**, e4.
125. Miroslaw,M. (2001) Officially Released Mutant Varieties – The FAO/IAEA Database. *Plant Cell Tissue Organ. Cult.*, **65**, 175–177.
126. Zheng,Y., Zhang,N., Martin,G.B. *et al.* (2019) Plant Genome Editing Database (PGED): a call for submission of information about genome-edited plant Mutants. *Mol. Plant*, **12**, 127–129.
127. Shikata,M., Hoshikawa,K., Ariizumi,T. *et al.* (2016) TOMATOMA update: phenotypic and metabolite information in the micro-tom mutant resource. *Plant Cell Physiol.*, **57**, e11.
128. McGill,B.J., Enquist,B.J., Weiher,E. *et al.* (2006) Rebuilding community ecology from functional traits. *Trends Ecol. Evol.*, **21**, 178–185.
129. Violle,V., Navas,M., Vile,D. *et al.* (2007) Let the concept of trait be functional! *Oikos*, **116**, 882–892.
130. Schneider,F.D., Fichtmueller,D., Gossner,M.M. *et al.* (2019) Towards an ecological trait-data standard. *Meth. Ecol. Evolut*, **10**, 2006–2019.
131. Allan,E., Manning,P., Alt,F. *et al.* (2015) Land use intensification alters ecosystem multifunctionality via loss of biodiversity and changes to functional composition. *Ecol. Lett.*, **18**, 834–843.
132. Diaz,S., Quetier,F., Caceres,D.M. *et al.* (2011) Linking functional diversity and social actor strategies in a framework for interdisciplinary analysis of nature's benefits to society. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 895–902.
133. Lavorel,S. and Grigulis,K. (2012) How fundamental plant functional trait relationships scale-up to trade-offs and synergies in ecosystem services. *J. Ecol.*, **100**, 128–140.
134. Ni,J., Pujar,A., Youens-Clark,K. *et al.* (2009) Gramene QTL database: development, content and applications. *Database (Oxford)*, **2009**, bap005.
135. Singh,K., Batra,R., Sharma,S. *et al.* (2021) WheatQTLdb: a QTL database for wheat. *Mol. Genet. Genomics*, **296**, 1051–1056.
136. Reich,P.B., Wright,I.J. and Lusk,C.H. (2007) Predicting leaf physiology from simple plant and climate attributes: a global GLOPNET analysis. *Ecol. Appl.*, **17**, 1982–1988.
137. Kissling,W.D., Walls,R., Bowser,A. *et al.* (2018) Towards global data products of Essential Biodiversity Variables on species traits. *Nat. Ecol. Evol.*, **2**, 1531–1540.
138. Peat,H.J. and Fitter,A.H. (1994) A comparative study of the distribution and density of stomata in the British flora. *Biol. J. Linn. Soc. Lond.*, **52**, 377–393.
139. Poschlod,P., Kleyer,M., Jackel,A.-K. *et al.* (2003) BIOPOP — A database of plant traits and internet application for nature conservation. *Folia Geobot.*, **38**, 263–271.
140. Garcia-Recio,A., Santos-Gomez,A., Soto,D. *et al.* (2021) GRIN database: a unified and manually curated repertoire of GRIN variants. *Hum. Mutat.*, **42**, 8–18.
141. Kühn,I., Durka,W. and Klotz,S. (2004) BiolFlor: a new plant-trait database as a tool for plant invasion ecology. *Divers. Distrib.*, **10**, 363–365.
142. Kleyer,M., Bekker,R.M., Knevel,I.C. *et al.* (2008) The LEDA Traitbase: a database of life history traits of the Northwest European flora. *J. Ecol.*, **96**, 1266–1274.
143. Tavsanoglu,C. and Pausas,J.G. (2018) A functional trait database for Mediterranean Basin plants. *Sci. Data*, **5**, 180135.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

144. Falster,D., Gallagher,R., Wenk,E.H. *et al.* (2021) AusTraits, a curated plant trait database for the Australian flora. *Sci. Data*, **8**, 254.
