

GENERAL QUESTIONS

(1) *Will the community fully embrace the Collaborative as its own?*

We are fully committed to making the Collaborative truly by, for, and of the community. We feel we will win community buy-in because the community itself will control the Collaborative and will make the key decisions regarding priorities and allocation of resources through the Board of Directors. The Collaborative will act as an open, service-oriented organization that exists to facilitate and enable the community's priorities for interdisciplinary, collaborative research, training, and outreach. Participation will be driven by the new opportunities created through the Collaborative for interdisciplinary advances in the plant, computer, information, social, and cyberinfrastructure sciences. Education, outreach and community building activities are targeted broadly across all education and research levels and will bring understanding of the Collaborative broadly to the community.

(2) *Will the Collaborative be successful in truly integrating plant, computer, information, social, and cyberinfrastructure sciences?*

The Collaborative's innovative Discovery Environment development process brings together researchers from the biological sciences, computer and information science and engineering (CISE), social sciences and statistics to work collaboratively on Grand Challenge problems. Solving these problems will require equal participation and sharing of expertise across disciplines and a community oversight process will ensure that this happens. Critically, from the outset, Grand Challenge problems will be identified, defined and analyzed *collaboratively* by plant, CISE, social, and statistical scientists.

(3) *Will the cyberinfrastructure innovations that are proposed be effective in meeting all of the goals of the Collaborative?*

The Collaborative's cyberinfrastructure innovations will be effective because a) they are primarily driven by community needs, b) they will be continually evaluated and improved through feedback and evaluation, c) we will use a contagion model for ensuring adoption, diffusion and absorption of the Collaborative through and into the community and d) research will be extensively interdisciplinary and will be integrated closely with training, education and outreach at all education levels.

(4) *Will the Collaborative be successful in adapting to a constantly changing technology landscape?*

The collaborative will be successful in adapting to a constantly changing technology landscape because of its open nature and responsiveness to community input and feedback, and constant evaluation of new technologies via a deliberate environmental scanning program. Substantial resources remain to be allocated throughout the project. By not making commitments to a large number of PI's at the outset, the Collaborative is free to use its resources wherever and whenever the new opportunities arise and/or the community wishes they be invested.

(5) *Will the structure and management of the Collaborative allow it to be an integrated resource that is more than just the sum of its parts?*

The structure and management of the Collaborative will make it more than the sum of its parts because of its unprecedented integration of community-building, infrastructure creation, and education. We help build diverse, interdisciplinary communities that create infrastructural tools that are in turn used to educate new community members in a self-propagating benevolent cycle. As the collaborative evolves, expertise and strategies for interdisciplinary communication will diffuse and propagate across the community.

SPECIFIC QUESTIONS

1) Vision and rationale for the proposed Collaborative.

1a. What are the measures of success for the Collaborative and how will these be evaluated?

The overall vision for the iPlant Collaborative (iPC) is to provide a virtual and physical environment where interdisciplinary teams can coalesce, form, and conduct scientific discourse across disciplines and collaborate to develop solutions to Grand Challenge questions in both plant biology and CISE. The iPC will enable new ways to explore biological questions via the deployment of Discovery Environments, custom designed to address a given problem and to engage the broader community including researchers, educators, and students at all levels. The iPC will be predominantly service oriented, providing infrastructure, services, and personnel to the community, which thereby will be enabled to address Grand Challenges that were formerly inaccessible. The community will control the iPC and in time will completely absorb it.

The project will be an unqualified success if it is able to:

- 1) *Generate a community of plant scientists, computer and information scientists, and social scientists that understands the nature of computational thinking in biology and values the importance of inter-disciplinary collaborations.* We will evaluate the success of this aim by measuring the extent of collaboration among the disciplines using metrics based on numbers of collaborative grant proposals submitted, collaborative publications accepted, and interdisciplinary graduate students and postdocs trained.
- 2) *Achieve robust community usage of the cyberinfrastructure developed by the iPC for novel research in plant biology.* Currently only a few elite plant researchers have access to sophisticated information management tools. The iPC aims to level the playing field by bringing these facilities, free of cost, to all plant biologists, regardless of their position and means. We will evaluate our success in this endeavor via usage statistics, user surveys, and citations, in order to precisely define who is using the infrastructure and how they are using it.
- 3) *Make measurable progress towards solving grand challenge problems in plant biology and CISE.* The Discovery Environments are developed in collaboration with goal-directed, community-organized Grand Challenge Teams. Ideally, this work should spur real progress in unraveling some of the mysteries of plant biology. We will evaluate our success in this goal by monitoring publications, awards, and other measures of academic success.
- 4) *Help plant biologists everywhere expand their teaching, training and educational outreach efforts to incorporate computational thinking in biology.* We will evaluate this by measuring number of teachers trained, and numbers of students exposed to computational thinking via physical and virtual classes, tutorials and workshops.
- 5) *Educate lay people in the plant sciences.* Our broadest goal is to increase the level of science knowledge among the general public, which we will achieve through online educational materials, K-12 curriculum-development programs and simplified versions of the Discovery Environments. We will measure success in this aspect by monitoring discussion of the project in the general press, online forums such as blogs, and the adoption of iPC-generated curricula in schools.

More specific information on how we will evaluate our success in these endeavors is provided in the detailed answers to site visit questions that appear later in this document.

1b. How will the proposed rationale, goals, and activities establish and sustain the vision of the proposed Collaborative?

To address goals 1, 2 and 3 (interdisciplinary community building and research), we have developed a community-driven process in which self-organized representatives of the biology, CS, IS and social science communities create Grand Challenge teams directed towards the solutions of significant problems in the plant sciences. These teams join with members of the iPC to design, develop and implement Discovery Environments that provide the information infrastructure needed to support the teams' research. The DEs ultimately become open, online resources for the broader community, who participate in the infrastructure as users, developers, data producers, curators, and integrators. This process is described in detail in the responses to questions 2-5.

To address goals 4 and 5 (education, training and outreach), we will embark on an innovative programme of curriculum development, outreach and training designed to bring the concepts of computational thinking in the plant sciences to students at all levels, and to raise the general level of science knowledge among motivated lay people. Details of this program are provided in our responses to questions 8 and 9.

Finally, we have deliberately created a process and an infrastructure that is self-sustaining. All components of the iPC infrastructure are open source and open access, meaning that they can be taken up, copied, modified and redistributed by anyone. It is our hope that this aspect of the infrastructure, coupled with the lasting effects of our community-building, educational efforts, and outreach, will long outlive iPC itself.

1c. If funded, what will this Collaborative enable that could not be accomplished by existing organizations, resources, and mechanisms? What features of the Collaborative are truly innovative?

We feel that our approach to building an interdisciplinary community based on a common infrastructure is unique in a large number of ways.

- 1) *It is truly interdisciplinary.* Nearly all existing laboratories are focused on a single research discipline or a small cluster of related disciplines. Our approach is truly interdisciplinary in that it integrates the different disciplines of biology, computer science, information systems, statistics, and the social sciences from the outset. It is driven by the needs of biologists to define not only the Grand Challenge Questions but also to articulate the new tools and techniques that will need to be designed and developed to solve these challenges.
- 2) *The community sets the agenda.* iPC staff do not select grand challenge questions or otherwise dictate research goals. A deliberate process of community engagement brings representatives of different research disciplines together to self-organize themselves into Grand Challenge teams with their own projects and research goals. These teams engage the iPC to design and develop Discovery Environments that address their information management needs, and, by extension, the needs of the broader community.
- 3) *Pervasive use of Web 2.0 technologies for broad community participation:* The Discovery Environments encourage users to contribute their own data sets, which are then integrated into a whole that is greater than the sum of its parts. This helps realize our vision of a community-driven infrastructure that is owned by the community, not by the iPC.
- 4) *The information infrastructure is designed by academics, but implemented by professionals using proven software-development methodologies.* Most academic software is implemented in an *ad hoc* way that lacks generality, is hard to port, and is poorly documented. We propose a two-step design of the DEs in which academic participants in the GC teams and the iPC create prototype software by following the fundamental principle of "systems analysis and design cycle" --

understanding the needs of end users, developing a blueprint for the design to support these requirements, prototyping tools, and testing and modifying them in response to user feedback. These prototypes are then passed on to a professional software engineering team for implementation as hardened, reliable, well-documented, stable software.

- 5) *The research & infrastructure development parts of the project are integrated with its educational goals:* The Discovery Environments will be integrated into the curricula and other teaching materials that we develop, creating a continuous connection between students, teachers and researchers that we feel will be much more effective than ghettoized “dumbed-down” online versions of these materials.
- 6) *Outreach based on sociological models of technology adoption and diffusion.* Our outreach efforts are informed by sociological models of technology adoption and diffusion that are widely used by marketing teams in the high-tech industry. These models allow us to target our outreach efforts to specific subpopulations of the research community using techniques that are most likely to be effective for that subpopulation.
- 7) *Integration of software tools and datasets:* The Discovery Environments focus on the seamless integration of data sets and data analysis tools, thereby merging the ideas of databases and data merging. Most biological software products focus on one or the other.
- 8) *Preparedness for the future:* Because technology is changing at a lightning place, our project has a dedicated “future shock” group whose responsibility is to continuously scan the horizon for promising technologies, tools and techniques and decide on which ones to explore further.
- 9) *Constant evaluation by an independent group:* An independent team of social scientists will be on hand throughout the project to evaluate the project using explicit metrics and to provide feedback to the project management. This allows us to monitor the effectiveness of our project in real time and respond in an agile fashion.

1d. What are the specific characteristics, features, and capabilities of the proposed Discovery Environments?

The Discovery Environments are shared, open source, web-based software platforms that address the grand challenge questions in plant sciences by allowing community members to integrate, visualize, analyze and annotate large-scale datasets and associated computational models.

