

The iPlant Collaborative: A Cyberinfrastructure-Based Community for a New Plant Biology

Vision and Rationale for *The iPlant Collaborative*

To understand how biological systems function, one must uncover the higher order principles by which biological systems self-organize, function, and evolve. Biologists need not only large datasets that describe the components and pairwise interactions of biological systems; they need to find higher order patterns within seeming disorder. The ability to detect reproducible patterns and to mathematically describe these patterns from novel organizing principles is central to truly understanding biological systems.

There are several barriers to addressing ‘grand challenges’ in biology. Biologists lack the ability to make complex queries across heterogeneous datasets that are usually incomplete and arbitrarily organized. Groups develop their own file formats, data models and application software without awareness of comparable needs elsewhere. Thus, it is almost impossible to combine data and models from multiple sources and utilize the most up-to-date tools to analyze and simulate complex inter-relations. More globally, computational thinking and collaboration is diffuse because the goals of computational scientists and biologists are often not well aligned. An effective cyberinfrastructure will encourage communication among disciplines and reuse of data models, file formats, application software and algorithms, while fostering the type of cross-disciplinary exchange of ideas that will advance both the specific case of plant science, and the broader fields of life and computer sciences.

To address these challenges we propose to develop *The iPlant Collaborative* (iPC), a new community of plant biologists, computer scientists, mathematicians and engineers organized around a core cyberinfrastructure. The iPC will empower these groups to integrate their research and collaboratively address major questions in biology, simultaneously advancing both the biological sciences and the computer and information sciences.

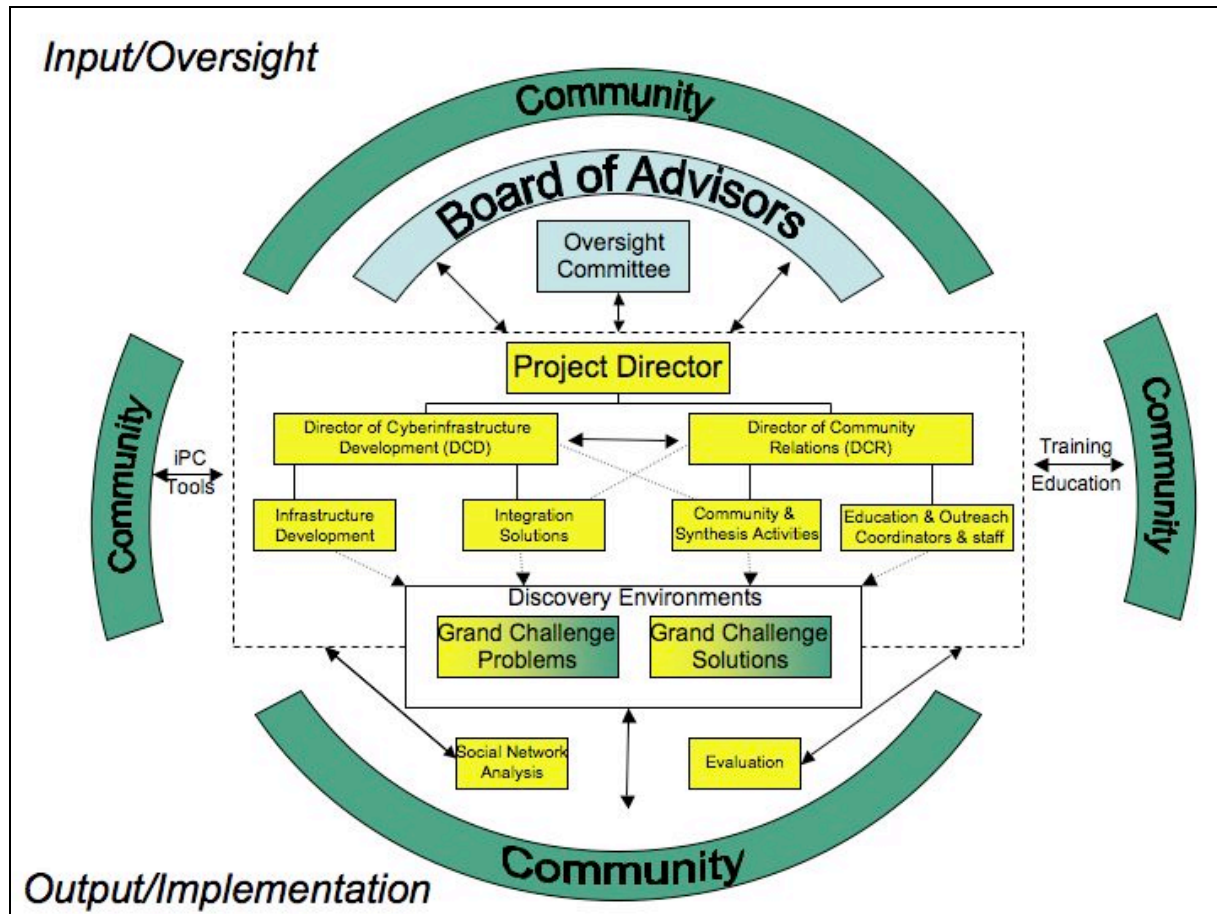
An organizing principle of the iPC will be Grand Challenge teams, cross-disciplinary, community-driven research groups that work collaboratively with iPC staff to design and develop ‘Discovery Environments’, software platforms custom designed to help the team address a Grand Challenge question. Discovery Environments will typically take the form of “mashup” applications which facilitate the integration of diverse types of data and tools, but beneath their surface simplicity discovery environments will support sophisticated systems for semantic integration, description, and manipulation of biological data types. Discovery Environments will be integrated into the growing infrastructure of the iPC, becoming in time an open source resource that is expanded and maintained by the community as a whole.

The scientific community will be encouraged to participate in, and take ownership of, the iPC via a wide range of synthesis activities, including smaller Discovery Environment design groups that are not necessarily associated with Grand Challenge Teams; a web-based virtual community center; outreach teams that train users, including those at undergraduate and minority-serving institutions, to use the iPC infrastructure to its best effect; and by partnering and developing synergistic, integrated ties with other centers, such as the ecology synthesis center (NCEAS) and the evolution synthesis center (NESCent). To guide its community relationships, the iPC will include resident social scientists who will work with the iPC to design, develop, and enable the implementation of social networking tools, resulting in strategies and mechanisms for building and nurturing collaborations.

