https://doi.org/10.1093/plphys/kiac468 PLANT PHYSIOLOGY 2023: 191: 35–46

Plant Physiology®

Update

Toward a data infrastructure for the Plant Cell Atlas

Noah Fahlgren [1,*] Muskan Kapoor [2] Galabina Yordanova [3] Irene Papatheodorou [3] Jamie Waese [4] Benjamin Cole [5] Peter Harrison [3] Doreen Ware [6,7] Timothy Tickle [8] Benedict Paten [9] Tony Burdett [3] Christine G. Elsik [10] Christopher K. Tuggle [2] and Nicholas J. Provart [4,*]

1 Donald Danforth Plant Science Center, Saint Louis, Missouri 63132, USA
2 Bioinformatics and Computational Biology Program, Department of Animal Science, Iowa State University, Ames, Iowa 50011, USA
3 EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK
4 Department of Cell and Systems Biology/Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario M5S 3B2, Canada
5 DOE-Joint Genome Institute, Lawrence Berkeley National Laboratory, 1, Cyclotron Road, Berkeley, California 94720, USA
6 Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor, New York 11724, USA
7 USDA ARS NAA Robert W. Holley Center for Agriculture and Health, Ithaca, New York 14853, USA
8 Data Sciences Platform, The Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, Massachusetts 02142, USA
9 UC Santa Cruz Genomics Institute, Baskin School of Engineering, 1156 High Street, Santa Cruz, California 95064, USA
10 Division of Animal Sciences/Division of Plant Science & Technology/Institute for Data Science & Informatics, University of Missouri, Columbia, Missouri 65211, USA

*Author for correspondence: nfahlgren@danforthcenter.org (N.F.), nicholas.provart@utoronto.ca (N.J.P.)

N.J.P. and N.F. wrote the first draft of the article. J.W. created the figures. M.K., G.Y., I.P., B.C., P.H., D.W., T.T., B.P., T.B., C.G.E., and C.K.T. all contributed to important discussions for subsequent iterations of the article and provided text for the Human Cell Atlas and Single Cell Expression Atlas sections. All authors provided editorial input.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plphys/pages/general-instructions) is: Nicholas J. Provart (nicholas.provart@utoronto.ca).

Abstract

We review how a data infrastructure for the Plant Cell Atlas might be built using existing infrastructure and platforms. The Human Cell Atlas has developed an extensive infrastructure for human and mouse single cell data, while the European Bioinformatics Institute has developed a Single Cell Expression Atlas, that currently houses several plant data sets. We discuss issues related to appropriate ontologies for describing a plant single cell experiment. We imagine how such an infrastructure will enable biologists and data scientists to glean new insights into plant biology in the coming decades, as long as such data are made accessible to the community in an open manner.

Introduction

The goal of the Plant Cell Atlas (PCA; Rhee et al., 2019; Rice et al., 2020), started in 2019, is to generate in a holistic way plant cell structure and organization data to help discover further cellular compartments, cell features, and cell types. The PCA was conceptualized from its inception as a community resource (www.plantcellatlas.org/) that will contain descriptions of the states of various kinds of plant cells. It will contain high-resolution spatiotemporal information of RNA and DNA molecules, proteins, and metabolites within plant cells. A recent roadmap paper outlines the steps to achieving the Plant Cell Atlas (Plant Cell Atlas Consortium,

Received July 14, 2022. Accepted September 18, 2022. Advance access publication October 6, 2022

ADVANCES

- High-throughput single-cell assays have led to efforts to produce comprehensive atlases of cell types and localization and organization of molecules, cells, and tissues.
- Databases, visualization, and modeling tools are being developed for exploring, analyzing, and visualizing multiscale and multimodal data.
- Data and metadata standards and vocabularies and consistent analysis pipelines are key for data sharing, annotation, curation, and integration.
- Cloud computing and cyberinfrastructure are enabling us to build community-based data infrastructure platforms.

2021). In this article, we dive into the data infrastructure that will be required to support the project, and cover aspects of data collection, curation, standardization, integration, and visualization. We also touch on funding such an infrastructure and explore some models used by other online resources. We expect that data from the Plant Cell Atlas-associated projects will consist initially of scRNA-seq matrices (i.e. transcript abundance for most genes in each cell of an scRNA-seq experiment), subcellular quantitative mass-spectrometry proteomics data, and fluorescent protein (i.e. GFP) localization images.

Databases and the PCA

A useful way of describing databases is to think about them being "infrastructure tier," "consolidation tier," or "investigator tier" databases (IAIC et al., 2010). Each tier has different levels of engineering, funding, species-specificity, scope, and management. In terms of repositories of gene expression data, large efforts like the Gene Expression Omnibus and Sequence Read Archive (GEO/SRA), run by the US National Center for Biotechnology Information (NCBI), and the Gene Expression Atlas effort run by the European Bioinformatics Institute (EBI), respectively, fall into the infrastructure tier category (Barrett et al., 2013; Papatheodorou et al., 2018; Cantelli et al., 2022; Moreno et al., 2022). Such efforts archive data from tens of thousands of studies. Most recently, the EBI has introduced a Single Cell Expression Atlas (SCEA;https://www.ebi.ac.uk/gxa/sc/home) section for single-cell expression profiling experiments (Papatheodorou et al., 2020), which contains data from four plant species. The Broad Institute in Massachusetts also offers a Single Cell Portal, at https://singlecell.broadinstitute.org/. These databases are great for retrieving data for follow-up studies or for providing interactive views of expression levels across metadata/QC metrics or mapped to cell ordinations, but in general do not focus on plant-specific functionality. The Single Cell Portal does not yet contain any plant data, but in theory would be capable of hosting plant data. It does, however, provide a uniform query interface for Human Cell Atlas (HCA) data, presented as curated "collections" (other collections, such as mouse anatomy and morphology single-cell datasets, are available). But the primary location for accessing HCA data is the HCA's Data Coordination Platform (see the "Infrastructure that could support the PCA" section).

Consolidation tier databases, such as The Arabidopsis Information Portal (TAIR; Lamesch et al., 2012), the Bio-Analytic Resource for Plant Biology (BAR; Toufighi et al., 2005; Waese and Provart, 2017), Gramene (Monaco et al., 2014; Tello-Ruiz et al., 2018, 2021), or MaizeGDB (Portwood et al., 2019) provide access to genomic data and to a lesser or greater extent to transcriptomic data. These databases, however, have yet to fully embrace single-cell RNA-seq data, with the exception of the BAR's root single-cell RNA-seq eFP Browser view (Waese-Perlman et al., 2021), based on data generated by Ryu et al. (2019).

Last, there are a small number of investigator tier databases for single-cell RNA-seq data from plants, notably the Plant Single Cell RNA-Sequencing Database from the Timmermans Lab (Ma et al., 2020) and the Wang Lab's Root Cell Atlas search tool (Zhang et al., 2019). These provide gene search functionalities. There are structural problems associated with keeping investigator tier databases online and up-to-date, and many end up going "dark" in the absence of funding and personnel who can keep web servers running; 62.3% of 326 databases listed in an early overview resource called DBcat (Discala et al., 2000) were considered "dead" when they were examined 18 years later (the first DBcat listing appear in May 1997; Attwood et al., 2015).