Specifically, the DEs are designed around the idea of the “mashup,” a system that provides a central data type whose common coordinate system acts as the framework for integrating other raw and computationally-derived information. In the case of genomic data, for example, the coordinate system is chromosomal position, all other datasets are layered on top of this common coordinate system, ranging from raw data such as mRNA transcription levels to highly-processed computational data such as binding site free energies. In other cases, the common coordinate systems will be geographic information, taxonomic trees, 3D anatomic maps, biological pathway graphs, or more abstract coordinate systems such as ontologies. The use of mashups is critical to the iPC’s integrative, cross-disciplinary, community-building goals, because mashups provide a powerful and intuitive way for diverse community members to combine their work and to see relationships among them.

DEs can be annotated, meaning that they encourage community members to use them as the building blocks for larger, integrative works. An annotation can be as simple as a one-sentence note attached to a data object, or as complex as an interactive review article which draws on graphics, computational models, and database tables from the DE as the figures and tables of the article. One very attractive model for this type of annotation is the WIKI; we envision joining open source community efforts to develop structured WIKIs in which free text is seamlessly merged with structured information.

DEs are user-friendly front ends to both databases and computational processes in a way that blurs the

distinction between the two. Users will be able to extract information from a variety of large local and remote datasets, and to access compute services including the type of very large-scale simulations that require tera- and peta-grid computing. Behind the scenes, we will use web services for DE extensibility, and strict provenance, versioning and semantic consistency checking to ensure the accuracy of what is presented to the user. The DE front ends will use the latest web technologies: including AJAX and the semantic web.

DEs are community-driven. We will design DEs within the context of the needs and requirements of specific grand challenge teams which are themselves composed of community members. As described in the timeline for DE development (see question 6), GCT community members participate in key ways throughout the design, implementation and testing of DEs. The exception to this rule is the iPlant “killer app” described in the answer to question (4a), on which we will begin work during the first year, before grand challenge teams are fully formed.

DEs are challenges for computer science and software engineering. The technology for integrating data and computational services across a diversity of platforms, protocols and formats is very much under development, and the theoretical frameworks for tracking provenance, managing semantic integration, and managing version changes, is a cutting-edge computer science research problem.

Finally, DEs are open. The APIs underlying the user-friendly front ends will be exposed to the world so that software developers, computer scientists and other computer-savvy users can access them directly. All DEs that we develop will be available for reuse, modification and redistribution under open source terms, thereby encouraging the community to take up and maintain the software long after the iPlant Collaborative has dissolved. Openness also means that the DEs will reuse existing technology; whenever possible we will make use of existing open source software packages. This allows us to leverage the excellent work already done by software developers both within and outside the biological community and to use our time to develop novel applications rather than reinventing wheels.

2) Grand Challenge identification processes to be used by the Collaborative

2a. For each of the proposed phases in the proposed grand challenge process, describe what activities will be undertaken and how these will proceed, who will be involved, who is responsible for organization/management, and who will make decisions. What are the respective roles of students and postdocs and of senior project personnel?

The grand challenge process aims to continuously engage the community in a process that leads to the development of a stable, community-supported, cross-disciplinary cyberinfrastructure. The means to this end are the Grand Challenge teams, which foster cross-disciplinary community connections, and the Discovery Environments, the software basis of the cyberinfrastructure.

Phase I: The Community Selects the iPC's Board of Directors

To ensure total control of the iPC by the community we propose a 'bootstrapping' process for forming the Board of Directors (BoD). The first step is to form a Nominating Committee comprised broadly of a small number of respected experts in the plant science and computing/information science communities. The NSF, working with the Collaborative Director and co-PIs (and/or members of the community selected by NSF), will consult with the community in the selection of the Nominating Committee. The Nominating Committee will then widely solicit nominations for the iPC's Board of Directors. Nominating committee members agree in advance not to be considered. The nominating committee then creates the Board in consultation with NSF and the Collaborative Director (both of whom will have to work closely with the BoD). The BoD selects its own chair, subject to NSF approval. The members of the BoD will have staggered terms, to ensure both continuity and renewal. The process should be complete by the end of September, 2007.

Phase II: Conference: "Grand Challenges"

Beginning in October, 2007, the BoD, working with the iPC's Executive Committee and assisted by the Director of Community Responsibilities and her/his staff, will widely solicit brief "white papers" from the broad community (including plant, computer, and information sciences and engineering) that propose topics, speakers, and workshops for the "kick-off" Conference. The BoD, assisted by the iPC, determines the agenda and chooses the invited speakers for the Conference, to be held in the first or second month of the project (early 2008; staff will be hired during the fall of 2007).

The BoD will also choose the venue. Two possibilities are Cold Spring Harbor Lab's Grace Auditorium and the Biosphere II conference facility in Oracle, AZ. We suggest that the first conference be held at CSHL to reduce the perception that the iPC is "that thing in Arizona." Follow-up conferences will be held annually in the second and subsequent years, with the venue rotating between CSHL, Arizona, and other locations selected by the BoD.

The iPC core staff, working under the direction of the Director of Community Responsibilities, will manage the logistics for the kick-off conference. However, the content of the program, such as the sessions and invited speakers, will be handled by the BoD in consultation with the community, assisted by staff as needed. Participation of iPC faculty, students, postdocs and staff would be extremely valuable, but would be subject to BoD approval to ensure an appropriate balance relative to the community they are serving. The initial recommended focus areas for the conference can be found in the Project Description. Information regarding these conferences will be widely distributed and advertised to ensure broad and diverse participation, including marker papers in major biology and CISE journals and through the professional societies.

Phase III: Focused Grand Challenge Symposia

The kick-off conference will be followed by 6 to 12 focused Grand Challenge Symposia held during the first year. These symposia have three main goals:

1. To discuss and define grand challenge problem(s), identify the data and tools currently available, and to outline the research needed to solve the problem.
2. To identify the cyberinfrastructure tools and research approaches that are needed to solve this problem.
3. To establish the nuclei of Grand Challenge Teams, cross-disciplinary community groups that will work with the iPC to design and develop the Discovery Environments that will provide the informatics infrastructure for their research.

GC symposia will be community-generated and driven, with iPC staff facilitating, but not directing. Following the kick-off conference, we anticipate that multiple groups will answer the call to organize GC symposia on topics of their choice. These self-organized groups will be asked to prepare short proposals describing the proposed symposia. The proposals will then be evaluated and selected by the BoD. The organizers of a given meeting will determine the agenda, participants, schedule and location, modified by the BoD as appropriate to ensure broad community and interdisciplinary participation.

It is possible that we will not receive sufficient proposals to fill out the target minimum of six symposia, or that the proposals we receive will be skewed towards a particular topic. In this case, the BoD will issue RFAs for GC symposia in particular areas in order to ensure that the symposia cover the full breadth of plant biology.

All symposia organizers will be encouraged to meet with project staff and faculty before the symposium begins, and to meet again after it is over to discuss community needs identified over the course of the meeting. Some symposia organizers may also wish to have iPC staff or faculty make a public presentation during the meeting in order to orient attendees to iPC resources and capabilities. All participants will be expected to provide substantive feedback for the community and iPC to evaluate the success of the meeting. All symposia organizers will be required to submit a final report one month after the meeting.

Phase IV: Formation of Grand Challenge Teams

The intent of the GC symposia is that they catalyze the formation of Grand Challenge Teams. The key stated deliverable for GC symposium organizers and attendees is a proposal to form one or more GC teams, interdisciplinary research collaborations that benefit from, and contribute to, the iPC infrastructure. The incentive to form a GC team is that community participants in the team receive “in kind” support from the iPC in the form of information systems management, the development of Discovery Environments tailored to their needs, and collaborative interactions with iPC faculty who are in the top academic tier of computer scientists, biostatisticians, software engineers and plant scientists. Community participants can also use the involvement of the iPC as leverage to obtain research grants to support the other aspects of their science. Rather than investing large amounts of money and effort into developing information systems, GC team project participants can focus on the science.

All GCT proposals that we receive will be evaluated by the BoD, in consultation with external peer reviewers. Criteria for review will include feasibility, scientific merit, innovation, potential for cross-disciplinary interactions, potential to advance “computational thinking,” and whether it will benefit from

cyberinfrastructure building. By the latter, we mean that if a proposed project requires several years of data gathering before any information processing or software development is needed, then it might be better to defer it in favor of another project that has more immediate information-management needs.

Each GCT will be led by a lead principal investigator chosen from participating community members, and will be assisted by a collaborating IST faculty member acting as an assistant principal investigator. As described in greater detail in the management plan, the lead principal investigators from each of the GCTs will become a member of the GCT Oversight Committee, whose responsibility is to monitor the activities of all the GCT projects.

We recognize that not every GC symposia will catalyze the formation of a Grand Challenge Team proposal. Symposium participants may decide, for instance, that the time is not yet ripe to attack the proposed problem in a concerted fashion. Even in this case, however, the symposium will have been valuable as a community building activity, because it brought together diverse experts across disciplines to discuss difficult problems and potential solutions. Participants in such symposia are likely to form new collaborations, develop new projects and new tools, and create new focused communities centered around specific, future grand challenge problems.

Phase V: Engagement between iPC and GC Teams to Develop Discovery Environments

The fifth phase of the process is engagement between members of the IST staff and GC teams to design and develop the Discovery Environments, work out iPC-facilitated data mining and data integration tasks, develop novel algorithms, and create other information resources that will form the information infrastructure for the GC teams' research. This process is described in detail in the answers to question 3.

Phase VI: Community Education and Outreach

The final and longest phase of the Grand Challenge Process is reaching out to the community to increase participation in the GC team projects, to encourage members of multiple communities to use the facilities provided by the DEs in their research and teaching work, and to educate students and motivated laypeople in the ways of computational thinking. This process is described in detail in the responses to questions 4, 5, 8 and 9.

2b. What are the expected outcomes/deliverables and the measures for success for the grand challenge process and how these will be evaluated?

