RCN: AN INTERNATIONAL ARABIDOPSIS INFORMATICS CONSORTIUM

PI:	Blake C. Meyers, University of Delaware
Co-PIs:	Erich Grotewold, The Ohio State University
	Doreen Ware, Cold Spring Harbor Laboratory
	Jim Carrington, Oregon State University
	Volker Brendel, Iowa State University
Other Senior Personnel:	Nicholas Provart, University of Toronto
	Ruth Bastow, GARNet, UK
	Jim Beynon, University of Warwick, UK
Proposed IAIC Coordinator:	Joanna Friesner, Univ. of California – Davis

Intellectual merit and objectives of the proposed activity

The aim of this proposal is to set up a Research Coordination Network (RCN) that will lead to the development of informatics tools and resources for plant biology, and in particular, for the reference plant Arabidopsis. This RCN will provide the framework to catalyze a radical change in the way in which US and international researchers coordinate data generation, analysis and visualization. It will be an open network that aims to draw in a broad cross-section of the international Arabidopsis community, particularly those working in the area of informatics. As we will explain, the Arabidopsis community is at a critical junction; a broad range of new data types are becoming available, and the community's ability to address important questions depends on the integration, visualization and analysis of these data. Diverse, international research groups are inventing new informatics approaches and thus the community needs a mechanism by which to identify, explore, promote, and enhance these resources. Meanwhile, consolidated resources developed by single groups are less able to address the ever-broadening range of community needs. This RCN will bring together an international community of Arabidopsis biologists and informaticians to develop a new, unique, multinational informatics resource that leverages expertise and funding on a global scale.

This proposal has the following specific aims:

1. Establishment of an International Arabidopsis Informatics Consortium (IAIC) that can build upon activities and resources that will be initiated under this NSF-sponsored RCN.
2. The RCN activities will facilitate development of Arabidopsis informatics tools and resources both in the US and among international participants.
3. The activities of the IAIC will lead to the identification and appointment of individuals within the IAIC to develop standards for Arabidopsis data, deposition and display.
4. Development of a 'white-paper' outlining plant informatics challenges and opportunities that can serve as a guideline for establishing future funding priorities. Included will be primary needs and challenges for informatics modules as well as technical and bioinformatics challenges.
5. Establishment of an IAIC website as a resource for interested scientists and the general public.
6. Formation of a Scientific Advisory Board (SAB) to oversee activities developed under the IAIC and to coordinate with the AIP.
7. Establishment of priorities, needs and specifications for additional IAIC modules.

Broader impacts of the proposed research

Through the coordination of Arabidopsis informatics efforts, the funding of this RCN proposal will advance plant biology, create novel opportunities for research and education, and lead to the establishment and growth of an International Arabidopsis Informatics Consortium (IAIC). It is likely that the problems that the Arabidopsis community is facing today will also be the challenges of other organismal communities tomorrow, and this type of coordinated international approach to leveraging informatics resources across borders may also help those organisms. Thus, experiences learned from this application will have a broad impact across biology, particularly for crop plants for which a similar expansion in data is beginning to occur and which typically build on efforts from Arabidopsis. The resources will be developed as a result of this consortium will be widely and freely available and are likely to be useful for students and researchers at educational institutions of all sizes, companies, and to the general public.

PROJECT DESCRIPTION

The aim of this proposal is to set up a Research Coordination Network (RCN) that will lead to the development of informatics tools and resources for plant biology, and in particular, for the reference plant Arabidopsis. This RCN will provide the framework to catalyze a radical change in the way in which US and international researchers coordinate data generation, analysis and visualization. As we will explain, the Arabidopsis community is at a critical junction; a broad range of new data types are becoming available, and the community's ability to address important questions depends on the integration, visualization and analysis of these data. Diverse, international research groups are inventing new informatics approaches and thus the community needs a mechanism by which to identify, explore, promote, and enhance these resources. Meanwhile, consolidated resources developed by single groups are less able to address the ever-broadening range of community needs. Through the coordination of Arabidopsis informatics efforts, the funding of this RCN proposal will advance plant biology, create novel opportunities for research and education, and lead to the establishment and growth of an International Arabidopsis Informatics Consortium (IAIC). It is likely that the problems that the Arabidopsis community is facing today will also be the challenges of other organismal communities tomorrow. Thus, experiences learned from this application will have a broad impact across biology.

Glossary of terms used in this proposal.

The RCN-funded network – an "all inclusive" group that encompasses the entire Arabidopsis community. The RCN PI is Blake Meyers, whose role is to coordinate the RCN and facilitate the formation of the consortium over the five years of requested funding. Beyond the senior personnel already associated with this proposal, the initial set of participants will include the 50+ workshop attendees (see Table 1) who recommended the organization that we seek to develop. Because RCN participation will be open to the community and thus a non-exclusive group, and because senior personnel on a proposal are limited, we did not attach biosketches for all 50+ individuals.

IAIC – International Arabidopsis Informatics Consortium. These are the contributors to core and non-core resources (see below). The IAIC director is yet to be named, to be elected by the IAIC based on recommendations from the SAB. The interim IAIC director is PI Meyers. The IAIC steering committee initially consists of the PI, co-PIs and senior personnel named on this RCN proposal, but is open to additions; committee members should be investigators (from any nation) contributing Arabidopsis modules.

AIP – Arabidopsis Informatics Portal. The primary interface providing dynamic access to core resources and key non-core resources, and links to "boutique databases". This is yet to be formed or funded, but will be a critical resource to be funded competitively in one or more of the nations with a significant Arabidopsis scientific presence. This will act as the central hub for coordination of Arabidopsis informatics, and will define standards for data storage and access and interconnectivity.

SAB – Scientific Advisory Board. Yet to be named; a group of seven international scientists that will actively oversee the IAIC and its activities. This will be named by the IAIC by year two of this RCN proposal, supported in years 2 & 3 by the RCN, and from then on, will be supported by the AIP.

SAP – Scientific Advisory Panel. Yet to be named; this panel will provide occasional, detailed, independent reviews of the progress made by the IAIC and AIP. This will require meeting only every 2 or 3 years, and it will be supported by the AIP, not by this RCN.

Core component – Module essential for the functioning of the IAIC including the central portal, the "gold standard" genome, literature curation, and stock center resources; currently provided in part by TAIR.

Non-core component – Module identified by the IAIC as an essential component once the core modules are in place; the definition of core/non-core is dynamic and will depend on the community needs .

Boutique databases – Resources developed for lab-specific projects that may be of significance to a small group of researchers, but still should be accessible from the AIP, perhaps as links if the data are not formatted according to AIP standards.