Arguably, species- or at least plant-specific consolidation tier resources, such as TAIR, the BAR, MaizeGDB, Gramene, and SoyKB (Joshi et al., 2014), are popular with biologists (as opposed to computational biologists) because they provide data from many different sources in an integrated manner. For instance, the BAR's ePlant tool (Waese et al., 2017) integrates natural variation data from the kilometer scale, expression data from the centimeter scale, subcellular localization from the submillimeter scale, protein–protein interaction data at the micrometer scale, and protein tertiary structure at the nanometer scale in a common interface to facilitate the navigation of such datasets. Several metadata views, such as annotation and gene structure, are also available. This is not to say that such consolidation tier resources are not useful for computational researchers. For example, the BAR maintains ThaleMine (Krishnakumar et al., 2015; Pasha et al., 2020), based on the InterMine framework (Kalderimis et al., 2014), which has an application programming interface (API) called BlueGenes (Yehudi et al., 2017) that has language bindings for Perl, Python, Ruby, and Java, allowing computational researchers to easily run programmatic queries of ThaleMine directly from scripts. TAIR and other consolidation tier databases do a great job of linking

Figure 1 The Plant Cell Atlas would sit at the nexus of an interconnected web of databases, in addition to housing data generated by single-cell approaches.

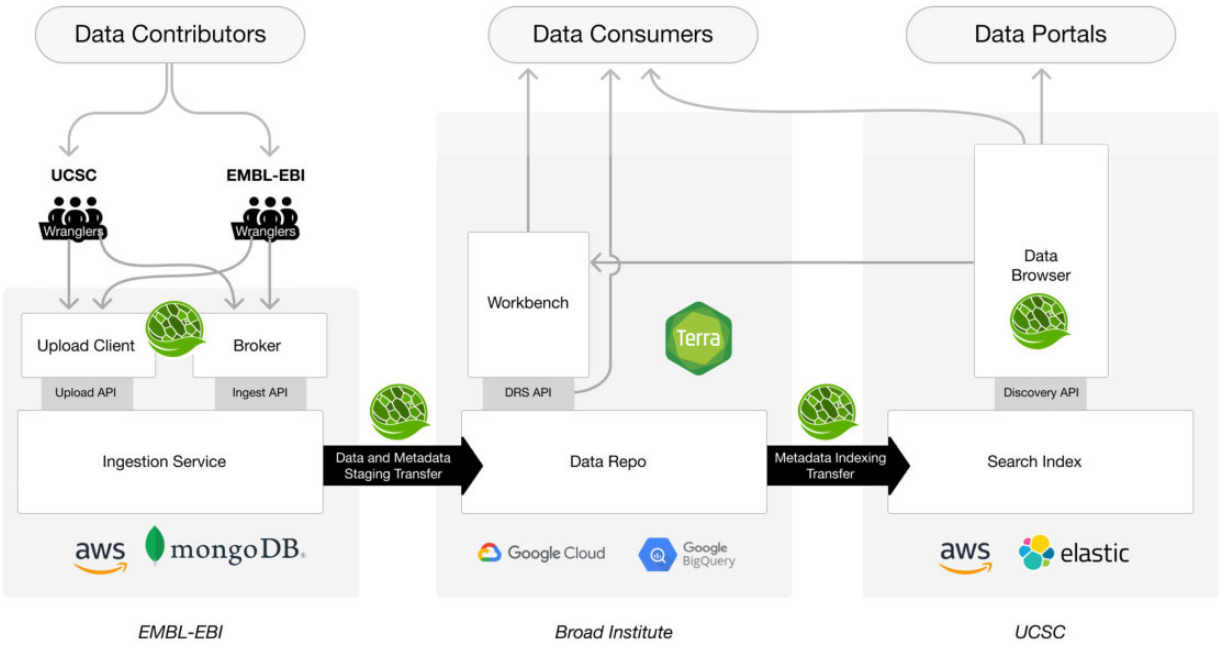

Figure 2 The current Human Cell Atlas Data Coordination Platform (HCA DCP) has evolved from the infrastructure proposed in the HCA white paper (Regev et al., 2018). The three boxes represent different aspects of the DCP (see text). The Plant Cell Atlas logo (depicted at the center of Figure 1) denotes parts of the infrastructure that would require plant-specific ontologies/metadata standards. API, application programming interface; DRS, data repository service; UCSC, University of California Santa Cruz; AWS, Amazon Web Services; EMBL-EBI, European Molecular Biology Laboratory—European Bioinformatics Institute.

to external resources and keeping their data as up-to-date as possible. The PCA would operate in a similar manner, warehousing where necessary and linking out to various resources where appropriate, see Figure 1. Several of these databases are part of the AgBioData Consortium (Harper et al., 2018) and it would make sense for the PCA to become part of this to help set metadata standards for plant single-cell data.

Infrastructure that could support the PCA

Two existing resources could be leveraged to provide infrastructure for the PCA: The Human Cell Atlas Data Coordination Platform or the EBI's SCEA. The goal of the HCA is to "create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease" (from the HCA website at

humancellatlas.org). The Data Coordination Platform (DCP) section of the HCA white paper (Regev et al., 2018) enumerates the scale and scope of the HCA project: "the Human Cell Atlas will contain petabytes of data on billions of cells and tissue sections across multiple modalities used by hundreds of labs around the world. A project of this scale and complexity demands an open, modular, and extensible approach to coordinating, standardizing, and sharing data." It took a team of engineers several years to set up the DCP and associated services, and many individuals now operate the system (see Figure 2, the current infrastructure). The main parts are a data ingestion service; a horizontally scalable synchronized data store residing on the Terra cloud platform (developed by the Broad Institute, Microsoft, and Verily); and a data browser and search index. Other different sites can act as tertiary portals for analyses, visualizations, and forms of data access, as denoted by the "Data Portals" tag in Figure 2. Currently, HCA data are ingested into Terra for scalable, secure storage, for which access is free for nonprofit organizations. Scientists have the option to use the cloud for subsequent storage or computational needs but here there are pass-through costs from the underlying cloud platform.

Currently, the HCA data portal (data.humancellatlas.org) permits the exploration of 26M cells from 113 anatomical organs, and 3.3K donors, encompassing 237 projects. Tools like the BioTuring Browser or Human Cell Atlas Galaxy instance can access the HCA data portal to permit seamless querying.

Discussions with members of the HCA DCP team, including Kathleen Tibbetts (Director, Data Engineering for the Data Science Platform, Broad Institute), Benedict Paten (Director, Computational Genomics Laboratory, University of California Santa Cruz), and Anthony Burdett (Technical Team Leader, Archival Infrastructure and Technology, EBI), provided insights into the possibility of leveraging components of the HCA's infrastructure for PCA data. Whereas the HCA Data Coordination Platform is open source, it is a complex ecosystem of services, and the PCA might not require all elements; it would be best to pick and choose pieces that are needed. For instance, the data ingestion service will require modification to be suitable for plant data, such as ontologies for environmental perturbations and plant parts (ENVO, Buttigieg et al., 2016; PO, Walls et al., 2019). The Data Browser would also need to be aware of such ontologies in order for researchers to be able to query PCA data residing in the Data Science Platform residing on Terra (https://app.terra.bio/). The ingestion service intakes the data while also performing basic quality assurance tests, such as metadata syntax, schemas, and file integrity, consistent with ontologies and controlled vocabularies. It is important to note that ingestion of HCA data is configured to a specific metadata schema and depends on that robust metadata schema. The PCA might need to develop its own metadata schema or extend the existing HCA metadata schema (see https://data.humancellatlas.org/metadata) with plant-specific ontologies as described below.

The EBI's SCEA project is simpler in its implementation than the HCA DCP. Both researchers and data curators at the SCEA use a tool called Annotare (Athar et al., 2019) to submit raw data and metadata to ArrayExpress (Parkinson et al., 2007; Athar et al., 2019), the container into which all types of expression data are deposited. Nicely, in the case of plant data, appropriate ontologies are available. Following submission, a data analysis pipeline is used to summarize expression values based on the latest EMSEMBL genome builds, and data are then integrated and may be visualized within the SCEA. One limitation of SCEA is that programmatic access is not very well supported, with the exception of being able to access SCEA within the Galaxy platform (Tekman et al., 2020). The SCEA group has worked with the HCA to develop minimum standards for reporting single-cell RNA-seq experiments (minSCe; Füllgrabe et al., 2020), and these standards in fact permit the SCEA group to use the Annotare tool to help with ingestion of HCA and other scRNA-seq data.