Ultimately, however, researchers cross-trained to apply computational thinking to biology are the real infrastructure. To help train a new generation of scientists who bring computational thinking to biological problems, the iPC will work with the community groups to develop innovative curricula and training programs for computational thinking in biology at the K-12 and university levels.

The Deliverables of the iPlant Collaborative will be: (a) a physical and social infrastructure that supports computational thinking and problem solving in biology; (b) software tools that enable cross-disciplinary research teams to exchange, integrate and compare ideas, data sets, and algorithms;

(c) biology community development, social networking tools, and teaching tools for a diverse variety of users, including researchers, K-12 educators and students, and university educators and students; (d) a diverse workforce with substantial representation by underrepresented minority students and their teachers; and (e) detailed understanding of the social networking and dynamic properties of the broader Collaborative community.



A Use Case for a Typical Grand Challenge Problem

To illustrate our vision of how the Collaborative's cyberinfrastructure will work, we pose a typical grand challenge question in plant developmental biology: How do leaf primordia develop into the great diversity of leaf morphologies seen in nature? A leaf primordium is a dome of cells that arise on the flank of a shoot apical meristem and gives rise to a mature leaf through cell divisions and cell expansions. This process is controlled by fields of information that vary dynamically in space and time to determine cellular differentiation and leaf development. Individual cells "know" their relationships to other cells via these fields and respond accordingly to produce a pattern of division and expansion that yields the morphology and anatomy of a mature leaf. Cell division and cell expansion decisions are made dynamically in response to cell-to-cell interactions and intrinsic and diffusible signals, including ions, small molecules, and RNA and protein molecules. To understand the organizing principles of this process, the Collaborative must help scientists build data systems and computational methods that integrate diverse information on (a) leaf primordium morphology, the spatial and temporal expression patterns of genes, proteins, small RNAs, hormones and other molecules and ions in the meristem, primordium and derived tissues, (b) communication among cells within a developing primordium (local and non-local), and (c) natural and induced variations that

affect primordium development. Here is how we see this process working:

1. *The synthesis*: A grand challenge symposium on plant development brings together plant biologists – representing developmental biology, evo-devo, molecular physiology, cell biology, genetics and genomics – with applied mathematicians and control theory researchers who are developing adaptive control models and computer scientists expert in analysis of visualization data. The symposium results in a partial consensus on how to proceed and this group decides to form a Grand Challenge team, along with several members of the Integrated Solutions group who will act as liaisons between the GC team and the core Infrastructure Development team (see Figure).

2. *An early Discovery Environment*. Maricela, a UA graduate student with a background in information systems (specifically, in web application development) creates a WIKI for the leaf primordium group. Initially, this WIKI contains review articles summarizing what is known about leaf primordium development, a comprehensive set of key literature references chosen and organized by experts in the various disciplines represented, and places where the group can upload and discuss datasets. In collaboration with the control theory group, Maricela creates a WIKI-based upload and visualization interfaces for the biologists' various datasets, such as 3-D visualization through development of cell walls, auxin and cytokinin levels, homeodomain proteins, microRNAs, and intracellular calcium levels. She also helps the modelers create an interface embedded in the WIKI that allows group members to launch adaptive control models and tinker with its parameters. In this way, researchers can compare predictions from control models to actual biological data.

3. *Exploratory data mining*. The Grand Challenge team is initially focused on Arabidopsis and tomato as model systems, but the team decides that it is important to obtain data from a monocot as well in order to explore both the specificity and generality of models. Xi, a bioinformatics postdoc from the Integrated Solutions group at CSHL, is asked to research the availability and quality of the appropriate datasets in monocot species. During this process, he narrows the field to *Zea mays*, which has an abundance of freely-available microarray data, confocal microscopy images of key protein distributions, and spontaneous and induced mutations affecting leaf morphology; Xi begins the process of consolidating and cleaning datasets needed to support the team's research. The team invites experts in maize leaf development to join the Grand Challenge team.

4. *Theory and reality collide*. The models developed by the control theory group to simulate cell division and growth have performance limitations that restrict their use to five neighboring cells at a time. Likewise most of the available biological data have resolution at only the 10-cell level, not single cells. Extensive video-conference discussions lead the group as a whole to realize that it needs to overcome several challenges in order to move forward and decides it should: (a) recruit additional expertise in image data processing to try to obtain better resolution closer to the cellular level from noisy data, (b) obtain higher quality microscopy data in the early stages of development when cells are small, and (c) generate more scalable control models.

5. *Computer science consultations*. To assist with these challenges, Lakshmi and John, computer scientists from IS, join the team to assist the control theory group in optimizing its model to handle more cells and develop a method for parallelizing the simulation across multiple nodes of a compute cluster, thereby allowing simulations to be run across ten distinct cells. Computer scientists with expertise in image data and digital signal processing are recruited by the Integrated Solutions team to extract additional information from diverse image types and accurately layer different data types (e.g., cellular boundaries upon the distribution of microRNA and auxin levels).

6. *From prototype to production software*. The algorithms developed by the Integrated Solutions team now need to be implemented as production-quality software. These tasks become the province of the Infrastructure Development (ID) team. Lakshmi, John, and Maricela prepare the requirements, analysis, specifications and other documents needed by the ID team and act as liaisons between the Grand Challenge team and the ID team to test, assess and deploy software into the developing Discovery Environment including integration with the large scale storage and compute systems

available to the Collaborative. Members of the ID team will test and refine the parallel algorithms to ensure that scale up to very large systems is possible.

7. *Widening community involvement.* The Discovery Environment becomes a public portal for accessing the data sets, algorithms, models and conclusions generated by the group. The Discovery Environment software itself is published in open source form suitable for running on iPC large-scale computation systems, on individual workstations or clusters, or for use at supercomputer centers such as those provided by NSF's Teragrid so that it can be used and reused by community groups independently of the iPC. Publication of parts or the whole of Discovery Environments will occur throughout the project.

8. *Charting the way forward.* Grand Challenge and Integrated Solutions team participants conduct regular web-based meetings to discuss project progress and to chart its course. These meetings use real-time collaboration facilities, including an electronic whiteboard, application-sharing, video-conferencing, and instant-messaging. Another source of communication is the Discovery Environment itself, which has now evolved from a simple WIKI into a place for executing queries on the shared data sets, visualizing the output of the control theory group's models, and exchanging protocols. The Discovery Environment is also used by the Grand Challenge team to prepare collaborative research proposals for producing new experimental datasets to test the group's hypotheses and improve the quality of models. During these collaborative meetings, Maricela, Xi, John, Lakshmi, their faculty mentors, and other Integrated Solutions participants work with Grand Challenge team members to prepare a framework for receiving and managing new data produced by the biologists, including the community at large. This framework will be prototyped by the Integrated Solutions team and implemented by the Infrastructure Development team, as described in step (6).