1. Introduction

The Multinational Arabidopsis Steering Committee (MASC) and the North American Arabidopsis Steering Committee (NAASC) hosted workshops in Nottingham, UK (April 15-16, 2010) and Washington DC, USA (May 10-11, 2010) to consider the future bioinformatics needs of the Arabidopsis community as well as other science communities that depend vitally on Arabidopsis resources. The outcomes of both workshops were presented and discussed at the International Conference on Arabidopsis Research (ICAR) in Yokohama, Japan (June 6 – 10, 2010). There were 36 and 37 participants at the first and second workshops respectively, including observers from several agencies (Table 1), with expertise in plant biology as well as database use, curation and development; workshop report describing these activities is in press in *The Plant Cell* [1]. The focus of the workshops was on Arabidopsis because of its unique and essential role as the reference organism for all seed plant species. The development of the highly annotated "gold standard" Arabidopsis genome sequence has been an invaluable resource for plant and crop sciences. This platform provides important information and working practices for other species, and for comparative genomic and evolutionary studies. Arabidopsis tools and resources for information storage, curation and retrieval have been developed over recent years primarily through the activities of TAIR (The Arabidopsis Information Resource), NASC (the Nottingham Arabidopsis Stock Centre) and ABRC (the Arabidopsis Biological Resource Center), among others. However, the Arabidopsis community and funding agencies recognize the need for a single data management infrastructure. The key challenge is to develop and fund this resource in a sustainable and transparent manner.

Table 1. Attendees at the two workshops who are potential IAIC members.

UK workshop attendees	US workshop attendees
Philip Benfey	Ewan Birney
Jim Beynon	Volker Brendel
Ewan Birney	Jim Carrington
Pascal Braun	Mike Cherry
Robin Buell	Joseph Ecker
Mario Caccamo	Janan Eppig
Mark Forster	Mark Estelle
Erich Grotewold	Erich Grotewold
Rodrigo Gutiérrez	Eva Huala
Pierre Hilson	Keith Lindsey
Eva Huala	Hong Ma
Manpreet Katari	Kathy Matthews
Paul Kersey	Sean May
Jörg Kudla	Klaus Mayer
Keith Lindsey	Blake Meyers
Sean May	Eric Mjolsness
Blake Meyers	Nicholas Provart
Harvey Millar	Paul Schofield
Basil J. Nikolau	Heiko Schoof
Magnus Nordborg	Julian Schroeder
Nicholas Provart	Taner Sen
Chris Rawlings	Dan Stanzione
Dan Stanzione	Todd Vision
Chris Town	Doreen Ware
Testuro Toyoda	
Sean Walsh	
Xiujie Wang	
Wolfram Weckwerth	
Weicai Yang	
Plus seven observers from funding agencies	Plus 13 observers from external agencies

Both meetings included the GARNet, MASC & NAASC coordinators, Ruth Bastow and Irene Lavagi; Joanna Friesner (NAASC coordinator) attended the US meeting.

Global challenges surrounding food and energy security require intelligent plant breeding strategies that will be strongly supported by a central Arabidopsis information resource to aid our understanding of gene function and associated phenotype in many different environments. The knowledge accrued in Arabidopsis informs our understanding of the genetic basis of plant processes and crop traits. To date, this has accumulated primarily through analysis of single genes. However, gene products do not act alone but rather in complex interacting networks. Thus, the challenge for the Arabidopsis community is to understand this higher level of complexity, to a significant extent through the application of new high throughput, quantitative experimental techniques. The goals of these efforts are to develop gene/protein/metabolite networks that will enable systems-level modeling of plant processes, and ultimately, to translate these findings to crop plants. To achieve these goals, we must develop novel approaches to data management, integration and access.

The UK workshop addressed three principal issues: the types of data generated by the Arabidopsis community; the types of data used by the community; and future needs of the community. The objective was to produce recommendations for 1) the type of infrastructure necessary to address the challenges and opportunities associated with the application of new

technologies, and 2) for a sustainable funding model to support this infrastructure. These recommendations were considered and expanded upon at the US workshop with the ultimate goal of generating solutions to the issues discussed in the first meeting. Specifically, the technological, financial, and organizational sustainability of a community-oriented international Arabidopsis informatics consortium were discussed. It was recognized that cohesive, cooperative, and long-term international collaboration will be critical to successfully maintain an Arabidopsis database infrastructure that is essential for plant biology research worldwide. The primary goal of this RCN is to initiate and build this consortium such that it becomes a self-sustaining, international organization for plant bioinformatics.

The workshop participants concluded that there is a continued need for a central Arabidopsis information resource based on the productivity of the Arabidopsis community and the critical importance of the findings generated by this community. For example, about 3,000 Arabidopsis publications are currently published in peer-reviewed journals each year, a nearly 10-fold increase since the early 1990s; and in 2009, TAIR was accessed by 335,692 unique visitors and had nearly 20 million page views. Furthermore, the importance of a current, well organized and carefully curated Arabidopsis genome to researchers studying other plants, including crops, cannot be overstated. In the future this resource should be part of a larger infrastructure that would be dynamic and responsive to new directions in plant biology research. This was envisioned to consist of a distributed system of data, tools and resources, accessed via a single point (information portal) and funded by a variety of sources, under shared international management and a scientific advisory board.

Table 2. Types of data deposited by Arabidopsis researchers.

Published literature
Genomes
Metabolome, catalogue of metabolites
Proteome
Protein sequence and structure
Protein subcellular localization
Protein modifications
Interactome
cDNA sequence
Expression maps
Genetic variation and accession genomes
SNPs and indels
QTL
eQTL
Alternative splicing
Phenomics and phenotypic data
Epigenetic data
Exogenous small molecules

1.1 Data types and use

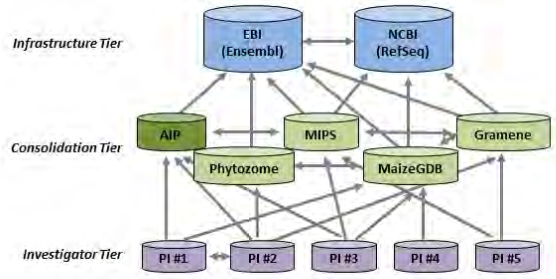
The kinds of data currently generated by Arabidopsis researchers are diverse and in a variety of formats (Table 2). They vary in volume and complexity, and although some of these data types are common among plant species, many have become available first in Arabidopsis, a pattern that is likely to be repeated for future technologies. Overall, the volume of data is dramatically increasing, particularly due to the exponential growth of next-generation sequencing (NGS) of genomes, chromatin, and RNA and, on a smaller scale, expanding proteome and metabolome datasets. The quantity of assembled data will require novel storage and display capabilities. In the future, we must deal with sophisticated new datasets including, but not limited to, high resolution microarray data, image data, cell-type specific or time-series expression profiles, protein localization data, protein activation and relocation data, protein-protein interaction data and promoter structure and transcription factor binding sites (both positional and temporal). All these datasets will be used to generate systems-level models that must also be stored in an accessible way. Because Arabidopsis has become the reference plant with unmatched tools and resources, it will be the plant system in which traditional and novel forms and quantities of data will first become available.