Ontologies and metadata

To build a practical data resource such as a biological database, data must be identified, collected, organized, and integrated into the database. It is useful to consider the approaches and challenges associated with these tasks in the context of the different types of databases described above. For infrastructure tier databases (e.g. GEO/SRA), investigators are typically responsible for depositing their own data and distributing the effort of curating research data into the community. To collect and integrate data from many independent investigators, infrastructure databases typically require standardized data formats and a minimum set of experimental and sample metadata descriptors (Barrett et al., 2013). Consolidation tier databases on the other hand are typically not data repositories and must identify datasets that should be included in the resource and may need to accept the data and metadata in a wider variety of formats, which places the effort of data curation on database curators (Harper et al., 2018).

Data curation and integration were major challenges identified for the Plant Cell Atlas during the community convening workshop (Rice et al., 2020). Development and community utilization of standardized data formats (e.g. FASTQ) and experimental metadata standards, e.g. Minimum Information about a high-throughput Nucleotide SEQuencing Experiment (MINSEQE; Brazma et al., 2012) and minSCe (Füllgrabe et al., 2020), are tools that can mitigate these challenges by distributing the effort of curating datasets onto data creators (Cock et al., 2010; Yilmaz et al., 2011; Sansone et al., 2019). Data standardization also promotes interoperability between resources that utilize the data, which when successful can lead to an ecosystem of tools that are modular and work in coordination, e.g. the wide range of tools that utilize data in FASTQ and SAM formats (Li et al.,

2009; Cock et al., 2010). However, data and metadata standardization alone does not promote data use and reuse, but are part of the FAIR (findable, accessible, interoperable, and reusable) guiding principles for data sharing (Wilkinson et al., 2016). The development and adoption of common community standards for data accessibility, format, and description will be key to the success of the Plant Cell Atlas data infrastructure.

For databases that provide services beyond data warehousing, datasets from disparate sources and of different types need to be integrated to create a knowledgebase that can be explored, queried, analyzed, and visualized. Two aspects are important here: standardized analysis pipelines and the use of ontologies. In terms of standardized analysis pipelines, recently developed algorithms like ComBat-seq (Zhang et al., 2020) can be used to permit comparisons between datasets generated in different laboratories. Ontologies are structured and controlled vocabularies that are used to represent knowledge by defining standardized terms and the relationships between them (Smith et al., 2007). Ontologies are used to annotate and label samples and associated datasets to create a layer of interoperability between datasets annotated with the same vocabularies, and provide a computational means of traversing the relationships between terms (Harper et al., 2018). For example, the Gene Ontology, one of the most widely used ontologies in biological databases, defines the set of possible gene functions and the relationships between them, which enables the computational analysis of gene function and gene families, both within and between species (Ashburner et al., 2000; Harper et al., 2018; The Gene Ontology Consortium, 2021). Similarly, the Plant Ontology defines terms related to plant anatomy and development and the relationship between structures and stages so that datasets annotated with these terms can be analyzed with a species-neutral approach (Walls et al., 2019). To meet the requirements of FAIR data standards, plant research communities will need three key components.

First, metadata standards that list the fields necessary for data interpretation from a specified experimental domain. Some examples of metadata standards include MINSEQE as well as MIAME (Minimum Information About a Microarray Experiment; Brazma et al., 2001) and a group of phenotyping databases such as BreedBase (Morales et al., 2022), GnpIS (Steinbach et al., 2013), PSB Interface for Plant Phenotype Analysis (PIPPA; https://pippa.psb.ugent.be/), and Plant Hybrid Information System (PHIS; Neveu et al., 2019). To enable interoperability between phenotypic domains, MIAPPE (Minimum Information About a Plant Phenotyping Experiment; Krajewski et al., 2015) was developed.

Second, ontologies or controlled vocabularies define metadata values, ensuring that they are objectively consistent and defined across datasets. For example, the Planteome project (Cooper et al., 2018) created three key ontologies: the Plant Trait Ontology (Jaiswal et al., 2005; Arnaud et al., 2022), which models species-independent plant traits under a broader scope; the Plant Ontology (Jaiswal et al., 2005; Cooper et al., 2013; Walls et al., 2019), which covers plant anatomical structures and development stages and allows interplant comparisons; and the Plant Experimental Conditions Ontology (Cooper et al., 2018), representing plant treatments. Other examples of ontologies include the Crop Ontology (Shrestha et al., 2012), an assorted species-specific ontology that depicts plant properties and techniques for analyzing them; the Agronomy Ontology (Aubert et al., 2017), which covers agronomic practices, techniques, and variables; the Environment Ontology (Buttigieg et al., 2013), which describes ecological environments; and the Statistics Ontology (Statistics Ontology Project, 2020), which describes statistical approaches. In the case of plant data in the EBI's SCEA, the Experiment Factor Ontology EFO (Malone et al., 2010) is able to encompass terms from other ontologies to describe samples in experiments (e.g. http://purl.obolibrary.org/obo/EO_0007404 uses the Environment Ontology term EO:0007404 to describe a "drought environment").

Third, machine-readable metadata exchange formats are important for data exchange. The standardized format for MIAME is often in the form of MAGE-TAB (MicroArray Gene Expression tabular; Rayner et al., 2006), and plant phenotyping and other wide range of fields commonly uses Investigation/Study/Assay tab-delimited (ISA-TAB) format (Rocca-Serra et al., 2010; Sansone et al., 2012).

Additionally, vocabularies can be linked or integrated to generate additional insights. For example, in work by Braun and Lawrence-Dill (2020), natural language processing was used to map text descriptions of plant phenotypes to formalized phenotype descriptions in the form of entity-quality statements, e.g. "entity: leaf" and "quality: increased length," where these statements were composed of ontology terms from the Plant Ontology, Gene Ontology, Phenotype and Trait Ontology, and Chemical Entities of Biological Interest ontology (Ashburner et al., 2000; Gkoutos et al., 2004; Cooper et al., 2013; Hastings et al., 2013). Using this approach, phenotype similarity networks were built using automated phenotype descriptions and could be successfully used to identify genes within and between species that function within a conserved pathway, even if they do not share sequence similarity (Braun and Lawrence-Dill, 2020). Similar approaches could enable comparisons between cell types within the Plant Cell Atlas or be used to make broader comparisons with other cell atlases (e.g. HCA). Databases like FungiDB in EuPathDB (Amos et al., 2022) provide comparative search options across species, which will be an important functionality of the PCA infrastructure.

Genome assembly, annotation, and curation concerns

A big issue for database curators is dealing with updated genome assemblies and constantly evolving annotations. This is a multifaceted problem with many levels. First, as genome sequencing technologies improve and assembly algorithms

get better, genome assemblies will change. Long-read sequencing and optical mapping can improve assemblies. Todd Michael and colleagues (Michael et al., 2018) used a MinION sequencer to appreciably improve the Arabidopsis Col-0 TAIR10 assembly, which still contained 29 larger misassemblies, had 117 gaps with unknown bases, and was missing about 25 Mb of repeat sequence, all in spite of the publication of the Arabidopsis genome more than 20 years ago (Arabidopsis Genome Initiative, 2000). The Michael et al. (2018) assembly covered 100% of the nonrepetitive genome space, with fewer gaps present than in the current TAIR10 assembly.

Once an assembly is updated, several downstream events need to happen. As a first step, gene model annotations need to be updated. Updates to gene model annotations can also happen independently of a new assembly. For instance, the Araport11 reannotation (Cheng et al., 2017) of the Arabidopsis Col-0 reference genome used the same underlying TAIR10 genome assembly as was used for the TAIR10 annotation, but gene models were revised based on transcripts generated from 113 published RNA-seq datasets. Any subsequent downstream applications, such as read mapping of RNA-seq data for expression quantification, should ideally be redone if an updated set of gene models is generated. In addition, it is likely that single-cell RNA-seq data could be used to inform gene structure predictions (Arzalluz-Luque and Conesa, 2018).

