The Plant Cell, Vol. 24: 2248–2256, June 2012, www.plantcell.org

COMMENTARY

Taking the Next Step: Building an Arabidopsis Information Portal [OA]

The International Arabidopsis Informatics Consortium[1,2]

The Arabidopsis Information Portal (AIP), a resource expected to provide access to all community data and combine outputs into a single user-friendly interface, has emerged from community discussions over the last 23 months. These discussions began during two closely linked workshops in early 2010 that established the International Arabidopsis Informatics Consortium (IAIC). The design of the AIP will provide core functionality while remaining flexible to encourage multiple contributors and constant innovation. An IAIC-hosted Design Workshop in December 2011 proposed a structure for the AIP to provide a framework for the minimal components of a functional community portal while retaining flexibility to rapidly extend the resource to other species. We now invite broader participation in the AIP development process so that the resource can be implemented in a timely manner.

INTRODUCTION

Increasing global demands for improved agricultural products including fiber, food, and fuel intensifies the need for a thorough understanding of the basic biology and ecology of plants. The growing challenges of climate instability, competition for arable land, and the need to transition to renewable and carbon-neutral energy sources will only be met through an expanded and energetic focus on basic plant biology. Research involving *Arabidopsis thaliana*, the primary reference and model plant, has increasingly impacted our understanding of other plants and underpins many major breakthroughs in plant science in the last 25 years. As we look to the next 25 years of plant biology research, it is clear that a shift from a descriptive to a predictive science will require major innovations in research approaches and information storage, integration, and management. It will be necessary to manage vast amounts of data effectively and present them dynamically to a globally distributed user group whose interests are broad and whose needs will constantly evolve. The *Arabidopsis* community can lead the way for the plant sciences through the integration of diverse data sets and the development of new computational tools and resources that will facilitate the transfer of knowledge to crops and other economically and agronomically important species. As in nearly all major scientific endeavors, the *Arabidopsis* community is global. However, in recent years, most of the financial support for the primary community informatics resource has been provided on a national level, specifically, the funding of The Arabidopsis Information Resource (TAIR) by the U.S. National Science Foundation. Simultaneously, many groups developing data sets and analysis tools in other countries are well funded to do their work and are eager and willing to contribute those resources broadly to advance science. Thus, we need a mechanism to leverage funding and resource sharing across borders.

The challenge for the community is to move from one primary public *Arabidopsis* database orbited by numerous, yet disconnected, smaller databases into a dynamic, modular, and distributed international consortium of databases with a single point of access for users. This integration will be provided by the Arabidopsis Information Portal (AIP). The AIP will be built to address the international nature of plant biology, requiring that it be designed, from the outset, to facilitate the coordination and integration of data and scientists from around the world. This AIP will serve as the primary user interface, providing dynamic access to all resources that are operationally compatible, and will function as the central hub for the coordination of *Arabidopsis* informatics. As a center point and portal for *Arabidopsis* information, it will define standards for data storage, access, and interconnectivity. It is worthwhile emphasizing that the *Arabidopsis* community in particular, and the plant community as a whole, needs the services currently provided by TAIR, and in the near future to be provide by the AIP, because innumerable non-*Arabidopsis* publications reference *Arabidopsis* genes. In the absence of such a resource, future advances in plant biology and bioinformatics would be confounded, resulting in a substantial loss in the quality of plant science and analyses of nonplant systems that build on the findings made in *Arabidopsis*.

The questions, therefore, that have driven the proposed design of the AIP are as follows. How can we develop a robust resource that integrates novel and evolving tools? How can the needs of the user community be most effectively addressed? How can the *Arabidopsis* informatics efforts be internationalized to maintain long-term stability for its vital bioinformatics resources? Can the AIP for the reference plant *Arabidopsis* serve as a model informatics interface for all plant species?

A PATH TOWARD A POSSIBLE SOLUTION

Step 1. The AIP: Functionality

The AIP will function as a centralized resource that will integrate distributed databases harboring genomic, transcriptomic, proteomic, metabolomic, functional, phenotypic,

[1]A list of participants and their affiliations is provided at the end of this article.
[2]Address correspondence to arabidopsisinformatics@gmail.com.
[OA]Open Access articles can be viewed online without a subscription.
www.plantcell.org/cgi/doi/10.1105/tpc.112.100669

and other diverse data types, some of which have yet to be described. These data would be generated internationally from experiments such as those involving different ecotypes, treatments, mutants, and cell types. In addition, the data will be stored and visualized using a variety of formats and tools developed by different groups. Thus, the major challenge for the AIP is the development of systems to connect these data smoothly on a common platform and develop new functionalities that would enable seamless queries across these data, resulting in new insights and transformative research.

As part of the Design Workshop organized by the Steering Committee of the International Arabidopsis Informatics Consortium (IAIC) in December 2011, participants were challenged to imagine features that researchers will need from a portal 10 years from now. We asked, what features will the AIP provide in the year 2021? This exercise stimulated many creative and specific examples of functions and uses, and we have distilled the more prominent of these to describe the portal's initial, essential characteristics.