Integration of these different data types will therefore be a key issue - both vertical integration, in which all available Arabidopsis information is accessible, and horizontal integration, whereby it is possible to move easily between different species. This horizontal integration process would begin with genome/ortholog alignment with plant/crop genomes, and extend to other datasets as the depth and complexity of the data from other plant species become sufficiently rich. As annotation and curation are increasingly inferred from several types of data, users will demand clear audit trails that indicate the provenance of the data pertaining to

genes and their products. Until now, TAIR has played a key role in providing an authoritative stamp for community-approved annotation (for example, defining a working complete set of gene models, and providing literature curation). However, it has become clear that an explosion in the amount of data produced by the community has surpassed the ability of TAIR (and similar data warehouses) to accommodate and distribute these data, resulting in a need to develop new models for data integration. It is important that data are readily available in convenient formats and via tools that are accessible to a range of users. Development of software based on an open source model should be a fundamental principle, as this approach most efficiently leverages expertise and capacity across the field of genomics and systems biology, and has been shown by experience to produce the most trusted and adaptable software tools. Most of the problems faced by the Arabidopsis community are not unique; cooperative tool development with researchers working on other species will ensure that widely useful software is developed in a cost-effective manner.

The highly curated and characterized gene/protein/metabolite networks developed in Arabidopsis will prove invaluable in systems biology approaches that seek to construct and constrain a range of models in Arabidopsis and other plants. These models will provide a framework for interpretation of a variety of complex results. The high standard of curation and data annotation in Arabidopsis makes these resources particularly important to researchers in other communities seeking to gain valuable functional insights into their own data. Examples include crop scientists as well as those studying model organisms and other less well-studied plant species. These wider applications underpin efforts to understand the molecular basis of plant growth and development, and ultimately, crop yield.

The high volume of data now generated in biological research increases the importance of efficient and flexible tools for data analysis, inspection and visualization. At present, the community's ability to access and analyze data is limited by the highly heterogeneous and often complicated (sometimes out of necessity) nature of bioinformatics tools. Traditionally, genome browsers have provided a basic framework through which additional annotation can be visualized. However, new data types are pushing the limits of visualization. For example, data on genomic variation, such as that generated by the 1001 Arabidopsis Genomes project [2], will help to link genotype to plant phenotype; yet the resources and tools needed to access and analyze these data are still in the early stages of development.



	Investigator Tier	Consolidation Tier	Infrastructure Tier
Main funding	Institutional and response mode	Strategic and response mode	Strategic and infrastructure
Scope	Specific topics	Area of scientific organization	All species
Time scale	3 – 5 years	5 years, with expectation of renewal	> 10 years
Engineering	*Ad hoc*	Mixed	High
Leadership	PI-led, independent	PI-led, with transitions; SAB oversight	Institution-led, part of administrative framework , independent of PI

Figure 1. Three tiers of data resources.
Proposed scope for an information resource to house interpreted biological resources. The lowest and most fundamental tier consists of local databases of specialized data resources run mainly by individual investigators (PIs). The second tier provides a layer of consolidation into more durable and useable forms for a larger defined community; an Arabidopsis community portal belongs in this level and is indicated as "AIP" (Arabidopsis Informatics Portal). The third tier enables cross-species comparisons; this requires an integrated set of diverse resources.

1.2 A continued need for an Arabidopsis community portal

The value of an Arabidopsis Information Portal (AIP) should be measured primarily through its ability to facilitate and stimulate high quality science. There is strong justification for such a resource that provides a vital service to what is a large and vibrant scientific community. This community comprises not only those working directly on Arabidopsis, but also researchers working on other plants and animals. In particular, scientists working on all the major crop plants look to Arabidopsis data to inform their research. Arabidopsis is likely to continue to play a nodal

role due to its well-annotated genome and its wealth of genetic and genomic resources, which make it unique amongst plant species in being well suited to systems biology research.

Clearly there is a need to define a manageable scope for any information resource. One division is between archives and interpreted resources. Archives (for relatively unprocessed data) can often be very broad in scope and for many data types, a specific Arabidopsis repository may not be needed; instead, some very raw data can be stored at the data-generating institution. Another set of resources can then provide interpreted views of the archived data for specific purposes. One can think of such interpreted resources as existing in three tiers (Figure 1). The first tier consists of local databases that feature novel or highly specialized data resources run mainly by individual researchers focused on a narrow biological question [3]. In the second tier, data are consolidated into forms that are more readily useable by a larger community (the group of federated IAIC components belongs in this level). In effect, a community trusts a resource of this type with custodianship of its data. For example production of 'gold standard' annotated genome is a key part of the function currently performed by TAIR. As different data types are brought into the community portal that forms the IAIC, the challenge will be to set priorities as to what should be consolidated and how data can and should be integrated. Input from the community both directly and through scientific advisory boards will be critical in setting these priorities, scope, and standards for quality control. The third tier enables cross-species comparisons of datasets, by integrating the outputs of differently focused resources; currently this is mainly feasible for genomes and gene expression. Work should be directed to developing common data formats and tools for interrogation, to facilitate exchange between databases for different species, and to ensure that Arabidopsis information can be fully exploited by bioinformatics resources being developed to serve communities for which Arabidopsis is a key model organism (e.g., crop science).

1.3 An International Arabidopsis Informatics Consortium (IAIC)

The Arabidopsis community has a strong tradition of international cooperation (e.g., multinational sequencing initiative, multinational steering committee, international stock centers, and annual international meetings). The development of a new international Arabidopsis bioinformatics initiative is a logical next step to manage the increasing amounts and types of data, and it will allow the leveraging of resources, knowledge, and collaborations. In our view there is a strong justification and incentive to expand the current informatics structure into an international organization, the International Arabidopsis Informatics Consortium (IAIC). The consortium will need to be dynamic and represent the evolving needs and capacities of the community while reflecting the funding interests of the respective countries.

We propose that the IAIC be made up of a distributed system of data, tools and resources that would be funded by a variety of sources under an international management and scientific advisory board. Participants at the two workshops in the spring of 2010 emphasized the importance of a unified front-end interface. We therefore envisage that the core of the IAIC would be the Arabidopsis Information Portal (AIP) that would interact with and link to resources across the globe, including Arabidopsis datasets generated in individual laboratories, information from other species, and other biological datasets. We propose that all data are accessed via the AIP, and that the AIP will combine outputs into a single user-friendly interface. The AIP will enable optimized use of data, tools and resources to maximize the return on public research investment for the wider scientific community. Currently, and among its other responsibilities, TAIR maintains a website (http://www.arabidopsis.org) that serves some of the functions that the AIP is expected to play, but it is clear that the AIP will need to provide a much more dynamic access to resources developed under the IAIC, than what TAIR's mission has been. TAIR handles gold standard genome, literature curation, and the information for the ABRC stock center, but there are many things that it is not prepared to do. TAIR was not built to provide dynamic access to new data structures; TAIR couldn't anticipate the amount of data that

the community is developing. Expanding or re-inventing a large database like TAIR to dynamically adapt to the rapidly changing needs of the community involves significant challenges and was not perceived in the discussions at the various workshops as a practical or optimal solution, given the complexity of new data, the internationally-distributed nature of expertise in these data types, and the need to leverage funding on an international basis. These are the issues that the workshop participants envisioned the AIP will address.