145. Houle,D., Govindaraju,D.R. and Omholt,S. (2010) Phenomics: the next challenge. *Nat. Rev. Genet.*, **11**, 855–866.
146. Hati,A.J. and Singh,R.R. (2021) Artificial intelligence in smart farms: plant phenotyping for species recognition and health condition identification using deep learning. *AI*, **2**, 274–289.
147. Saleem,M.H., Potgieter,J. and Mahmood Arif,K. (2019) Plant disease detection and classification by deep learning. *Plants*, **8**, 468.
148. Zhang,C., Zhou,L., Xiao,Q. *et al.* (2022) End-to-end fusion of hyperspectral and chlorophyll fluorescence imaging to identify rice stresses. *Plant Phenomics*, **2022**, 9851096.
149. Sandhu,K.S., Mihalyov,P.D., Lewien,M.J. *et al.* (2021) Combining genomic and phenomic information for predicting grain protein content and grain yield in spring wheat. *Front. Plant Sci.*, **12**, 613300.
150. Araus,J.L., Kefauver,S.C., Zaman-Allah,M. *et al.* (2018) Translating high-throughput phenotyping into genetic gain. *Trends Plant Sci.*, **23**, 451–466.
151. Steinbach,D., Alaux,M., Amselem,J. *et al.* (2013) GnpIS: an information system to integrate genetic and genomic data from plants and fungi. *Database*, **2013**, bat058.
152. Pommier,C., Michotey,C., Cornut,G. *et al.* (2019) Applying FAIR Principles to Plant Phenotypic Data Management in GnpIS. *Plant Phenomics*, **2019**, 1671403.
153. Brookes,A.J. and Robinson,P.N. (2015) Human genotype-phenotype databases: aims, challenges and opportunities. *Nat. Rev. Genet.*, **16**, 702–715.
154. Cobo-Simón,I. (2022) Cartograplant: cyberinfrastructure to improve forest health and productivity in the context of a changing climate. In *Plant and Animal Genome XXIX Conference*, San Diego (CA)
155. Sansone,S.A., McQuilton,P., Rocca-Serra,P. *et al.* (2019) FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol.*, **37**, 358–367.
156. Bulow,L., Schindler,M., Choi,C. *et al.* (2004) PathoPlant: a database on plant-pathogen interactions. *Silico. Biol.*, **4**, 529–536.
157. Bulow,L., Schindler,M. and Hehl,R. (2007) PathoPlant: a platform for microarray expression data to analyze co-regulated genes involved in plant defense responses. *Nucleic Acids Res.*, **35**, D841–845.
158. Wu,W., Wu,Y., Hu,D. *et al.* (2020) PncStress: a manually curated database of experimentally validated stress-responsive non-coding RNAs in plants. *Database*, **2020**, baaa001.
159. Global Burden Of Disease Cancer,C., Fitzmaurice,C., Abate,D. *et al.* (2019) Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2017: a systematic analysis for the global burden of disease study. *JAMA Oncol.*, **5**, 1749–1768.
160. Dhondt,S., Wuyts,N. and Inze,D. (2013) Cell to whole-plant phenotyping: the best is yet to come. *Trends Plant Sci.*, **18**, 428–439.
161. Diaz,B.P., Knowles,B., Johns,C.T. *et al.* (2021) Seasonal mixed layer depth shapes phytoplankton physiology, viral production, and accumulation in the North Atlantic. *Nat. Commun.*, **12**, 6634.
162. Adak,A., Murray,S.C., Calderon,C.I. *et al.* (2023) Genetic mapping and prediction for novel lesion mimic in maize demonstrates quantitative effects from genetic background, environment and epistasis. *Theor. Appl. Genet.*, **136**, 155.
163. Hill,D.P., D'Eustachio,P., Berardini,T.Z. *et al.* (2016) Modeling biochemical pathways in the gene ontology. *Database*, **2016**, baw126.
164. Poux,S. and Gaudet,P. (2017) Best practices in manual annotation with the gene ontology. *Methods Mol. Biol.*, **1446**, 41–54.
165. Chibucos,M.C. and Tyler,B.M. (2009) Common themes in nutrient acquisition by plant symbiotic microbes, described by the Gene Ontology. *BMC Microbiol.*, **9**, S6.