The primary outcome of the Grand Challenge Identification Process is the formation community-driven dynamic communicative and focused plans and teams to attack significant GC questions in plant biology and CISE. The outcomes and deliverables below are based on the phases described in the previous question. For organizational purposes, the information is presented in tabular format. The outcomes will be evaluated based on the indicators of success using the evaluation methods previously described. This includes, aligning the outcome with the evaluation matrix, collecting data from specific project populations using methods such as surveys and document analysis, analyzing the data, and reporting the results in formative reports to the project management.

Outcome	Indicators of Success
Phase I: Community selects iPC's BoD	<ul style="list-style-type: none"> ▪ Formation of BoD follows 'bootstrapping' process described in project documents
Phase II: Conferences (kick-off/annual) are organized and held	<ul style="list-style-type: none"> ▪ White paper solicitation attracts 25 proposals ▪ Locations are chosen and secured

	<ul style="list-style-type: none"> ▪ 150 in-person attendees, 250 web-based attendees (kick-off) ▪ 200+ participants attend annual conference
Phase III: 6-12 GC Symposia held during first year	<ul style="list-style-type: none"> ▪ 10 proposals from self-forming groups ▪ 6-12 symposia organized and meet
Phase IV: GCTs form	<ul style="list-style-type: none"> ▪ 2 Symposia groups form GCTs
Phase V: Development of DEs by GC teams and iPC staff	<ul style="list-style-type: none"> ▪ DEs are designed and developed by IST and GCTs ▪ Novel algorithms are developed ▪ Information resources that form the information infrastructure for the GC teams' research are created
Phase VI: Increased community participation through specified project segments	<ul style="list-style-type: none"> ▪ Participation in GCT projects increases ▪ Community uses DEs in research and teaching ▪ Students and laypeople have greater understanding of computational thinking

3) Cyberinfrastructure (CI) plan

3a) How will the CI resources be linked together into a functional whole?

The nature of the CI resources developed by iPlant are dictated by the needs of the Grand Challenge Teams (GCTs) and the needs of the broader communities of biologists and computer scientists from which the GCTs arise. In practice, CI resources will be accessed via Discovery Environments consisting of:

- Core datasets, either maintained locally or in offsite locations
- Compute services that manipulate the data;
- Shared ontologies that describe the datasets, the compute services, and the relationships among them;
- A low-level API suitable for use by bioinformatics and software developers that provide access to the datasets, compute services and ontologies at the semantic level;
- A high-end visualization and manipulation environment that provides an intuitive front end for Discovery Environment end users.
- Collaborative WIKI-like "mashup" facilities for drawing conclusions from the DE and sharing those conclusions with other members of the community.

Although each Discovery Environment is custom-tailored for its community, the underlying technology will be recycled from one DE to the next. We will use an N-tiered architecture in which remote and local datasets are stored in the format most suitable for the access patterns under which they will be used, typically a relational or object-relational database. The "bus" connecting datasets and compute services to the low-level API will use ontology-aware semantic web technologies such as REST-based access to OWL-DL data streams. The top tier of visualization and collaboration facilities will typically use AJAX-based applications that run on common web browsers.

By providing a software developer's API that gives access to the data at the semantic level, we will facilitate the development of multiple tools that build upon the DE infrastructure. By reusing existing open source technologies and by pursuing an aggressive open source policy ourselves, we will encourage the community to build on top of the iPlant infrastructure.

3b) What is the anticipated user community?

We define three general user groups for iPlant, the "core" constituency, the "broad" constituency, and the "extended" constituency. The core constituency consists of members of the community-organized Grand Challenge Teams, with whom we engage intensively in a goal-directed fashion. The broad constituency consists of the general community of plant biologists, translational researchers (e.g. breeders), computer scientists, software engineers, mathematicians, and graduate and postgraduate students, with whom we engage via the online Discovery Environments. The extended constituency is other interested groups, including the lay public, students at the K-12 and college levels, science writers, lawyers, environmental health workers, and so forth. We will engage the extended community via education and outreach activities, parties, and layperson-directed material available from the DEs.

Our core constituency consists of members of the GCTs. As described in the research plan and in answers to earlier questions, we envision developing 3-4 GCTs during the first year of the project. The GCTs are composed primarily of outside members of the research community, who interact with a few well-chosen iPlant faculty and staff to design and implement the DEs needed to support their grand challenge questions.

Based on our experience with recent grand challenge questions in genomics (the HapMap project www.hapmap.org, and the modENCODE project www.genome.gov/25521166/), we envision GCTs as consisting of teams of 30-40 community members in the following rough distributions:

- 1 lead faculty level plant biologist
- 1 co-lead computer scientist, mathematician, bioinformatician or statistician
- 3-5 additional faculty-level scientists
- 10-15 postdoctoral level scientists
- 10-15 graduate students

The distribution of biologists, computer scientists, mathematicians, bioinformaticians and statisticians will vary from one GC project to the next. In addition, these non-experimental scientists will be interacting with many more experimental scientists; for instance, we have observed that a ratio of 5 bench scientists per "in silico" scientist is typical of many of the omics projects that we have participated in. Thus, our core constituency during the early phases of the iPlant project will be 100-200 researchers, distributed among experimental and non-experimental scientists in a 5:1 ratio.

We will "leaven" each GCT with 4-6 iPlant staff, who will intensively work with community members as project facilitators throughout the course of the GCT project. The leavening will consist of at least one faculty-level bioinformatician, computer scientist or statistician from the iPlant IST team, two full time PhD-level postdocs or CS staff, and 1-3 graduate students who will be cross-trained in experimental biology and computer science. When appropriate, GCT teams will also involve graduate student consultants in statistics brought in through the UA Graduate Interdisciplinary Program in Statistics and through our partnership with the Purdue STATCOM program (see question 10). These facilitators will be assisted by a variable number of professional software engineering staff members from the iPlant infrastructure core.

The backgrounds and expectations of the core constituency will vary considerably from computer-savvy and biology-naive to biology-savvy and computer-naive. This is unavoidable -- and probably desirable -- in an interdisciplinary project of this scope. The DEs that we develop will have multiple layers of access to satisfy many levels of expertise. Team members with extensive software development experience will be able to access the DEs using the low-level API, while the experimental biologists will most likely interact with the system via the user-friendly front ends.

Our hardest task will be to manage team member's expectations by sharing realistic estimates of the time it takes to identify and cleanse data sets, assemble the infrastructure for their semantic integration, and creating stable and compelling user interfaces.

Beyond our core constituency is the broad general community of plant biologists, interested computer scientists, mathematicians and software engineers, who will be encouraged to participate in iPlant as users and contributors to the DEs as described in the answers to section 4. The DEs will be open to the general community at both the front end and API levels. General community members will be welcome to browse through the datasets made available through the DEs, annotate and integrate information via the DE WIKI and mashup facilities, and contribute their own primary or derived data. Community members who become strongly engaged with one of the DEs will be encouraged to become core participants in the corresponding GCT.

There are more than 6,000 registered members of the American Society of Plant Biologists (www.aspb.org/publicaffairs/news/aspbpr.cfm), but our general constituency is potentially much larger than that, encompassing taxonomists, ecologists, and plant breeders, as well as the communities of plant biologists outside of the United States.

Our potential broad constituency also includes computer scientists, software engineers, and mathematicians. To engage computer scientists, the DEs will include materials that describe why the problem is interesting from an algorithmic standpoint. For software engineers, we will provide pointers to the source code, installation documentation, and hints on how to repurpose the DE infrastructure for their own needs.

For lay people, K-12 and college students, and other members of the extended constituency, the iPlant staff, working collaboratively with GCT community members, will incorporate education and outreach material into the DEs by providing them with background reading on the problem toward which the DE is directed, the social and scientific significance of the problem, and the techniques being used to address the problem. This will be supplemented with curriculum development, iPAT outreach, and the teacher training activities described later in this document. Together, this background material will help maintain a high level of transparency for iPlant, engage the community, and provide the broader community with a sense of inclusiveness and ownership.

3c. How will users interface with the Collaborative, including access, application for resources, and technical assistance?

Access. Most users will access iPlant resources via Discovery Environments, as described in the answer to question 3a. There are two levels of access: 1) interactive access, which takes place via the web portal, mash-up and WIKI tools, and the visualization tools provided; and 2) “expert” access, which takes place through the low-level Application Programming Interfaces (APIs) and Service Oriented Architecture (SOA) to allow direct and automated access to data and computational resources.

Application for Resources. Many iPlant resources will be directly accessible via Web interfaces. For "hardware" resources, such as computational cycles, we will follow the model that has been successfully used for many years at the NSF-funded supercomputer centers. In particular, any affiliated user can apply for a basic level of access to storage and computing resources. Users and GC projects with more significant needs will submit proposals describing the research they will do and the resources they require. These will be reviewed by a resource allocation subcommittee appointed by the BoD. Users with significant computational needs that cannot be fulfilled directly by the iPC will be directed to an appropriate national facility such as the Teragrid, NPACI at San Diego, or NCSA at Illinois. Limited number of life scientists at present make use of supercomputing facilities, but this is likely to change over time. The iPlant collaborative will provide a stepping stone for plant scientists by providing a "medium scale" set of resources on which they can develop experiments and simulations, thereby preparing them to write competitive applications for more significant computational resources when they become required.

Technical Assistance. For “support” interactions, we will provide online and phone access to a help desk facility. We will use a Customer Relations Management (CRM) framework or “ticketing” system, such as the open source RT system, to manage this process. The Collaborative will implement a suitable system similar to the successful ticketing system used in the NSF Teragrid. The open source RT (request tracking) software will form the core of this system.