9. *A wealth of new data.* As new data from other funded projects arrive, the Grand Challenge team visualizes and analyzes it using the now public Discovery Environment, created by the combined efforts of Grand Challenge and Integrated Solutions team members and usefully embellished by independent contributions from the community. The Discovery Environment allows users to integrate and visualize information from RNA and protein expression patterns, anatomy, and molecular physiology through time and space, and to compare these data sets to the predictions of models. Users also control which features of the environment they choose to use, such as 'competing' algorithms from the team and the broader community. A statistician from the Integrated Solutions group assists in preparing new data for analysis, identifying outliers and systemic errors, performing the data analysis, and facilitating interpretation of how the results impact the team's cell growth models. Future studies are then designed to expand upon the discoveries.

Synthesis Activities: Bringing Research Communities Together

Synthesis activities team will be the 'community face' of the collaborative. The iPC's Community and Synthesis Oversight Team (ComSOT) will be charged with ensuring inclusiveness, diversity, and effectiveness. Its role, assisted by the Director of Community Relations (see Management Plan), will be to facilitate the community's identification of grand challenge problems and prioritization of iPC resources and activities and to foster productive interactions between Grand Challenge teams and the Integrated Solutions team. Clear and effective communication will be essential in order for the community to understand the nature and scope of opportunities presented by the Collaborative's capabilities. We will use a variety of means to introduce and explain opportunities, such as announcements to plant and computational scientists in academic institutions in the US, and informational workshops at major meetings, *e.g.*, PAG, ASPB, BSA, ICIS, and ESA, and project-organized annual conferences to explain the multiple ways community members may participate. With oversight from an external Board of Advisors (BoA), the iPC will develop policies and procedures to ensure that the priorities and perspectives of the community, as well as inclusiveness and diversity, are considered and addressed throughout the infrastructure design

process. The effectiveness of iPC communications, policies, and practices will be assessed on an ongoing basis by the Social Networking and Evaluation teams.

Facilitating Discussion, Analysis, and Choice of Grand Challenge Problems. We will organize a “kick-off” conference on ‘Grand Challenge Problems in Plant Biology’ in early 2008 that will host up to 150 attendees spanning all relevant disciplines and will be publicly webcast to foster the broadest possible community participation. We will schedule an equal balance of invited speakers, discussion time, and break-out groups (about 1/3 each). Speakers will not only be invited, but also drawn from registrants, and subject to BoA oversight and approval. A video of the main proceedings will be accessible on the web throughout the project.

Major focus areas of the kick-off conference will include:

- A. *What is a grand challenge and what are the most important ones in plant biology?* Presentations by leaders in the plant biology community.
- B. *What are the data management, mining, discovery environment, and other infrastructural needs?* Presentations by members of the Integrated Solutions and Infrastructure teams, as well as members of the community, including examples of how collaborations involving computational thinking lead to new insights and discoveries, and ways to overcome barriers to collaboration.
- C. *Training the next generation.* Presentations by members of the community (including project team leaders) on “teaching computational thinking in biology” and how to use the iPC to facilitate this.
- D. *Invitation of proposals for Grand Challenge teams.* Proposals will be invited from the community at large through a variety of communication tools.
- E. *Discussion of the diverse means for community participation and learning.* Smaller more focused faculty teams will be supported with visits from iPC members.

A major focus of the Collaborative will be a series of small, focused symposia held at the Biosphere 2 conference facility in Oracle, AZ, or CSHL’s Banbury Center to create a retreat-like atmosphere. Each symposium will consist of 25-40 invited leaders from experimental and theoretical plant biology, computer scientists, information systems specialists, statisticians and engineers, and will be focused on broad areas of plant biology, e.g., plant development. These symposia will be designed specifically to encourage delegates to go beyond their standard, canned talks to challenge experimentalists to enunciate the types of information resources and theoretical models they need to interpret the biological system under discussion. Theorists will be challenged to list the datasets they would need in order to create working models. In addition to community researchers, symposia will be attended by representative members of the Integrated Solutions, Infrastructure Development, and Synthesis Activities teams. Typically, the first half day of a workshop will consist of a meeting with the Integrated Solutions team at BIO5 on the UA campus to learn about the iPC’s cyberinfrastructure capabilities and plans. This will be followed by a 3-day intensive workshop at Biosphere 2 to address Grand Challenge Problems. A final half day will be used for a follow-up meeting with the Integrated Solutions team at BIO5 to discuss the data analyses and cyberinfrastructure needs that have arisen during the retreat. (For retreats held at the Banbury Center in New York, meetings with the teams in Arizona will be held via web-conferencing, and follow-up visits to UA will be scheduled as needed). Up to twelve symposia will be held during the first year, and up to four in each subsequent year. From some of these symposia will arise the seeds of Grand Challenge teams, whose plans will be developed in greater detail in follow-up virtual and in-person meetings.

We anticipate at least two types of Grand Challenge symposia: one organized to assess the nature of grand challenges in a given discipline (such as plant-microbe interactions or signal transduction networks) and another that is driven by a self-forming group with a specific interest in a Grand Challenge Problem. Proposals of each type will be evaluated with the assistance of a nationally representative group of reviewers/advisors with oversight by Board of Advisors. Symposia teams

will be configured to be diverse and multi-disciplinary. Symposia may lead to one or more proposals to create Grand Challenge teams, again to be vetted by representative reviewers and the BoA. Selected Grand Challenge teams will come to BIO5 to work with project staff as needed to plan and to develop a custom Discovery Environment for the project. A broadly representative community oversight committee will be selected for each Grand Challenge team to review progress. UA faculty counterparts will be assigned as appropriate to provide expertise and necessary local advising, mentoring, and oversight of students and postdocs working with Grand Challenge teams. Visiting scientists may work in residence at BIO5 or CSHL for extended periods of time. To the extent possible, “common denominators” among GCP projects will be identified in order to help prioritize efforts. In addition, we anticipate that symposia will be held that focus on problems in biological data management and algorithms, as well as symposia that focus on higher education and K-12 education/outreach.