The AIP clearly needs to be engaging to its users. This will require interactive tools and ways to facilitate communication among scientists, enhancing the international nature of the project. Interaction with the AIP will take place via elegant and appealing user interfaces that will include novel visualization tools that capture the richness and breadth of the underlying data and convey them to biologists clearly and simply. These tools will be highly customizable, allowing the users to define the data displayed, including the researcher's own data, the scale or type of view, the types of comparisons, etc. Some visualization tools may be primarily web-based (i.e., operate in the users' web browsers), while others may be non-browser-based applications that would operate on the user's desktop but retrieve data from the portal. Users would be able to access data using web browsers as well as more specialized visualization and analysis tools, such as Integrated Genome Browser (Nicol et al., 2009), R (R Development Core Team, 2012), TableView (Johnson et al., 2003), Cytoscape (Smoot et al., 2011), and others. Diverse modes of providing access to data would be supported, maximizing the impact of international investments in *Arabidopsis* data sets. One of the portal's most important contributions will be easy accessibility of numerous, well-curated data sets, and this accessibility, combined with the powerful and advanced tools contained within the site, will enable in silico predictive experiments. Most importantly, to allow the integration of many data sets, to permit the evolution of these tools, and to grow with the needs of the community, the portal must be extensible, which will be achieved via a modular design. To facilitate the development and improvement of AIP components by different groups, the underlying architecture will be open, robust, dynamic, reliable, easy to understand, and scalable.

Step 2. The AIP: Framework and Requirements

As described previously (International Arabidopsis Informatics Consortium, 2010), the AIP will serve as a hub for informatics resources, linking both core and noncore components. From the user's point of view, the AIP will provide a front-end, primarily web-based graphical user interface exposing functionality provided by the main AIP components as well as data, tools, and resources provided by the noncore components (modules). The AIP will also provide the framework and interface to support the development of new modules, thus encouraging developers to contribute new functionalities over time. The main portal components will be developed by the AIP team and will build upon a set of platform services that enable access to low-level computer, data, visualization, and instrumentation resources. These platform services could be provided by one or more entities such as the iPlant Collaborative project (Goff et al., 2011) or other international cyber infrastructure provider. Building upon an available project leverages many years of effort and funding already invested in developing a production platform and facilitates cross-project synergies and collaborations. Furthermore, given the existing international investment in such platforms, recreating such technology is not within the scope of the AIP and should only be obtained through active engagement with existing providers.

The primary user needs to be addressed by the AIP will provide basic functionality for the initial release of the portal (Figure 1). These services include the following: Curation: verifies and validates data using social, manual, and automated mechanisms; Visualization: provides mechanisms for viewing and interacting with data in graphical formats; Execution: executes supported operations on data, such as data searching and custom bulk data downloads; Registry: publishes a directory of available third-party modules; Discovery/Search: searches and discovers data across all providers; Metadata: manages metadata documenting files, images, etc.; Data management: basic data management and browsing; Notification: manages notifications between applications and individuals; Collaboration: manages access and permissions for objects across all providers; Authentication/Authorization: handles federated authorization/authentication for all users; Identity: handles mapping of federated identities for all users.

The collection of databases and resources that the AIP must serve include core components, described below, such as the Gold Standard Genome Annotation, Curation of Functional Data, and the databases for the stock centers (International Arabidopsis Informatics Consortium, 2010). A key design goal is for the AIP to provide useful functions to the majority of users from the start, requiring that the data corresponding to these core components be accessible from the first day of operation. In addition to these core components, there are a number of important data sets and analysis tools that currently can be defined as "noncore components." This definition is dynamic and contingent on the rapidly evolving needs of the community. Looking ahead to subsequent releases of the AIP, third-party module examples such as those seen in Figure 2 and described below will expand the capabilities of the AIP to more users and grow the overall accessibility and value of the AIP to the global community. The modular structure of the AIP will coordinate and facilitate the simple and rapid development of these components, allowing the community to contribute

COMMENTARY

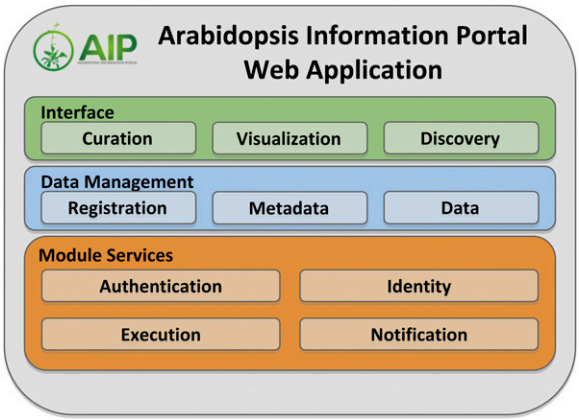

Figure 1. Main Components of the AIP, Version 1.0.

An overview of the user needs that the AIP will be designed to satisfy, including the basic functionalities to be built into the initial release of the portal.

noncore components and even develop potential replacements for core components as needed. All components should be able to provide data downloads in commonly used formats; currently, these include formats such as BAM, BED, VCF, GFF3, or XGMML, but new formats would need to be supported as they develop. A working draft of AIP requirements proposed by participants at the Design Workshop can be found at the IAIC Resources webpage (http://www.arabidopsisinformatics.org/resources/).

Step 3. The AIP: Integration of Currently Available Modules

Core modules that should be immediately integrated into the AIP include the following.

