Review

Agricultural Sciences in the Big Data Era: Genotype and Phenotype Data Standardization, Utilization and Integration

Cecilia H. Deng *, Sushma Naithani, Sunita Kumari, Irene Cobo-Simon, Elsa H. Quezada-Rodriguez, Maria Skrabisova, Nick Gladman, Melanie J. Correll, Akeem Babatunde Sikiru, Olusola O. Afuwape, Annarita Marrano, Ines Rebollo, Wentao Zhang, Sook Jung *

Posted Date: 14 June 2023

doi: 10.20944/preprints202306.1013.v1

Keywords: Genotype; phenotype; sequencing; phenomics; data integration; metadata; standardization

Review

Agricultural Sciences in the Big Data Era: Genotype and Phenotype Data Standardization, Utilization and Integration

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

[1] The New Zealand Institute for Plant and Food Research Limited, Auckland, New Zealand
[2] Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA
[3] Cold Spring Harbor Laboratory, 1 Bungtown Rd, Cold Spring Harbor, New York, USA
[4] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, USA
[5] Departamento de Producción Agrícola y Animal, Universidad Autónoma Metropolitana-Xochimilco, Ciudad de México, México
[6] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Ciudad de México, México.
[7] Department of Biochemistry, Faculty of Science, Palacky University, Olomouc, Czech Republic
[8] U.S. Department of Agriculture-Agricultural Research Service, NEA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, New York, USA.
[9] Agricultural and Biological Engineering Department, University of Florida, Gainesville, FL, USA
[10] Federal University of Agriculture Zuru, Nigeria
[11] University of Lagos, Nigeria
[12] Phoenix Bioinformatics, Newark (CA), USA
[13] Universidad de la República, Uruguay
[14] National Research Council Canada, Canada
[15] Department of Horticulture, Washington State University, USA
* Correspondence: author e-mail: Cecilia.Deng@plantandfood.co.nz
† These authors contributed equally to this work: Cecilia H. Deng, Sushma Naithani, Sunita Kuman, Sook Jung.

Abstract: The Genotype-Phenotype Working Group was established in November 2021 as part of the AgBioData Consortium (https://www.agbiodata.org) with the goal of identifying current challenges in annotating and integrating large-scale genotype and phenotype data. Over the course of the year, the members of this working group identified different types of data sets, explored experimental platforms and methods for data generation, and examined how these data are annotated including the metadata requirements. We conducted a thorough review of publicly funded repositories for raw and processed data for each data type. We also examined several secondary databases and knowledgebases that enable the integration of heterogeneous data types in the context of the Genome Browser, Pathway Networks and tissue-specific gene expression. The review revealed a need for additional infrastructural support, standards, and tools to connect Genotype to Phenotype data and enhance data interoperability for knowledge synthesis and to foster translational research.

Keywords: Genotype; phenotype; sequencing; phenomics; data integration; metadata; standardization

1. Introduction

Genotype-phenotype (G2P) integration is the process of linking genetic data to measurable qualitative and quantitative phenotypes and traits. Historically, linking genetic markers or genes

 doi:10.20944/preprints202306.1013.v1

associated with desirable traits have led to the development of improved cultivars with higher yields and quality, enhanced disease resistance or climate resilience. In past two decades, the generation of high-throughput Omics or "big data" including plant genomes and pan-genomes, genetic variation data including Single Nucleotide Polymorphisms (SNPs) and structure variations (SVs), 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 the large-scale phenotype data collection is peaking by. Often, a particular set of big data is being analyzed to bridge specific knowledge gaps identified by the projects and remain underexploited for synthesis of new knowledge. Going forward, the different data types produced in various experiments can be reutilized for synthesizing new knowledge, developing data-driven hypotheses, and for experimental research. The integration of large-scale datasets from diverse sources, however, can be challenging and typically involves quality check, data re-formatting, curation, and re-analysis. 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) (3-5). To meet the FAIR standards, any dataset should include metadata providing the standard terms and details necessary for data interpretation using plant ontologies or controlled vocabularies. Data and metadata standardization can be achieved by developing common community standards of data formats, and description so that diverse datasets from different sources can be accessible and interoperable 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 Genetic Genomic Breeding (GGB) database tools and operations, with the goal towards increased data interoperability for future research (4). 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 to big data agricultural research. A AgBioData consortium white paper (4) has previously identified challenges facing GGB Databases and suggested common guidelines for bio-curation, ontologies, metadata, GGB database platforms, programmatic (machine) access to data, communication among various partners and stakeholders, and sustainability of genomic resources/databases. 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 NSF RCN project, working groups were formed around major data-related challenges and needs. The Genotype-Phenotype working group (G2P-WG) was formed in November 2021 with the goal of identifying current challenges in annotating and integrating large-scale genotype and phenotype data. The efforts and work of the AgBioData GP-WG brought to the current white paper, which summarizes the current status of FAIR practices of phenotype and genotype data (see Figure 1). For common genotype and phenotype data, we report (*i*) a brief introduction of the diverse data type 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 re-use and reanalysis of omics data, and (*v*) limitations of data re-use.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2023 doi:10.20944/preprints202306.1013.v1