The support process is initiated with either an e-mail or a phone call from a user. This creates a new ticket in the system that is then delegated to the most appropriate domain expert on the team. The ticketing software assigns a case number, tracks all interactions with the user, and logs the process to a database. Using the ticketing software, we will regularly review the open ticket queue in order to escalate any long-open tickets to the appropriate supervisory personnel and to identify any persistent problems that need to be addressed systemically. These records, along with summary information such

as average problem resolution time, will be provided as part of regular reporting to NSF to measure how effectively Collaborative staff are responding to the needs of the user community.

The Arizona State collaborators on this proposal have experience in running responsive customer support operations with the RT software, and will be running a distributed phone support operations for the “Ranger” system on the Teragrid supporting a broad national community. The University of Arizona Biotechnology Computing Facility (BCF) is well versed and equipped to assist life scientists with support requests arising during usage of computational resources and analysis tools; providing hands on assistance utilizing cross platform remote desktop sharing tools for problem resolution that involve complex GUI (Graphical User Interface) interactions.

While e-mail support will be available 24/7, we don’t feel that providing 24/7 human staff at a help desk justifies the significant expense that would be incurred. Therefore, phone support would likely be available during normal business hours only. For e-mail support, initial responses will be provided to users within one business day.

The “first responder” responsibilities for the virtual helpdesk will be shared by all iPC sites to reduce total cost. Independent of where the ticketing system resides, iPlant staff at all sites will respond to tickets, and the “on-call” duty can also be distributed when necessary via call-forwarding. All sites (UA, ASU, CSHL) will also participate in supporting users. However, the community will see a single e-mail address and 800 number for all support requests.

3d. How will the Collaborative anticipate and adapt to new technologies, concepts, strategies, and user expectations for cyberinfrastructure in the future?

The iPC will address these issues in two ways. First, core staff at all sites will keep track of technology changes as they are doing now at UA, ASU, and CSHL. This includes hardware trends and software developments. We have also been in contact with external groups such as Google and LionShare at Penn State, as described in Appendix A3 of the proposal. Second, upgrades and expansions to the core infrastructure will be driven by the needs of the community, in particular the needs of the Grand Challenge teams. One of the outcomes of each Grand Challenge workshop will be a specification of the infrastructure required to tackle a specific GC problem. This includes both hardware requirements (computing cycles, data storage, visualization, and networking) as well as software requirements, in particular building the appropriate Discovery Environment. Decisions on acquisitions and allocations of personnel will be driven by the needs of GC teams, and they will be made after review and prioritization by the Board of Directors, as described in Appendix A2 of the proposal.

The iPC has taken a deliberate strategy of not making a large investment in “big iron”. Rather, we have chosen to leverage the computation, storage, and visualization resources available at the partner campuses, as well as such national resources as the Teragrid. There are several advantages to this approach. Tremendous capabilities are made available to the iPC, without the project having to bear the full cost for the deployment and maintenance of these services. More importantly, this approach allows the Collaborative to remain nimble, without large “sunk costs” in a single technology platform. By reducing the investment in iron, the iPC’s cyberinfrastructure resources will instead focus on the *people* required to adapt to ever-changing technology targets. As the deployed technologies available both nationally and as the partner institutions evolve, the iPC will have the flexibility to evolve with them.

Other groups are better equipped to develop optimized and possibly cutting edge creative solutions. Via our partnership process (see question 10) we will connect with them in order to bring our cutting edge needs to their attention. This can be facilitated by stressing the importance of maintaining relationships with peer projects throughout the Discovery Environment design and development process. An additional advantaging of leveraging and building on top of existing institutional resources is that we can benefit from the

technology refresh cycle that is already built into that infrastructure. In addition to providing hardware refresh, the significant cost of training, setup, integration, and transitioning is not a burden on the collaborative.

Our iterative and agile approach to software development and deployment will allow the collaborative to integrate and adopt novel methods, algorithms and user requirements as they evolve. The software architecture and underlying computational infrastructure is intended to be service oriented, affording the ability to swap components and modules as more suitable and optimized solutions are identified. For instance, the scientific community is pushing the frontiers for application-specific, CPU-assisted reconfigurable computing in order to accelerate compute-intensive applications by several orders of magnitude compared to traditional processor-based architectures. Reconfigurable computing enables the end user to exploit both hardware and software levels of optimizations. The flexible nature of this emerging technology affords the possibility of meta-architecture; "morphing" hardware configurations with software as needed, improving efficiency, robustness, security and capabilities on-the-fly. The active involvement of the community at various levels in the DE and formulation of GC questions will ensure that user needs are addressed not just during the initial stages but also as the application evolves. Supporting data from the request tracking system coupled with community feedback and direction from the social scientists and evaluators will play a key role in anticipating and addressing user expectations.

3e. How will the cyberinfrastructure resources be maintained over time?

This question can be considered in three parts: 1) How will we keep facilities operational in spite of failures? 2) How will we keep facilities current/state of the art? 3) Or how will we keep the infrastructure operating after the grant expires? All aspects are important. We answered the second interpretation above in 3d. Below we consider ongoing maintenance and future operation.

Ongoing Maintenance. As the cyberinfrastructure comprises both physical and software resources, the maintenance plan will differ for each. For hardware, as mentioned in the response to 3d, the Collaborative's approach is to make fairly limited direct investment in hardware and instead to focus on adding capacity to existing large scale computing, storage, and visualization systems. Both UA and ASU commit substantial institutional resources to maintaining these systems, both in terms of supporting existing systems and providing technology refresh to acquire new systems as needed. The Collaborative's tools will be developed in such a way that they can also take advantage of remote resources, such as Teragrid systems which are maintained by NSF. This leaves the Collaborative itself with a fairly small set of servers to maintain, primarily to host the various services associated with the DE. While a number of servers will be available for development, we are exploring other options for long term care of the production systems. In the near term, solutions such as ASU's "virtual server" offering may make sense; it provides long term contracts for hosting a virtual machine (VM), independent of underlying hardware (the VM runs on a server farm maintained by ASU's IT services, and is periodically updated). In the longer term, the Collaborative is pursuing a number of potential commercial partnerships to supplement the "hard" services provided. The collaborative is in active discussion with Google and Amazon (Simple Storage Services and Elastic Computing cloud) to potentially be the long term storage site for the datasets of the collaborative. UA also utilizes an equipment life cycle approach along with virtualization to phase hardware resources in and out. As new equipment is procured (i.e. that is not commodity off the shelf) it is maintained under service contract provided by the manufacturer. Over the duration of its productive life cycle these contracts are maintained; usually after crossing the productive threshold, this equipment is utilized for applications that are deemed as "not mission critical" and without any service contracts. Ultimately the economics of housing the equipment such as electrical consumption, space, HVAC (Heating, Ventilation, Air Conditioning) dictate the phase out timeline.

For software maintenance, we mostly use open source tools which are maintained by their development

community at no charge to users; we anticipate the constituents of the collaborative will contribute back to these projects in form of patches and enhancements made for fine tuning these applications for use in the DE. The iPC will need to acquire some commercial products, e.g., for database administration, large scale visualization etc.; the costs of software licenses, which include upgrades, are included in the project budget.

Future Operation. The collaborative infrastructure will be able to remain operational for at least some number of years after the grant expires, as follows.

- *Hardware* for the base systems at partner sites will continue to be maintained and refreshed and will be available to iPC users, but of course new capacity needs will not be easily fulfilled. However, commercial enterprises such as Amazon and Google are starting to make computing and data resources available fairly inexpensively, and we anticipate that the new interdisciplinary communities and research projects that the iPC inspired will be able to take over management of their hardware needs.
- The DEs and other *software* that we develop will be open source, and will be built on top of open source tools. This allows the user community to take over its future development and maintenance. As needed, we will define an exit strategy to facilitate the access to and replication of services that are valuable to the community but not easily supported by the open source process itself. Appropriate use of virtualization technology will allow interested participants to recreate instances of these services at their institutions.
- For long term *data storage*, we expect to partner with Google to provide long term storage, which would enable data sets to be preserved and accessed long past the lifetime of the project.

3f. What are the provisions for security, accountability, and assurance for the resources and information to be provided by the Collaborative?

Security:

The iPC will benefit from the services provided by the UA Security Incident Response Team (SIRT) that actively monitors the university network using IDS (Intrusion Detection Systems) and manages the firewall. They also assist software developers with conducting vulnerability and penetration testing using open source tools like NESSUS, OSSEC and other 3rd party packages from offsite locations.

While most resources will be public, some features in the DEs (e.g. personalized workflows, individual online lab notes) will require secure access. All iPC users will be issued a credential when requesting initial accounts for iPC systems. This will be implemented using single sign on/identity management system based on LDAP (Lightweight Directory Access Protocol) and will cohesively tie into all sub systems of the CI including the web access, storage, analysis components. We are carefully tracking the progress of federated identity management system technology (e.g. Shibboleth) to allow users to gain access using the credentials issued by their home institutions. The state of Arizona is currently experimenting with cross-university credentialing systems, and some of the personnel who will be part of the iPC infrastructure team are part of the development effort on this project.

The basic security requirements for all CI computing equipment on the network will include ability to transmit event logs to a centralized log server (using OSSEC); this will leverage from the existing accounting infrastructure and provide measures for utilization and usage. Where applicable tools like Ganglia will provide supplemental data for accountability and capacity planning.

Assurance:

Maintaining the integrity of datasets for which the Collaborative acts as custodian will be of paramount importance to the project. This concern was one of the reasons the Collaborative chose to partner with

the supercomputer centers on the ASU and UA campuses. All data will be stored on professionally managed, enterprise-class storage systems. In addition to maintaining local snapshots and mirrors, the Collaborative will have the capability to provide additional data security by mirroring critical datasets between the UA and ASU campuses, providing protection against large-scale disasters. Since the CI will be leveraging on existing computing infrastructure it will also benefit from the defined disaster recovery procedures in place

We are also in discussions with a number of potential academic and commercial partners about serving as off site mirrors.