To ensure that the IAIC is built on strong foundations, workshop participants proposed that the IAIC have a core initially consisting of four parts: 1) the AIP as outlined above; 2) the "gold-standard" genome annotation, i.e. a finished genome (no gaps), annotated with protein and non-protein coding genes and gene models that are revised by curation or targeting programming based on feedback and new data; 3) genome/sequence curation, that provides functional information on each gene, its product(s) and associated regulatory landscape in a genomic context; and 4) Stocks and Resources Database(s). The core of the IAIC could be one unit located in one physical location and managed by one PI/Group. Alternatively and perhaps more in line with the distributed and federated concept of the IAIC, the core could be distributed amongst a set of providers. Distributing the core allows for specialization; a single core component can focus on a particular area of expertise rather than expend effort (and funds) to develop capability in all the areas necessary to maintain the core as a whole. This latter approach generates a level of adaptability whereby each core component can evolve separately to take advantage of expertise as it emerges across the globe. The distribution of core functions also spreads the risk and responsibility between a numbers of locations.

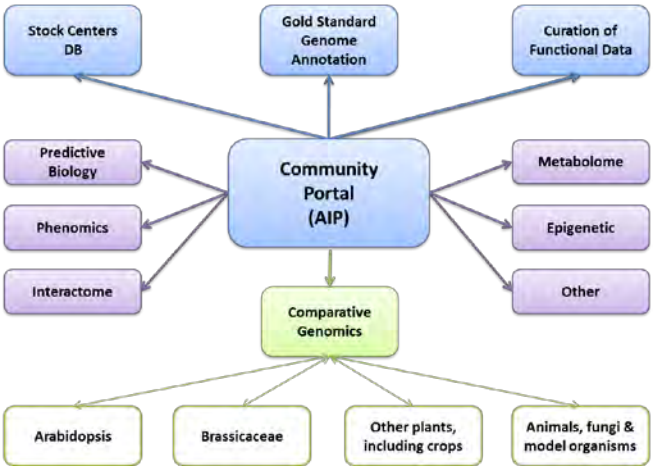


Figure 2. The structure of the International Arabidopsis Informatics Consortium (IAIC)
The initial IAIC will consist of four components, depicted in blue 1). The Arabidopsis Information Portal acts as the central hub of the IAIC, provides a single user interface to access to all the constituent parts of the consortium, set standards and provides training 2) "gold standard" genome annotation 3) Curation of functional data 4) Stock center database(s) to enable rapid access to resources. Future potential modules are illustrated in purple; those listed in the figure are just examples and are not meant to be an exhaustive list. The Comparative Genomics module (in green) provides one example of how the IAIC will link out to other plant species.

Using the four 'core' components as the basis for the IAIC, additional 'non-core' but important informatics modules can then easily be added to form the IAIC as illustrated in Figure 2. Indeed, such a model as proposed here with a clearly defined set of standards, allows for any data, resource or tools generated worldwide to become part of the IAIC. By utilizing a distributed approach, the user does not have to face a dispersed landscape of data; instead, a federated approach accessed through the AIP gives the user the impression of a seamless whole. Further, a distributed model allows the workload, human expertise, innovation and costs to be shared across many sites that are internationally located. The proposed model for the IAIC produces additional resilience and flexibility by providing opportunities to bring together creativity and energy from many places. As highlighted above, a federated approach also has the advantage of specialization with each module being able to focus on a particular area of expertise. Examples of such a distributed informatics model exist for other organisms, such as WormBase for *Caenorhabditis* and FlyBase for *Drosophila*, or for specific topics like PDB (The Protein Data Bank).

The proposed modular structure provides an ideal opportunity for the IAIC to link out and interact with other plant species and grow into an International Plant Informatics Consortium in

the future, if this was deemed useful by the wider crop and plant sciences communities. In fact, workshop participants noted that an essential function of the IAIC would be to ensure that the distributed set of resources that make up the IAIC could easily be leveraged to benefit those communities. We propose that the most effective way to achieve this is to develop a non-core module in comparative genomics that would allow integration of data from other species as it reaches sufficient depth and quality. The module could then grow at varying rates depending on the datasets available, ease of integration and interoperability. Workshop participants envisaged that such a module could consist of four layers: 1) Arabidopsis – natural variation and genome evolution; 2) other Brassicaceae – nearest relatives enabling wider genome associations, orthology, natural variation, evolution, crop traits; 3) crop genomes – evolution, orthology, crop traits; and 4) other species. Such a module would not only allow other plant and crop researchers to access Arabidopsis information but would also enable Arabidopsis researchers to link out to appropriate orthologs and associated data in other plant species. To ensure that there is interoperability between data and resources generated in other communities, it will be essential for the IAIC to establish strong links with other plant data providers, to allow exchange of information, best practices, and to help build a common framework.

1.4 Ensuring the sustainability of an International Arabidopsis Informatics Consortium

Management and Operations

To ensure the IAIC fulfills the objectives outlined above, one of the goals of the RCN-initiated IAIC will be the establishment of two of three planned oversight and advisory boards. This will include the IAIC Committee (initially named here as PIs and senior personnel, but with more members added as the IAIC grows) and an international Scientific Advisory Board (SAB); the Scientific Advisory Panel (SAP) that is described in more detail below will be developed later and under separate funding. Developing the IAIC Committee and the SAB will be a significant aspect of this RCN.

The role of the SAB will be to: 1) direct future activities of the IAIC, both core components and non-core modules; 2) help to encourage compliance with the standards set out by the AIP; 3) liaise with funding agencies in the respective countries involved in the IAIC; 4) act as a point of contact for PIs/Groups wishing to contribute to the IAIC; and 5) liaise with the community to ensure that the IAIC continues to anticipate and serve the needs of the community. The SAB will be formed with an initial group of seven international scientists, by a minimum of one scientist from each of the countries involved in supporting the IAIC. The SAB will be selected in consultation with the Multinational Arabidopsis Steering Committee (MASC) and the funding agencies supporting the IAIC. It will be essential for SAB members to have the appropriate expertise in technical implementation and community needs**.** Members of the funding agencies supporting the IAIC would be invited to be observers at SAB meetings. Workshop participants proposed that the SAB meet twice a year, once at ICAR and once in a virtual meeting.