Depending on the plant species, different pipelines for updating genome annotations exist. In Arabidopsis, TAIR is the primary curator of the Arabidopsis Col-0 "reference" genome. And while long-read genome assemblies (e.g. Michael et al., 2018, but also from others) for other *Arabidopsis thaliana* ecotypes have been loaded into the Genome Context Viewer (Cleary and Farmer, 2017), released as part of Araport's resuscitation (Pasha et al., 2020), the creation of a pan-genome for the Arabidopsis species is still awaiting funding. Community-based annotations as enabled by tools like the Generic Online Annotation Tool (GOAT, https://goat.phoenixbioinformatics.org/) can facilitate this Herculean task by distributing the work among many researchers. For maize, an alternate strategy was used for the latest version of the genome. Here, the B73 reference variety was sequenced and assembled along with a set of 25 maize inbred lines known as the NAM founder lines by the National Science Foundation (NSF)-funded NAM Sequencing Consortium using long reads and a mate-pair strategy to create the RefGen_v5 assembly, released in January 2020 (Hufford et al., 2021). This assembly included a pangenome analysis showing that of the ~100,000 genes found in any of these lines, roughly only a third are present in all genotypes.

In contrast to this consortium-based approach, the predecessor B73 RefGen_V4 assembly and annotation effort was led by a smaller group of researchers in the Ware Laboratory at USDA ARS/Cold Spring Harbor Laboratory, focused on utilizing emerging long single molecule technologies (Jiao et al., 2017). Thus, even for a single species, annotation strategies can change over time. Enticingly, however, RNA-seq datasets that are part of EBI's expression atlases are reprocessed when there is a new genome release for any species that is part of ENSEMBL Plants.

The last aspect of annotation is often called "functional annotation," and this involves ascribing a function to a given gene/gene product. Typically, this is done based on arduous literature curation, whereby a gene is identified in a mutant screen and then characterized using molecular methods, that story is published and then a curation group like TAIR captures the details, or by "lifting over" the functional annotations associated with homologs of a given gene. Updates to Gene Ontology terms occur quite frequently, and thus, enrichment tests for differentially expressed genes should be considered only a snapshot at a given moment in "annotation time."

In the case of the Plant Cell Atlas, the possibility of updated genome assemblies and new genome annotations would need to be built into the data infrastructure. At the very least, which genome version and GFF file (general feature format file—used to describe genome annotations, such as where exons and introns start and stop) version was used for a particular analysis will need to be captured in metadata associated with a cell profiling experiment. How many curators/data wranglers will be necessary to ensure experiments are represented faithfully? This is a separate question from data infrastructure needs. Data wranglers play an important role in "ingesting" HCA data into the HCA Data Coordination Platform and into EBI's SCEA (see the "Infrastructure that could support the PCA" section).

Data infrastructure, sharing, and interconnectivity

As discussed above, database resources often need to consider at least two types of users. One set of users will benefit from accessing resources through a user-friendly graphical user interface that is designed to guide the users to the subset of data they are interested in, e.g. the TAIR, BAR, Gramene, and MaizeGDB web applications (Lamesch et al., 2012; Waese and Provart, 2017; Portwood et al., 2019; Tello-Ruiz et al., 2021). Another set of users will benefit from access to the underlying datasets and knowledgebase to do computational analyses at larger whole-genome, pangenome, cell atlas, and other dataset/multidataset levels. Access to the underlying data infrastructure can also support data sharing between databases. However, without standardized, machine-readable methods for accessing a data infrastructure, data use and sharing are impeded by the need to create resource-specific methods for each infrastructure (Harper et al., 2018). An API is a structured specification that defines how software applications can request information from a data infrastructure. APIs can be thought of as analogous to the structured vocabularies discussed above because APIs define the kinds of data that can be accessed, the relationships between data, and the data formats and methods that are required to both access and interpret data

(Harper et al., 2018). Whereas APIs can be specific to a data infrastructure (for instance, the EBI Search API accesses only EBI databases), APIs such as the Breeding API have been developed to support interoperability between databases by defining an infrastructure-independent specification (Selby et al., 2019). The adoption of APIs by database providers can help build an ecosystem of interconnected data, create mutual benefits between projects, and support FAIR data-sharing principles.

The Plant Cell Atlas "vision" white paper specifies that data infrastructure to support the PCA should incorporate and collaborate with existing tool and data platforms to leverage these resources and avoid duplication of effort (Plant Cell Atlas Consortium, 2021). Whereas data sharing and syndication of resources are powerful tools for building an ecosystem of platforms, the Plant Cell Atlas will still undoubtedly require its own storage and computing infrastructure. For long-term sustainability, the Plant Cell Atlas platform should be built with modularity, flexibility, and service-agnostic principles in mind. Virtualization and container-based systems such as Docker (https://docker.com) and Singularity (https://sylabs.io) allow software and services to be packaged into portable environments that can be moved between infrastructures (da Veiga Leprevost et al., 2017). Management of software and service containers can also be done using tools, such as Kubernetes, that are commonly available on local and cloud-based infrastructure (Novella et al., 2019). Additionally, APIs such as Tapis can be used to create a programmatic interface to multiple infrastructure resources so that management of a platform can be independent of the underlying resources it uses (Cleveland et al., 2020). Together, these tools can be used to build a platform that can quickly adapt to new technologies, services, and infrastructure platforms.

By developing the Plant Cell Atlas platform software and service stack independently from any particular infrastructure service provider, the Plant Cell Atlas could utilize multiple types of infrastructure platforms and adapt to changes in the technology landscape over time. Like the HCA DCP, the Plant Cell Atlas could similarly build a platform across multiple commercial cloud service providers, but it is also worth considering utilizing publicly-funded and community-based infrastructure resources such as CyVerse, KBase, XSEDE, Jetstream, Open Science Grid, ELIXIR, Galaxy, and others (Altunay et al., 2011; Towns et al., 2014; Fischer et al., 2017; Arkin et al., 2018; Swetnam et al., 2018; Drysdale et al., 2020; Tekman et al., 2020). For example, both the Legume Federation and SoyKB platforms utilize CyVerse services to publicly store data or for user authentication, respectively (Joshi et al., 2014; Dash et al., 2016; Swetnam et al., 2018). Whereas the EBI's SCEA does not offer an API, the SCEA is available as a Galaxy instance, facilitating high-throughput analyses (Papatheodorou et al., 2020). Science Gateways are also successful examples of community-based data and analysis portals that utilize the XSEDE national (US) cyberinfrastructure services (Wilkins-Diehr, 2007; Towns et al., 2014). Building a platform that utilizes the best features of a variety of resources will increase the resiliency of the Plant Cell Atlas, as technologies and funding resources change over time.

Visualizations and simulations

The OpenWorm project (openworm.org) might be a useful effort as inspiration for visualizing Plant Cell Atlas data, and for developing simulations based on such data. OpenWorm (Szigeti et al., 2014) was started in 2011 and aims to create models and frameworks for visualizing and simulating the biology of the roundworm *Caenorhabditis elegans* (Sarma et al., 2018). An early effort of OpenWorm, WormSim, created an environment for simulations, such as the results of virtual muscular activity driven by virtual neuronal potentials. The neuronal circuits are represented in models, e.g. c302 (Gleeson et al., 2018), and simulations may be run in the Sibernetic framework (Palyanov et al., 2018). Such models faithfully translate to a wriggling model worm on a computer screen. Researchers are also using OpenWorm resources to e.g. "paint" their own data onto OpenWorm's anatomical models of *C. elegans* to aid in visualization.