The iPlant Collaborative will develop specific guidelines and standards for depositing datasets (either physically or virtually as links) with detailed metadata documentation in consultation with the community. We will ensure correct attribution for ownership of data and provide provenance to help users gauge the quality of datasets. Software tools developed in the iPC will also be made available for widest possible dissemination using open source distribution policies with the clear indicators of copyright, and intellectual property ownership.

Accountability:

We will ensure proper tracking, capture and storage of all interactions and users for all users, data, tools and other resources using a structured model of provenance. Detailed tracking of all interactions will enable audit trails, error checking and correction, and proper attribution of resources in the iPC. It will also allow resolution of problems that may occur in the use, storage, and management of data, tools and other resources in the iPC.

4) *A community resource*

4a) *The review panel identified as a significant risk the relatively centralized structure of the Collaborative, which it felt could impede community buy-in. How will this risk be addressed to ensure the Collaborative is viewed as a community resource?*

Community-building pervades every aspect of the iPC plan. Our goal is to provide the community with the infrastructure it needs to do great science while playing a supporting behind-the-scenes role as facilitators, consultants, and help-desk operators.

The Grand Challenge symposia, conferences and workshops are organized by self-selected community members and reviewed and approved by a community-led Board of Directors (BoD). These meetings bring together members of the biological, mathematical and computational science communities who might never meet in the ordinary course of events. We see little risk that these meetings will fail for lack of community buy-in, provided that we maintain transparency in how the meeting topics are chosen, ensure that the organizers are respected representatives of their communities, and trust the organizers to set the meeting format and agenda. To ensure diversity and aid in the community-building aspects of these meetings, we will provide travel funding and meeting registration scholarships targeted to students, women, and members of underrepresented minorities.

The Grand Challenge Collaborative Teams, which arise from contacts forged during meetings, are reviewed and approved by the community-led BoD, and directed by community members. The iPC staff's role in GCTs is to provide consultation, support and software development services. While we will exercise due diligence (through external review and approval of the BoD) before committing large-scale infrastructural resources to a proposed Grand Challenge project, this due diligence will be based largely on the team's own feasibility studies, exploratory activities and software prototyping. iPC staff will assist with the feasibility study phase, but will not direct it.

The Discovery Environments are developed to assist the Grand Challenge teams, and are designed specifically to encourage members of the community to contribute their ideas, insights, syntheses, data sets and data models. DEs are classic community infrastructure products whose framework is built by iPC staff, but whose content is owned by grand challenge teams and members of the community at large. Consider the WormBase database, a community genome database for *C. elegans*, written in part by Lincoln Stein: its genome browser features annotation tracks contributed by hundreds of members of the *C. elegans* community. Research groups routinely submit their data to WormBase in advance of submitting their papers for publication, and grouse if the release of their data is delayed by as little as a week. The *C. elegans* community owns WormBase and has only a vague notion of who are the people responsible for its infrastructure.

To make the Discovery Environments even more community-driven, we plan to build Google Maps and Wikipedia-like “mashup” facilities into each of them, such that researchers can upload, annotate, integrate and synthesize their own data sets with data sets contributed by others.

The educational and outreach components of the iPC, with their emphasis on curricula development at the K-12 and undergraduate levels, target the training of teachers in a way that will result in the decentralized, self-propagating teaching of computational thinking in biology. In addition, by adapting the web-based Discovery Environments and iPC tools for use in the classroom, iPlant can potentially reach into schools with far greater penetration than can be accomplished with in-person workshops.

As we see it, the main risk to iPlant is not so much alienating the community as being ignored by it. Symposia, workshops and grand challenge teams go only so far towards achieving community buy-in.

Using the language of software marketers, we need a “killer app” that will attract attention from “early adopters” and then spread “virally” to members of the plant and computer science communities. The killer app must be intuitive, attractive and responsive, and most critically must allow researchers to understand complex data sets and make biologically-meaningful inferences that they could not otherwise achieve. The iPlant killer app is intimately tied to the Discovery Environments; indeed it is our hope that the Discovery Environments will become one or more killer apps. However, in order not to leave this to chance, the IST group will spend a considerable portion of its first years’ effort in designing and implementing an application that addresses a plant sciences problem not currently well served by existing applications. Among the ideas we have been kicking around are:

- An “Allen Brain Atlas for Plants:” 3D anatomical representations of various plant species, with facilities that allow for overlaying it with data sets (e.g. gene expression patterns) and manual annotations.
- “Virtual Linnaeus:” This is an application which, given the 5S RNA or mitochondrial DNA sequence from a plant species, will match it to a database of existing fingerprints and identify the species. If the species is not in the database, collect the taxonomic information from the user, if available, and add it to the database, to create an ever-growing community resource for plant identification. A more sophisticated version of this could be used to identify heterogeneous sequences from a meta-genomics study.
- “Plant Pathways:” A browseable collection of genetic and physical pathways in plant species, with facilities allowing community members to add their own pathways and annotate existing ones, along with analysis facilities for interpreting functional data sets. The application incorporates and unifies PlantCyc, AraCyc, RiceCyc, Arabidopsis Reactome and other existing pathway databases.
- “Diversity Integrator:” A web-based application for allowing community members to upload, integrate and analyze QTL and genotypic diversity information using the most up-to-date algorithms. These algorithms are currently spread out among multiple incompatible software packages and often inaccessible to the users who need them most.
- “Multi-Genome Browser:” A genome browser front end to comparative mapping tools, that allows researchers to build comparative maps on the fly, as well as to upload and analyze private sequencing sets.
- “Web2.0Biocrawlers:” Software to automatically browse the internet to locate, classify, and manipulate bio-relevant datasets. The idea here would be to start the process of automatic analysis of the existing and constantly increasing multidimensional datasets of biological information, whether it be DNA sequences, microarray expression and genotyping datasets, data relating to proteomics and metabolomics, and so-on. The ultimate goal is to provide in silico “leads” for the experimental biologist to address in terms of wet-lab experiments. Phase I of this work would simply be the identification of datasets containing simple identifiers (for example arabidopsis locus numbers). Phase II would involve the automated parsing of the identified datasets, to classify their likely information content. Phase III would be the return of processed data from these datasets, including data processed across different dataset sources

We realize that these ideas are ambitious. However, the iPlant software development team has the background, expertise, and resources to create a killer application to engage the community and win buy-in from skeptics, and we are confident of success. Our target is to have a compelling application up and available for community use no later than the end of the first year of the project.

To monitor how well we are succeeding in winning community buy-in, we will have two independent teams monitoring the project. Susan Brown's Social Networking Analysis team will monitor the adoption of Discovery Environments and measure the number and nature of social interactions that derive from Grand Challenge projects, meetings, iPAT teams, and other synthesis activities. Barbara

Heath's external evaluation group at EMEC will apply strict performance metrics to the project; most metrics will focus on community building targets. Both groups will report regularly to the board of advisors and the GC oversight committee, thereby giving us early warning and allowing us to make needed course corrections.

We end by noting that while it is true that 80% of the budget is allocated to the University of Arizona (10% to CSHL; 4% to Arizona State; 4% to EMEC and <1% to UNC), a large portion (12%) of the Arizona funding flows directly to outreach, educational and community building activities, including grand challenge conferences, symposia, Grand Challenge Team travel, community workshops and training activities, high school, undergraduate and teacher research opportunities, and other educational activities. An additional 25.5% of UA funds are reserved to support Grand Challenge Team research activities as they arise some or all may of which may occur at institutions beyond the core group (UA, CSHL, ASU, UNCW, Purdue). As described elsewhere, we are also pleased to have brought Rebecca Doerge and her team from Purdue into the project, which increases the decentralization of the project by bringing in another core participating site. .

4b) How will the Collaborative define the “community” it will serve?

We define the community broadly as:

- Plant biologists who need computational techniques to address the problems they are working on. Examples include geneticists, cell and developmental biologists, physiologists, genomicists, taxonomists, ecologists, molecular biologists, and evolutionary biologists.
- Translational biologists who will use the tools developed by the Collaborative to create improved plant stocks and agricultural products. Examples include plant breeders who will use phenotypic and genotypic diversity information to improve breeding stocks, and agricultural chemists who will develop a new generation of targeted products to fight invasive weed species based on an understanding of their distinctive biological pathways.
- Theoretical biologists, including systems biologists and synthetic biologists, who seek to create mathematical models of complex biological processes, or to forward engineer custom systems.
- Computer scientists, applied mathematicians, and statisticians who are seeking challenging new problem domains and the large-scale data sets to test them on.
- Physicists and chemists, who are looking for new challenges in the fields of sensor design, imaging, and electrophysiology.
- Students, in any the fields mentioned above, who are looking to be cross-trained in one or more of the others.

Naturally, it is harder to build a community than to identify it. Our community-building plan is informed by the Rogers model for the adoption and diffusion of innovators (www.valuebasedmanagement.net/methods_rogers_innovation_adoption_curve.html), which was first used to describe the adoption of new maize seed stocks, and later used extensively by marketers in the tech industries. In this model, the potential community is divided into Innovators, Early Adopters, Early Majority, Late Majority and Laggards.

Innovators are the brave few (classically 2.5% of the target audience) who are eager to try out new ideas, even if the road to adoption is rocky. The iPlant synthesis activities, including the symposia and workshops, target these innovators by recruiting them to form grand challenge teams and contribute their energy and creativity to the joint endeavor.

Early Adopters, who typically account for 10-15% of the community, are high-profile community

opinion leaders who are cautiously open to new ideas. It is critical to win the regard of Early Adopters because they can quickly sway the community one way or another. The Discovery Environments will be our main tool for capturing this segment of the community. Our primary strategy, as described in the answer to the previous question, is to engage the Early Adopters via a compelling “killer app” during the first year of the project. To help get the word out to potential Early Adopters, we will identify key members of the community in consultation with Grand Challenge Team collaborators and the project Advisory Board. We will then target these individuals by:

- 1) Offering them the opportunity to have an iPlant staff member visit their labs and give a seminar or demo;
- 2) Inviting them to UA, CSHL or Purdue to test drive the application and advise us on how we can improve it; or,
- 3) Inviting them to participate in an interactive webcast in which we demo the application to a group of potential Early Adopters.