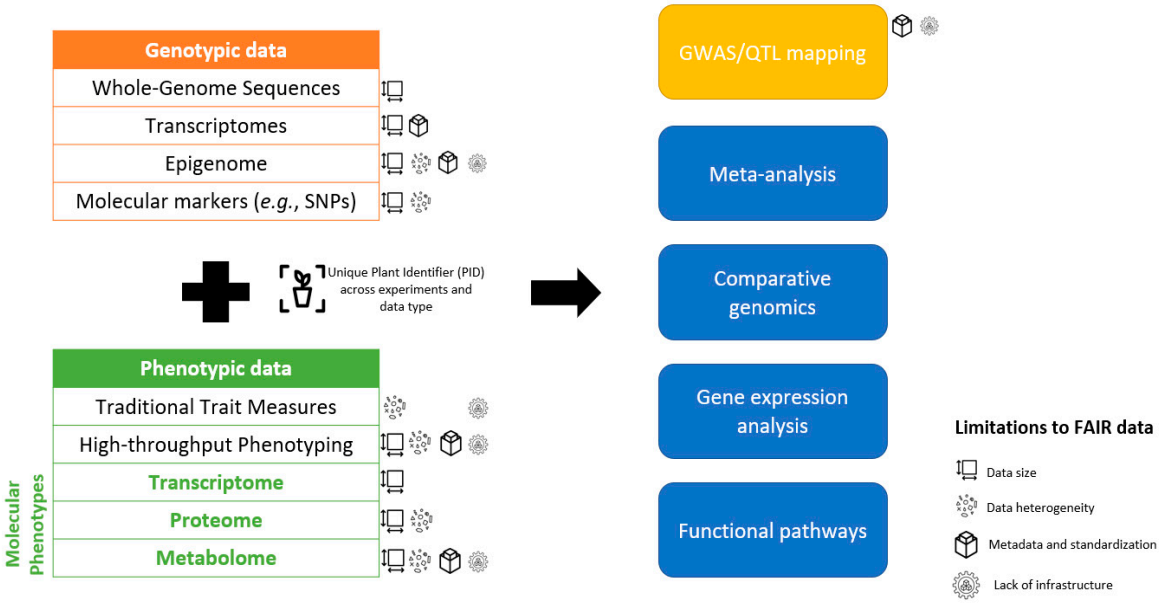

Figure 1. Current status of genotype to phenotype data integration. A summary of diverse genotype and phenotype data is on the left, while on the right, a list of potential integrative analyses that can be carried out by plant researchers using the various data types.

2. Genomics and Transcriptomics data

2.1. 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 the Whole Genome Sequencing (WGS) and transcriptome sequencing. There are basically three general methods of DNA/cDNA sequencing : (i) Sanger chain termination sequencing and Maxam Gilbert sequencing; (ii) short-read sequencing known as Next Generation Sequencing (NGS) (6) including Ion Torrent, Solexa/Illumina, Roche/454 pyrosequencing; and (iii) more recent long-read Third Generation Sequencing (3GS), primarily single molecule real time sequencing from Pacific Bioscience (PacBio), and nanopore sequencing from Oxford Nanopore Technologies (ONT). At present, Illumina is the dominant and most popular platform in NGS for both genomes and transcriptome sequencing because of high accuracy, low cost, and global distribution. PacBio and ONT are gaining popularity and becoming affordable for high-quality long-read/ full-length sequences. Similarly, DNBSeq from MGI Tech, a subsidiary of Beijing Genomics Institute (BGI) group and Ion Torrent Systems (7,8) are making advances. There are several file formats used in WGS and the most common is the compressed FASTQ format that is used for both sequencing platforms, NGS and 3GS. The original file formats for 3GS include legacy h5 format for PacBio, the industry-standard BAM format, and the FAST5 format for ONT that is based on the hierarchical data format (HDF5) used for ONT data storage sequencer. There are numerous basecallers available for conversion to FASTQ format (9) and in general we find that sequencing data has achieved standard data formatting.

2.2. Genome Sequencing Strategies for Genotyping

Genotyping is a crucial component in linking genotype to phenotype. The first-generation genotyping marker was Restriction Fragment Length Polymorphisms (RFLPs), which relied upon underlying differences in base pair sequences to create an autoradiographic fingerprint after DNA regions were digested with known restriction enzymes (10). These techniques progressed with technological advancements in PCR and other DNA sequencing techniques to include genotyping

 doi:10.20944/preprints202306.1013.v1

via microsatellite markers, specifically simple sequence repeats (SSR) or short tandem repeats (11). High throughput low- and high- density SNP arrays provide a cost-effective genotyping solution for studies such as population structures, genomic diversity, gene discovery and molecular breeding. Array technology can genotype a large number of samples in a short period of time, and data analysis is much 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.

Various genome sequencing strategies can be employed based on research aims and funding.

Sequencing of sub-sampled loci (12) has been widely used in phylogenomics studies for cost-effective large-scale genotyping. Skim sequencing (13) is a low coverage whole genome sequencing approach. Target enrichment sequencing investigates specific genomic elements via pre-defined probe sequences (14). Exome sequencing is a common type of target sequencing that focuses on protein-coding regions of genes (15). Amplicon sequencing is a highly targeted approach addressing specific genome loci. Genotyping-by-sequencing (GBS) (16,17) and restriction-site associated DNA marker sequencing (RAD-seq) (18,19) are two popular cost-effective sequencing strategies for shearing the genome via restriction enzyme(s). This advent of high-throughput sequencing has generated immense amounts of data that plays into several areas of genomic concern. 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 structural variation between two or more genomes at a given locus (20). 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 GWAS-VCF (21) and GVCF (tinyurl.com/5f8wpmhr), and accommodates variant information of polyploid genomes. In addition to low coverage genome sequence, transcriptome sequence is routinely used for genotyping and identification of useful genetic markers. More recently, integration of single cell genome sequencing and single cell transcriptome sequencing tools have facilitated quantifying genetic and expression variability between individual cells (22). Like sequence data, genotyping data has standardized formats.

2.3. Public repositories for genomics and transcriptomics data

Regardless of the sequencing platform or strategy used, raw sequencing data in compressed fastq.gz format is submitted to a public data repository such as National Center for Biotechnology Information (NCBI) GenBank, Sequence Read Archive (SRA) and/or Gene Expression Omnibus (GEO) 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), Sequence Read Archive (SRA) via DDBJ submission navigation website, or the European Nucleotide Archive (ENA) through BioStudies portal. DDBJ, ENA and NCBI GenBank (see Table 1 and Supplementary Table S1) form the International Nucleotide Sequence Database Collaboration (INSDC) 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 (23). Additional public platforms that host the sequence data includes the US Department of Energy (DOE) Joint Genome Institute (JGI) 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 (24), the Plant Comparative Genomics portal, for genome accessing, comparison and visualization (see 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 both at 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), Harvard Dataverse (https://dataverse.harvard.edu), etc. accepts data submission in any file format.

 doi:10.20944/preprints202306.1013.v1

Outside 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 (25). The Indian Biological Data Center (IBDC, https://ibdc.rcb.res.in) is a public repository in India to host various life science data. For sequencing data, IBDC provides the INSDC-compatible Indian Nucleotide Data Archive (INDA, https://inda.rcb.ac.in/home) with data synchronized to NCBI/ENA/DDBJ; and the Indian Nucleotide Data Archive-Controlled Access (INDA-CA, https://inda.rcb.ac.in/indasecure/home) for private data. In New Zealand, the Aotearoa Genomic Data Repository (AGDR) hosts genomics data, especially for native *taonga* ('treasure' in Maori language) species.