In subsequent years, we will hold a large Grand Challenge Problems in Plant Biology conference (200-300 participants), alternating between CSHL and Arizona (or other sites specified by the community), inviting the top symposia participants and Grand Challenge team members of all disciplines to present their perspectives and their results to date. We intend to select as many new attendees for the conference as possible in order to welcome and introduce more people into the Collaborative; again, the proceedings will be webcast to permit maximum participation.

iPC WIKI. The process of proposing, reviewing and organizing Grand Challenge Problem symposia will be aided by a project-wide WIKI, which will be open to all community members. In a similar fashion, we will use an internal (closed) WIKI to manage applications from prospective postdoctoral fellows and graduate students. This will help to coordinate the project so that there is a good balance between emerging Grand Challenge teams and a pool of talented and qualified postdocs and graduate students. Postdoc candidates for Grand Challenge team(s) will be selected through a national competition; they will have backgrounds in biology and/or CISE and be cross-trained in multiple areas and will work in teams with graduate students and Integrated Solutions developers, and will participate in K-12 outreach activities.

Partnerships. Partnerships with organizations such as NCEAS and NESCENT, as well as the biotechnology industry, will be addressed and promoted through liaisons: Brian Enquist for NCEAS, Michael Sanderson for NESCent, Steven Goff for industry, Lincoln Stein for international interactions, and Jean-Philippe Vielle-Calzada for interactions with developing countries.

Create/Support iPlant Action Teams (iPATs). There are now many examples of individual partnerships between computational scientists and plant biologists; smaller universities have typically provided fertile ground for these chance collaborations. We propose to increase the number of these individual collaborations, develop training materials, and provide access to iPC tools by organizing iPlant Action Teams (iPATs), small interdisciplinary teams of plant biology, computer science, and statistics / mathematics faculty and students, who will operate by going out into the community to teach and train interested groups in the community. These iPAT-trained groups will be able to contribute to one or more public Discovery Environments, which will also serve as a resource for the development of teaching tools. A selection committee will evaluate applications by small groups of scientists for consulting visits by iPlant core team members, who will assist in design of data analyses that build true collaboration and have the potential to fit with future grand challenge questions. The Algorithms subgroup of Integrated Solutions will work with the selection committee to choose proposed analyses with the broadest potential for reuse. Each iPAT community node will construct and contribute a software module to the public Discovery Environment(s), and one member of each iPAT team will then become an iPAT trainer and train at least one new team to leverage the development of cyberinfrastructure collaborations. The Social Networking and Evaluation teams will assist in tracking the performance and quality of the iPAT path to collaboration.

The Integrated Solutions Team (IS Team): Bridging Biology and Infrastructure

The IS team is led by co-PIs Stein and Ram, who bring together many years of experience in biology, computer science, enterprise data management, bioinformatics, genomics and software engineering. Other IS key personnel, including Barnard, Ware, and Snodgrass, add expertise in image analysis, machine learning, workflow management, cluster computing, statistical analysis, and large-scale data management. Two types of staff members will contribute to IS: the research staff, consisting of postdoctorals, graduate students, and their mentors, participants in a Grand Challenge team, and the software engineering staff, which helps the research staff design and prototype the software applications that make up the Discovery Environments needed to support GCP projects.

IS research team members are drawn from diverse backgrounds, including bioinformatics, computational biology, computer science, information systems, statistics, physics, and mathematics. The majority of research team members are trainees at the graduate student or postdoctoral levels. We believe that having students and postdocs at the center of IS research, contributing directly to the GCP projects is the most direct way to foster a new generation of plant scientists skilled in quantitative, computational, and integrative thinking. Team members will attend grand challenge symposia and will become members of the resulting GC teams providing expertise in exploratory algorithm development and data management and mining. They will also participate in the design of Discovery Environments to support their grand challenge projects and work with IS software engineers to prototype the Discovery Environments. When the prototyped Discovery Environments are sufficiently stable, IS research team members will hand over the problem to the core Infrastructure Development team who will turn the ideas developed by the GC teams into production-quality, portable software, databases, and visualization engines. IS research team members will continue to liaise between the infrastructure staff and grand challenge collaborators to ensure that the software gets into the hands of the collaborators and does what it is intended to do.

IS software engineering staff will be professional software engineers, whose role is to provide support to research staff for data mining, algorithm implementation, data management, and application development. These software engineers are distinguished from those who work in the infrastructure core by having skills in agile software development, a paradigm that emphasizes rapid development, flexible requirements analysis, and extensive early prototyping and testing. Software engineers with more traditional training easily get frustrated when dealing with biological applications due to the fluid and underspecified nature of the problems. Agile software developers, who often come from open source software development backgrounds, are more temperamentally suited to the fluid environment of the GCP projects, but typically poorly suited to the task of creating finished, hardened software that is the domain of the infrastructure core.

Discovery Environments are the main deliverables of the IS Team. A discovery environment is a software system that allows GC team participants to access the relevant data sets, integrate across them to identify connections, visualize them in ways that allow the ‘big picture’ to appear, manipulate the data with analytic tools, and share results by facilitating computational steering. Our model for Discovery Environments are Internet ‘mashups’, also known as Web 2.0 applications, which allow community members to build content in a democratic way, to make and label connections between different types of content, and to integrate a variety of different types of information in a single user interface. The wildly successful Wikipedia project (www.wikipedia.org) is one such application. It allows users to create and interconnect knowledge using the shared metaphor of an encyclopedia. Google Maps is another well-known mashup application. It provides the community a common reference system, a detailed geographic map of the world, and encourages people to link in their own data sets indexed by GPS location. Data sets contributed by independent groups, for example average housing prices compiled by one group and mean SAT scores of students enrolled in school districts compiled by another, become dramatically more useful when linked together by a common coordinate system. Mashups reveal new patterns among data and allow one to

make hypotheses of causality that would be impossible if the data sets were examined separately.

Discovery Environments are the cyber equivalent of the iPC physical meeting spaces. During the formative phases, they will provide a way to exchange ideas and prototypes and collaboratively create and refine the approach. During the production phase, they will provide a collaborative environment in which to exchange ideas, integrate data sets, share protocols and explore algorithmic approaches. Ultimately, they will be a way to publish the project's research findings to the world and to invite participation from the wider community.