An important asset in the iPlant Collaborative is that a significant number of the iPlant core faculty are themselves well-respected Early Adopters in their fields. Examples include Richard Jorgensen, Vicki Chandler, Steve Goff, David Galbraith, Carolyn Napoli and Robert Martienssen in experimental plant biology, Rod Wing in genomics, Brian Enquist in ecology, Lincoln Stein, Doreen Ware and Michael Sanderson in bioinformatics and phylogenetics, Sudha Ram and Kobus Barnard in computer science, and Rebecca Doerge in biostatistics. iPC faculty will lead the community by example, and use the tools that they have helped create.

The Early and Late Majority, accounting for two thirds of the community, look to the Early Adopters for guidance. We will also seek to engage this large group with broad outreach efforts, including the iPAT outreach teams, as well as talks, tutorials, workshops and printed brochures at major scientific meetings, publications describing iPlant resources in high-impact journals, webinars, and online tutorials.

Finally, the Laggards, estimated to amount to 16% of potential users, can't easily be convinced to adopt new tools no matter how aggressively they are pursued; we will consider ourselves spectacularly successful if we manage to engage the other 84%.

4c. How will this community be engaged and what strategies will be used to ensure full community buy-in to the project?

We will use models of innovation diffusion and adoption borrowed from social science to ensure that the iPC is properly and adequately diffused into the community. This will include developing mechanisms to survey, characterize, and classify potential users into categories such as innovators, early adopters, early and late majority. The characteristics of these types of users will be further studied and analyzed to help develop specific strategies for ensuring adoption of the resources in the iPC. For example, early adopters need to be addressed as individuals or smaller groups at the early stages of development to ensure that their needs are properly understood and catered. These early adopters have different needs and interests than those who will follow. Even though their needs are different, early adopters earn the role of “opinion leaders” in their respective communities and will help spread the use of the iPC via a “contagion” model.

Relying on what is known about adoption and diffusion, we will leverage characteristics of early adopters to identify them and take advantage of their interests in encouraging other adopters. The very earliest of adopters – innovators – will want to participate because the iPC is new. Our task will be to communicate to these individuals regarding the newness of the iPC. They can tolerate uncertainty, and will thus be quite beneficial to the early development stages of the project. Early adopters are the

individuals who take on the role of opinion leadership. They benefit from the experiences of innovators, but this group of individuals is respected in the community and willing to learn from the experiences of the innovators. The early adopters are essential to the success of the project, as they are the ones most likely to influence the majority. Early adopters will play a key role in the ongoing diffusion process.

A number of different strategies can be taken to train individuals in the use of the system. One very successful approach is to have individuals who understand the system and its uses train others. Again, we will draw on the innovators and early adopters to facilitate this training process. In addition, there is evidence to suggest that different people respond differently to training and education strategies. A survey can be administered to understand individual learning preferences, and the results of the survey can be used to design appropriate training experiences for participants.

We envision two types of users of the iPC broadly – these include the “producers” of resources and “consumers” of resources. Producers and consumers are not necessarily mutually exclusive, a producer of one type of resources (e.g. A software tool for statistical analysis) may be a consumer of another type of resource (eg. One or more datasets or compute cycles). Adoption and diffusion strategies will thus be specifically tailored to address both potential producers and consumers from multiple disciplines including CS, IS, Biology, and others.

4d. What are the measures of success for achieving community acceptance and how will these be evaluated?

Again, the evaluation team has selected portions from the evaluation framework to provide an outline of the “community-driven” outcomes that align with community involvement and acceptance efforts in iPC.

Outcome	Project Activity	Indicators of Success
iPC operates as an extension of the research community	<ul style="list-style-type: none"> ▪ Symposia, conferences, workshops organized by community ▪ Education/Outreach offer research opportunities to K-20 	<ul style="list-style-type: none"> ▪ Meetings are implemented as described in project documents, attendance meets benchmarks ▪ Research slots are filled yearly
iPC actively attracts input and participation from all elements of the community	<ul style="list-style-type: none"> ▪ iPC applies adoption/diffusion models as engagement strategies ▪ GCTs approved by community-led BoD ▪ DEs seek community contributions through mashups 	<ul style="list-style-type: none"> ▪ Models are implemented as described in project documents, participation meets benchmarks ▪ BoD represents the community in decision-making ▪ Mashups provide avenue for community contributions beyond GCT members
iPC responds to the needs and opportunities that the community identifies	<ul style="list-style-type: none"> ▪ Symposia, conferences, workshops organized by community ▪ iPC provides travel funding, meeting scholarships to targeted 	<ul style="list-style-type: none"> ▪ Meetings are implemented as described in project documents, attendance meets benchmarks ▪ Scholarships

	students, women, underrepresented groups	opportunities are filled with diverse population
--	---	---

4e) How would an “outsider” with an idea for a working group or resource engage the Collaborative?

The iPlant infrastructure will offer multiple online resources to help newcomers form collaborative projects, including WIKIs, online discussion forums, a SourceForge-style software development environment, mailing lists, and blog facilities. In some cases these facilities alone will help newcomers get collaborative projects started and they will need no more assistance.

In other cases, the newcomer will have an idea that requires more active engagement with iPlant. The process of engagement will begin with the newcomer sending a message to the iPlant help desk, or contacting one of the iPlant faculty members directly. The idea will be discussed at the next meeting of the Grand Challenge Team Oversight committee. In some cases, the idea will be a natural fit for an ongoing grand challenge project, iPC project, or discovery environment, in which case we will encourage the newcomer to join the ongoing project.

In other cases, the newcomer’s project idea will be a novel one that does not fit into an ongoing activity. In this case we will encourage the newcomer to use the community White Paper process described in the answer to question (2a) to identify the community his or her project serves, the scientific problem it addresses, and the resources it needs. Depending on the nature of the idea, we might also offer to help the newcomer organize a meeting or working group using the iPlant conference mechanism. We will then evaluate the project idea using the criteria described in the answer to (2a) to determine whether the project idea is sufficiently within the scope of the PSCIC mission (interdisciplinary, computationally challenging, community-oriented). If so, we will ask the newcomer to lead a new community project, and designate one or more iPlant faculty to become its liaisons and facilitators. We emphasize that at all of the stages where go/no go decisions are made, these decisions are made by the BoD composed of community members.

4f) How would an “outsider” community (i.e. a community, such as science journalism or environmental planning) be able to participate in and benefit from the Collaborative?

Cross-disciplinary alliances happen when two communities’ interests are aligned; how can each one benefit from the other? The answer to this question therefore depends strongly on what particular scientific questions community participants in the iPlant Collaborative undertake.

To take the first hypothetical, let’s say that one of the grand challenge projects that emerges is the modeling of the spread of invasive species and their effect on communities of native species. In this case, there is common ground with the environmental planning community: the grand challenge team benefits from the data collection methodologies of the environmental planning community, and the environmental planning community benefits from the predictions of the grand challenge team’s models.

It is also easy to see mutual needs between the iPlant Collaborative and science journalists. The iPC’s various educational and research projects will make a good stories; in addition the extensive open source material generated by the DEs and grand challenge projects will make good background source material. For our part, we will have a strong desire to engage journalists to get our participatory message out to the community. We can easily envision collaborating with a group of motivated journalists to sponsor a workshop on writing for the plant sciences. Nor is it outside the realm of possibility that the iPC could one day help a group of journalists create an online science-oriented news magazine or journal.

An outsider community will engage the iPC in the same way that any individual does. Representatives of the community contact us through the help desk or individual faculty members, the contact is discussed by the GCT Oversight Committee, and the community members are invited to begin a gradual “getting to know you” process of meetings, white papers, and follow-on workshops. If we find sufficient common ground for a collaboration, then we’ll either add augment an existing project, or create a new one, as described in the answer to the previous question.

5) *Integration across disciplines.*

5a. *What are the anticipated roles of plant scientists, computer, computational, and information scientists, cyberinfrastructure researchers and developers, and social scientists in the Collaborative?*

iPC research, outreach and education activities are interdisciplinary at every level.

- Members of all the above fields will work closely together to develop DEs in response to needs identified through the Grand Challenge teams. As described in 3b, each grand challenge team will be interdisciplinary and led by co-directors from different fields. While GC teams are being established, interdisciplinary groups will begin interfacing immediately through the planning and execution of the large symposium and through development of killer apps (described in 2a and 4a, respectively).
- The iPATs will include CISE and plant biology faculty and students who interface with core iPC software engineering support and educational teams.
- Social scientists will be intimately involved throughout to assist with establishing the measures to assess interaction and adoption and diffusion of products. (See 5b for additional ways social scientists will be involved).
- iPC will provide a fertile ground for state-of-the-art biological research, and as discussed in 5c below, also for cutting-edge computational, informational and social science investigations, many of which will be interdisciplinary.
- Faculty and students from the breadth of plant biology will participate in the symposia and grand challenge teams, where they will interface extensively with CISE faculty and students and with the DE teams of computer, computational, and information scientists as described in response to questions 1 and 2. The DE and GC teams will interface continually with the core infrastructure teams.
- Community scientists with expertise in all of the above fields will be asked to be external evaluators for the white papers for Grand Challenge teams, also made up of community scientists, and the community members serving on the BoD will be from all representative fields.
- Interdisciplinary teams will be overseeing student, teacher, and faculty mentor training and will develop educational and K-12 outreach materials.

5b. *Will social scientists be involved in research, planning, implementation, and evaluation/assessment activities of the Collaborative and, if so, how?*

Social scientists will be involved in all activities of the collaborative.