Data submission to cloud storage is also gaining popularity. Amazon Web Services (AWS) offers Open Data (https://aws.amazon.com/opendata) for unregistered users to find and use publicly available datasets, and allows subscribed customers to search and access third-party data (https://docs.aws.amazon.com/data-exchange/index.html). In addition, it provides Amazon Omics (https://aws.amazon.com/omics/) and Plant & Animal Genomics (https://aws.amazon.com/solutions/agriculture/plant-animal-genomics/) platforms to facilitate omics data analysis and integration. Other options include Google Cloud Life Sciences (https://cloud.google.com/life-sciences) and Microsoft Genomics (https://azure.microsoft.com/en-in/products/genomics/).

We compiled a list of public repositories for genome, genotyping and transcriptome sequence data (see Table 1) that are active, maintained and updated. Detailed information about metadata availability, data file formats related to these repositories are described in Supplementary Table S1.

Table 1. A list of public repositories for genomic, genotyping and transcriptom data that are active, maintained and updated. The "+" and "-" symbols indicate the presence and absence, respectively, of the supported data type and data format. Databases that support any data type beyond the specified most common types are marked by "¥". Out of the INSDC source data bases were established and maintained by a, National Genomics Data Centre, China, and China National Center for Bioinformation; b, The Indian Biological Data Center; c, New Zealand Ministry for Business Innovation and Employment; d, University of North Carolina at Chapel Hill, California Digital Library; d, CERN; f, Digital Science. Holtzbricnck Publishing Group, Macmillan Publishers Limited. "†" indicates that data is available upon request. "‡" recommended by FAIRsharing.org. "?" means that the information is not available. "+/-" means that this data type can be submitted only through command line or programmatic approach but not by interactive interface. Detailed information about metadata requirements and database URLs are available in Supplementary Table S1.

Database name	NCBI	DRA	ENA	GSA	IBDC	AGDR†	DRYAD¥	Zenodo‡,¥	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 metadata associated with sequence and genotype data promotes a dataset's discoverability and reusability. We note here that many specific secondary public repositories exist that exclusively host data on promoters, transcription factors, proteomes, various RNA types, epigenomic data and Pangenomes (see Supplementary Table S2). However, here we limit our discussion to primary

 doi:10.20944/preprints202306.1013.v1

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.

2.3.1. The metadata requirements on genomics and transcriptomics data set

The metadata associated with the genome/genotyping/transcriptome sequencing is crucial for its re-use and interoperability. To maximize the implementation of FAIRstandards, the metadata should be described with accurate Gene Ontology (GO) and Plant Ontology (PO) terms with proper evidence codes wherever applicable. Project and sample level metadata typically include 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 are directly related to the preparation of biological materials undergoing the assay including method detail(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 (i.e., name, affiliation and email of the data submitter, and other authors), BioProject goals (i.e., genome sequencing and assembly; raw sequence reads, epigenomics, exome, proteome, variation, etc.), and BioSamples information (i.e., organisms 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 data release date, project title, and public description of the study goals are the minimum general information required for a project. 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 information (number, title, grantee), research consortium name and URL, data provider and URL (if different from the submitting organization), publication information.

Optional but useful metadata for BioSamples include sample title, BioProject accession, biomaterial provider (lab 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 Phenotypic quality Ontology (PATO) terms at http://bioportal.bioontology.org/visualize/44601), population (filial generation, number of progeny, 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.

The mandatory attributes for library construction metadata are BioSample name, library ID, a title, data type and method information (eg. WGA, WGS, RNA-Seq, EST, ChIP-Seq, and so on), source (GENOMIC, TRANSCRIPTOMIC, GENOMIC SINGLE CELL, METAGENOMIC, etc.), selection (PCR, RANDOM, RT-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 submission of metadata but encourage data submitters to provide as much details as possible. In this category AGDR (https://repo.data.nesi.org.nz/DD) requires submitter ID, project ID, project code, project name, programme 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, RIN, adapter name and sequence, barcoding, base caller name and version, experiment name, flowcell barcode, fragment sizes, instrument model, lane number, library name, 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, any legal or

 doi:10.20944/preprints202306.1013.v1

ethical considerations, etc.), and research domain. Optional metadata recommended are 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 that is similar to INSDC repositories but does not enforce it as a requirement. The storage quota for a free account is 20GB and up to 100 projects.

2.3.2. 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) (26), but a newer repository has also arisen in the Genome Variation Map (GVM) (27). NCBI does host the dbSNP and dbVar databases, but those are intended for human data. All repositories strive to adhere to FAIR practices, but additional recommendations have been put forth by others (28). The EVA repository accepts VCF file structures, including hapmap formatted files (29) and SNP genotyping arrays, that are validated through a 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 are 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 and 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.

2.3.3. Crop/clad Community GGB Databases

Whole genome, transcriptome, and genotype data can also be submitted to most of the GGB databases such as GDR (30-32), CottonGen (33,34), SoyBase (35,36), LIS (37,38), SGN (39,40), MaizeGDB (41,42), TreeGenes (43,44), TAIR (45,46), KnowPulse (47), and InterMine (48,49) databases (Table 2). Some of these databases, such as Gramene (50-52), SorghumBase (53) and InterMine (48,(54), 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 QTL, GWAS, markers, alleles, genetic maps, and cultivar/germplasm phenotyping data, and integrate them with whole genome, transcriptome, and genotyping data. These GGB databases standardize various names, associate the data with various ontologies to integrate data from various sources and of various types. This integration of different types 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 enables researchers to discover novel associations between different types of data, potentially leading to valuable insights and breakthroughs. For example, combining SNP genotype data and phenotype data from multiple locations of the same germplasm can reveal how particular genotypes manifest specific phenotypes in distinct environments. However, achieving such insights requires meticulous integration of data from multiple sources.

 doi:10.20944/preprints202306.1013.v1

Table 2. List of Crop/clad Community GGB Databases that integrate various types of data 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 genomic database used by multiple communities		
A comparative genomic database for ~300 plant species	Phytozome	https://phytozome-next.jgi.doe.gov/
A comparative genomic 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/

2.3.4. Uses and Applications

WGS data can be reused in genome assembly, pan-genome construction, single nucleotide variation (SNV), copy-number variation (CNV), and structure variation (SV) discovery, phylogenomics, comparative genomics, and other genome research to study genome structure, genome diversity, the evolution of gene families or organisms, and crop domestications. Genotyping data in VCF format can be used for numerous purposes: storage of the location of given variants

 doi:10.20944/preprints202306.1013.v1

(including GWAS-associated variants); to identify targets of molecular markers for genotyping purposes; evaluating the effects of given base pair and structural variants on gene function; comparative genomics and evolutionary studies; and computational breeding approaches via machine learning and other methods. Data extraction and manipulation of VCF files is easy with the use of existing software toolkits such as VCFtools and SAMtools and can be utilized in conjunction with existing and ad hoc bioinformatic pipelines due to its command line functionality. By integrating VCF data with RNA-Seq and phenomics data, researchers can utilize these data sets quantitative genetic studies including genome-wide association studies (GWAS), quantitative trait loci (QTL) analysis, marker discovery, and genome selection (GS), to accelerate modern breeding techniques. Integrating transcriptomics data with metabolomics data can help in predicting biomarkers, which are 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 copy number variations can be used to categorize samples into groups based on their similarity to two datasets.