Gold Standard Genome Annotation

This proposed module is part of the minimal data set critical for proper AIP functioning and would be based on data provided by the current TAIR release, perhaps with minor updates, that would start to be updated again once it was modularized in the AIP. As should be the case for other modules as well, this module could represent a fundable unit, and support would be needed to update the static genome data presented in the AIP. A single standardized release of the genome and genome annotations is essential, not only to provide reliable data to users but also to provide map positions and Arabidopsis Genome Initiative codes that will be central for data integration across the AIP. This module will provide an easy-to-use, web-based genome browser that displays gene model diagrams, similar to the current TAIR design, but it will also provide access to the same data via easy-to-use web services (Jenkinson et al., 2008) needed to support desktop tools used in bioinformatics and data analysis. Another important facet will be to provide convenient but robust mechanisms for researchers to update gene models as new information becomes available. Logical extensions to this module would include integrating related data from the 1001 Genomes Project (Weigel and Mott, 2009) and undertaking efforts to map the probe sets from microarray platforms.

Stock Center Databases

The ABRC (https://abrc.osu.edu/), Nottingham Arabidopsis Stock Centre (http://arabidopsis.info/), and RIKEN Biological Resource Center (http://www.brc.riken.jp/inf/en/index.shtml) are the main providers of *Arabidopsis* seed and DNA stocks. It is crucial that the *Arabidopsis* community and other scientists can easily order these stocks that are typically based on a stock center identifier. There should be no change in the way the current centers partition the orders among themselves to obviate any loss of order numbers and decrease vital revenue needed to maintain the resources. One possible extension of this module could be a list of "tools and services" whereby providers of, for example, microarray hybridization and ultra-high-throughput sequencing services could post links for fee-for-service operations. AIP would also provide sequence coordinates and other relevant information related to stocks, such as annotations of T-DNA insertion sites. Seamless integration of information between the globally located stock centers and presentation in the AIP would be a useful improvement for users; currently, users may need to access the databases separately to gather information.

Curation of Functional Data

This module would maintain a comprehensive list of gene, allele, and germplasm names and Gene Ontology (GO) annotations describing gene product function and localization, Plant Ontology annotations describing gene expression patterns, and phenotype descriptions. In the future, the use of controlled vocabularies may be extended to other data types, such as phenotypes and experimental conditions, for better data integration across experiments and species. The GO Consortium (Ashburner et al., 2000) is funded to carry out literature curation and ontology development for a set of reference genomes, including *Arabidopsis*, and can serve as the source of these annotations for the AIP. It may be necessary to supplement the GO Consortium effort with a community annotation approach to ensure that all information is captured. In addition, the AIP should provide a way for researchers to "tag" gene records with newly published articles describing gene function or comment on GO annotations that their own research indicates may be incorrect or incomplete. Automated GO annotation for genes lacking experimental data could be performed by the group(s) responsible for updating the *Arabidopsis* gene models (see Gold Standard Genome Annotation above). Tools such as AgriGO (Du et al., 2010) or GOrilla (Eden et al., 2009) could be integrated to provide the ability to perform GO term functional enrichment tests. Responsibility for providing expression patterns could be

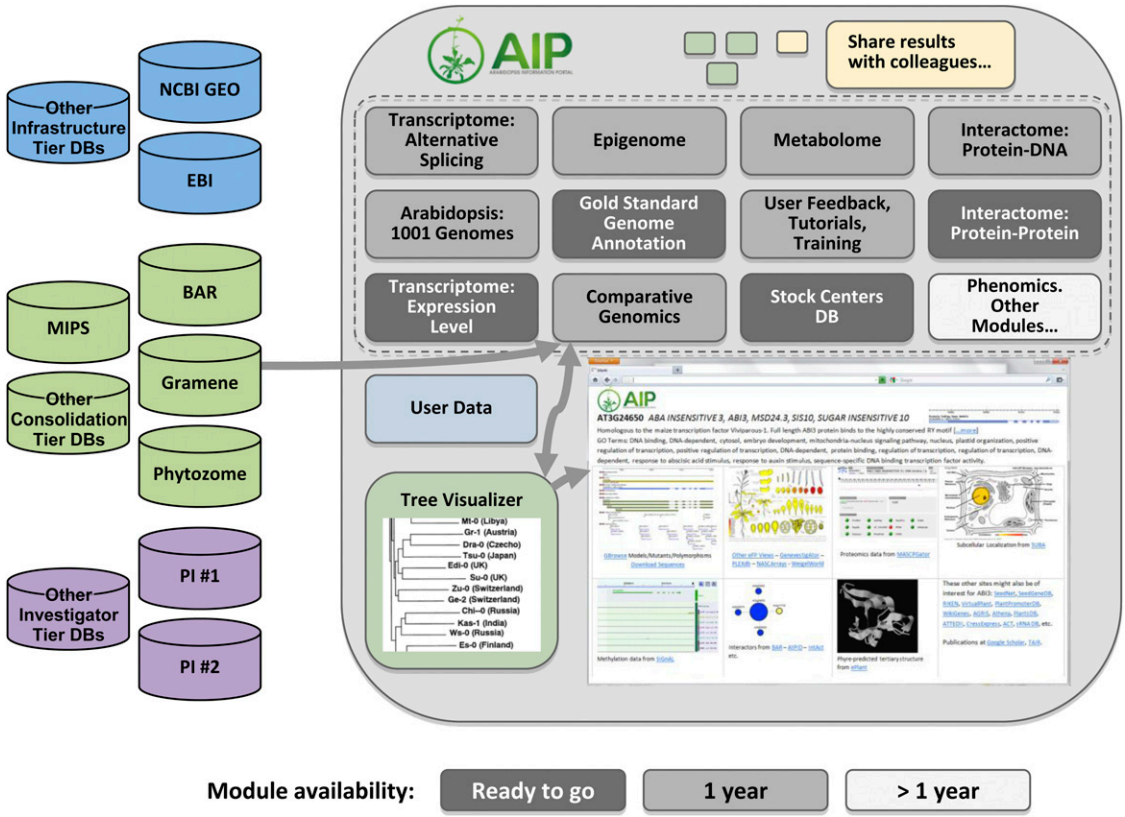

Figure 2. Examples of Module Contributions to the AIP.