Another example of visualization is provided by ePlant (Waese et al., 2017), as mentioned earlier, which provides a unified platform for traversing a conceptual hierarchy of biological data from big (kilometer scale natural variation data) to small (nanometer scale protein tertiary structure data).

Visualization of single-cell data is often done using dimensionality-reducing t-distributed stochastic neighbor embedding (t-SNE) or uniform manifold approximation and projection (UMAP) plots (Moon et al., 2019), whereas URD plots (named after one of a trio of Norse goddesses who decide the fate of people) can be used reconstruct inferred cell lineage maps based on scRNA-seq data (Farrell et al., 2018). Perhaps a combination of such visualization tools, including newer ones like Azimuth (Hao et al., 2020), will provide access to reference Plant Cell Atlas datasets. Having such data integrated into a larger framework, as depicted in Figure 3, would be welcome. It would be a useful exercise to imagine use-case scenarios for being able to query Plant Cell Atlas data (see the PCA "Vision" paper; Plant Cell Atlas Consortium, 2021). One such use case is outlined in Figure 3.

Funding

A big question surrounds funding for a Plant Cell Atlas data infrastructure. At the infrastructure tier level, funding is stably provided by the government. In the case of the EBI, funding comes from its parent, the European Molecular Biology Laboratory (EMBL), whose funding in turn comes from the governments of EMBL's member states. In the case of NCBI, its funding is provided by Congress via bills introduced starting in 1987 by Senator Claude Pepper (Smith, 2013). Other databases have not been so fortunate. The saga of TAIR is illustrative but provides a possible model in its resolution. From TAIR's creation in 1999 until 2013, it was funded for the most part by the NSF in the U.S.A.,

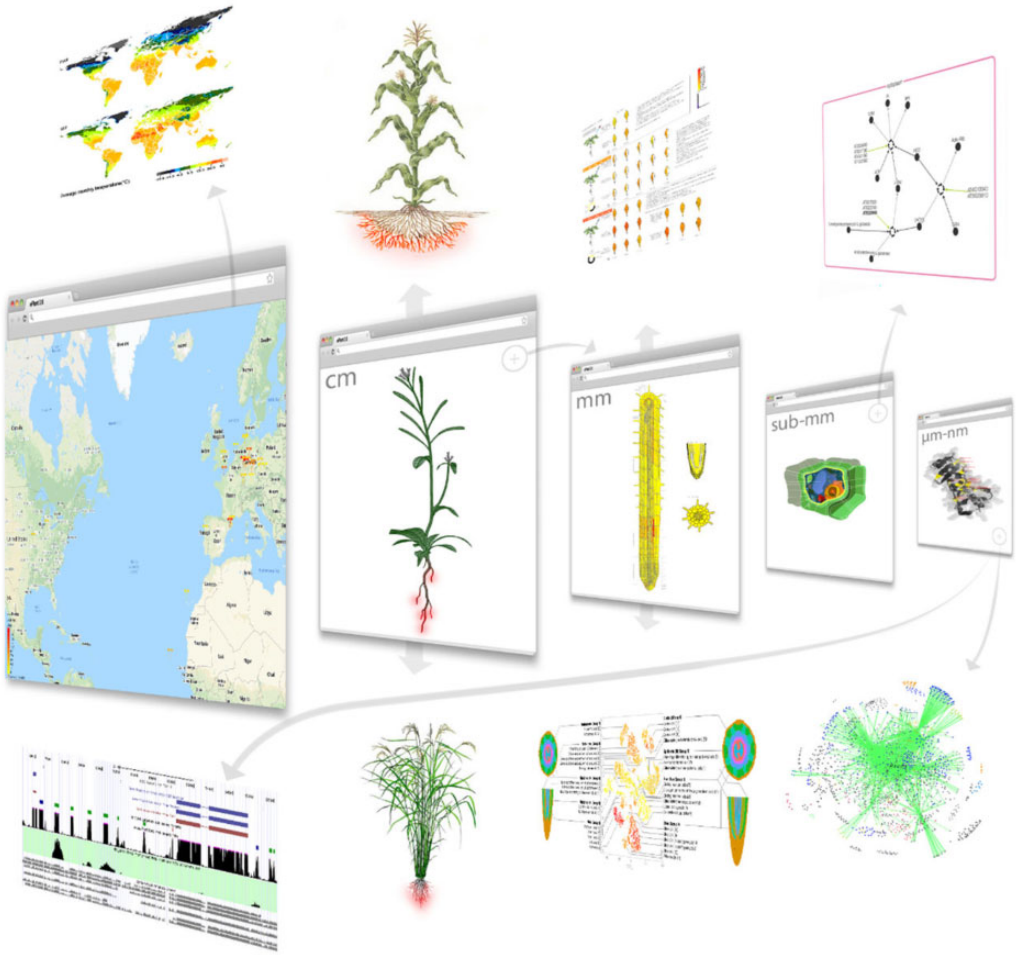

Figure 3 Traversing Plant Cell Atlas and other data in an integrated manner, across species. A hypothetical browser would access different data sources, including PCA data, to provide a unified browsing experience for plant biologists who might wish to explore, for example, how equivalent cells in different species respond to the same environmental stimulus.

through its Division of Biological Infrastructure. In 2009, TAIR was informed by the NSF that its grant would be reduced by 25% each year, ending in 2013. Another initiative, Araport, was created to ostensibly take on the task and cost of collecting and maintaining Arabidopsis datasets and tools, distributed across groups and countries (IAIC et al., 2010). Whereas the Araport framework was also funded by the NSF and by the Biotechnology and Biological Sciences Research Council (BBSRC) in the UK, ultimately its funding too was eliminated, necessitating a rescue (Pasha et al., 2020). Ironically, it was TAIR that helped to rescue Araport, and it was able to do so by implementing a subscription-based model starting in 2014 (Reiser et al., 2016). The manual curation aspect of TAIR was clearly valued enough by the community to pay for it, and the Araport project was never intended to do manual functional curation. Thus, the demand for high-quality annotations could be met by a subscription-based TAIR. Another interesting funding model is that used by the OpenWorm project. Early parts of that project were funded by a Kickstarter campaign that welcomed small contributions from the community to pay programmers. Further work is supported by volunteers who receive virtual "badges" for their efforts, under an open-source framework. In the case of the Plant Cell Atlas, the data infrastructure needs are substantial, and it will take considerable resources to build them and, importantly, to maintain them. One option might be to consider funding for a Plant Cell Atlas Data Synthesis Center from the Molecular and Cellular Biosciences division of the NSF. The National Center for Ecological Analysis and Synthesis (NCEAS) was the first synthesis center to be established in 1995. Since then, four more synthesis centers have been funded (Baron et al., 2017). "Synthesis centers do not support the collection of new data; instead, they add value to the data already collected across a diverse and extensive suite of research projects spanning a range of disciplinary and subdisciplinary domains" (Rodrigo et al., 2013), and thus the curation, databasing, and tool development aspects of a Plant Cell Atlas digital ecosystem would meet the criteria for a synthesis center.

Another opportunity to explore the possibility to use existing HCA DCP infrastructure components to create a prototype PCA platform was recently funded by the United States Department of Agriculture's Agricultural Genome to

OUTSTANDING QUESTIONS

- To what extent can existing infrastructures and platforms be used to build a Plant Cell Atlas data infrastructure?
- Will community-based curation help accelerate the development of biological databases and permit the use of artificial intelligence approaches to advance plant genomics and biology?
- To what extent will artificial intelligence generate further knowledge from the Plant Cell Atlas and other biological data?
- What kinds of databases, tools, and resources will be needed to integrate data at substantially different scales or resolutions? And how can temporal, environmental, and other metadata be integrated?