3. Phenotype and Phenomics

3.1. 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 (55). The classical genetic and functional genomics studies in model and crop plants have identified numerous mutants that show distinct morphological and anatomical mutants and associated the individual mutant phenotypes with one or more genes, pathways and molecular processes. Table 3 lists databases that host the mutant collections and description of phenotype of individual mutants and associated genes, including MaizeDIG (42), RIKEN Arabidopsis Genome Encyclopedia (56), Mutant Variety Database (57), Plant Genome Editing Database (58), Tomato mutant Archive TOMATOMA (59), and Plant Editosome Database (60).

In addition, complex phenotypic traits (i.e., morphological and physiological) related to the fitness and performance of an organism are often quantitative in nature and have multiple genetic determinants (61,62). Examples of traits that are determined by multiple genes (known as Quantitative Trait Loci, 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 (63), as they provide a general understanding of a wide range of ecological and evolutionary phenomena such as impact of climate change, and anthropogenic land use on biodiversity (64-66). In Table 3, we provide a list of a key databases (or portal of bigger databases) that host information related to traits, QTLs, and associated data including the Gramene QTL database (67), QTL database for wheat (68), GLOPNET (69), TRY (70), a database of Ecological Flora of the Britain and Ireland (71), BIOPOP (72), GRIN (73), the USDA PLANTS Database, BiolFlor (74), LEDA Traitbase (75), BROT database of plant traits for Mediterranean basin species (76), and AusTraits (77). 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 (78). Phenomics can be used for species recognition and biodiversity characterization (79), for stress quantification (79-81), and for crop yield prediction (82,83). Thus, phenomics data sets are very large and have different formats (e.g., JSON file). Some of the databases that host phenomics data include GnpIS (84,85), PGP (86), Cartograplant (87), AgData commons (https://data.nal.usda.gov/; (88), PathoPlant (89,90), PncStress (91), OSRGD (92).

Despite its analogy to genomes, it is not possible to fully characterize phenomes due to heterogeneity and multifaceted nature of phenotypic data with added layers reflecting complexities

 doi:10.20944/preprints202306.1013.v1

at the cell, tissue, and whole plant level that have further variations according to development stages, and growth environment (78,93). 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 under 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 images or time-series RGB images and remote sensing to monitor the algal blooms in the ocean (94). As phenomics data can be extremely variable in nature, necessary metadata includes information about plant species, tissue, developmental stage, environmental conditions, experimental design, data collection, processing, and analysis.

In addition to traditional phenotypes, molecular phenotypes include changes in the chromatin organization, transcripts, proteins, metabolites and ions (95-97). 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, vegetables; (ii) the quality of bio products 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 (98-102).

Proteome and metabolome datasets allow the 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 (103). Once the original proteomic datasets and associated metadata/manuscript have been submitted to public data repositories such as PRIDE (103-105), MassIVE (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp), JPOST (106,107), IProX (108,109), Panorama (110), and Peptide Atlas (111,112), 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 Tsugawa et al., 2021 (113). The types of data collected for metabolomics depends 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 (RT), 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 are MZmine, XCMS, MSdial, metaMS, Progenesis QI and MetAlign. For annotation for unknown metabolites, popular software tools include MS-FINDER, MetDNA, MetFamily, and GNPS among others. Raw file formats generated by the machines include d, raw, idb, cdf, wiff, scan, dat, cmp, cdf.cmp, lcd, abf, jpf, xps, mgf. Derived file formats are mzml, nmrml, mzxml, xml, mzdata, cef, cnx, peakml, xy, smp, scan. Due to the complexity of metabolomic data, several initiatives were undertaken. The Chemical Analyses Working group started the **Metabolomics Standard Initiative (MSI)** to develop metabolomic standards (114,115) with revisions suggested by (116). 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' (**COSMOS) (117), and MetaboLights (118), for developing tools to ease submission of metabolomic data (119). ProteomeCentral and Omics DI serve as central repositories for these datasets, which are then re-used in protein knowledge bases (Uniprot and NeXtProt), genome browsers (Ensembl and UCSC), proteomics resources and other bioinformatics resources (ex. 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

 doi:10.20944/preprints202306.1013.v1

Data Platform (PoDP) (120) 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 NIH Common Fund's National Metabolomics Data Repository (NMDR) portal and the Metabolomics Workbench, and MetaboLights.

Some gene expression and metabolic phenotype often culminate in visible phenotypes, which can be described using the Plant Ontology terms (121-123). More recently, Plant Ontology terms have been extended to large scale phenomics data from a single species (124) to support the comparative phenomics in plants (125) and describe trait phenotypes expressed under specific developmental stage or specific environment and stress (126). For covering the genotype-phenotype gap, we need integration of multiple types of data including genotypic, large-scale phenome, gene expression, proteome and metabolome data, described using defined and standardized ontologies.

After collecting and generating phenotypic and phenomics data, it is recommended that they are formatted using community guidelines and submitted to primary data repositories, along with well-described metadata. The primary repositories serve as a source of primary or raw data (with base annotations) to the secondary databases for their visualization on genome browser (127) or for synthesizing new information by integrating them to other data types like plant metabolic networks (128,129), system-level plant pathways (130-132), expression Atlas, metabolic models, etc. These secondary knowledge bases are of primary importance to the plant researchers for formulating data-driven hypothesis for experimental and translational research and for analyzing the high-throughput omics data in the overall context of a species genome, systems-level pathway networks (133), and for gaining evolutionary insights by conducting intraspecies and interspecies comparisons. The implementation of standards and the development of infrastructure of public repositories are crucial for FAIR phenotypic data, even if many public repositories are currently not supporting the submission of the phenotype data (see Table 3).