- 1) The collaborative provides an excellent research environment due to its size and scope, the complexity of the technology and interactions associated with it, and its longevity. Issues associated with adoption and diffusion of innovations can be studied, particularly addressing questions regarding how the collaborative unfolds over time. In addition, the development and evolution of social networks in a research environment is an important research topic. The scale of the collaborative will provide interesting insights into information systems implementation questions. Finally, social scientists can research the interactive nature of ongoing assessment and its impacts on systems enhancements and subsequent impact on assessment.
- 2) For planning and implementation, social scientists will examine issues associated with system usability. Usability assessments will be conducted, with the results provided to the development team. In addition, prior research in innovation diffusion will be leveraged to identify individuals most likely to embrace the collaborative. These same individuals are most likely to be the opinion leaders for future members of the collaborative, and thus they will be leveraged during the implementation process to increase acceptance.

- 3) A number of system metrics will be used to evaluate the system. For example, members of the collaborative will be surveyed regarding the ease of using the system, its usefulness to them in their research, the influence of colleagues' opinions regarding the system, and the conditions that enable (or hinder) their use of the system. These key characteristics have been found to be predictive of actual system use across a variety of systems. Additional metrics will include monitoring of use logs to understand who is using the system and how.
- 4) From an external evaluation perspective, social scientists will be collecting data from all project populations to determine the overall successes and shortcomings of the project. The external evaluation team will collaborate with other social scientists who are researching specific aspects of the Collaborative. Data collection efforts will be streamlined by identifying areas of overlap, thus reducing redundancy in collection methods. The external evaluators will be involved in planning and implementation in two ways, i) providing an external view of the project based on project documents and collected data that is shared verbally during meetings and ii) providing written reports with recommendations for keeping project activities intact or making changes based on the data from a variety of project activities and populations.

5c. How will the Collaborative engage computer, information and cyberinfrastructure scientists in research in their respective fields?

The iPC will enable exploration of a number of very interesting computer science, information systems, social sciences, statistics, economics, and legal questions. One vehicle for selling this message to those communities will be a list---maintained by the community---of “grand challenge questions” in **other** disciplines where the iPC provides an ideal environment for discovery. The iPC will be very proactive in making itself known and attracting those working in other areas, especially those not yet seriously engaged in interdisciplinary work. Our vision is that the iPC will be co-driven and co-constructed with those disciplines rather than just “engaging” them. The iPC recognizes that involvement of multiple disciplines at the outset is essential. For example, the iPC will ensure that the GC discovery process will involve relevant other communities, with special care taken to solicit involvement from those that would be under-represented without the iPC catalyst. A critical element for real progress in plant biology is new ideas and new thinking, and researchers and communities external to the existing Plant Science research teams will be an essential source of that. Hence iPlant will campaign relevant communities via conferences, mailing lists, and personal contacts, making use of initial seed connections. For example, GC imaging issues will be sold to the computer vision community as inherently relevant to that field. As an example, hosting an effort to model plant structures for recognition from images, with serious involvement from the vision community, would be a major benefit to **both** disciplines.

Examples of areas that the iPC provides a uniquely attractive platform for cutting edge research in non-Plant Science disciplines include scientific workflow automation, algorithms, data management, statistics and data mining, image analysis, economic models for “pricing” and sustainability, intellectual property, adoption and diffusion of DEs, social networking and evaluation. Specific questions will emerge from the interdisciplinary GC teams. However, we provide some examples here:

- 1) *Workflow automation*: Studying the nature of scientific biological analysis workflows, how they differ from structured business, manufacturing and other kinds of workflows, how they can be decomposed and assembled to support collaboration among diverse biologists and different levels of analyses. These workflows need to be curated and connected so they can be repeatable, modifiable, and accessible for others to use to solve GC problems.
- 2) *Data management*: Generating new information or “semantic” models for representing diverse biological data and the links among them, mining for new links, generating new links, modifying old ones as they are discovered via the DEs, semi-automated semantic integration of biological data, designing new techniques for harvesting and using provenance, and versioning of schema

and data. Being able to integrate the specifications of associated yet dissimilar experiments pushes into another grand challenge computer science problem of knowledge representation and fusing.

- 3) *Algorithm design, analysis, and evaluation*: Designing data structures and algorithms that are efficient in the context of external memory (e.g. non-cache memory or disk).
- 4) *Image analysis and representation of structure*. Automated (and thus high-throughput) image analysis, particularly in the modeling of form with appropriate representation so that it can be linked to function.
- 5) *Statistical modeling of multi-modal data*: Modeling and mining complex data that is sparse (e.g. microarray data) and thus require relatively strong models, linked data of other modalities, and/or domain knowledge.
- 6) *Data visualization*: Displaying complex data for enhanced understanding, and discovering patterns that automated methods cannot find. A second important issue is how can people best interact with such data to provide input into the process, both for understanding and for working with automated methods for discovering patterns.
- 7) *Economic models for sustainability*: Designing new and hybrid economic models for “pricing” of services and tools as well as sustaining the development of new capabilities.
- 8) *Adoption and Diffusion*: Applying models of product adoption and diffusion developed for other fields such as consumer products and technological innovations to interdisciplinary biological research is an important topic for social science research. Due to the scope and diversity of individuals involved in the collaborative, this environment provides interesting opportunities to address unanswered questions in this area.
- 9) *Social networking*: Understanding how and why various connections among plant scientists and across fields develop and evolve has important implications for the field, as well as for understanding the impact of other large-scale system implementations.

Scalability is a theme that runs across all of the areas mentioned above. Any algorithms, data management techniques and models designed in the iPC have to be scalable to address large numbers and sizes of datasets, users, and computational complexity.

A second theme is large scale repeatable evaluation. The iPC will be used by a wide variety of researchers on a wide variety of data sets with established evaluation protocols. Hence methods will be routinely evaluated much more extensively than is typical now.

5d. What is the proposed balance between enabling cross-training of individuals and promoting interdisciplinary teams?

iPC research and education activities utilize a holistic approach, acknowledging that each discipline represents an area of expertise required for the final project. Each discipline will be required to participate in all of the activities raising awareness between the disciplines, providing the opportunity to pool their respective expertise to contribute equally to the process and the product. The process will empower individuals by providing resources and ideas that would normally be outside of their respective disciplines. Cross-training will be a natural product of the synthesis activities.

6) *Budget and timeline*

Provided as separate document.

7) Management plan and Organizational Structure

7a) Describe the responsibilities of the director and explain whether a half-time or other appointment level is appropriate for this position. What criteria will be used in recruiting a Director of Cyberinfrastructure Development and a Director of Community Relations?

The Collaborative Director is responsible (a) for realizing the PSCIC's broad vision and keeping the Collaborative focused on meeting its responsibilities to the community, (b) for building community and acting as a broker, facilitator, and advocate for the community, and for coordinating among the Board of Directors, Management Team / Oversight Committee, and the Directors of Cyberinfrastructure Development and Community Interactions. The Directors of CD and CI are responsible for *implementation* of the vision and the directives of the BoD, M/OC, and Collaborative Director, and report directly to the latter. The two full-time Directors jointly function as the Chief Operations Officer, whereas the Collaborative Director functions as the Chief Executive Officer. Splitting the COO function into two full-time positions was judged to be essential for the Collaborative Director to meet his responsibilities to the iPC/community and to meet other commitments, such as managing his own research lab and professional service. Importantly, the Collaborative Director will have the active support of the Management Team and the Oversight Committee, comprised of the coPIs and other key personnel, who will assume responsibility for many management and implementation tasks as discussed in the Management Plan, Appendix A2.

The Collaborative Director has significant experience managing large research projects in both industry and academia. He fully understands the necessary degree of commitment for a project such as the iPC, and, importantly, that effective management requires delegation of major responsibilities to key individuals, who in this case include mainly the Directors of CD and CI and the Management Team. The Collaborative Director is committed to investing as much time as is needed to meet his responsibilities and ensure the success of the iPC, and the iPC will be his highest priority, just as his role as Editor-in-Chief of The Plant Cell has been for the past 4 years. That position required significant time involving oversight and mentoring of more than 30 coeditors, scattered around the world, and 5 staff in Maryland; being responsive to authors, readers, and the community, which involved responding promptly to concerns of authors, readers, and other stakeholders; reviewing, revising, and setting journal policy; and organizing and running editorial board meetings. A commitment of 50% time to the iPC likely represents at least 35 hours/week, similar to the time devoted to The Plant Cell as EIC. If the NSF and the site visit team have questions about the EiC's performance, they are encouraged to contact the Executive Committee of the editorial board (Cathie Martin, Sally Assmann, Bill Lucas, David Smyth, and Nick Talbot), as well as Dr. Crispin Taylor, the Executive Director of the ASPB, the publisher of TPC. ASPB is currently interviewing for the next EiC, whose term is scheduled to begin January 1, 2008, following a transition during the last 6 months of 2007.