The IAIC Committee will consist of the PIs leading the modules of the IAIC, although in the beginning, it will consist of the PI, co-PIs, senior personnel of this RCN proposal and everybody from Table 1 interested. To ensure that membership of the IAIC Committee does not cross over with membership of the SAB, SAB members should not lead modules of the IAIC. Workshop participants recommended that a Chairperson that is not involved in any part of the IAIC be appointed by the SAB to oversee the IAIC Committee; initially, PI Meyers will serve in this role until the IAIC grows to the point at which an appointment can be made. The Committee would report to and interact with the SAB. The Committee would meet at least twice a year, once at ICAR and once in a virtual meeting.

A Scientific Advisory Panel (SAP) will also be formed to review the progress of the IAIC and eventually the AIP. SAP members will be selected from the Arabidopsis and wider research communities, and consist of a set of advisors that are distinct from the SAB and IAIC committee. The SAP will ultimately meet once every 2 or 3 years to assess the IAIC. On these cycles, the SAP could assist with mid-term review and end-of-grant reviews. However, a major challenge would be in coordinating this with the different funding cycles of international funding agencies. The SAP will be formed after the SAB, and thus its activities will be funded by the AIP – and not via this RCN proposal

In addition to the three bodies outlined above, it would be very useful if a committee of the funding agencies involved with or interested in supporting the IAIC could also be formed with named representatives from each country, to help facilitate a clear dialogue between the funding agencies and the SAB. However, this will be up to the funding agencies.

The managerial structure of the IAIC is outlined in Figure 3, but it should be emphasized that the Arabidopsis and plant informatics communities will have direct input into IAIC activities via IAIC-sponsored workshops at the annual Arabidopsis meeting. In the context of this RCN proposal, the senior personnel on this proposal represent the initial IAIC steering committee; it is likely that this committee will add members as additional components of the AIP are developed and those investigators will be invited to join. The SAB will be formed out of the activities of this committee via interactions with the community, and the initial two years of their work will be funded by this RCN proposal, with their work beyond this proposal funded by the AIP. The SAP will be formed later in conjunction with the AIP, and we anticipate that the funding for that resource will include funds for the SAP activities.

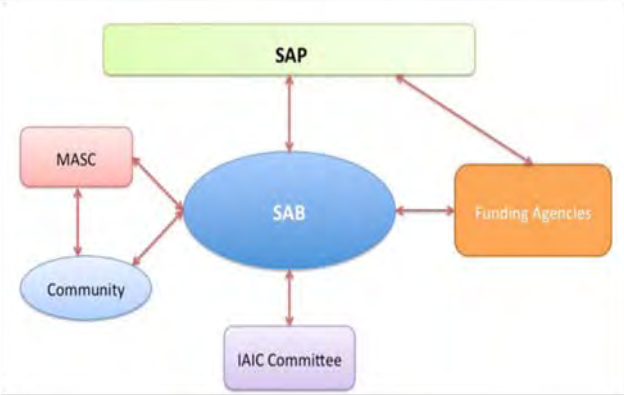


Figure 3. Management structure of the International Arabidopsis Informatics Consortium (IAIC)
The management of the IAIC is split into three levels.

1) IAIC Committee consisting of the PIs leading the modules of the IAIC (initially, all senior personnel on this proposal). This committee would report to and interact with the SAB.
2) Scientific Advisory Board (SAB), consisting of scientists from countries involved in the IAIC. The SAB would oversee the development of the IAIC and interact with the funding agencies, MASC and the community. The SAB will be developed under this RCN proposal.
3) Scientific Advisory Panel (SAP) that would review the progress of the IAIC, and would be developed later under the AIP.

Since the funding streams supporting both core components and non-core modules are expected to come from different international funding sources, efficient operation of the IAIC will require careful planning. It is therefore proposed that the establishment of the IAIC is divided into two phases: 1) development of the IAIC and 2) operation of the IAIC. In phase one, and as one of the goals of this RCN proposal, the SAB would be appointed and begin liaising with funding agencies to determine possible mechanisms for setting up the core components and non-core modules. In some cases, this might require the establishment of specific calls for proposals, while in other cases existing funding schemes may already be in place. Irrespective of the mechanism(s) that funding agencies are able to provide, workshop participants strongly recommended that funding for the initial modules be secured. During the first phase of the IAIC, the SAB will also develop a suggested list of 'non-core' modules and appoint the IAIC director (PI Meyers will serve as the interim director until this time). There are likely to be many examples of projects that currently exist that could easily be adapted to become part of the IAIC. The SAB would help identify and liaise with such projects and provide information regarding the funding mechanisms available to adapt or establish these modules to become a part of the IAIC.

PIs will be encouraged to apply for funds in specific countries to adapt or establish components of the AIP.

While there may appear to be an overlap of functions between the SAP (reporting to the funding agencies) and the SAB (liaising with the funding agencies and reporting to the SAP), experience in other areas has shown that these two boards can fulfil very different roles. In particular, the SAB can have a more private and direct interaction with the scientists and PIs overseeing work within the consortium; thus the SAB has the opportunity to be more constructively critical of these scientists and the project.

Funding

During the workshops, there were wide ranging discussions of the current and future funding mechanisms for informatics and cyberinfrastructure and it was concluded, for the reasons that are clearly articulated in Chandras *et al* [4], that commercial, semi-commercial and cross-subsidy models are not feasible approaches for funding the IAIC. Instead, since the use, development and contribution of data, tools and resources are international, a transnational funding structure appears to be the most common sense mechanism for providing support for the IAIC, providing good value for money for scientists and funding agencies alike. Coordinated, international support for the IAIC would increase the number of financial stakeholders, spread the burden of long-term funding, and because the whole will be greater than the sum of its parts, we envisage that a distributed model that is internationally funded would encourage a variety of funding bodies to become involved and support this endeavor. The goal of this RCN proposal is to get the IAIC "off the ground" and to launch these initiatives.

Given the critical nature of the core components to the success of the project, a greater stability and therefore financial commitment from the funding agencies involved is required for the core of the IAIC in contrast to non-core modules of the IAIC. While there may be some turnover of the non-core components, driven either scientifically or financially, a stable core means that the resource remains sustainable over time.

Workshop participants proposed that the central components of the IAIC, in particular the AIP, should be stably funded on a five-year rolling basis with the appropriate review and renewal at time points consistent with the funding body/bodies supporting the core components. We envisage that each of the 'non-core' modules will be funded nationally or through consortia of national/international funding agencies with shared policy priorities. An internationally distributed funding model in the long term for the IAIC provides the plurality of funding, spreads the costs and the risks and generates added value for both core components and non-core modules investment. The separation of funding priorities between the core components and non-core modules allows financial sustainability to be prioritized and distributed between these activities thus providing greater stability for the core. This separation also provides considerably more flexibility in the spectrum of models, which might be adopted simultaneously across the IAIC.