166. Fox,S.E., Geniza,M., Hanumappa,M. *et al.* (2014) De novo transcriptome assembly and analyses of gene expression during photomorphogenesis in diploid wheat Triticum monococcum. *PLoS One*, **9**, e96855.
167. Vining,K.J., Romanel,E., Jones,R.C. *et al.* (2015) The floral transcriptome of Eucalyptus grandis. *New Phytol.*, **206**, 1406–1422.
168. Fennell,A.Y., Schlauch,K.A., Gouthu,S. *et al.* (2015) Short day transcriptomic programming during induction of dormancy in grapevine. *Front. Plant Sci.*, **6**, 834.
169. Gupta,P., Geniza,M., Naithani,S. *et al.* (2021) Chia (Salvia hispanica) gene expression atlas elucidates dynamic spatio-temporal changes associated with plant growth and development. *Front. Plant Sci.*, **12**, 667678.
170. Godoy,F., Kuhn,N., Munoz,M. *et al.* (2021) The role of auxin during early berry development in grapevine as revealed by transcript profiling from pollination to fruit set. *Hortic. Res.*, **8**, 140.
171. Perez-Riverol,Y., Xu,Q.W., Wang,R. *et al.* (2016) PRIDE Inspector Toolsuite: moving toward a universal visualization tool for proteomics data standard formats and quality assessment of ProteomeXchange datasets. *Mol. Cell. Proteomics*, **15**, 305–317.
172. Kosova,K., Vitamvas,P., Urban,M.O. *et al.* (2018) Plant abiotic stress proteomics: the major factors determining alterations in cellular proteome. *Front. Plant Sci.*, **9**, 122.
173. Jarnuczak,A.F. and Vizcaino,J.A. (2017) Using the PRIDE Database and ProteomeXchange for submitting and accessing public proteomics datasets. *Curr. Protoc. Bioinfor.*, **59**, 13 31 11–13 31 12.
174. Okuda,S., Watanabe,Y., Moriya,Y. *et al.* (2017) jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Res.*, **45**, D1107–D1111.
175. Moriya,Y., Kawano,S., Okuda,S. *et al.* (2019) The jPOST environment: an integrated proteomics data repository and database. *Nucleic Acids Res.*, **47**, D1218–D1224.
176. Chen,T., Ma,J., Liu,Y. *et al.* (2022) iProX in 2021: connecting proteomics data sharing with big data. *Nucleic Acids Res.*, **50**, D1522–D1527.
177. Ma,J., Chen,T., Wu,S. *et al.* (2019) iProX: an integrated proteome resource. *Nucleic Acids Res.*, **47**, D1211–D1217.
178. Sharma,V., Eckels,J., Taylor,G.K. *et al.* (2014) Panorama: a targeted proteomics knowledge base. *J. Proteome. Res.*, **13**, 4205–4210.
179. Desiere,F., Deutsch,E.W., King,N.L. *et al.* (2006) The Peptide Atlas project. *Nucleic Acids Res.*, **34**, D655–658.
180. Deutsch,E.W. (2010) The PeptideAtlas Project. *Methods Mol. Biol.*, **604**, 285–296.
181. Tsugawa,H., Rai,A., Saito,K. *et al.* (2021) Metabolomics and complementary techniques to investigate the plant phytochemical cosmos. *Nat. Prod. Rep.*, **38**, 1729–1759.
182. Members,M.S.I.B., Sansone,S.A., Fan,T. *et al.* (2007) The metabolomics standards initiative. *Nat. Biotechnol.*, **25**, 846–848.
183. Sumner,L.W., Amberg,A., Barrett,D. *et al.* (2007) Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*, **3**, 211–221.
184. Vinaixa,M., Schymanski,E.L., Neumann,S. *et al.* (2016) Mass spectral databases for LC/MS- and GC/MS-based metabolomics: state of the field and future prospects. *TrAC*, **78**, 23–35.
185. Salek,R.M., Neumann,S., Schober,D. *et al.* (2015) COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. *Metabolomics*, **11**, 1587–1597.