Modules can be built on existing infrastructure, consolidation, and investigator tier databases. Some modules, denoted with dark gray backgrounds, are ready to be integrated into the AIP immediately, while those with medium to light gray backgrounds would require further development to be part of the AIP. As an example, arrows denote lines of communication activated by adding a Comparative Genomics module visualized with a Tree Visualizer app to a user's workspace for a gene of interest. Other modules would communicate with each other or with the depicted external databases if a user selected it to appear on his/her workspace. Note that the databases listed do not constitute a complete list; many that could be added were omitted due to space limitations.

assumed by the Transcriptome and Proteome modules, and phenotype data could be provided by a Phenomics module. Collection of gene, allele, and germplasm names may be accomplished through a coordinated effort including the GO Consortium, the stock centers, a nomenclature committee, and individual researchers, facilitated by a name registry tool hosted at the AIP.

Additional available modules that could be integrated into the AIP include the following.

Transcriptome: Expression Level

This proposed module, as well as the two following, were not originally envisioned as one of the core components (International Arabidopsis Informatics Consortium, 2010) but are clearly of importance to the community. The Transcriptome: Expression Level module would be based on existing (primarily microarray) data available in Gene Expression Omnibus/ArrayExpress/NASCArrays, perhaps from a consolidation tier-level provider such as Genevestigator (Zimmermann et al., 2004; Hruz et al., 2008) or the Bio-Array Resource (Toufighi et al., 2005; Winter et al., 2007). Additionally, coexpression neighbors should be readily available, such as those available immediately from the VirtualPlant project (Katari et al., 2010) or ATTEDII (Obayashi et al., 2009). It will be essential that the original .CEL or .SRA format data sets are easily accessible to AIP users. Such a module will also need to encourage data providers to undertake proper curation of gene expression data sets.

Interactome: Protein-Protein

This module could be sourced from an infrastructure tier database such as IntAct (Kerrien et al., 2012) or from a consolidator tier database such as AtPID (Cui et al., 2008) or the Arabidopsis Interaction Viewer database (Geisler-Lee et al., 2007). Ideally, the interactions could be rendered as an interaction network dynamically, for example, using CytoscapeWeb (http://cytoscapeweb.cytoscape.org/). Some consolidation-level interactome databases have predicted

COMMENTARY

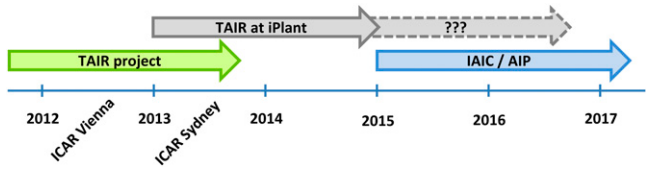

Figure 3. The Planned Transition from TAIR to AIP, Including a Transitional Period in Which TAIR Resources Are Housed at iPlant.

The coming years are indicated on the horizontal axis, including the next two planned International Conferences on Arabidopsis Research (ICAR). The timing of the transition of TAIR to the iPlant servers, and the development of the AIP, is indicated above.

interactions; therefore, the evidence used to construct an interaction must be made clear in any AIP interactome module.

Proteome

This module would be based logically on the extant MASCPGator (Joshi et al., 2011). This tool provides protein expression data by aggregating protein data from 10 proteomics databases covering phosphorylation, phosphomodulation, subcellular localization, fluorescent protein subcellular localization, and organ spectral count in a (currently) small set of sampled organs.

Step 4. The AIP: Examples of Modules That Could Be Integrated into the AIP in the Near Future

Transcriptome: Alternative Splicing and Other Forms of Transcript Variation

Many plant genes produce multiple transcripts encoding different protein products due to alternative splicing, alternative promoters, and alternative transcription termination. New methods for ultra-high-throughput sequencing of cDNA (RNA-Seq) are making it possible to assess alternative splicing and how splicing patterns vary across treatments (Dugas et al., 2011; Gulledge et al., 2012), developmental stages (Li et al., 2010), or cell and organ types (Zhang et al., 2010; Davidson et al., 2011). To help researchers assess alternative transcription in their genes of interest, the AIP will provide access to RNA-Seq data sets in standard formats, such as BAM (Li et al., 2009) and Tabix (Li, 2011). The goal of the AIP will be to provide these data in formats that will support an individual laboratory's research as well as the development of new methods and software for analysis and visualization. It is envisioned that this module would feed data into the Gold Standard Genome Annotation module.

Interactome: Protein-DNA

IntAct is currently archiving protein–DNA interactions (Kerrien et al., 2012), and position weight matrices are likely to be the preferred data format for describing protein–DNA interactions. Several other groups would also be able to contribute to the establishment of this module, such as AGRIS (Yilmaz et al., 2011).