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

Category	Databases	URLs
Species-specific mutant collections	Database of image 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 (RARGE)	http://rarge-v2.psc.riken.jp/line
	TOMATOMA	https://tomatoma.nbrp.jp/index.jsp
	Plant Editosome	https://ngdc.cncb.ac.cn/ped/
Traits and QTL	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

 doi:10.20944/preprints202306.1013.v1

	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 (ICPD)	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://www.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	MetaboLights	https://www.ebi.ac.uk/metabolights/index
	PoDP	https://pairedomicsdata.bioinformatics.nl/

 doi:10.20944/preprints202306.1013.v1

Knowledgebase	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 (BAR)	https://bar.utoronto.ca
	The protein-protein interaction database for Maize (PPIM)	https://mai.fudan.edu.cn/ppim/

3.2. Phenotype data formats, standards and metadata

The structure and characteristics of data types, along with any additional metadata, is crucial for enabling future data re-use and re-analysis by other researchers. The most relevant metadata shared across the various data types (generated by a diverse set of methods and platform) include taxonomic identification of the plant, the individual or cultivar name or accession ID, geo-references 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) (70,134). These metadata can be entered as simple text format during the submission of the raw data to any primary repository and are easily exported from one database to another as TXT files.

Furthermore, plant phenotype/traits can be classified as categorical (qualitative and ordinal) or quantitative (continuous) traits (135). Some phenotypes are rather stable within species (mostly categorical traits), and some of these can be systematically compiled from species checklists and floras (e.g., (136). 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 have to 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 (70). Thus, the collection of these quantitative traits and their essential environmental covariates is of vital importance. While trait measurements themselves 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 not only limits the number of measurements but also 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 data set 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 (134). Hence, an accurate and careful collection of data, their associated meta-data and ancillary data, is key to correctly preserve this valuable information, as well as to perform a suitable data integration across studies, species and data types.

 doi:10.20944/preprints202306.1013.v1

4. Association mapping (GWAS) and linkage mapping (QTL)

Genome-Wide Association Study (GWAS) and QTL mapping are statistical methods used to identify marker-trait associations and candidate genes (causative mutations) controlling traits of interest. Both approaches rely on the linkage disequilibrium (LD) between the tested markers and the functional polymorphisms at the causative genes. However, they differ on the type of genetic populations used for the study: GWAS relies on 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). Regardless of the fact that they are both analytical methods, their results can be used as data inputs for other types of analysis (e.g. meta-analysis, estimation of polygenic scores) (137). 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 as 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 also key to contributing to link 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 (138). One of the key metadata for GWAS/QTL data is the type of statistical method used to calculate and correct the *p*-values (GWAS/QTL). Regarding the SNPs, the most important metadata include the model species and the version of the reference genome against which these SNPs are mapped (see Genotype data section). The metadata required to make the traits interoperable and reusable is explained in the section lab/field traits. 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 (139).

Unlike the human and animal GWAS and QTL data open access resources such as the NHGRI-EBI GWAS Catalog, GWAS Atlas (140), OpenGWAS, Animal QTL database, and Animal Genome Informatics resources (USDA national infrastructure NRSP-8: 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 (140) contains recomputed GWAS results using a standardized GWAS pipeline on all publicly available phenotypes from AraPheno (141).

Meta-analysis is the widely used analysis for integrating the summary statistics from multiple GWAS/QTL studies. It is a set of methods that allows the quantitative combination of data from multiple studies, and the evaluation of the consistency, inconsistency, or heterogeneity of the results across multiple datasets. Meta-analysis of GWAS/QTL datasets can improve the power to detect association signals by increasing sample size and by examining more variants throughout the genome than each dataset alone (142). However, in order to integrate datasets coming from different studies in meta-analysis, a standardized data and meta-data collection among the studies is needed. 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, linking genomes to complex phenotypes across species.

5. Data reusability limitations and challenges

Accessing, reusing and integrating analytic data from various data types remain difficult (143). 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. Even if the data is submitted to public repositories, it often remains inaccessible or unusable due to missing fundamental metadata or improper formatting. The lack of community-based guidelines of data sharing, coupled with the complexity of data size and type generated in GGB research are likely

 doi:10.20944/preprints202306.1013.v1

the biggest limitations for data reuse in this field. Another matter arises from various levels of attitude of research journals/editorials for data sharing requirements and policy. Here we discuss limitations to data reuse in genotype-to-phenotype studies in three aspects.

5.1. Challenges

Data diversity and data format heterogeneity. Agriculture and horticulture research involves a wide range of genotypic, phenotypic, and environmental data, which often come from different experimental protocols and data generation technologies, and data processing workflows. As a result, data formats can be highly heterogeneous, making it difficult to integrate data from different sources and reuse in future studies (70). This issue is even more significant for phenotypic data, especially with the new 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 variety of data and metadata generated by these technologies can be highly variable in terms of file size, format, and content. The heterogeneity of data analysis pipeline also contributes to the complexity of standardization in phenomics.

Data size, quality and versioning. Most genomics, transcriptomics, epigenetics, and phenomic data are extremely large in file size and computationally intensive. For example, whole-genome sequencing data used for variant calling or VCF files that collates 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, which can prevent their reuse.

Object identification. Data submitted to a public domain often lacks a unique data object identifier (e.g., 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. To improve data findability and reuse, it would be desirable to have a universal DOI associated with its PID. However, most data used and generated in research studies are not shared or are inaccessible or not reusable because of missing fundamental metadata or improper format of the data.

Metadata and data standardization.

Metadata are 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 (MIGS) from the Genomic Data Standards Consortium, the Plant Phenotype Ontology (PPO), the Minimum Information About a Plant Phenotyping Experiment (MIAPPE). For GWAS data, GWAS-VCF format (21) has been proposed. However, the promotion and consistent application of these standards across different research groups and databases remain a challenge. For instance, if there are standards for how to collect and describe trait measurements, they are organism-specific (*e.g.*, International Organization of Vine and Wine (OIV); www.oiv.int) or based on model species.

The metabolomic research community faces similar challenges. An initiative to identify the grand challenges of metabolomic research was coordinated by the USA Plant, Algae and Microbial Metabolomics Research Coordination Network (PAMM-NET; (144)). As noted, the data obtained from metabolomic analyses can often result in different chemical features values even in the same biological treatments due to the variability associated with biological systems, differences in the equipment, differences in the protocol and reagents. Therefore, identifying metabolites with confidence and the limited metabolome depth of coverage are the key grand challenges in metabolomic research (144). A recent review of LC-MS literature found a lack of details reported on methodology and level of confidence for metabolites in most of the reviewed research articles (145). To address these challenges, multi-dimensional analyses methods, the use of standard libraries for

 doi:10.20944/preprints202306.1013.v1

metabolite characterization, and tools that simplify the submission of metadata and data are being developed (119,146).