186. Steinbeck,C., Conesa,P., Haug,K. *et al.* (2012) MetaboLights: towards a new COSMOS of metabolomics data management. *Metabolomics*, **8**, 757–760.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024

187. Considine,E.C. and Salek,R.M. (2019) A tool to encourage minimum reporting guideline uptake for data analysis in metabolomics. *Metabolites*, **9**: 43.
188. Schorn,M.A., Verhoeven,S., Ridder,L. *et al.* (2021) A community resource for paired genomic and metabolomic data mining. *Nat. Chem. Biol.*, **17**, 363–368.
189. Cooper,L. and Jaiswal,P. (2016) The Plant Ontology: a tool for plant genomics. *Methods Mol. Biol.*, **1374**, 89–114.
190. Cooper,L., Walls,R.L., Elser,J. *et al.* (2013) The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant Cell Physiol*, **54**, e1.
191. Avraham,S., Tung,C.W., Ilic,K. *et al.* (2008) The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Res.*, **36**, D449–454.
192. Warman,C., Sullivan,C.M., Preece,J. *et al.* (2021) A cost-effective maize ear phenotyping platform enables rapid categorization and quantification of kernels. *Plant J.*, **106**, 566–579.
193. Oellrich,A., Walls,R.L., Cannon,E.K. *et al.* (2015) An ontology approach to comparative phenomics in plants. *Plant Methods*, **11**, 10.
194. Cooper,L., Meier,A., Laporte,M.A. *et al.* (2018) The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Res.*, **46**, D1168–D1180.
195. Tello-Ruiz,M.K., Naithani,S., Gupta,P. *et al.* (2021) Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. *Nucleic Acids Res.*, **49**, D1452–D1463.
196. Naithani,S. and Jaiswal,P. (2017) Pathway analysis and omics data visualization using pathway genome databases: FragariaCyc, a case study. *Methods Mol. Biol.*, **1533**, 241–256.
197. Naithani,S., Raja,R., Waddell,E.N. *et al.* (2014) VitisCyc: a metabolic pathway knowledgebase for grapevine (Vitis vinifera). *Front. Plant Sci.*, **5**, 644.
198. Gupta,P., Naithani,S., Preece,J. *et al.* (2022) Plant reactome and PubChem: the plant pathway and (Bio)Chemical Entity Knowledgebases. *Methods Mol. Biol.*, **2443**, 511–525.
199. Naithani,S., Gupta,P., Preece,J. *et al.* (2020) Plant Reactome: a knowledgebase and resource for comparative pathway analysis. *Nucleic Acids Res.*, **48**, D1093–D1103.
200. Jaiswal,P. and Usadel,B. (2016) Plant Pathway Databases. *Methods Mol. Biol.*, **1374**, 71–87.
201. Kattge,J., Ogle,K., Bönisch,G. *et al.* (2011) A generic structure for plant trait databases. *Meth. Ecol. Evolut.*, **2**, 202–213.
202. van Kleunen,M., Pysek,P., Dawson,W. *et al.* (2019) The Global Naturalized Alien Flora (GloNAF) database. *Ecology*, **100**, e02542.
203. Manolio,T.A., Collins,F.S., Cox,N.J. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
204. Visscher,P.M., Brown,M.A., McCarthy,M.I. *et al.* (2012) Five years of GWAS discovery. *Am. J. Hum. Genet.*, **90**, 7–24.
205. Visscher,P.M., Wray,N.R., Zhang,Q. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.*, **101**, 5–22.
206. Uffelmann,E., Huang,Q.Q., Munung,N.S. *et al.* (2021) Genome-wide association studies. *Nat. Rev. Methods Primers*, **1**, 1–21.
207. Falconer,D.S. and Mackay,T.F.C. (1996) *Introduction to Quantitative Genetics*. Longmans Green, Harlow, Essex, UK.
208. Kearsey,M.J. (1998) The principles of QTL analysis (a minimal mathematics approach). *J. Exp. Bot.*, **49**, 1619–1623.
209. Lynch,M. and Walsh,B.V. (1998) *Genetics and Analysis of Quantitative Traits*. *Sinauer Associates*, Sunderland, MA.
210. Sallam,A., Eltaher,S., Alqudah,A.M. *et al.* (2022) Combined GWAS and QTL mapping revealed candidate genes and SNP network controlling recovery and tolerance traits associated with drought tolerance in seedling winter wheat. *Genomics*, **114**, 110358.