Phenome Initiative (AG2PI), to Chris Tuggle, Christine Elsik, Nicholas Provart, and Peter Harrison, along with collaborators including co-authors Tony Burdett, Tim Tickle, and Ben Cole. Efforts in this "seed funding" will be to test ingestion of representative plant and livestock scRNA-seq data with existing or newly developed metadata standards into components of the HCA DCP, and develop tools for a prototype data browser by Christine Elsik and teams from the EBI and UCSC. Finally, it should be pointed out that the EBI's SCEA is supported by the Wellcome Trust to encompass data from all species.

Data infrastructure as a data science platform

The field of data science formed from the combination of statistics and computer science in response to the massive growth of data and the need to turn "big data" into a resource for producing knowledge (Blei and Smyth, 2017). Data science approaches, including artificial intelligence approaches, are increasingly used in plant genomics, phenomics, and other areas, and artificial intelligence has emerged as a promising technology for accelerating plant breeding (Harfouche et al., 2019; Wang et al., 2020). It is tempting to envision that large consolidation tier databases that bring together disparate and multimodal datasets can be used as a platform for data science applications that identify patterns and produce knowledge that might otherwise remain unseen. However, it is important to recognize that there is "no free lunch" (Wolpert and Macready, 1997). Data management and integration, particularly for large multidimensional and multimodal datasets from multiple sources and produced with different approaches, are two major challenges for unlocking the potential of artificial intelligence in plant science (Williamson et al., 2021). Similarly, for single-cell datasets, integration of data between samples, experiments, and heterogeneous data modalities is challenging due to noise, sparsity, the lack of benchmarks, and other challenges (Lähnemann et al., 2020; Argelaguet et al., 2021). Methods such as ComBat-seq (Zhang et al., 2020), as mentioned earlier, might help, but these advances are likely insufficient to permit full PCA data integration. However, reason for optimism exists. For example, the ATTED-II plant coexpression network database has demonstrated a successful approach to integration of diverse gene expression datasets to enable multispecies comparisons (Obayashi et al., 2018). Artificial intelligence approaches have also been used to successfully leverage large, multispecies datasets to predict promoter activity and design highly active synthetic promoters (Jores et al., 2021). These successes highlight the feasibility of achieving the Plant Cell Atlas vision and encourage bold thinking about the fundamental questions in plant biology that such a resource can address. The development of community-based data management and data sharing standards, and the development of data infrastructure to support these activities and form the basis of a reference framework, are critical actions to address these challenges (Lähnemann et al., 2020; Williamson et al., 2021).

Concluding remarks

In assembling this review, it has become apparent that the Plant Cell Atlas should not reinvent the wheel. Although a new portal will need to be created, it would be highly beneficial to use existing architectures and frameworks, such as the Human Cell Atlas Data Coordination Platform or the EBI's SCEA system. The benefits of enabling such a platform will help address the Outstanding Questions we posed above. It is exciting to imagine an ecosystem of tools and analyses tapping into a Plant Cell Atlas digital infrastructure helping researchers advance plant biology in the coming decades, with stable funding provided through a synthesis center grant.

Acknowledgments

We apologize to those colleagues whose work was not cited due to space constraints. We are grateful to colleagues at the Broad Institute, the EBI, and UCSC for helpful discussions and insights into the HCA DCP. We have mentioned these individuals in the manuscript.

Funding

N.J.P. was supported by an National Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant. C.K.T., P.H., C.G.E., N.J.P., T.B., T.T., and B.C. received a United States Department of Agriculture (USDA) Agricultural Genome to Phenome Initiative (AG2PI) grant to explore possible single-cell repositories for plant and animal data.

Conflict of interest statement. None declared.

References

Altunay M, Avery P, Blackburn K, Bockelman B, Ernst M, Fraser D, Quick R, Gardner R, Goasguen S, Levshina T, et al. (2011) A science driven production cyberinfrastructure—the open science grid. J Grid Comput **9**: 201–218

Amos B, Aurrecoechea C, Barba M, Barreto A, Basenko EY, Bażant W, Belnap R, Blevins AS, Böhme U, Brestelli J, et al. (2022) VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. Nucleic Acids Res **50**: D898–D911

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408**: 796–815

Argelaguet R, Cuomo ASE, Stegle O, Marioni JC (2021) Computational principles and challenges in single-cell data integration. Nat Biotechnol **39**: 1202–1215

Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez F, Canon S, et al. (2018) KBase: The United States department of energy systems biology knowledgebase. Nat Biotechnol **36**: 566–569

Arnaud E, Cooper L, Shrestha R, Menda N, Nelson RT, Matteis L, Skofic M, Bastow R, Jaiswal P, Mueller L, et al. (2022) Towards a reference plant trait ontology for modeling knowledge of plant traits and phenotypes. 4th International Conference on Knowledge Engineering and Ontology Development **2012**: 220–225

Arzalluz-Luque Á, Conesa A (2018) Single-cell RNAseq for the study of isoforms—how is that possible? Genome Biol **19**: 110

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. (2000) Gene ontology: tool for the unification of biology. Nat Genet **25**: 25–29

Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, Snow C, Fonseca NA, Petryszak R, Papatheodorou I, et al. (2019) ArrayExpress update - from bulk to single-cell expression data. Nucleic Acids Res **47**: D711–D715

Attwood TK, Agit B, Ellis LBM (2015) Longevity of biological databases. EMBnet J **21**: 803

Aubert C, Buttigieg PL, Laporte M-A, Devare M, Arnaud E (2017) CGIAR agronomy ontology. http://purl.obolibrary.org/obo/agro.owl

Baron J, Specht A, Garnier E, Bishop P, Campbell CA, Davis FW, Fady B, Field D, Gross LJ, Guru SM, et al. (2017) Synthesis centers as critical research infrastructure. BioScience **67**: 113

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. (2013) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res **41**: D991–D995

Blei DM, Smyth P (2017) Science and data science. Proc Natl Acad Sci USA **114**: 8689–8692

Braun IR, Lawrence-Dill CJ (2020) Automated methods enable direct computation on phenotypic descriptions for novel candidate gene prediction. Front Plant Sci **10**: 1629

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet **29**: 365–371

Brazma A, Ball C, Bumgarner R, Furlanello C, Miller M, Quackenbush J, Reich M, Rustici G, Stoeckert C, Trutane SC, et al (2012) MINSEQE: Minimum Information about a high-throughput Nucleotide SeQuencing Experiment - a proposal for standards in functional genomic data reporting. https://zenodo.org/record/5706412

Buttigieg PL, Morrison N, Smith B, Mungall CJ, Lewis SE, the ENVO Consortium (2013) The environment ontology: contextualising biological and biomedical entities. J Biomed Semant **4**: 43

Buttigieg PL, Pafilis E, Lewis SE, Schildhauer MP, Walls RL, Mungall CJ (2016) The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. J Biomed Semant **7**: 57

Cantelli G, Bateman A, Brooksbank C, Petrov AI, Malik-Sheriff RS, Ide-Smith M, Hermjakob H, Flicek P, Apweiler R, Birney E, et al. (2022) The European Bioinformatics Institute (EMBL-EBI) in 2021. Nucleic Acids Res **50**: D11–D19

Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD (2017) Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J **89**: 789–804

Cleary A, Farmer A (2017) Genome Context Viewer: visual exploration of multiple annotated genomes using microsynteny. Bioinformatics **34**: 1562–1564

Cleveland SB, Jamthe A, Padhy S, Stubbs J, Packard M, Looney J, Terry S, Cardone R, Dahan M, Jacobs GA (2020) Tapis API development with python: best practices in scientific REST API implementation: experience implementing a distributed Stream API. *In* Practice and Experience in Advanced Research Computing, PEARC '20. Association for Computing Machinery, New York, NY, USA, pp 181–187