Other barriers for FAIR data include considerations of data privacy and confidentiality, legal and ethical issues, concerns of ownership, lack of incentive if not credited for sharing data, lack of awareness of existing data standards or data repositories, or lack of resources to implement data standards.

5.2. 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 make 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, as well as sufficient user support. For some types of plant data such as QTL and GWAS, there are basically no primary repositories where researchers can submit their data. Community GGB databases (Table 2) addresses this issue by collecting, curating, and integrating various data types from different sources and related species, playing a key role in data integration. Not all plant GWAS data, however, are timely stored in databases due to either 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.

5.3. 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 (3). These principles apply to every type of scholarly digital object archived in a repository, and their implementation has started in many different research fields (147-149). In summary, these principles recommend that, when data are submitted, they are very well described using richly 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 they can be easily interpreted and exchanged by human and computer systems. The use of controlled vocabularies and ontologies is strongly encouraged to facilitate data interoperability across database resources. FAIR data, however, do not mean open-access or free, but refers to clarity and transparency about the conditions governing access and reuse (e.g., credential system to access and download data; (150). All these principles together aim to increase data transparency and improve data reuse for new research purposes, enhancing data value across time.

The implementation of FAIR data policy, however, can be challenging due to several reasons. Firstly, making data FAIR requires additional efforts and time commitment from researchers, which can be a barrier to implementation. Secondly, many scientists are not aware of the FAIR principles, community-based standards, and ontologies available to make their data FAIR. Thirdly, there is also a need for long-term sustainability of database resources, which requires ongoing funding and infrastructure support. Many databases struggle to secure funding and may face difficulties in maintaining FAIR data quality and accessibility over time.

 doi:10.20944/preprints202306.1013.v1

6. Recommendations

With the latest advances in DNA sequencing and phenotyping technologies, the analysis of large datasets can be used to study the genetic instructions from either a single gene or the whole genome to be translated into the full set of phenotypic traits of an organism. The phenotypic data must be high quality digital phenotypic data with robust metadata that can be used for further downstream analysis and mathematical modeling of the phenotypic traits such as development, stress tolerance etc. To improve the data collection and data sharing of genotypic to phenotypic data, here are the several recommendations that can be taken to ensure their interoperability and reproducibility.

1. **Standardization of data collection protocols**: Standardizing data collection protocols and using common data formats can help to ensure that data is collected in a consistent and comparable way. Use of metadata standards and the requirement of new ontology terms will make it easier to share and compare data across different studies.
2. **Centralized data sharing platform**: Developing and using centralized data sharing platforms, the use of standardized data models and exchange formats and the deployment of existing and emerging software components can help to facilitate the sharing of genotypic and phenotypic data among researchers. It includes the use of online databases and repositories that are specifically designed for the storage and sharing of the plant genetic and phenotypic data.
3. **Consistent data annotation**: Consistently annotating data with relevant information such as the genotype, phenotype, and experimental treatments can help to make the data more easily searchable and usable by other researchers.
4. **Data quality control**: More automated management of data flows and implementing data quality control such as data curation and validation can help to ensure that the data is accurate, reliable and can be used to make valid conclusions.
5. **Data integration**: Adoption of new database technologies and the development of robust data standards can facilitate the global integration of G2P data in future. Data integration from different resources such as genomics, transcriptomics, proteomics and metabolomics can help to better understand the complex relationship between genotype and phenotype.
6. **Community driven efforts**: Community driven efforts such as open-source projects, workshops and collaborations can help to promote the sharing and use of data among researchers, which in turn will lead to better understanding of the G2P relationship. There should be encouragement on integrated science training plans that enable biologists to think quantitatively and facilitate collaboration with experts in physical, computational and engineering sciences. It can help the scientists to get familiar with the development of computational pipelines and workflows that will be essential for researchers to acquire, analyze and critically interpret G2P data.
7. **Data storage infrastructure, data management software and data curation tools** are necessary to handle the large volumes of data in diverse formats.
8. **A concerted effort to make multi-omics data sets interoperable** by automated biocuration with controlled ontology terms will help address this issue. Community databases address some of this issue 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.

9. **Data security, backup and recovery** must be considered and implemented for sustainability.
10. **Data compliance** with data sharing policies, privacy regulations and laws should be enforced.

Acknowledgements: We acknowledge the funding to AgBioData Consortium through the National Science Foundation (NSF) for Research Coordination Network (RCN) project (award abstract #2126334). We gratefully acknowledge Dr. Catrin Guenther, Farhana Pinu and Mareike Knaebel at The New Zealand Institute for Plant and Food Research Limited for proofreading and providing feedback.

References

1. Scossa, F., Alseekh, S., 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. 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.
4. Harper, L., Campbell, J., Cannon, E.K.S., *et al.* (2018) AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. *Database (Oxford)*, **2018**.
5. Adam-Blondon, A.F., Alaux, M., Pommier, C., *et al.* (2016) Towards an open grapevine information system. *Hortic Res*, **3**, 16056.
6. Ekblom, R., Wolf, J.B. (2014) A field guide to whole-genome sequencing, assembly and annotation. *Evol Appl*, **7**, 1026-1042.
7. Kanzi, A.M., San, J.E., Chimukangara, B., *et al.* (2020) Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. *Front Genet*, **11**, 544162.
8. Patterson, J., Carpenter, E.J., Zhu, Z., *et al.* (2019) Impact of sequencing depth and technology on de novo RNA-Seq assembly. *BMC Genomics*, **20**, 604.
9. Wick, R.R., Judd, L.M., Holt, K.E. (2019) Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol*, **20**, 129.
2. Grodzicker, T., Williams, J., Sharp, P., *et al.* (1975) Physical mapping of temperature-sensitive mutations of adenoviruses. *Cold Spring Harb Symp Quant Biol*, **39 Pt 1**, 439-446.
3. 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.
4. McKain, M.R., Johnson, M.G., Uribe-Convers, S., *et al.* (2018) Practical considerations for plant phylogenomics. *Appl Plant Sci*, **6**, e1038.
5. Kumar, P., Choudhary, M., Jat, B.S., *et al.* (2021) Skim sequencing: an advanced NGS technology for crop improvement. *J Genet*, **100**.
6. 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.
7. 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.
8. Deschamps, S., Llaca, V., May, G.D. (2012) Genotyping-by-Sequencing in Plants. *Biology (Basel)*, **1**, 460-483.
9. 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.
10. 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.
11. 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.
12. Danecek, P., Auton, A., Abecasis, G., *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156-2158.
13. 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.
14. Bronner, I.F., Lorenz, S. (2019) Combined Genome and Transcriptome (G&T) Sequencing of Single Cells. *Methods Mol Biol*, **1979**, 319-362.
15. (2020) Promoting best practice in nucleotide sequence data sharing. *Sci Data*, **7**, 152.