Comparative Genomics: Arabidopsis 1001 Genomes, Other Plants, and Crops

Various research studies examine within-species comparisons of *Arabidopsis* natural variation and population genetics, while others focus on cross-species comparative genomics. However, the ability to retrieve a specific region of the genome across a large set of *Arabidopsis* ecotypes is not a simple task to execute, which hinders these investigations. Capturing genetic diversity is a major challenge, and integrating and managing the 1001 Genomes data (Weigel and Mott, 2009) will be essential to any progress in this area. Magnus Nordborg and colleagues from the European Union transPLANT project (http://www.transplantdb.eu/) are currently evaluating mechanisms to support these activities, and the AIP will need to build upon these outcomes. In terms of comparisons with other species, Phytozome (Goodstein et al., 2012), Gramene (Ware et al., 2002; Youens-Clark et al., 2011), PLAZA (Van Bel et al., 2012), or CoGe (http://genomevolution.org/CoGe/) could be appropriate module providers. TAIR also has computed orthologs across several species, which could be used to populate an early iteration of a comparative genomics AIP module.

Metabolomics

Wolfram Weckworth and colleagues are working to develop an aggregator tool, MASCMGator, for collating metabolomics data from several international sources. The recently funded Department of Energy Systems Biology Knowledgebase (http://genomicscience.energy.gov/compbio/) will be developing infrastructure to support metabolic models, and the AIP will need to take account of this and other similar projects. To provide useful information to the community, the module should also provide information on pathways such as those available in Arabidopsis Reactome (web services available) (Tsesmetzis et al., 2008), AraCyc (Zhang et al., 2005), KEGG (Kanehisa, 2002), Model Seed (http://www.theseed.org), and others, which are expected to be included in the MASCMGator.

Epigenomics

The Epigenomics of Plants International Consortium (https://www.plant-epigenome.org/) is aiming to have a browser for all available epigenomic data sets (including small RNAs) established within a relatively short period of time (The EPIC Planning Committee, 2012).

Phenomics: Via Literature and Linked Data

Powerful search engines are needed that integrate information across species, including genomes, transcriptomes, proteomes, localizomes, and phenomes. PosMed-plus (Makita et al., 2009) ranks candidate functional genes by connecting phenotypic keywords to the genes through linked data of biological interactions.

Other modules, such as a Phenomics module to provide information on all available

phenotypes associated with a particular germplasm or variation, would take longer to develop or will be created as new methods of measuring entities become available. The modules listed above are provided as examples to give an indication of what is possible rather than to provide an exhaustive list. In addition to modules that will be extensively utilized across the community, many investigators will continue to develop project-specific data sets and analysis tools that are likely to be of greater significance to a set of their close collaborators in their field of expertise. Such highly specialized databases are unlikely to become modules but would be encouraged to make their data available in a format that could be accessible through the AIP. The development of modules and data made accessible via the AIP will be a dynamic process that is continually evolving to meet the needs of the community.

FROM CONCEPT TO IMPLEMENTATION

What Will It Take to Build the AIP? Lessons Learned from the Design Workshop

The first and most critical tangible component of the IAIC will be the AIP. The portal's success will rely on significant and cohesive community effort and strong backing by funding agencies. The December 2011 Design Workshop challenged participants to creatively think of several possible ways to build the AIP through iterative rounds of small discussion groups that rotated members in order to maximize diverse interactions and outputs. Each small group presented its findings to workshop participants for comment and analysis. On the workshop's final day, participants self-organized into four groups charged with the task of imagining and articulating what they envisioned as the best portal design. The groups considered key technical and user-based elements, community involvement, challenges and possible solutions, feasibility of design, transformative aspects, costs, etc. Examples of key elements that were identified include the following: reuse (where feasible, seek to reuse existing code, previous investment, or existing infrastructure); openness (open data, open access, and open source principles); compatibility/interoperability (ensuring that software or architecture components can "talk" to one another across scientific domains, user communities, or geographic boundaries); unified web-based interface (effective links to federated modules); and support of community annotation.

Similarly, key challenges to the AIP identified include the following: achieving community buy-in, community agreement on data standards and their adoption, the expense of software development, achieving discrete fundable projects, redirecting to the community the development of applications ("apps"), leveraging existing cyber infrastructure, and managing data. Clearly, there are significant challenges to undertaking an endeavor such as this. It will require the synergistic activities of biologists, bioinformaticians, and computer scientists to achieve such an ambitious goal. Success will depend heavily on strong funding support from traditionally science-supportive governments. To meet this challenge head on, it will also be important to consider new collaborations, such as with commercial endeavors that share interests with the plant biology community, as well as to build upon and link activities across plant biology.

Noncentralized data repositories, including many of the noncore modules mentioned here, are by nature more specialized in focus, and for this reason have often failed to capture the type of broad user community that has characterized TAIR, for example. This does not represent in any way a failure or shortcoming of these resources, as they have often been developed with the goal to serve a particular community need. Yet, by having these resources integrated through the AIP, their visibility and utility to the community have the potential to be significantly increased. TAIR has itself been tremendously successful in consolidating and delivering information and providing access to *Arabidopsis* resources, and for this reason, it has been greatly appreciated and highly utilized by the community. However, a globally integrated information portal would include TAIR's resources and many others and would be significantly more useful than the sum of the parts of currently available resources.