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res **38**: 1767–1771

Cooper L, Walls RL, Elser J, Gandolfo MA, Stevenson DW, Smith B, Preece J, Athreya B, Mungall CJ, Rensing S, et al. (2013) The plant ontology as a tool for comparative plant anatomy and genomic analyses. Plant Cell Physiol **54**: e1

Cooper L, Meier A, Laporte MA, Elser JL, Mungall C, Sinn BT, Cavaliere D, Carbon S, Dunn NA, Smith B, et al. (2018) The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Res **46**: D1168–D1180

Dash S, Campbell JD, Cannon EK, Cleary AM, Huang W, Kalberer SR, Karingula V, Rice AG, Singh J, Umale PE, 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–D1188

Discala C, Benigni X, Barillot E, Vaysseix G (2000) DBcat: a catalog of 500 biological databases. Nucleic Acids Res **28**: 8–9

Drysdale R, Cook CE, Petryszak R, Baillie-Gerritsen V, Barlow M, Gasteiger E, Gruhl F, Haas J, Lanfear J, Lopez R, et al. (2020) The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences. Bioinformatics **36**: 2636–2642

Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF (2018) Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science **360**: eaar3131

FischerJ, Hancock DY, Lowe JM, Turner G, Snapp-Childs W, Stewart CA (2017) Jetstream: a cloud system enabling learning in higher education communities. In Proceedings of the 2017 ACM SIGUCCS Annual Conference, SIGUCCS'17. Association for Computing Machinery, New York, NY, USA, pp 67–72

Füllgrabe A, George N, Green M, Nejad P, Aronow B, Fexova SK, Fischer C, Freeberg MA, Huerta L, Morrison N, et al. (2020) Guidelines for reporting single-cell RNA-seq experiments. Nat Biotechnol **38**: 1384–1386

Gkoutos GV, Green EC, Mallon A-M, Hancock JM, Davidson D (2004) Using ontologies to describe mouse phenotypes. Genome Biol **6**: R8

Gleeson P, Lung D, Grosu R, Hasani R, Larson SD (2018) c302: a multiscale framework for modelling the nervous system of Caenorhabditis elegans. Philos Trans R Soc B Biol Sci **373**: 20170379

Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al. (2020) Integrated analysis of multimodal single-cell data. Cell **184**: 3573–3587

Harfouche AL, Jacobson DA, Kainer D, Romero JC, Harfouche AH, Mugnozza GS, Moshelion M, Tuskan GA, Keurentjes JJB, Altman A (2019) Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. Trends Biotechnol **37**: 1217–1235

Harper L, Campbell J, Cannon E, Jung S, Poelchau M, Walls R, Andorf C, Arnaud E, Berardini TZ, Birkett C, et al. (2018) AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. Database **2018**: bay088

Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, et al. (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res **41**: D456–D463

Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, Ricci WA, Guo T, Olson A, Qiu Y, et al. (2021) De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science **373**: 655–662

Internation Arabidopsis Informatics Consortium (2010) An international bioinformatics infrastructure to underpin the Arabidopsis community. Plant Cell **22**: 2530–2536

Jaiswal P, Avraham S, Ilic K, Kellogg EA, McCouch S, Pujar A, Reiser L, Rhee SY, Sachs MM, Schaeffer M, et al. (2005) Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. Comp Funct Genomics **6**: 388–397

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin CS, et al. (2017) Improved maize reference genome with single-molecule technologies. Nature **546**: 524–527

Jores T, Tonnies J, Wrightsman T, Buckler ES, Cuperus JT, Fields S, Queitsch C (2021) Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. Nat Plants **7**: 842–855

Joshi T, Fitzpatrick MR, Chen S, Liu Y, Zhang H, Endacott RZ, Gaudiello EC, Stacey G, Nguyen HT, Xu D (2014) Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding. Nucleic Acids Res. **42**: D1245–D1252

Kalderimis A, Lyne R, Butano D, Contrino S, Lyne M, Heimbach J, Hu F, Smith R, Stěpán R, Sullivan J, et al. (2014) InterMine: extensive web services for modern biology. Nucleic Acids Res **42**: W468–W472

Krajewski P, Chen D, Ćwiek H, van Dijk AD, Fiorani F, Kersey P, Klukas C, Lange M, Markiewicz A, Nap JP, et al. (2015) Towards recommendations for metadata and data handling in plant phenotyping. J Exp Bot **66**: 5417–5427

Krishnakumar V, Hanlon MR, Contrino S, Ferlanti ES, Karamycheva S, Kim M, Rosen BD, Cheng CY, Moreira W, Mock SA, et al. (2015) Araport: the Arabidopsis information portal. Nucleic Acids Res **43**: D1003–D1009

Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A, et al. (2020) Eleven grand challenges in single-cell data science. Genome Biol **21**: 31

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res **40**: D1202–D1210

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079

Ma X, Denyer T, Timmermans MCP (2020) PscB: a browser to explore plant single cell RNA-sequencing data sets. Plant Physiol **183**: 464–467

Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H. (2010) Modeling sample variables with an Experimental Factor Ontology. Bioinforma Oxf Engl **26**: 1112–1118

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. (2018) High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat Commun **9**:541

Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, Kumari S, Amarasinghe V, Youens-Clark K, Thomason J, Preece J, et al. (2014) Gramene 2013: comparative plant genomics resources. Nucleic Acids Res **42**: D1193–D1199

Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, Yim K, Elzen A, van den Hirn MJ, Coifman RR, et al. (2019) Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol **37**: 1482–1492

Morales N, Ogbonna AC, Ellerbrock BJ, Bauchet GJ, Tantikanjana T, Tecle IY, Powell AF, Lyon D, Menda N, Simoes CC, et al. (2022) Breedbase: a digital ecosystem for modern plant breeding. G3 GenesGenomesGenetics **12**: jkac078

Moreno P, Fexova S, George N, Manning JR, Miao Z, Mohammed S, Muñoz-Pomer A, Fullgrabe A, Bi Y, Bush N, et al. (2022) Expression Atlas update: gene and protein expression in multiple species. Nucleic Acids Res **50**: D129–D140

Neveu P, Tireau A, Hilgert N, Nègre V, Mineau-Cesari J, Brichet N, Chapuis R, Sanchez I, Pommier C, Charnomordic B, et al. (2019) Dealing with multi-source and multi-scale information in plant phenomics: the ontology-driven Phenotyping Hybrid Information System. New Phytol **221**: 588–601

Novella JA, Emami Khoonsari P, Herman S, Whitenack D, Capuccini M, Burman J, Kultima K, Spjuth O (2019) Container-based bioinformatics with Pachyderm. Bioinformatics **35**: 839–846

Obayashi T, Aoki Y, Tadaka S, Kagaya Y, Kinoshita K (2018) ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. Plant Cell Physiol **59**: e3

Palyanov A, Khayrulin S, Larson SD (2018) Three-dimensional simulation of the Caenorhabditis elegans body and muscle cells in liquid and gel environments for behavioural analysis. Philos Trans R Soc B Biol Sci **373**: 20170376

Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, Burke M, Füllgrabe A, Fuentes AM-P, George N, et al. (2018) Expression Atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Res **46**: D246–D251

Papatheodorou I, Moreno P, Manning J, Fuentes AM-P, George N, Fexova S, Fonseca NA, Füllgrabe A, Green M, Huang N, et al. (2020) Expression Atlas update: from tissues to single cells. Nucleic Acids Res **48**: D77–D83

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, et al. (2007) ArrayExpress–a public database of microarray experiments and gene expression profiles. Nucleic Acids Res **35**: D747–D750

Pasha A, Subramaniam S, Cleary A, Chen X, Berardini TZ, Farmer A, Town C, Provart NJ (2020) Araport lives: an updated framework for Arabidopsis bioinformatics. Plant Cell **32**: 2683–2686

Plant Cell Atlas Consortium (2021) Vision, challenges and opportunities for a Plant Cell Atlas. eLife **10**: e66877