16. Goodstein, D.M., Shu, S., Howson, R., *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*, **40**, D1178-1186.
17. Members, C.-N., Partners (2021) Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2021. *Nucleic Acids Res*, **49**, D18-D28.
18. 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.
19. 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.
20. 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. *Scientia Horticulturae*, **303**, 111199.
21. International HapMap, C. (2003) The International HapMap Project. *Nature*, **426**, 789-796.
22. Jung, S., Jesudurai, C., Staton, M., *et al.* (2004) GDR (Genome Database for Rosaceae): integrated web resources for Rosaceae genomics and genetics research. *BMC Bioinformatics*, **5**, 130.
23. 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.
24. Jung, S., Staton, M., Lee, T., *et al.* (2008) GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Res*, **36**, D1034-1040.
25. 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.
26. Yu, J., Jung, S., Cheng, C.H., *et al.* (2021) CottonGen: The Community Database for Cotton Genomics, Genetics, and Breeding Research. *Plants (Basel)*, **10**.
27. 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.
28. 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.
29. 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.
30. 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.
31. 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.
32. 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**.
33. Lawrence, C.J. (2007) MaizeGDB. *Methods Mol Biol*, **406**, 331-345.
34. 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.
35. Wegrzyn, J.L., Lee, J.M., Tearse, B.R., *et al.* (2008) TreeGenes: A forest tree genome database. *Int J Plant Genomics*, **2008**, 412875.
36. Falk, T., Herndon, N., Grau, E., *et al.* (2019) Growing and cultivating the forest genomics database, TreeGenes. *Database (Oxford)*, **2019**.
37. Garcia-Hernandez, M., Berardini, T.Z., Chen, G., *et al.* (2002) TAIR: a resource for integrated Arabidopsis data. *Funct Integr Genomics*, **2**, 239-253.
38. Poole, R.L. (2007) The TAIR database. *Methods Mol Biol*, **406**, 179-212.
39. 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.
40. 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.
41. Kalderimis, A., Lyne, R., Butano, D., *et al.* (2014) InterMine: extensive web services for modern biology. *Nucleic Acids Res*, **42**, W468-472.
42. Tello-Ruiz, M.K., Jaiswal, P., Ware, D. (2022) Gramene: A Resource for Comparative Analysis of Plants Genomes and Pathways. *Methods Mol Biol*, **2443**, 101-131.
43. Ware, D.H., Jaiswal, P., Ni, J., *et al.* (2002) Gramene, a tool for grass genomics. *Plant Physiol*, **130**, 1606-1613.
44. Ware, D. (2007) Gramene. *Methods Mol Biol*, **406**, 315-329.
45. Gladman, N., Olson, A., Wei, S., *et al.* (2022) SorghumBase: a web-based portal for sorghum genetic information and community advancement. *Planta*, **255**, 35.
46. Lyne, R., Sullivan, J., Butano, D., *et al.* (2015) Cross-organism analysis using InterMine. *Genesis*, **53**, 547-560.
47. Fasoula, D.A., Ioannides, I.M., Omirou, M. (2019) Phenotyping and Plant Breeding: Overcoming the Barriers. *Front Plant Sci*, **10**, 1713.

48. 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.
49. Miroslaw, M. (2001) Officially Released Mutant Varieties – The FAO/IAEA Database. *Plant Cell, Tissue and Organ Culture* **65**, 175-177.
50. 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.
51. 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.
52. 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.
53. McGill, B.J., Enquist, B.J., Weiher, E., *et al.* (2006) Rebuilding community ecology from functional traits. *Trends Ecol Evol*, **21**, 178-185.
54. Violle, V., Navas, M., Vile, D., *et al.* (2007) Let the concept of trait be functional! *Oikos*, **116**, 882-892.
55. Schneider, F.D., Fichtmueller, D., Gossner, M.M., *et al.* (2019) Towards an ecological trait-data standard. *Methods in Ecology and Evolution*, **10**, 2006-2019.
56. 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.
57. 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.
58. Lavorel, S., Grigulis, K. (2012) How fundamental plant functional trait relationships scale-up to trade-offs and synergies in ecosystem services. *Journal of ecology,*, **100**, 128-140.
59. Ni, J., Pujar, A., Youens-Clark, K., *et al.* (2009) Gramene QTL database: development, content and applications. *Database (Oxford)*, **2009**, bap005.
60. Singh, K., Batra, R., Sharma, S., *et al.* (2021) WheatQTLdb: a QTL database for wheat. *Mol Genet Genomics*, **296**, 1051-1056.
61. Reich, P.B., Wright, I.J., Lusk, C.H. (2007) Predicting leaf physiology from simple plant and climate attributes: a global GLOPNET analysis. *Ecol Appl*, **17**, 1982-1988.
62. 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.
63. Peat, H.J., 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.
64. Poschlod, P., Kleyer, M., Jackel, A.-K., *et al.* (2003) BIOPOP — A database of plant traits and internet application for nature conservation. *Folia Geobotanica*, **38**, 263-271.
65. 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.
66. Kühn, I., Walter Durka, Klotz, S. (2004) BiolFlor: a new plant-trait database as a tool for plant invasion ecology. *Diversity and Distributions*, **10**, 363-365.
67. 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. *Journal of ecology,*, **96**, 1266-1274.
68. Tavsanoglu, C., Pausas, J.G. (2018) A functional trait database for Mediterranean Basin plants. *Sci Data*, **5**, 180135.
69. Falster, D., Gallagher, R., Wenk, E.H., *et al.* (2021) AusTraits, a curated plant trait database for the Australian flora. *Sci Data*, **8**, 254.
70. Houle, D., Govindaraju, D.R., Omholt, S. (2010) Phenomics: the next challenge. *Nat Rev Genet*, **11**, 855-866.
71. Hati, A.J., 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.
72. Saleem, M.H., Potgieter, J., Mahmood Arif, K. (2019) Plant Disease Detection and Classification by Deep Learning. *Plants (Basel)*, **8**.
73. 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.
74. 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.
75. 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.
76. Steinbach, D., Alaux, M., Amselem, J., *et al.* (2013) GnpIS: an information system to integrate genetic and genomic data from plants and fungi. *Database (Oxford)*, **2013**, bat058.
77. Pommier, C., Michotey, C., Cornut, G., *et al.* (2019) Applying FAIR Principles to Plant Phenotypic Data Management in GnpIS. *Plant Phenomics*, **2019**, 1671403.
78. Brookes, A.J., Robinson, P.N. (2015) Human genotype-phenotype databases: aims, challenges and opportunities. *Nat Rev Genet*, **16**, 702-715.