Since the Design Workshop, IAIC Steering Committee members and workshop participants have begun discussions with teams interested in developing a portal prototype. The process involves facilitating team formation and inviting broader participation from teams not represented at the workshop with key goals to (1) have team(s) in place to seek AIP development funding and (2) start dialogs on core standards and module examples, both by the International Conference on Arabidopsis Research in July 2012. Teams may find the draft AIP requirements document written by workshop participants to be helpful in guiding AIP development (http://www.*Arabidopsis*informatics.org/resources/). The recently appointed Scientific Advisory Board (SAB; described below) will provide crucial community guidance to teams that advance with portal development.

IAIC Community Governance

The SAB consists of scientists from countries actively involved in the IAIC. The SAB will oversee the development of the IAIC and its activities and interact with funding agencies, the IAIC Steering Committee, the Multinational Arabidopsis Steering Committee, and others in the research community. SAB members collectively represent a breadth and depth of backgrounds, including experimental researchers, those with appropriate expertise in technical implementation, and those with long-term experience with *Arabidopsis* community needs and a demonstrated commitment to furthering community progress. Importantly, the SAB will help articulate the broad and transformative community vision and technical requirements of the resource and can gauge completeness in the context of community needs. Key roles for the SAB include to direct and shape future activities of the IAIC, encourage compliance with the standards set out by the AIP, liaise with funding agencies in the various countries involved in the IAIC, act as a point of contact for principal investigators or groups wishing to contribute to the IAIC, and liaise with the community to ensure that the IAIC continues to anticipate and serve the needs of the community.

COMMENTARY

The inaugural SAB was appointed in February 2012 following a 3-month process involving solicitation of community nominations and recommendations from Multinational Arabidopsis Steering Committee and other community members. Appointed members are Gloria Coruzzi (New York University, United States), Kazuki Saito (RIKEN, Japan), Magnus Nordborg (Gregor Mendel Institute of Molecular Plant Biology, Austria), Mark Estelle (University of California San Diego, United States), Mark Forster (Syngenta, United Kingdom), Paul Kersey (European Bioinformatics Institute, United Kingdom), and Xuemei Chen (University of California Riverside, United States). The SAB will serve for staggered, 3-year terms, with replacement members appointed from the community as needed.

Transition and Timeline

To maintain access to essential resources during the transition to the new AIP, the existing TAIR website will be moved from servers at the Carnegie Institution for Science to the iPlant Collaborative project. This will allow TAIR to remain accessible during the AIP development phase and provide continuous availability of *Arabidopsis* data to the plant research community until the new portal is operational. The iPlant project is well positioned both to effectively host essential *Arabidopsis* data and associated resources in the short term and to act as a potential long-term solution for the AIP (Goff et al., 2011).

Initial planning meetings of the transition committee held in March 2012, including TAIR, ABRC, and iPlant staff, have determined that the most efficient approach will be to move all components of the TAIR site to virtual machine servers at iPlant. It is estimated that the new site will be operational approximately 6 to 9 months after the start of work, which is proposed to begin in June 2012 and to be completed by February 2013. This timetable allows several additional months of buffer time before the end of the TAIR project in August 2013 in case of unforeseen difficulties. The proposed work will result in a fully functional copy of TAIR hosted by iPlant. As soon as the new site is fully functional, all TAIR traffic will be directed to it. For the 6 months of the TAIR funding period following the transfer, the remaining TAIR budget will be used to carry out software maintenance and data updates.

The end of TAIR funding in August 2013 will mark the start of a transition period during which all data currently stored in TAIR will continue to be available through iPlant (Figure 3). This can provide a satisfactory solution for a period of several months; however, if a longer transition period is needed, some level of funding to continue updates of gene structure and function data within TAIR will be necessary to avoid the problem of increasingly stale data. Although the transition plan does not currently include any data updates after the end of the current TAIR funding, the software pipelines for importing new gene function data will be in place.

CONCLUSIONS

The overarching goal of the effort that we have described is to develop a portal and facilitate access to data that will allow the exploration of gene/protein/metabolite networks in *Arabidopsis*. The current challenge for the *Arabidopsis* community is to fully integrate available data while accommodating the increasing complexity of new data sets. This challenge is not unique to *Arabidopsis*; it is common to most other organisms. The *Arabidopsis* community, through the IAIC, is leading the way to develop what we hope will become a new flexible, distributive, sustainable, and robust mechanism to integrate data and information resources. At the core of the IAIC is the AIP. The end result of the planning and development of the AIP should have a significant impact on both science and scientists to enable important scientific goals, such as predicting the outcome of the perturbation of signaling networks, enabling systems-level modeling of plant processes, strengthening of the interactions and communications among scientists, and increasing the translation of work performed in *Arabidopsis* to crop plants. The challenges of global food and energy security require a broad set of strategies; the efforts of the AIP will support these via the improved understanding of gene function and genotype–phenotype interactions. The standardization of data representations, shared analysis tools, and coordination across plant science, led, in part, by the community-supported AIP, will make this effort both visible and attractive to scientists and the funding agencies that support their work.