Although it's beyond the scope of this RCN proposal (possibly the purview of the Scientific Advisory Panel), the workshop participants discussed and recommended that funding agencies consider providing specific programs for funding informatics and cyberinfrastructure that are separate from hypothesis-driven science and provide longer-term grant cycles. Ideally there should be a commitment in principle to a 5- to 10-year funding period with review and renewal possibilities at appropriate time points consistent with funding agency policy. In addition, new and more appropriate metrics for assessing the impact, quality and usefulness of data, tools and resources urgently need to be developed and utilized by funding agencies. Taken together, these recommendations would allow for long-term planning, recruitment and career development as well as the scientific development of informatics and cyberinfrastructure projects such as the IAIC.

Technology and Standards

The technological sustainability of the IAIC will depend on several features including openness, standards, intelligent new web-based solutions, widely applicable tools, and a centralized body to enforce standards. Openness in the context of data means that none are proprietary or subject to use restrictions and that raw data are easily downloadable. Openness in the context of database tools means that the underlying code for these is developed following an open source, collaborative model.

In utilizing a distributed model for the IAIC, whereby data from geographically dispersed sites are accessed and linked through one portal (the AIP), the development of clear standards to allow archiving, exchange, and mining of data will be critical. For the data contained in the AIP to be easily accessed and utilized, adherence to community standards for metadata will also become increasingly important. Examples of such standards for microarray expression data (MIAME) and for proteomics data (MIAPE) already exist, while others such as those for metabolomics data still need to be developed. In order for dispersed sites to feed data to the AIP on the fly, intelligent web-based solutions such as web services could be employed. Again, standards will need to play a role to make sure that the most current data are available via the AIP and also to ensure that there is interoperability across the IAIC.

To meet these challenges, the AIP will help develop and establish standards for existing data, tools and resources. These would assist current projects to be adapted to become part of IAIC and ensure interoperability between all parts of the IAIC. The AIP would also ensure that future resources conform to the necessary standards if they wish to become a module of the IAIC. To be effective, the IAIC will need to interact with and learn from the wealth of research communities that are also tackling the challenges of archiving, exchanging and mining data to ensure that the IAIC is part of a common technological framework whereby the information in IAIC can be brought to other communities and vice versa.

It will also be essential for the AIP to provide training for researchers wishing to access data in the IAIC as well as for those generating data, tools and resources and wishing to interact with or become part of the IAIC.

1.5 Conclusions

This is a critical moment for Arabidopsis informatics in particular and for plant informatics in general; the current model for the curation and delivery of Arabidopsis data is being perceived as insufficiently capable of meeting the community's expanding needs, as the amount of data is accumulating at a rapidly increasing rate. This presents a challenge to the community to review its needs and priorities. These should be articulated clearly and appropriately to national funding agencies that support major users and generators of Arabidopsis data. There is now an opportunity for plant biologists to develop a new international approach to informatics and cyberinfrastructure that will meet new needs for data integration, access and analysis. The workshop participants concluded that the development and maintenance of plant data, tools and resources, including those of Arabidopsis, would require significant support by funding agencies. However, the IAIC would leverage funding from a variety of sources, develop richer tools than a single group, and help to establish and set standards for informatics resources. As proposed, a federated, international model could facilitate inclusion of data and resources developed by, and for, other plant communities. Our recommendations are not without risks, and other model organisms face similar issues in sustaining their informatics resources and may well come to different conclusions about the best path forward. In the context of Arabidopsis and the tightly knit, yet global, group of researchers that study it, a well-executed implementation of these recommendations should establish a sustainable informatics platform to serve the broad range of needs and applications that we, and scientists studying other species, have for Arabidopsis data.

2. The rationale for a new network focused on Arabidopsis informatics

As described above, this proposal requests funding to develop a new organization, the IAIC, to establish the SAB, and to involve international research groups in the coordination of resource development for Arabidopsis bioinformatics. All senior personnel listed on this proposal will be involved in at least one of these activities. Thus, the RCN funding will allow us to meet, to coordinate informatics activities, to broadly advertise the mission and aims of this new network of scientists, to identify groups that can provide either core or non-core functions, and to coordinate international funding requests.

Part of the rationale for this proposal is the immediacy of the need to address the needs of Arabidopsis scientists. TAIR is charged with the mission of collecting information and maintaining a series of databases of genetic and molecular biology data for Arabidopsis. As described above, and while this could change, TAIR currently serves as a single source for core activities. TAIR has built a remarkable reputation for quality, and is housed and operated by the Carnegie Institution of Washington Department of Plant Biology, at Stanford University. For more than ten years, TAIR's funding has come predominantly from the National Science Foundation (awards DBI-9978564 and DBI-0417062). As has been discussed in a number of public forums [5], TAIR's most recent proposal was funded at a reduced rate relative to the full funding required to continue with their original plan of work. Most importantly, while TAIR has performed excellent work in meeting the objectives with which it was charged, it was not designed to be a dynamic distribution system for the myriad data types now available and planned for Arabidopsis; this is an objective that we believe can be met via the new AIP and modules developed under its umbrella by the international consortium. Thus, it is critically important that Arabidopsis scientists rapidly develop a new plan to meet the expanding informatics needs of the Arabidopsis community as well as other science communities that depend vitally on Arabidopsis resources. Not responding swiftly to this challenge is likely to result in the Arabidopsis community at best falling behind other organisms and at worst lacking a central information resource, either of whcih would have a dramatic impact in the value of Arabidopsis as a reference organism. The most effective approach will involve all relevant stakeholders, including new domestic and international partners, to develop new resources that better reflect the priorities of these communities.

3. Formation of the RCN and proposed activities

The first step in the effort to define the future of Arabidopsis informatics was the pair of meetings held in the UK and US in the spring of 2010. Participants included members of the Arabidopsis community as well as scientists working in other systems (e.g. *Drosophila,* mouse, yeast, *C. elegans*, maize), representing North America, Europe, and Asia. The focus of these discussions was the potential to reorganize Arabidopsis informatics using novel approaches. A tangible outcome of the meetings was a plan to organize, affiliate, and coordinate international informatics groups in order to establish a new Arabidopsis informatics consortium; this is the content of section 1 of this proposal. The objectives of the current proposal represent the next step toward realizing this plan.

The main aims of this IAIC RCN are: (1) to identify key informatics needs and challenges (both current and developing) of the Arabidopsis community and ways to coordinate international efforts to address them; (2) to bring together groups that will develop standards for plant bioinformatics data, deposition and display; (3) to form a new consortium to coordinate Arabidopsis bioinformatics in the US and abroad that will provide a robust framework to continue well beyond this RCN funding; (4) to “kick-start” this process so that no time is wasted in developing the resources needed by the community.