79. Cobo-Simón, I. (2022) Cartograplant: Cyberinfrastructure to Improve Forest Health and Productivity in the Context of a Changing Climate., *Plant and Animal Genome XXIX Conference*, San Diego (CA)
80. 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.
81. Bulow, L., Schindler, M., Choi, C., *et al.* (2004) PathoPlant: a database on plant-pathogen interactions. *In Silico Biol*, **4**, 529-536.
82. Bulow, L., Schindler, M., 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.
83. 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 (Oxford)*, **2020**.
84. 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.
85. Dhondt, S., Wuyts, N., Inze, D. (2013) Cell to whole-plant phenotyping: the best is yet to come. *Trends Plant Sci*, **18**, 428-439.
86. 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.
87. Hill, D.P., D'Eustachio, P., Berardini, T.Z., *et al.* (2016) Modeling biochemical pathways in the gene ontology. *Database (Oxford)*, **2016**.
88. Poux, S., Gaudet, P. (2017) Best Practices in Manual Annotation with the Gene Ontology. *Methods Mol Biol*, **1446**, 41-54.
89. Chibucos, M.C., Tyler, B.M. (2009) Common themes in nutrient acquisition by plant symbiotic microbes, described by the Gene Ontology. *BMC Microbiol*, **9 Suppl 1**, S6.
90. 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.
91. Vining, K.J., Romanel, E., Jones, R.C., *et al.* (2015) The floral transcriptome of Eucalyptus grandis. *New Phytol*, **206**, 1406-1422.
92. 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.
93. 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.
94. 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.
95. 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.
96. 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.
97. Jarnuczak, A.F., Vizcaino, J.A. (2017) Using the PRIDE Database and ProteomeXchange for Submitting and Accessing Public Proteomics Datasets. *Curr Protoc Bioinformatics*, **59**, 13 31 11-13 31 12.
98. Okuda, S., Watanabe, Y., Moriya, Y., *et al.* (2017) jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Res*, **45**, D1107-D1111.
99. 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.
100. 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.
101. Ma, J., Chen, T., Wu, S., *et al.* (2019) iProX: an integrated proteome resource. *Nucleic Acids Res*, **47**, D1211-D1217.
102. Sharma, V., Eckels, J., Taylor, G.K., *et al.* (2014) Panorama: a targeted proteomics knowledge base. *J Proteome Res*, **13**, 4205-4210.
103. Desiere, F., Deutsch, E.W., King, N.L., *et al.* (2006) The PeptideAtlas project. *Nucleic Acids Res*, **34**, D655-658.
104. Deutsch, E.W. (2010) The PeptideAtlas Project. *Methods Mol Biol*, **604**, 285-296.
105. 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.
106. Members, M.S.I.B., Sansone, S.A., Fan, T., *et al.* (2007) The metabolomics standards initiative. *Nat Biotechnol*, **25**, 846-848.
107. 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.

108. 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 Trends in Analytical Chemistry*, **78**, 23-35.
109. 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.
110. Steinbeck, C., Conesa, P., Haug, K., *et al.* (2012) MetaboLights: towards a new COSMOS of metabolomics data management. *Metabolomics*, **8**, 757-760.
111. Considine, E.C., Salek, R.M. (2019) A Tool to Encourage Minimum Reporting Guideline Uptake for Data Analysis in Metabolomics. *Metabolites*, **9**.
112. 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.
113. Cooper, L., Jaiswal, P. (2016) The Plant Ontology: A Tool for Plant Genomics. *Methods Mol Biol*, **1374**, 89-114.
114. 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.
115. 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.
116. 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.
117. Oellrich, A., Walls, R.L., Cannon, E.K., *et al.* (2015) An ontology approach to comparative phenomics in plants. *Plant Methods*, **11**, 10.
118. 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.
119. 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.
120. Naithani, S., Partipilo, C.M., Raja, R., *et al.* (2016) FragariaCyc: A Metabolic Pathway Database for Woodland Strawberry Fragaria vesca. *Front Plant Sci*, **7**, 242.
121. Naithani, S., Raja, R., Waddell, E.N., *et al.* (2014) VitisCyc: a metabolic pathway knowledgebase for grapevine (Vitis vinifera). *Front Plant Sci*, **5**, 644.
122. 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.
123. 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.
124. Jaiswal, P., Usadel, B. (2016) Plant Pathway Databases. *Methods Mol Biol*, **1374**, 71-87.
125. Naithani, S., Jaiswal, P. (2017) Pathway Analysis and Omics Data Visualization Using Pathway Genome Databases: FragariaCyc, a Case Study. *Methods Mol Biol*, **1533**, 241-256.
126. Kattge, J., Ogle, K., Bönisch, G., *et al.* (2011) A generic structure for plant trait databases. *Methods in Ecology and Evolution*, **2**, 202-213.
127. Kattge, J., Bonisch, G., Diaz, S., *et al.* (2020) TRY plant trait database - enhanced coverage and open access. *Glob Chang Biol*, **26**, 119-188.
128. van Kleunen, M., Pysek, P., Dawson, W., *et al.* (2019) The Global Naturalized Alien Flora (GloNAF) database. *Ecology*, **100**, e02542.
129. Lee, Y.H. (2015) Meta-analysis of genetic association studies. *Ann Lab Med*, **35**, 283-287.
130. Dehghan, A. (2018) Genome-Wide Association Studies. *Methods Mol Biol*, **1793**, 37-49.
131. 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**.
132. 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.
133. 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.
134. Kraft, P., Zeggini, E., Ioannidis, J.P. (2009) Replication in genome-wide association studies. *Stat Sci*, **24**, 561-573.
135. 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**.
136. 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.
137. 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.

138. 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**.
139. Jeliazkova, N., Apostolova, M.D., Andreoli, C., *et al.* (2021) Towards FAIR nanosafety data. *Nat Nanotechnol*, **16**, 644-654.
140. Pacheco, A.R., Pauvert, C., Kishore, D., *et al.* (2022) Toward FAIR Representations of Microbial Interactions. *mSystems*, **7**, e0065922.
141. 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.
142. Mons, B., Neylon, C., Velterop, J., *et al.* (2017) Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, **37**, 49-56.