genomeweb | BioInform
The Integrated Informatics News Source

# Arabidopsis Data Gets New Home at JCVI's AIP as TAIR Turns to Subscriptions to Stay Afloat

September 20, 2013

By Uduak Grace Thomas

The J Craig Venter Institute has received a National Science Foundation grant to create a new free genetic data resource for the *Arabidopsis thaliana* plant that will house data from The Arabidopsis Information Resource, or TAIR, which officially lost its funding last month.

JCVI has received $2.4 million of a $4.5 million award from the NSF that will fund the first two years of a planned five-year project to develop the Arabidopsis Information Portal, or AIP. Besides hosting TAIR's current datasets, AIP will contain information about the model plant such as gene expression and protein interaction data, and genetic variation data from the 1001 Genomes Project — an online catalog of polymorphisms in 1,001 *Arabidopsis* strains.

Responding to the announcement, Eva Huala, TAIR's principal investigator, said she was "very happy" to hear that the JCVI had been funded to build a new *Arabidopsis* resource. "I look forward to building a productive partnership between TAIR and the new project over the next few years, both as a subcontractor on the new JCVI award and as director of a collaborating resource," she told *BioInform* in an email.

TAIR will continue operating in parallel under a newly adopted subscriptions model, offering well-curated data to industry and academia for a fee. In contrast, AIP will do some curation but it will not be as extensive as TAIR's. In an open letter posted on the website last month, TAIR staff said that starting this October, it will charge companies a yet-to-be-disclosed fee to access its content. Academic subscriptions, however, won't go into effect until the spring or summer of 2014 and will include a free limited data access option for academics that cannot afford to pay the subscription fee.

Having two resources to choose from isn't a bad thing, according to Huala. "Arabidopsis researchers will benefit from having two major informatics resources which together will be able to provide a broader array of data sets and tools than either project could do alone," she said.

Initially resistant to the idea of a subscription model, TAIR's organizers tried other avenues to raise the funds needed to keep the resource afloat following the NSF's 2009 decision to phase out its funding over the next few years — a move that drew the ire of many in the *Arabidopsis* community (***BI*** 12/4/2009).

TAIR's fundraising efforts included a corporate sponsorship program (***BI*** 3/19/2010) that received support from companies such as Dow AgroSciences and Syngenta Biotechnology (***BI*** 8/6/2010). But those efforts don't seem to have borne much fruit and with no funding forthcoming from the NSF, there is "no other sustainable way to continue providing high-quality curated data. For that reason, and with strong encouragement from NSF, we have decided to move to subscription-based support," TAIR staff wrote in the letter.

In response to a question from *BioInform* about the decision to fund AIP's activities but not TAIR's, Peter McCartney, an NSF program director, said in an email that "as with all award decisions we make, a proposal was presented to us, we got peer input via the merit review process, and then made a decision to award based on our best assessment of our portfolio and what we think would advance the community." He added that "the last award to TAIR and the recent award for AIP occurred three years apart, and resulted from independent cycles of this review process."

The AIP grew out of a series of discussions and workshops that were organized by the International Arabidopsis Informatics Consortium, or IAIC, a community of researchers that mobilized following the NSF's 2009 announcement to discuss new ways of storing and accessing *Arabidopsis* data. They decided to develop a replacement system that would enable broad data access to the data if TAIR was discontinued but would move away from its single source approach to data storage and management.

Following a series of community workshops in 2010 and 2011, the IAIC decided on a federated approach where genomic, transcriptomic, proteomic, and other kinds of information stored in disparate databases globally would be accessible from a single interface through a series of modules. This would move the community away "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," they wrote in a 2012 paperpublished in *The Plant Cell* that provides details about AIP's proposed structure.

It would also bring together "ever-increasing amounts of *Arabidopsis* data into a single, user-friendly location using the latest web technologies and services," according to JCVI's grant abstract. Under this model, individual data providers are responsible for generating and maintaining their own databases, and the burden of supporting these resources is distributed across a wider range of funding agencies and countries which should help JCVI and its collaborators save on operating costs.

Planned portal features include a "modular, community-extensible web-based interface that will include user work spaces that can be configured with data retrieval, analysis, and visualization applications," the abstract states. It will also include "an implementation of an *Arabidopsis*-specific instance of InterMine, a data integration platform that is widely accepted in the animal model organism database community; and a web services layer that facilitates data access and integration" with resources from groups like the iPlant collaborative as well as other data providers.

"We envisage this portal as having a three-tier architecture," Christopher Town, head of the plant genomics group at JCVI and principal investigator on the NSF grant, told *BioInform*. He said that it will have a user interface, a database infrastructure, and a web services layer that will pull information from JCVI-hosted data warehouses and external databases through third party applications.

Town has been involved in *Arabidopsis* research projects for about 27 years and helped lead JCVI's efforts to sequence the model plant's genome. Over the next two years, his team of five along with collaborators at the Texas Advanced Computing Center, TACC, and the University of Cambridge will transfer data from TAIR's database into AIP, and develop and test the first data modules.

They've sub-contracted TAIR's staff to migrate data from their existing repository into a new Chado database that will underlie AIP — Chado is an open source relational database schema that was built to handle complex biological knowledge representations. TAIR's data is currently sitting on virtual machines on

the iPlant consortium's servers. Staff members moved the data there earlier this summer so that the community could still be able to access it even after its funds ran out and while the AIP is being developed.

It's not clear at the moment whether the AIP will take on all of TAIR's data or not, Town said. For example, it has some old expressed sequence tag data that may no longer be needed thanks to newer technologies like RNA-sequencing, he said. There's also some uncertainty about how long the *Arabidopsis* Biological Resource Center, which offers DNA and seed stocks, will continue to use TAIR data for its stock ordering and when it might transition to AIP. "If it turns out that [ABRC] is going to run off AIP then we will want to take in any data types that are relevant for the stock center to continue functioning," Town said.

Other planned activities for the first two years will be to test the workability of the modular approach with groups like the developers of the Bio-Analytic Resource at the University of Toronto. Town said that the UToronto team is developing modules for the AIP that will give users access to gene expression and protein interaction data contained in the bioarray database. The team is also giving JCVI the complete datasets to store locally.

"Although the [IAIC] envisaged that we would do all this via web services, it turns out that not many *Arabidopsis* data resources at this time host functionally adequate web services," he explained.

This means that the first phase of the project will be a hybrid approach "where we do some data warehousing while at the same time we develop the ability both to accept and put out data via web services," Town said. "Having access to the data by both mechanisms also makes it possible to "do a side-by-side prototyping of the relative merits of warehousing this data en masse and serving it out that way versus trying to serve it on the fly via web services," he said.

JCVI is working with researchers at the University of Cambridge to create a version of the InterMine database that will serve as an aggregator for information in the AIP's data warehouse and will provide tools for running complex queries to find links between genes, gene expression, and protein interactions, for example. InterMine was developed by researchers at Cambridge and is used to integrate and query data for a number of model organisms including flies, mice, and yeast.

For its part, TACC will provide the hardware needed to power the portal and will be responsible for developing its graphical user interface. TACC has experience with providing infrastructure for plant genomes because it developed code and tools that are used to support the iPlant consortium's activities.

Town's team expects to have a functioning prototype of the portal by the end of the year. Ultimately, he said, when it's time to apply for renewed funding in two years, the researchers hope to have a system that demonstrates to the NSF and the *Arabidopsis* community that AIP is "actually something that really is going to be worthwhile to grow and develop," he said.