As will be discussed below, the IAIC RCN will host a series of activities to achieve our goals. These activities will be coordinated by national and international steering committees, with substantial input requested from the Arabidopsis and plant bioinformatics communities. We

have already started to coordinate with international groups such as efforts taking place in Europe (see attached letter from Willi Gruissem), and we are working to identify counterparts in Asia as well.

3.1 Meetings of the IAIC RCN

The IAIC steering committee named on this RCN will coordinate meetings, with an initial burst of meetings every six months, and then continued on an annual basis with members invited to attend in person or via video conference. The agenda for each meeting will be set by the IAIC steering committee. We plan to solicit advice from representatives of existing bioinformatics projects; examples include (among others) iPlant, BAR, TAIR, NASC, and GEO, as well as from individual computational biologists, for development of guidelines for data collection, data deposition, and data display (e.g. browsers). Meetings will be held annually at the International Conference on Arabidopsis Research (ICAR, held one out of three summers in the US, typically in Madison, WI, or at changing international locations), with an initial meeting held in January, 2011 around the Plant and Animal Genome meeting (PAG) in San Diego, and a second "interim" meeting held in January, 2012 also at the PAG meeting. A timeline of the RCN objectives is shown in Figure 4.

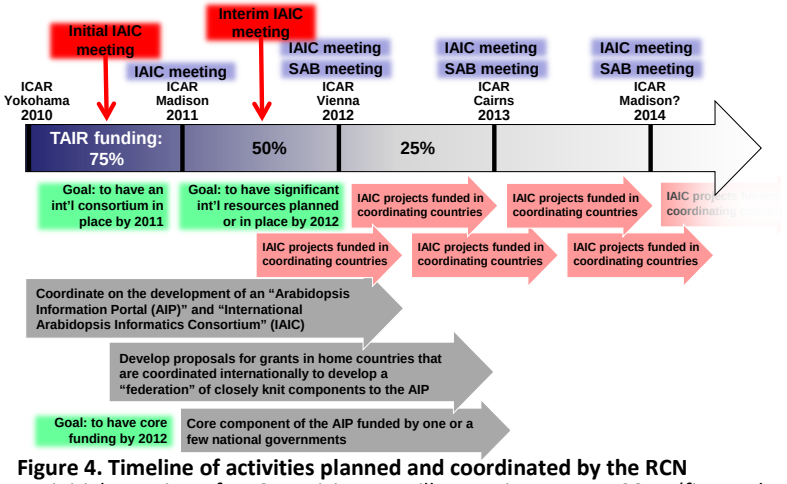


Figure 4. Timeline of activities planned and coordinated by the RCN
An initial meeting of IAIC participants will occur in January, 2011 (first red box), probably at the PAG meeting (San Diego). A second meeting will occur at the ICAR (International Conference on Arabidopsis Research) at Madison, and annually thereafter at successive ICAR meetings. A third meeting will be held in winter of 2012, most likely again in San Diego during the PAG conference. The first AIP board meeting will occur at the ICAR meeting in 2012. Also indicated are important goals (green), and the approximate proportion of funding that TAIR has budgeted relative to their 100% funding in 2010.

3.2 A draft plan of RCN-sponsored meeting objectives

The initial meetings of RCN participants will focus intensively on developing the goals and priorities of the IAIC, and on establishing community interest and involvement in the network/consortium. Importantly, we want to ensure participation of individuals in the IAIC from members of the Arabidopsis community as well as from broader disciplines including crop researchers, computational biologists and others. We will facilitate information exchange in a bidirectional manner to 1) inform researchers of opportunities to participate (e.g. develop module funding proposals and link existing and proposed modules to the AIP), and 2) incorporate valuable feedback on community priorities and needs. Although the objectives are sure to evolve as groups form and resources emerge, we outline a preliminary set of objectives for the planned network meetings which are as follows:

1) January, 2011 (San Diego): Held in conjunction with the Plant & Animal Genome (PAG) conference, but with the meeting likely held on the UCSD campus. Two planning and organizational sub-meetings will take place: A) involving the RCN PI and Co-PIs and B) involving also participants from the previous US and UK workshops (attending PAG) and several 'Subject Area Experts' to provide valuable expertise on areas such as database management, international collaborations, bioinformatics, data standards, etc. During the time between this meeting and the July meeting, we will build a website (see below for

more details) and start to recruit individuals into IAIC who are interested in contributing components. Expected outcomes of the first meeting include:

a. The identification of suggested names for the Scientific Advisory Board (which will be developed in consultation with MASC and funding agency representatives).
b. Preliminary objectives and framework for the IAIC to be discussed with the broader Arabidopsis community at the June, 2011 Arabidopsis Conference (ICAR) in Madison, Wisconsin, including a list of essential modules that need to be developed very early and those that can come later.
c. A plan for publicity and outreach to the Arabidopsis and other communities via websites, including an IAIC website (developed as part of this proposal), TAIR, NASC, and others; via invitation to participate in the 2011 ICAR public forum; and through development of an email list to keep 'interested community members' informed on IAIC news and progress.
d. Development of a plan to communicate with funding agencies and effectively convey relevant funding opportunities to the community.

2) June, 2011 (Madison): We will sponsor a workshop at the ICAR meeting, the first of a series to be held at all successive ICAR conferences. This will be a forum for selected IAIC contributors to outline their modules, including at least one that will demonstrate an example of a module that can connect with the AIP, once it's developed. We will invite funding agencies to talk about how they'll fund modules and how applications will be handled. The IAIC Steering Committee will sponsor a workshop at the ICAR to engage community members in the development of the IAIC and solicit feedback and input. The primary goals of this workshop are to solicit feedback and provide information, and in particular to facilitate the development of module funding proposals by community members. During this meeting, and in consultation with MASC, which also holds its annual meeting at ICAR, the IAIC Steering Committee will develop a list of candidates for the SAB, with the aim to have the SAB in place by July, 2012 (Vienna).
3) January, 2012 (San Diego): Held again during the PAG conference. By this meeting, invitations to join the SAB will have been sent to the candidates, and the responses will be evaluated. We will discuss the funding of the major components of the AIP and progress on the contributions by the IAIC members. The PAG meeting represents an opportunity for us to present the goals of the IAIC to the larger plant community, so a workshop will be held with this aim in mind.
4) At each of the next four ICAR meetings, we will have an IAIC steering committee working meeting, organize an SAB meeting, and sponsor a community workshop to highlight the progress of the consortium and the resources and tools that have been developed. This is an excellent opportunity to obtain feedback from the community on their needs and emerging challenges and opportunities. We envision that the role of the RCN PI and senior personnel will gradually diminish as the community (the emerging IAIC, stimulated by this RCN) and the SAB takes on more of the responsibilities and the IAIC grows. We anticipate that the AIP will be taking shape and growing during this time, and will begin to assume a larger central role in organizing informatics, including interacting with the SAB and forming the SAP. The planned meetings include the following ICAR dates and locations:

a. July, 2012 (Vienna)
b. July, 2013 (Cairns)
c. July, 2014 (tentatively Madison, WI)
d. July, 2015 (int'l location, TBD)