The AIP will provide a mechanism by which disparate databases, and the data that they house, can be compared and connected, linking global scientific resources, scientists, and analysis tools into a cohesive and user-friendly portal for *Arabidopsis*. Data sets will include those generated in increasing size and scope by individual laboratories on a global scale, large-scale data from major initiatives, comparative information from other species, and other biological data sets as they become available. The AIP will enable optimized use of data, tools, and resources to maximize the return on public research investment for the wider scientific community. The modules that we envision will be an important part of the AIP and will extend its functionalities and usefulness by allowing small and large projects to link to the AIP and share their tools and resources.

In the future, we envision incorporating connections between the AIP and other plant systems, for example, with the expanding resources for *Brassica* species (http://www.brassica.info/). As proposed, a federated, international model would facilitate research in other plant systems by establishing a framework for global integration and the inclusion of data and resources. While a generic system derived from the AIP would require specialization for different plant species, this customization could easily fit within the modular system that we have described for the AIP. In addition, the global leveraging of scientific data will maximize funding and resources by sharing these efforts across borders, which benefits the science while addressing the directives of the many funding agencies that support our work.

COMMENTARY

ACKNOWLEDGMENTS

The IAIC and its activities, including the AIP Design Workshop, are supported by the National Science Foundation (MCB Award 1062348, made to the U.S. members of the IAIC Steering Committee). Additional support for participation in the Design Workshop was provided by the Deutsche Forschungsgemeinschaft and the Biotechnology and Biological Sciences Research Council.

CONTRIBUTORS, PRIMARILY INCLUDING IAIC PARTICIPANTS AT THE DESIGN WORKSHOP

Katja Baerenfaller, Eidgenössisch Technische Hochschule Zürich, Switzerland; Ruth Bastow, Genomic Arabidopsis Resource Network, United Kingdom; James Beynon, University of Warwick, United Kingdom; Siobhan Brady, University of California, Davis, United States; Volker Brendel, Iowa State University, United States; Sylva Donaldson, University of Toronto, Canada; Rion Dooley, University of Texas-Austin, United States; Mark Forster, Syngenta, United Kingdom; Joanna Friesner, North American Arabidopsis Steering Committee, United States; David Gifford, RIKEN, Japan; Erich Grotewold, Ohio State University/ABRC, United States; Rodrigo Gutierrez, Pontificia Universidad Católica de Chile; Eva Huala, Carnegie Institution for Science, United States; Pankaj Jaiswal, Oregon State University, United States; Hiren Joshi, Lawrence Berkeley National Laboratory, Joint BioEnergy Institute, United States; Paul Kersey, European Bioinformatics Institute, United Kingdom; Lei Liu, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China; Ann Loraine, University of North Carolina at Charlotte, United States; Eric Lyons, University of Arizona, United States; Sean May, University of Nottingham and Nottingham Arabidopsis Stock Centre, United Kingdom; Klaus Mayer, Munich Information Centre for Protein Sequences/Institute of Bioinformatics and Systems Biology, Germany; Dan MacLean, Sainsbury Laboratory, United Kingdom; Blake Meyers, University of Delaware, United States; Lukas Mueller, Cornell University, United States; Robert Muller, TAIR, United States; Hans-Michael Muller, California Institute of Technology, United States; Francis Ouellette, Ontario Institute for Cancer Research, Canada; J. Chris Pires, University of Missouri-Columbia, United States; Nicholas Provart, University of Toronto, Canada; Dorothee Staiger, University of Bielefeld, Germany; Dan Stanzione, University of Texas-Austin, United States; James Taylor, Emory University, United States; Crispin Taylor, American Society of Plant Biologists, United States; Chris Town, J. Craig Venter Institute, United States; Tetsuro Toyoda, RIKEN, Japan; Matt Vaughn, University of Texas-Austin, United States; Sean Walsh, Eidgenössisch Technische Hochschule Zürich, Switzerland; Doreen Ware, U.S. Department of Agriculture Agricultural Research Service and Cold Spring Harbor Laboratory, United States; Wolfram Weckwerth, University of Vienna, Austria.

AUTHOR CONTRIBUTIONS

Primary contributors, including the IAIC steering committee and significant contributors to the publication, are listed in alphabetical order: Ruth Bastow, Jim Beynon, Volker Brendel, Rion Dooley, Joanna Friesner, Erich Grotewold, Eva Huala, Ann Loraine, Blake Meyers, J. Chris Pires, Nicholas Provart, Dan Stanzione, Chris Town, and Doreen Ware. Secondary contributors include the set of all workshop attendees.

Received May 17, 2012; revised June 5, 2012; accepted June 13, 2012; published June 29, 2012.

REFERENCES

Ashburner, M., et al.; The Gene Ontology Consortium. (2000). Gene Ontology: Tool for the unification of biology. Nat. Genet. **25:** 25–29.

Cui, J., Li, P., Li, G., Xu, F., Zhao, C., Li, Y., Yang, Z., Wang, G., Yu, Q., Li, Y., and Shi, T. (2008). AtPID: Arabidopsis thaliana Protein Interactome Database—An integrative platform for plant systems biology. Nucleic Acids Res. **36:** D999–D1008.

Davidson, R.M., Hansey, C.N., Gowda, M., Childs, K.L., Lin, H., Vaillancourt, B., Sekhon, R.S., de Leon, N., Kaeppler, S.M., Jiang, J., and Buell, C.R. (2011). Utility of RNA sequencing for analysis of maize reproductive transcriptomes. Plant Genome **4:** 191–203.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). AgriGO: A GO analysis toolkit for the agricultural community. Nucleic Acids Res. **38:** W64–W70.