The IS team will create Discovery Environment mashups for plant biology by (1) identifying long-lived data sets that will serve as shared coordinate system frameworks for integrating disparate data sets and (2) providing the community with software services that enable the layering of data sets on top of these frameworks, in a distributed, community-controlled manner. The particular data set frameworks identified by IS staff will depend on the GCP's that are chosen by the community, but illustrative examples include annotated genomes, named sets of genes, their aliases and cross-species orthology relationships, phylogenetic trees, protein structures, anatomical descriptions of plant tissues and/or developmental stages, annotated collections of microscopic images, machine-readable descriptions of biochemical or regulatory pathways, and geographical descriptions of species distributions. For example, for a GCP project that requires extensive cross-species gene comparisons, we might build a Homology Registry that allows community members to assert (and dispute) phylogenetic relationships among members of gene families based on different types of evidence such as sequence conservation and synteny. For a GCP project that involves dissection of signal transduction pathways, we might provide a WIKI-like environment that allows GC team members to assemble a comprehensive description of plant G-protein coupled receptor kinases that combines written text with embedded media that show the position of the kinases on several genomes, the evolutionary trees that relate the kinases across species, kinetic models of signaling cascades driven by the family, and a map showing the geographical distribution of allelic variants of plant kinases.

Whenever possible, Discovery Environments will be based on existing software products and will be coordinated with groups performing similar work. For example, if a Discovery Environment requires a common coordinate system based on an ontology, we will use an existing ontology such as the Plant Ontology (Ilic et al. 2007), if feasible, and coordinate the effort with the National Center for Biomedical Ontologies. Likewise, Discovery Environments based on a genome assembly and annotation will leverage interfaces developed by existing repositories such as TAIR (Garcia-Hernandez et al. 2002), Gramene (Jaiswal et al. 2006), and NCBI (Wheeler et al. 2007), rather than attempt to replace the functionality of those resources. We are also aware of numerous efforts in the bioinformatics and broader web development communities to create mashup systems, several of which would make good foundations for specific Discovery Environments, including QEDWiki (<http://services.alphaworks.ibm.com/qedwiki/>), the Taverna workflow management system (Oinn et al. 2004), AJAX GBrowse (<http://biowiki.org/view/GBrowse/WebHome>; for genome-based collaboration), and Galaxy2 (Giardine et al. 2005).

Data Management is another key role of the IS team, which will be responsible for developing state-of-the-art data management capabilities for the iPC. Grand challenge research teams will need access to large existing data sets and, in many cases, will be generating novel data sets of their own. Unfortunately, many of the existing datasets have incompatible formats, are difficult to locate, or pose other barriers to effective utilization. For this reason, the IS data management subteam will assist in identifying, managing and utilizing existing and novel data sets within the context of Discovery Environments. In order to support a diversity of grand challenge problems, the iPC will have to support a broad variety of data types including sequences, gene expression profiles, publications, proteomics, metabolomics, phenotypes, genetic maps, morphology, and genotypic diversity. The iPC has to accommodate data in a startling variety of formats, data with differing spatial and temporal granularities, and data that are rife with synonyms and homonyms. Lastly, as

most data will be stored remotely in third-party maintained repositories, the iPC will have to provide the same quality access to these datasets as to Grand Challenge team-generated local data.

To meet these requirements, iPC will provide (a) a meta data management system allowing data to be described with community-developed standards; (b) capability to browse, query, download and/or use data from multiple repositories; (c) services for the plant sciences community to deposit data into the iPC; (d) mechanisms to link databases as plant scientists use the data and discover interesting relationships among the data (e) a framework for defining data sharing policies and service level agreements among the community; (f) mechanisms to resolve naming, format and other types of heterogeneities among the various data sets; and (i) the ability to store large temporary data sets for use in Discovery Environment workflows. Details on data access, protection and preservation policies are provided in Appendix A4.

While several of the capabilities described here can be supported by adapting and integrating existing tools and techniques, there are some fundamental information representation, integration and other data management solutions tailored to plant biology that will emerge from this group. These solutions will include mechanisms for modeling the semantics of plant sciences data, tracking the provenance, frameworks for resolving semantic conflicts by incorporating ontologies specific to plant sciences, and schema/data versioning to support evolution in data sets. These will be developed by the computer science, statistics, and information systems research members of the IS team, working in close conjunction with the GC teams. We will also coordinate and partner with other NSF funded projects that provide useful tools in this arena. These include the CiteSeer, ChemXSeer and Lionshare projects at Penn State University (citeseer.ist.psu.edu/ ; www.lionshare.edu) that have developed tools for secure data deposition, sharing and information extraction and mining based on metadata schemes (See letter of support from L. Giles, P. Mitra and M. Halm).

Algorithmic Infrastructure is vital to the iPlant mission of accelerating discovery through advanced computational methodology. Hence, IS research staff will be on hand to evaluate, adjust, integrate, index, improve performance, and broker the implementation of new algorithms, simultaneously serving the needs of GC teams and advancing the fields of information and computer science. For this goal to be more than a fervent hope, it needs to be done through a tight collaboration between plant scientists, computational scientists, and other members of the research community, with iPC acting more as a midwife than a mother.

An Algorithms subgroup of the IS team will work proactively to determine what is needed by multiple GCP projects at the domain level, oversee the evaluation of the state-of-the-art algorithms, consider the costs and benefits of developing algorithms that are not yet available, set up an ongoing evaluation of those algorithms that are incorporated into DE's, and broker the development of new methods where there is need. A key charge would be to ensure efficient transfer between state-of-the-art math, computer science and biology.

Although we cannot know in advance what types of techniques will be needed to address GCP's, the IS algorithms subgroup participants do have expertise in a wide range of techniques that are likely to be relevant, including Bayesian methods, biometry/biostatistics, bioinformatics, ecoinformatics, biological sequence analysis, classification and regression trees, cluster analysis, computational biology, data mining and visualization, ecoinformatics, hidden Markov models, image analysis and machine vision, machine learning, meta-analysis, Markov chain Monte Carlo sampling, multi-modal modeling, network analysis, pattern recognition, population and quantitative genetics, principle components and correspondence analysis, research design, signal detection, spatial and spatio-temporal analysis, and stochastic optimization and modeling.

Workflow Management. As GC teams mature, repetitive computational workflows will emerge; for instance, it may be necessary to run a sequence similarity search, followed by Markov clustering, followed by a Gene Ontology labeling of the clusters, every time a source sequence database changes. The IS team will assist GC teams in managing such workflows by prototyping and testing

each step of the workflow and then handing the prototype off to the infrastructure group in order to make it fully error-resistant, portable and parameterized (e.g., abstracting away the dependence on a particular sequence database). An often ignored aspect of workflow management is that large scale high-end computing systems already have in place mandatory resource management and scheduling software. The IST will build workflow management software that can cooperate seamlessly with the most common resource management and data migration software available at supercomputing centers (e.g., LSF, PBS, and Torque). To assist in the development of reusable workflows, we will instantiate the workflows as assemblages of web services whenever possible, using domain-aware technologies such as bioMOBY (Wilkinson et al. 2005), semantic MOBY, also called SSWAP (semantimoby.org/), DAS (Dowell et al. 2001) and caGRID (Saltz et al. 2006). As an aid to the design and execution of workflows, we will assess and adopt a suitable workflow front end application such as Taverna (Oinn *et al.* 2004).