211. Hayes,B.J., Gjuvsland,A. and Omholt,S. (2006) Power of QTL mapping experiments in commercial Atlantic salmon populations, exploiting linkage and linkage disequilibrium and effect of limited recombination in males. *Heredity*, **97**, 19–26.
212. Joiret,M., Mahachie John,J.M., Gusareva,E.S. *et al.* (2019) Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. *BioData Min.*, **12**, 11.
213. Hartl,D.L., Clark,A.G. and Clark,A.G. (1997) *Principles of Population Genetics*. *Sinauer Associates*, Sunderland, MA.
214. Lee,Y.H. (2015) Meta-analysis of genetic association studies. *Ann. Lab. Med.*, **35**, 283–287.
215. Dehghan,A. (2018) Genome-wide association studies. *Methods Mol. Biol.*, **1793**, 37–49.
216. Buniello,A., MacArthur,J.A.L., Cerezo,M. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
217. Togninalli,M., Seren,U., Freudenthal,J.A. *et al.* (2020) AraPheno and the AraGWAS Catalog 2020: a major database update including RNA-Seq and knockout mutation data for Arabidopsis thaliana. *Nucleic Acids Res.*, **48**, D1063–D1068.
218. Zeggini,E. and Ioannis,J.P.A. (2009) Meta-analysis in genome-wide association studies. *Pharmacogenomics*, **10**, 191–201.
219. Soriano,J.M., Colasuonno,P., Marcotuli,I. *et al.* (2021) Meta-QTL analysis and identification of candidate genes for quality, abiotic and biotic stress in durum wheat. *Sci. Rep.*, **11**, 11877.
220. Kraft,P., Zeggini,E. and Ioannidis,J.P. (2009) Replication in genome-wide association studies. *Stat Sci.*, **24**, 561–573.
221. Li,P., Zhang,Y., Yin,S. *et al.* (2018) QTL-by-environment interaction in the response of maize root and shoot traits to different water regimes. *Front. Plant Sci.*, **9**, 229.
222. Lowry,D.B., Lovell,J.T., Zhang,L. *et al.* (2019) QTL × environment interactions underlie adaptive divergence in switchgrass across a large latitudinal gradient. *Proc. Natl. Acad. Sci.*, **116**, 12933–12941.
223. Pinu,F.R., Beale,D.J., Paten,A.M. *et al.* (2019) Systems biology and multi-omics integration: viewpoints from the metabolomics research community. *Metabolites*, **9**, 76.
224. Pacheco,A.R., Pauvert,C., Kishore,D. *et al.* (2022) Toward FAIR Representations of Microbial Interactions. *mSystems*, **7**, e0065922.
225. Sumner,L.W., Styczynski,M., McLean,J. *et al.* (2015) Introducing the USA plant, algae and microbial metabolomics research coordination network (PAMM-NET). *Metabolomics*, **11**, 3–5.
226. Kodra,D., Pousinis,P., Vorkas,P.A. *et al.* (2022) Is current practice adhering to guidelines proposed for metabolite identification in LC-MS untargeted metabolomics? A meta-analysis of the literature. *J. Proteome Res.*, **21**, 590–598.
227. Schroeder,M., Meyer,S.W., Heyman,H.M. *et al.* (2019) Generation of a collision cross section library for multi-dimensional plant metabolomics using UHPLC-Trapped Ion Mobility-MS/MS. *Metabolites*, **10**, 13.
228. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
229. Jeliazkova,N., Apostolova,M.D., Andreoli,C. *et al.* (2021) Towards FAIR nanosafety data. *Nat. Nanotechnol.*, **16**, 644–654.
230. Iturbide,M., Fernandez,J., Gutierrez,J.M. *et al.* (2022) Implementation of FAIR principles in the IPCC: the WGI AR6 Atlas repository. *Sci. Data*, **9**, 629.
231. Mons,B., Neylon,C., Velterop,J. *et al.* (2017) Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European open science cloud. *Inform. Serv. Use*, **37**, 49–56.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baad088/7469248 by guest on 29 April 2024