Dugas, D.V., Monaco, M.K., Olsen, A., Klein, R.R., Kumari, S., Ware, D., and Klein, P.E. (2011). Functional annotation of the transcriptome of *Sorghum bicolor* in response to osmotic stress and abscisic acid. BMC Genomics **12:** 514.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics **10:** 48.

Geisler-Lee, J., O'Toole, N., Ammar, R., Provart, N.J., Millar, A.H., and Geisler, M. (2007). A predicted interactome for Arabidopsis. Plant Physiol. **145:** 317–329.

Goff, S.A., et al. (2011). The iPlant collaborative: Cyberinfrastructure for plant biology. Front. Plant Sci. **2:** 34.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res. **40:** D1178–D1186.

Gulledge, A.A., Roberts, A.D., Vora, H., Patel, K., and Loraine, A.E. (2012). Mining *Arabidopsis thaliana* RNA-seq data with Integrated Genome Browser reveals stress-induced alternative splicing of the putative splicing regulator SR45a. Am. J. Bot. **99:** 219–231.

Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W., and Zimmermann, P. (2008). Genevestigator v3: A reference expression database for the meta-analysis of transcriptomes. Adv. Bioinforma. **2008:** 420747.

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

Jenkinson, A.M., et al. (2008). Integrating biological data—The distributed annotation system. BMC Bioinformatics **9 (suppl. 8):** S3.

Johnson, J.E., Stromvik, M.V., Silverstein, K.A., Crow, J.A., Shoop, E., and Retzel, E.F. (2003). TableView: Portable genomic data visualization. Bioinformatics **19:** 1292–1293.

Joshi, H.J., et al. (2011). MASCP Gator: an aggregation portal for the visualization of Arab-

COMMENTARY

idopsis proteomics data. Plant Physiol. **155:** 259–270.

Kanehisa, M. (2002). The KEGG database. Novartis Found. Symp. **247:** 91–101.

Katari, M.S., Nowicki, S.D., Aceituno, F.F., Nero, D., Kelfer, J., Thompson, L.P., Cabello, J.M., Davidson, R.S., Goldberg, A.P., Shasha, D.E., Coruzzi, G.M., and Gutiérrez, R.A. (2010). VirtualPlant: A software platform to support systems biology research. Plant Physiol. **152:** 500–515.

Kerrien, S., et al. (2012). The IntAct molecular interaction database in 2012. Nucleic Acids Res. **40:** D841–D846.

Li, H. (2011). Tabix: Fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics **27:** 718–719.

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

Li, P., et al. (2010). The developmental dynamics of the maize leaf transcriptome. Nat. Genet. **42:** 1060–1067.

Makita, Y., et al. (2009). PosMed-plus: An intelligent search engine that inferentially integrates cross-species information resources for molecular breeding of plants. Plant Cell Physiol. **50:** 1249–1259.

Nicol, J.W., Helt, G.A., Blanchard, S.G., Jr., Raja, A., and Loraine, A.E. (2009). The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics **25:** 2730–2731.

Obayashi, T., Hayashi, S., Saeki, M., Ohta, H., and Kinoshita, K. (2009). ATTED-II provides coexpressed gene networks for Arabidopsis. Nucleic Acids Res. **37:** D987–D991.

R Development Core Team. (2012). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics **27:** 431–432.

The EPIC Planning Committee. (2012). Reading the second code: Mapping epigenomes to understand plant growth, development, and adaptation to the environment. Plant Cell **24:** 2257–2261.

Toufighi, K., Brady, S.M., Austin, R., Ly, E., and Provart, N.J. (2005). The Botany Array Resource: e-Northerns, expression angling, and promoter analyses. Plant J **43:** 153–163.

Tsesmetzis, N., et al. (2008). *Arabidopsis* reactome: A foundation knowledgebase for plant systems biology. Plant Cell **20:** 1426–1436.

Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., and Vandepoele, K. (2012). Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiol. **158:** 590–600.

Ware, D., Jaiswal, P., Ni, J., Pan, X., Chang, K., Clark, K., Teytelman, L., Schmidt, S., Zhao, W., Cartinhour, S., McCouch, S., and Stein, L. (2002). Gramene: A resource for comparative grass genomics. Nucleic Acids Res. **30:** 103–105.

Weigel, D., and Mott, R. (2009). The 1001 genomes project for *Arabidopsis thaliana*. Genome Biol. **10:** 107.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V., and Provart, N.J. (2007). An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. PLoS ONE **2:** e718.

Yilmaz, A., Mejia-Guerra, M.K., Kurz, K., Liang, X., Welch, L., and Grotewold, E. (2011). AGRIS: The Arabidopsis Gene Regulatory Information Server, an update. Nucleic Acids Res. **39:** D1118–D1122.

Youens-Clark, K., et al. (2011). Gramene database in 2010: updates and extensions. Nucleic Acids Res. **39:** D1085–D1094.

Zhang, G., et al. (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res. **20:** 646–654.

Zhang, P., Foerster, H., Tissier, C.P., Mueller, L., Paley, S., Karp, P.D., and Rhee, S.Y. (2005). MetaCyc and AraCyc. Metabolic pathway databases for plant research. Plant Physiol. **138:** 27–37.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., and Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol. **136:** 2621–2632.