The Infrastructure Core: Hardened, production quality software for stable cyberinfrastructure

The physical iPC infrastructure will contain computational facilities to support software development and the computing and visualization requirements of scientists doing computational modeling, analysis, data discovery, and other computing-intensive experiments. It will also contain large storage systems to provide persistent, reliable, and effectively unbounded storage for plant science data. The repository will ensure that key data sets are preserved beyond the lifetime of the projects that produced them. It will also support reproducibility of experimental results by providing mechanisms to archive snapshots of experimental configurations, including all software and data used to generate a given set of results.

The iPC Infrastructure Development team will have a full-time staff to install, develop, document, and maintain software tools in support of Grand Challenge teams, administer the physical infrastructure, and provide help-desk support for users. The staff will also have a small research and development team to design, prototype, and eventually deploy software systems that are needed but not available elsewhere. For example, creating a ‘reproducible experiment’ archive as specified above is a hard, as yet unsolved problem. One key function of the infrastructure staff will be to insure scalability of the software to both large numbers of processors and large datasets. Several members of the team have significant experience in developing applications for extreme-scale systems.

Appendix A3 describes the proposed iPC infrastructure in detail and explains how the infrastructure will be developed and managed so that it is kept at the leading edge of technologies required to solve grand challenge problems in plant science.

Education and Outreach: Training the Next Generation of Plant and Computer Scientists

Biologists are moving rapidly from working on simple systems of a few interacting genes or species to complex webs of thousands of interacting partners. To understand this complexity, biologists must adopt the paradigms of ‘computational thinking’, using problem-solving techniques developed by computer science and mathematics. To prepare the next generation in these new ways of thinking and working together, the iPC will provide a range of activities for students and teachers from K-12 through graduate education. These activities will strive to bridge communication gaps across disciplines, provide multidisciplinary research opportunities at all levels, increase participation by members of under-represented groups, and provide new models for education across disciplines.

Computational Thinking in K-12 Education (iPC-K12). We will develop a strong, innovative K-12 outreach program which serves and engages, at some level, all participants of the K-12 education community including teachers, administrators, students and parents. Our program will provide a means for teachers to prepare the next generation of scientifically literate citizens who understand and use computational thinking and will provide opportunities for teachers and students to participate in grand challenge projects to understand how computational thinking will enhance an understanding

of plant biology. We will form partnerships with administrators to involve schools as active participants, especially to develop strategies to engage parents in exciting, new dimensions of their children's education. K-12 outreach has a strong presence at UA and its partner institutions in this project. Of particular importance are existing K-12 programs, GK-12 grants, teacher internships programs, summer opportunities for independent study for high school seniors, Masters Programs for teachers, and the nationally recognized Dolan DNA Learning Center at CSHL. We will devise ways to engage not only partner institutions' K-12 programs, but nationally recognized programs, to create synergy, communication and leadership among all these resources.

As an example of the type of an iPC-K12 activity, we will develop a summer teacher program to teach basic concepts of computational thinking integrated with plant biology. Teams of nationally selected interns will be formed for K-5, middle school, and high school, with each group developing the appropriate content for their age groups. The K-5 team will focus on augmenting widely used kit-based materials, such as the FOSS system developed by Lawrence Hall of Science, and middle and high school teachers will explore ways to integrate computational thinking/plant biology into existing state and national standards. Each three-member team will work separately yet interact to create continuity among the age groups. A team can be based at the UA or a partner institution. Each team will include expertise in computer and information sciences, mathematics and biology; each team will include at least one teacher from a predominantly minority school to ensure that we reach underrepresented populations. Ideally, teachers will commit to the program for several years and become community mentors for knowledge dissemination.

Deliverables will be produced in the form of teaching modules and tutorials that target a national and international audience of educators and students. iPC team members will provide learning opportunities and guidance, but ultimately teachers must customize the learning modules for their specific age group. Of particular importance is the exploration of gaming as learning tools. As a brief example of tutorials, for Year 1 middle and high school teachers could focus on computational thinking and plant genomics/bioinformatics, but as grand challenge projects develop in subsequent years, teacher teams will take advantage of these projects to produce teaching modules. K-5 project development will await input from teachers.

We will make use of an existing high school student summer research experience program organized by co-PI Napoli through the UA Honors College will support two teams of high school students with interests in biology, computer sciences, statistics, and math to attend the UA Summer of Excellence program. Students will be recruited from a national pool of applicants. The student teams will interact with team members on iPC projects. We will focus on preparing them for new college programs that integrate computational thinking and biology. We will track these students throughout the grant period to understand the impact of mentoring and preparation to success in future years.

The broad goal of iPC-K12 is to lead the nation in outreach. To encourage more participation and input from the national education community, as well as nationally-recognized K-12 outreach programs, and to broadened exposure to iPC-K12; we will develop a yearly K-12 Computational Thinking Workshop that includes the educational advisory board, as well as additional K-12 education leaders, plant biologists committed to furthering K-12 education, and a group of nationally selected teachers with expertise in computer sciences, mathematics, and plant biology. While this proposal presents examples of outreach activities, the ultimate program will depend on input from the workshops.

Computational Thinking and Plant Biology in Higher Education. In higher education, we will provide mechanisms to integrate computational thinking with biology across all levels from community colleges through graduate school with the ultimate goal of developing broad-based, interdisciplinary programs in Computational Thinking in Plant Biology. Our approach is twofold: 1) Directly involve undergraduate and graduate students from multiple disciplines in the core activities

of the iPC as members of the IST and GC teams. By participating in integrative research, this new generation of scientists will learn how to tackle and solve grand challenge problems in plant biology through multidisciplinary teamwork. Students at non-iPC institutions will have access to iPC through iPlant Action Teams to access and contribute to designing Discovery Environments. Students may also participate in summer research opportunities. 2) Engage educators with experience or interest in teaching computational thinking in biology and facilitate discussions on how best to integrate various disciplines to expand opportunities for students, building on the work of many established programs nationwide. This will be done through symposia held at UA and CSHL, workshops held at the annual meetings of professional societies, such as ASPB, BSA, and SIGCSE, and community discussions via Wikis. These activities will lead to the development of new courses, learning modules, curricula, and web-based learning tools and games to engage students at all levels. The results of these activities and their effectiveness will be disseminated through various education conferences and journals in the participating disciplines.