3.3 IAIC community involvement and web site

It is vitally important that we involve the Arabidopsis community in the activities of the RCN, to advertise our objectives, recruit IAIC members, receive guidance and suggestions, and promote our successes. Every effort will be made to ensure that the community doesn't perceive the IAIC as a "private club", but rather as an integral engine of the Arabidopsis society. While some of these interactions will come via members of IAIC, we propose to develop a website for both scientists and the general public that describes the activities of the RCN, incorporates community feedback, and provides access to information on IAIC participants. This site will be a component of the AIP. The steering committee members will design and set up the IAIC RCN website; the site will become the online forum for the International Arabidopsis Informatics Consortium and will be maintained under the auspices of this consortium. Sections on this site will describe: (A) IAIC members and subgroups; (B) IAIC RCN activities including news, IAIC meeting reports, guidelines for participation, major informatics needs and technical challenges, the time/location/agenda of the next annual meeting, upcoming workshops, international collaborations, funding opportunities and award announcements, etc.; (C) Standards for data collection, deposition, and display; (D) Educational, teaching and outreach opportunities.

3.4 Outreach and Training

Outreach and research training activities will be strongly encouraged for all projects that are designed under the auspices of the IAIC. One of the central goals of the IAIC is to promote public access and use of Arabidopsis data. However, most specific activities are likely be incorporated as part of individual grants that address research priorities identified via this RCN.

4. Management and Coordination of Activities

4.1 Project PI and senior personnel (including co-PIs): the initial IAIC steering committee

We have designated eight individuals as senior personnel, including the PI and four co-PIs of this project. Initial responsibilities for guiding and developing the IAIC are to be shared among these individuals, with the aim to share these responsibilities as more members of the community join the IAIC via their role in contributing informatics modules. PI Meyers will have the responsibility of heading the RCN, and he will work with a subset of members to organize the meetings described in section 3. The senior personnel (expected to include new members of the IAIC that will join as the consortium builds) will coordinate IAIC activities with existing funded initiatives whose missions intersect those of the IAIC RCN. These individuals will contribute to the annual meeting, will help gather and consolidate input from the Arabidopsis and larger plant communities, and will lead initiatives important to the IAIC that are relevant to their areas of expertise. They will also help write annual meeting reports for the IAIC website. With this set of senior personnel, we have assembled an international group of outstanding researchers who are leaders in the Arabidopsis community and in bioinformatics, with diverse backgrounds and expertise. Biosketches for this initial set of IAIC steering committee members (e.g. the senior personnel) are attached to this proposal. The University of Delaware will serve as the lead institution for the administration of the award.

4.2 The IAIC Coordinator

A coordinator, charged primarily with assisting the RCN PI and other members of the IAIC, will ensure appropriate and timely action is taken by the globally-dispersed members and is central to the success of the IAIC.

In addition to a solid background in Arabidopsis biology, and equally important with scientific knowledge, the individual must have demonstrated excellent (1) verbal and written communication skills, and (2) organizational skills, since most of the individual's responsibilities will be communicating information and facilitating interactions including workshops and meetings. It will be particularly helpful if the coordinator is already familiar with the Arabidopsis

community. To this end, we have identified Dr. Joanna Friesner (current coordinator of NAASC, the North American Arabidopsis Steering Committee) as a candidate for this position; her biosketch is attached.

Duties of the proposed IAIC RCN coordinator include the following:

- Facilitate development of the IAIC and advisory board(s) via electronic, written, visual, verbal communication
- Coordinate meeting organization including meeting materials, site selection, lodging, food/beverage, audio-visual, travel, meeting rooms, etc.
- Facilitate IAIC meetings with on-site assistance including, e.g., set-up, speaker assistance, communications with meeting staff, session facilitator/note-taker, etc.
- Develop and/or contribute to drafting of meeting reports
- Publicize/distribute outcomes of meetings; liaise with the international community to facilitate two-way information sharing (e.g. via MASC, NAASC, stock centers, ICAR, Arabidopsis bulletin boards, other plant communities); post relevant information on the IAIC website.
- Liaise with funding agencies to recommend to the community what grant opportunities should be considered as part of building the resources
- Develop and/or contribute to drafting annual reports; facilitate timely grant reporting to NSF
- Manage participant reimbursements including acting as point-of-contact for information exchange, receipt of supporting documents, accounting, feedback, etc.

4.3 The IAIC Scientific Advisory Board (SAB)

As described above, the SAB will be formed early in the RCN, with the goal to provide scientific oversight to the activities of the IAIC and the AIP. SAB members will receive a small honorarium and will meet annual at the ICAR meeting; both of these expenses are budgeted as part of this proposal for the first two years, with the assumption that the AIP will be funded by the fourth year of this RCN and will then take over the funding of the SAB activities.

5. Anticipated outcomes resulting from the formation of the IAIC RCN.

The IAIC RCN has several interlinked goals that will enhance international research coordination and collaboration. These goals, outlined below, will be achieved through the annual meetings and additional discussion formats described in the preceding sections:

1. Establishment of an International Arabidopsis Informatics Consortium (IAIC) that can build upon activities and resources that will be initiated under this NSF-sponsored RCN.
2. The RCN activities will facilitate development of Arabidopsis informatics tools and resources both in the US and among international participants.
3. The activities of the IAIC will lead to the identification and appointment of individuals within the IAIC to develop standards for Arabidopsis data, deposition and display.
4. Development of a ‘white-paper' outlining plant informatics challenges and opportunities that can serve as a guideline for establishing future funding priorities. Included will be primary needs and challenges for informatics modules as well as technical and bioinformatics challenges.
5. Establishment of an IAIC website as a resource for interested scientists and the general public.
6. Formation of the SAB to oversee activities developed under the IAIC and to coordinate with the AIP.
7. Establishment of priorities, needs and specifications for additional IAIC modules.

References

1. International Arabidopsis Informatics Consortium. (2010) *An international bioinformatics infrastructure to underpin the Arabidopsis community* Plant Cell. **TBD**.
2. http://www.1001genomes.org/
3. Parkhill, J., E. Birney, and P. Kersey. (2010) *Genomic information infrastructure after the deluge.* Genome Biology. **11**: 402.
4. Chandras, C., T. Weaver, M. Zouberakis, D. Smedley, K. Schughart, et al. (2009) *Models for financial sustainability of biological databases and resources.* Database (Oxford). **2009**: bap017.
5. http://www.nature.com/nature/journal/v462/n7271/full/462252a.html