Portwood JL II, Woodhouse MR, Cannon EK, Gardiner JM, Harper LC, Schaeffer ML, Walsh JR, Sen TZ, Cho KT, Schott DA, et al. (2019) MaizeGDB 2018: the maize multi-genome genetics and genomics database. Nucleic Acids Res **47**: D1146–D1154

Rayner TF, Rocca-Serra P, Spellman PT, Causton HC, Farne A, Holloway E, Irizarry RA, Liu J, Maier DS, Miller M, et al. (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics **7**: 489

Regev A, Teichmann S, Rozenblatt-Rosen O, Stubbington M, Ardlie K, Amit I, Arlotta P, Bader G, Benoist C, Biton M, et al. (2018) The Human Cell Atlas White Paper. https://arxiv.org/abs/1810.05192

Reiser L, Berardini TZ, Li D, Muller R, Strait EM, Li Q, Mezheritsky Y, Vetushko A, Huala E (2016) Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a

case study of a subscription-based funding model. Database J Biol Databases Curation **2016**: baw018

Rhee SY, Birnbaum KD, Ehrhardt DW (2019) Towards building a Plant Cell Atlas. Trends Plant Sci **24**: 303–310

Rice S, Fryer E, Jha SG, Malkovskiy A, Meyer H, Thomas J, Weizbauer R, Zhao K, Birnbaum K, Ehrhardt D, et al. (2020) First plant cell atlas workshop report. Plant Direct **4**: e00271

Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, Field D, Harris S, Hide W, Hofmann O, et al. (2010) ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinforma Oxf Engl **26**: 2354–2356

Rodrigo A, Alberts S, Cranston K, Kingsolver J, Lapp H, McClain C, Smith R, Vision T, Weintraub J, Wiegmann B (2013) Science incubators: synthesis centers and their role in the research ecosystem. PLOS Biol **11**: e1001468

Ryu KH, Huang L, Kang HM, Schiefelbein J (2019) Single-cell RNA sequencing resolves molecular relationships among individual plant cells. Plant Physiol **179**: 1444

Sansone S-A, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, Fang H, Neumann S, Tong W, Amaral-Zettler L, et al (2012) Toward interoperable bioscience data. Nat Genet **44**: 121–126

Sansone S-A, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, Thurston M. (2019) FAIRsharing as a community approach to standards, repositories and policies. Nat Biotechnol **37**: 358–367

Sarma P, Lee CW, Portegys T, Ghayoomie V, Jacobs T, Alicea B, Cantarelli M, Currie M, Gerkin RC, Gingell S, et al. (2018) OpenWorm: overview and recent advances in integrative biological simulation of Caenorhabditis elegans. Philos Trans R Soc B Biol Sci **373**: 20170382

Selby P, Abbeloos R, Backlund JE, Basterrechea Salido M, Bauchet G, Benites-Alfaro OE, Birkett C, Calaminos VC, Carceller P, Cornut G, et al. (2019) BrAPI—an application programming interface for plant breeding applications. Bioinformatics **35**: 4147–4155

Shrestha R, Matteis L, Skofic M, Portugal A, McLaren G, Hyman G, Arnaud E (2012) Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. Front Physiol **3**:326

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol **25**: 1251–1255

Smith K (2013) A Brief History of NCBI's Formation and Growth (National Center for Biotechnology Information, USA) https://www.ncbi.nlm.nih.gov/books/NBK148949/

Statistics Ontology Project (2020) Statistics Ontology

Steinbach D, Alaux M, Amselem J, Choisne N, Durand S, Flores R, Keliet A-O, Kimmel E, Lapalu N, Luyten I, et al. (2013) GnpIS: an information system to integrate genetic and genomic data from plants and fungi. Database **2013**: bat058

Swetnam TL, Walls R, Devisetty UK, Merchant N (2018) CyVerse: a ten-year perspective on cyberinfrastructure development, collaboration, and community building. AGU Fall Meet. Abstr

Szigeti B, Gleeson P, Vella M, Khayrulin S, Palyanov A, Hokanson J, Currie M, Cantarelli M, Idili G, Larson S (2014) OpenWorm: an open-science approach to modeling Caenorhabditis elegans. Front Comput Neurosci **8**:137

Tekman M, Batut B, Ostrovsky A, Antoniewski C, Clements D, Ramirez F, Etherington GJ, Hotz H-R, Scholtalbers J, Manning JR, et al. (2020) A single-cell RNA-sequencing training and analysis suite using the Galaxy framework. GigaScience **9**: giaa102

Tello-Ruiz MK, Naithani S, Stein JC, Gupta P, Campbell M, Olson A, Wei S, Preece J, Geniza MJ, Jiao Y, et al. (2018) Gramene 2018: unifying comparative genomics and pathway resources for plant research. Nucleic Acids Res **46**: D1181–D1189

Tello-Ruiz MK, Naithani S, Gupta P, Olson A, Wei S, Preece J, Jiao Y, Wang B, Chougule K, Garg P, et al. (2021) Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. Nucleic Acids Res **49**: D1452–D1463

The Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res **49**: D325–D334

Toufighi K, Brady SM, Austin R, Ly E, Provart NJ (2005) The botany array resource: e-northerns, expression angling, and promoter analyses. Plant J **43**: 153–163

Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, Hazlewood V, Lathrop S, Lifka D, Peterson GD, et al. (2014) XSEDE: accelerating scientific discovery. Comput Sci Eng **16**: 62–74

da Veiga Leprevost F et al. (2017) BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics **33**: 2580–2582

Waese J, Fan J, Pasha A, Yu H, Fucile G, Shi R, Cumming M, Kelley LA, Sternberg MJ, Krishnakumar V, et al. (2017) ePlant: visualizing and exploring multiple levels of data for hypothesis generation in plant biology. Plant Cell **29**: 1806–1821

Waese J, Provart NJ (2017) The bio-analytic resource for plant biology. Methods Mol Biol Clifton NJ **1533**: 119–148

Waese-Perlman B, Pasha A, Ho C, Azhieh A, Liu Y, Sullivan A, Lau V, Esteban E, Waese J, Ly G, et al. (2021) ePlant in 2021: new species, viewers, data sets, and widgets. bioRxiv: 2021.04.28.441805

Walls RL, Cooper L, Elser J, Gandolfo MA, Mungall CJ, Smith B, Stevenson DW, Jaiswal P (2019) The plant ontology facilitates comparisons of plant development stages across species. Front Plant Sci **10**: 1–17

Wang H, Cimen E, Singh N, Buckler E (2020) Deep learning for plant genomics and crop mprovement. Curr Opin Plant Biol **54**: 34–41

Wilkins-Diehr N (2007) Special issue: science gateways—common community interfaces to grid resources. Concurr Comput Pract Exp **19**: 743–749

Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. SciData **3**: 160018

Williamson HF, Brettschneider J, Caccamo M, Davey RP, Goble C, Kersey PJ, May S, Morris RJ, Ostler R, Pridmore T, et al. (2021) Data management challenges for artificial intelligence in plant and agricultural research. F1000Research **10**: 324

Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput **1**: 67–82

Yehudi Y, Butano D, Chadwick M, Clark-Casey J, Contrino S, Heimbach J, Lyne R, Sullivan J, Micklem G (2017) Forever in BlueGenes: a next-generation genomic data interface powered by InterMine. F1000Research **6**, https://doi.org/10.7490/f1000research.1114527.1

Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, et al. (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol **29**: 415–420

Zhang T-Q, Xu Z-G, Shang G-D, Wang J-W (2019) A single-cell RNA sequencing profiles the developmental landscape of Arabidopsis root. Mol Plant **12**: 648–660

Zhang Y, Parmigiani G, Johnson WE (2020) ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics Bioinforma **2**: lqaa078