Three-day education symposia will be held approximately twice a year at UA/Biosphere2 and CSHL/Banbury Center with about 25-30 participants drawn from a broad range of disciplines and higher education institutions, including community colleges, PUIs and minority serving institutions. A broadly representative Education Advisory Board be responsible for identifying and selecting participants. The first symposium will provide an introduction to the capabilities and resources of iPC and how they can be accessed. Sessions will identify key issues and new approaches for preparing young scientists. Working groups will form to discuss specific issues. The final day will be used to bring ideas from the working groups together, identify common themes, and prepare an advisory statement for curriculum development. Subsequent symposia will focus on developing curricula, research topics, integration and contributions of multiple disciplines, etc. Three faculty (chosen nationally) will be supported each summer to develop courses, projects, and work with students at the iPC center in BIO5.

With guidance from symposia participants, we will start in Year 1 to develop curricula for inclusion of Computational Thinking in Biology into traditional programs in biology and computer and information sciences as a minor or emphasis. The eventual goal is to create new degree programs at the graduate and undergraduate levels at the UA. Tailored curricula will be established in conjunction with affiliated departments for students with particular interest in specific disciplines, with a goal of training them to be multidisciplinary scientists. Many disciplines will participate, including biology (Plant Sciences, EEB, and MCB) and computer and information sciences and engineering (CS, ECE, MIS), as well as Linguistics, Applied Math, and Statistics. To build common knowledge, courses that bridge gaps of understanding between different disciplines, for example Computational Thinking for non-Computer Scientists and Biological Systems for non-Biologists will be identified and developed as needed. The new interdisciplinary graduate program will offer PhD, MS, and PSM (Professional Science Masters) degrees. A cornerstone of all programs will be participation in iPC multidisciplinary research teams for students at each level to develop and apply their skills. UA's LTC will provide support to make courses and educational materials available online for a broad audience and all iPC workshops will be recorded and made available as soon as possible.

Undergraduates and exchange students at UA participating in iPC research teams will have access to UA's successful UBRP (Undergraduate Biology Research Program) and BRAVO (Biomedical Research Abroad Vistas Open) programs. UBRP is an established program that teaches undergraduate students how research is done by placing them in research groups. In 2005, an Interdisciplinary UBRP program was established to target undergraduates majoring in mathematics, computer science, engineering or physics and providing them with a biologically related research experience. The BRAVO program enables students to forge lasting relationships with foreign scientists at a formative time in their lives through a three-month summer abroad experience. This

grant will sponsor up to 10 UBRP students each year starting in year one and will award three BRAVO scholarships to UBRP students starting in year two. While these programs are focused at UA, they are models for national implementation.

Social and Behavioral Aspects of the iPlant Collaborative

It is important to acknowledge that not all plant scientists are familiar with emerging technology tools that can support and enhance their work. Collaborative tools associated with virtual meetings, blogs, WIKIs and discussion forums are examples of technology applications that can support collaborative work over geographic distances. To achieve the goal of facilitating interaction among plant scientists, it is essential that the iPC design our software systems with usability in mind, and that we actively reach out to the community to promote the adoption of the tools. In order to assess whether we have achieved our goals, we will conduct social network analysis to quantitatively assess the degree to which the system has contributed to new ways of collaborating among plant scientists.

Usability has been defined in a variety of ways, but essentially it focuses on the ability of users to engage with a system in order to achieve desired outcomes effectively and efficiently, and ultimately to be satisfied with the experience. Usability assessments are useful both during the development of the system and as a way to enhance adoption (e.g., Agarwal and Venkatesh 2002). For example, during development, a set of potential users can be recruited to evaluate the system – the results can be rapidly fed back to the development team so that adjustments can be made to improve the system.

Because the value of a collaborative technology increases in proportion to the number of users, it is important to achieve a critical mass of users as quickly as possible. We will use technology adoption and diffusion life-cycle theories (e.g., Moore 1999; Rogers 1995) to identify biologists who would be early adopters. In addition to their fascination with novelty, early adopters tend to be less critical from a usability perspective because they can envision what is possible and tend to provide valuable information regarding system enhancements that will ultimately have a positive influence on later adopters. Typically, early adopters are also the information providers for later adopters. Thus, once a system is successfully established with these early adopters, others will follow.

In addition to usability, many other elements contribute to system adoption, including social influences, compatibility, and external incentives. If a system is compatible with the way work is done, individuals are more likely to adopt it. However, we recognize that the iPC's cyberinfrastructure is likely to change how work is done by plant biologists. Thus, the influence of peers and incentives will be necessary to increase adoption. To address this, information scientists, biologists and social scientists will work closely to develop mechanisms that work best in biology.

Social networking analysis involves understanding the current social network via surveys and co-citation analyses, monitoring its evolution over the course of the project, and leveraging it to enhance adoption and diffusion. The social scientists involved in this project will examine the existing social networks among biologists using social network analysis and bibliometric techniques. Consistent with the suggestions of Berman and Brady (2005), these networks will be monitored over time to understand how and why the community of users changes with system use. This process will include developing metrics to evaluate the size and quality of existing social networks for research collaborations and examining how these collaborations change and evolve over time as more tools are developed. Such an evaluation will also point out the types of (new) tools that may need to be developed to enable new types of collaborations. Finally, the evaluation can be used to identify areas of support that are needed so that collaboration will continue, even after funding has ceased.

In summary, the purpose of the social science arm of the project is two-fold. First, social science research can be leveraged to enhance development and increase adoption and diffusion of the plant science cyberinfrastructure. Second, the plant science cyberinfrastructure collaborative provides a unique opportunity to collect data in order to study how a collaborative such as this evolves and, more importantly, sustains itself over time.

External Evaluation of the iPlant Collaborative. The evaluation of the iPC will be conducted by East Main Educational Consulting, LLC of Wilmington, NC. EMEC has drafted an evaluation plan that is responsive to the needs of the project team and the constraints set by the funding agency (Stake, 1975), and which has the built-in flexibility to adjust the plan as the project evolves, while maintaining its focus on communication with project stakeholders. The plan includes a mix of quantitative and qualitative tools and methods which together allow for cross-checking and triangulation of measurements, thereby providing a more accurate insight into the intricacies of the project than any single measurement technique alone (Lincoln & Guba, 1985). The Evaluation Team will work closely with the Social Networking Team, integrating evaluation and research efforts to increase efficiency and reduce redundant data collection.

The primary framework for evaluation of the iPC will follow an outcomes measurement scheme (United Way of America, 1996) that focuses the evaluation on how the project “makes a difference in the lives of people” (p.4). This framework also provides the opportunity to use the data to strengthen existing services, target effective services for expansion, identify participant and team training needs, justify budgets, prepare long-range plans, and focus stakeholder’s attention on programmatic issues (p.5). Outcomes, indicators, data sources, and data collection methods selected for the evaluation of the iPC have been drafted but are not included due to space limitations.

Formative and Summative Reports. The formative evaluation process will begin during the iPC planning phase and will continue through to the completion of the project in 2018. The ongoing data collection and analysis along with comparisons to baseline data will inform the evaluators and project team of the successes and shortcomings of the project. We will prepare annual reports for project personnel and the funding agency. If the evaluation data warrant, the project team may modify the project plan in order to achieve the iPC goals, in which case we may need to make compensatory changes to the evaluation plan itself. However, all changes will be recorded with accompanying rationale. When necessary, we may produce interim reports to inform the project team of particular project activities that require immediate attention or alteration.

The final, summative phase of the evaluation will occur during and beyond the final months of the project. We will use this time to collect the final datasets and begin analysis of the project as a whole. Similar to the formative report, we will evaluate the progress toward each requirement, and will measure how closely the center came to achieving its goals and meeting its stated deliverables. All data will be aggregated so that commentary on the overall successes and shortcomings can be reported. We will include best practices and lessons learned in the summative report, along with responses to key questions that develop over the course of the implementation. The summative evaluation will be provided to the project team and funding agencies in the form of a written report with electronic copies of annual and formative reports, datasets, and analyses.

Prior Results of NSF Funding (directly relevant to this proposal)

Lincoln Stein: NSF award #0321685; VCA Gramene: A Platform for Comparative Cereal Genomics; www.gramene.org) is a comprehensive web-based community database of cereal genomes, genetic and physical maps, and genetic variants. Gramene is synergistic with the Plant Ontology project (NSF award #0321666; www.plantontology.org), a community-managed ontology of terms describing the developmental stages and anatomic structures of flowering plants that is used to semantically integrate the properties of genes, pathways, mutants and cultivars across multiple independently-managed databases. A core source of information for Gramene is “The Oryza Map Alignment Project”: NSF award #DBI-0321678; www.omap.org), a project to develop and align the physical maps of 11 wild rice species, shedding light on the evolutionary history of this genus (Jaiswal et al 2006, Ware et al. 2002a, Ware et al. 2002b; Ilic et al. 2007; Pujar e al. 2006). [Two other projects in the Stein lab synergistic with the mission of PSCIC: the Generic Model Organism Database Project (funded by the USDA, NIH, NSF and DOE), a collaboration among roughly 30

model organism databases to make open source software, ontologies, file formats and protocols available to biological researchers. It uses a modular relational database schema capable of representing a wide variety of biological data types (Drysdale and Crosby 2005) as the foundation for a series of visualization engines, editor tools, analysis tools, workflow management systems, and web site display tools (Stein et al. 2002; www.gmod.org). Stein is also the principle designer of the Distributed Genome Annotation System (NIH), an XML-based system for exchanging and comparing annotations among genome databases (Dowell et al. 2001; Eilbeck et al. 2006).

Sudha Ram: NSF #IIS0455993: Investigating Data Provenance in the Context of New Product Design and Development. This project has resulted in an ontological model of provenance called the **W7 model** that captures the semantics of data provenance as a combination of seven interconnected elements (Ram & Liu, 2006a-b). We have implemented a software system based on this model and have tested it in various application domains (Ram & Liu, 2006c, 2007). In coordination with the UA Bio Computing Laboratory, we are using this system to harvest provenance for biological images (Ram et. al., 2005) The harvested provenance permits biologists to interpret the images in context and also to replicate and validate the procedures that created or processed the images. In other synergistic research, Ram's Advanced Database Research Group has developed heterogeneous database integration ontologies (Ram & Park, 2004), a software system for integrating heterogeneous databases (Park & Ram, 2004), a semantic model for biological sequence data and 3D protein structures (Ram & Wei, 2003, 2004a), integration techniques for biological data (Ram, 2005) and mechanisms for identifying and labeling the links among multiple biological database (Ram & Wei, 2004b, 2005).

Gregory Andrews was PI of NSF grant EIA-0080123, CISE Research Infrastructure, Optimization of Distributed and Networked Systems: A Spectrum of Techniques, \$1,004, 979. The project explored a variety of complementary techniques for optimizing distributed and networked systems—from client interfaces, through middleware and servers, to the communication infrastructure—and a variety of optimization criteria—including time, space, power consumption, and quality of service. The grant supported the acquisition of major computing infrastructure, including a high-performance cluster, shared-memory multiprocessors, graphics and visualization processors and displays, a mobile/wireless computing laboratory, and storage and networking upgrades. Research projects that used the infrastructure produced 53 journal publications and 90 selective conference publications during the period of the grant.

Rich Jorgensen/Vicki Chandler: NSF DBI #9975930. Functional Genomics of Chromatin: Global Control of Plant Gene Expression \$10.5M: R. Jorgensen; co-PI: V. Chandler (+8 Co-PIs). Renewal: DBI#0421619. Functional Genomics of Maize Chromatin Genes \$6.6M; 4 yrs: 10/04 – 9/08; K. Cone, PI; coPI, V Chandler (+6 Co-PIs). The project identified and functionally characterized hundreds of genes in maize and Arabidopsis contributing to chromatin-based control of gene expression; 19 publications from these grants to date. An outcome of this project was the Plant Chromatin Database (ChromDB; www.chromdb.org), which led to award # DBI-0421679 (PI Napoli, co-PI Jorgensen; publications: Tuskan et al. 2006; Palenik et al. 2007). NSF DBI #0321663. Microarray resources for maize research \$3.66M; 3 years 10/03-9/06; PI: V. Chandler. 3 Co-PIs. Project provided >3500 maize oligonucleotide arrays to the community, performed >300 hybridizations for the community and developed Zeamage relational database that stores and disseminates expression data generated by project participants and array users (Gardiner et al. 2005).