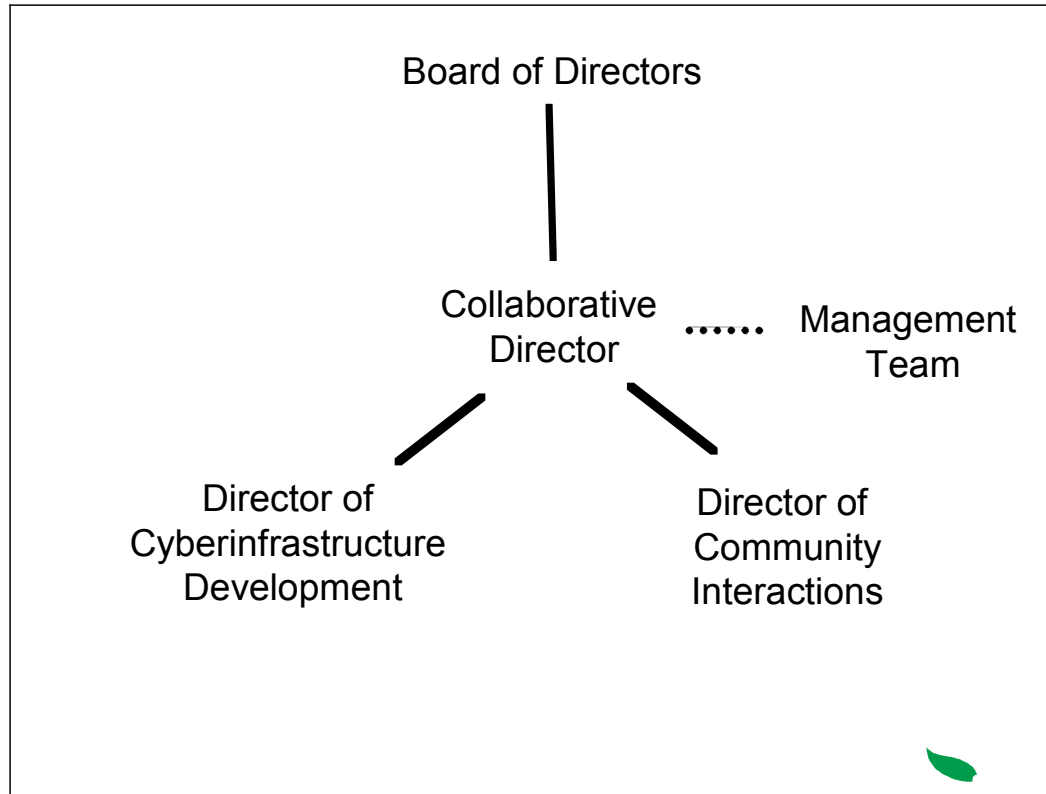
The qualifications for the Director of Cyberinfrastructure Development are as follows. Required: Degree in computer science or related field and at least 5 years experience in managing the computing infrastructure of a medium to large sized organization. S/he must have the experience of having managed teams of programmers and systems staff and must be familiar with technologies including database systems, networking, storage systems, and software engineering. Desirable: Work experience in managing the computing infrastructure for a biology-related organization. Familiarity with one or more of the following is also desirable: work with biology data, biological applications, and biological or other scientific workflows.

The qualifications for the Director of Community Interactions are as follows. Required: Advanced degree in the biological sciences and at least 5 years managing a medium-sized group in research or

development in academia or industry. Strong skills in personnel management, team building, and community interactions. Desirable: Experience in plant biology and bioinformatics.

7b) Provide an organizational chart that clearly illustrates the relationship (chain of commands) between the Collaborative Director, Oversight Committee, Board of Advisors, and the Directors of Cyberinfrastructure Development and Community Relations.

As indicated in the appendix, the Oversight Committee assists the Management Team, but does not have a direct management role. The Management Team was erroneously left out of the management diagram in the appendix. The diagram below shows the direct relationships between the Management Team to the Collaborative Director and the Directors of Cyberinfrastructure Development and Community Interactions.



7c) The centralized nature of the proposed Collaborative was identified by the review panel as a potential management risk, noting that this structure may result in the Collaborative being viewed as a University of Arizona entity as opposed to a community resource. Are there plans for implementing the management plan and structuring the organization in ways that could mitigate this risk?

There are two distinct parts to this question. First, it is important to recognize that a centralized organization is much easier to manage than an unnecessarily distributed organization. We have chosen an initially centralized structure for certain aspects of the project precisely for the reason that it reduces management risk in the implementation of community needs and priorities. The second part of the question deals with the risk that the community might fail to perceive the Collaborative as a community resource. This risk is mitigated by community control of the Board of Directors, who will make all major decisions about resource allocation and priorities. A major responsibility of both the Collaborative Director and the Director of Community Interactions will be to ensure that the community's perception

is that the Collaborative belongs to them. This is a major reason why we feel the DCI position is essential to the Collaborative.

7d) What measures will be used to assess management success and how will they be evaluated?

The external evaluation has selected sections from the Request for Proposals that describe the expectations set forth by the funding agency in the areas of Expected Characteristics, Primary Responsibilities, and Core Capabilities. These items stand as the framework for the evaluation effort. One of the six Expected Characteristics is that the project will be effectively managed. Further, the management should provide for:

- effective leadership
- efficient implementation
- reliable oversight accountability
- quick response to new opportunities and gauging community needs
- inclusion of external advisory groups for guidance, identifying frontier opportunities, and allowing for breadth of input and advice

A series of tables has been prepared that outlines the sections from the RFP and matches them to the iPlant project activities. Additionally, success measures and data sources have been outlined. The segment of Table 1 that addresses effective management is included below. Data will be collected from appropriate sources, dependent upon which indicator is being measured. Data collection methods for this segment will include document analysis, surveys, focus groups, and interviews.

Effectively managed by...	Project Activities	Indicators of Success
Providing for effective leadership	Management structure is organized and roles are clearly defined DCD and DCR are hired through competitive process BoA is qualified and diverse	Leadership at all levels fulfill roles defined in project documents DCD and DCR positions are filled by qualified personnel BoA is organized and functions as described in project documents
Providing for efficient implementation	ComSOT ensures inclusiveness, diversity and effectiveness through clear communication practices	Project completes project tasks as outlined in timeline
Reliable oversight accountability	BoA oversees development of policies and procedures that ensure community priorities/perspectives are addressed Social Networking and Evaluation teams measure effectiveness of multiple management aspects	Policies and procedures are developed and disseminated to the Community, Community reports satisfaction with priorities and perspectives addressed by iPlant Social Networking and Evaluation provide formative information to project management that is used for decision-making
Enabling quick response to new opportunities and gauging community needs	ComSOT team employs multiple methods for distributing and collecting information from the	Multiple methods are employed for distribution/collection of community information, Information is organized

	<p>community to be organized and shared with the GC and IS teams</p> <p>BoA oversees development of policies and procedures that ensure community priorities/perspectives are addressed throughout infrastructure design process</p> <p>IS team will be professional software engineers with skill set that is agile and flexible</p> <p>ID team will include help-desk support</p>	<p>provided to GC and IS teams, GC and IS teams use information in decision-making and design</p> <p>Policies and procedures are developed and disseminated to the Community, Community reports satisfaction with iPlant products</p> <p>IS team fulfills roles defined in project documents</p> <p>Help-desk support is provided, Community reports satisfaction with help-desk support</p>
<p>Inclusion of external advisory groups for a) guidance, b) identifying frontier opportunities, c) allowing for breadth of input and advice</p>	<p>Project management meets with Oversight Committee, BoA, evaluators, and social scientists at regular intervals</p> <p>Conferences and symposia provide venue for brainstorming</p> <p>iPlant website provides community with feedback capability beyond social networking and evaluation efforts</p>	<p>Advisory groups are selected and organized</p> <p>Meetings are scheduled and well-attended</p> <p>Conferences/symposia are scheduled and well-attended, attendees report satisfaction with experience, Conferences result in products useful to iPlant</p> <p>Community provides information and feedback via interactive segments of the iPlant Collaborative website</p>

8) Education and training: the next generation

8a: How will the Collaborative promote the integration of research and education?

The research, education and outreach goals and activities of the iPC are tightly integrated to meet the objective of diffusing computational thinking and grand challenge ideas into K-20 education. This is the first time in the history of science that students can work with the same data, using the same tools, and in the same time frame as researchers. The challenges are: 1) to get these tools and datasets into the hands of the scientists and educators who are the “gatekeepers” at the cusp of research and science education; 2) to broadly train the next generation of scientists to work effectively within interdisciplinary research teams; and 3) to increase the number and diversity of students pursuing science careers. To meet these challenges, the iPC will target many groups: undergraduate and graduate students; faculty at two-year, four-year and research-intensive institutions; K-12 teachers and high school students. Our approaches include:

- adapting elements of *Discovery Environments* and *iPC* tools for use in inquiry learning and research projects
- disseminating these tools beyond the immediate audience of high-level plant biology and CISE researchers, of particular interest are educators at non-research universities, smaller colleges, historically Black and Hispanic institutions, tribal colleges, and two-year colleges
- teaching the teachers through a variety of mechanisms such as 1.5 day workshops, think tank symposia, summer research internships, and iPlant Action Team (iPAT) experiences
- build a community developing and disseminating curricular ideas and materials to integrate computational thinking into biology and computer science education at all levels
- Grand Challenge research opportunities through participation in DEs, iPATs, through graduate fellowships, undergraduate research programs (UBRP, BRAVO! as models), and summer internships for high school students and teachers
- A partnership with STATCOM, a Purdue-based program in statistics which promotes statistical thinking at the K-12 and graduate levels.

The various activities will be executed by key personnel at UA (Napoli, Westbrook, UA Learning Technology Center), Cold Spring Harbor Labs (Micklos, Dolan DNA Learning Center) and iPATs (coordinated by Stapleton), as summarized in the attached table.

8b (part 1): How will the Collaborative promote innovative use of cyberinfrastructure in education and training?

Through participation in Grand Challenge Teams and iPATs students, teachers, and researchers will contribute to iPC activities and have access to resources, no matter where they are located. In addition, a variety of web-based learning materials (courses, modules, games, interactive websites, etc.) will be developed and made freely available. Importantly, internet technology allows students a real opportunity to participate in unsettled, “revolutionary” science – **using the same data and same tools at the same time as research biologists**. Both the Dolan DNA Learning Center and UA's Learning Technology Center have extensive experience in developing web-based educational materials. Although the integrative tools and datasets developed for the *iPC* will be freely available, we will need to work to disseminate them broadly. One key mechanism will be 1.5 day workshops that will target faculty in the “2+2+2” continuum of high schools, two-year colleges, and four-year colleges. These biologists need a basic understanding of how to use the *Discovery Environment* and educational interfaces – to apply to their own research and to use with classes they teach. Dolan DNALC has the capability and experience in providing workshops and classes to both students and teachers at the high school and college levels throughout the country; to date 5,300 precollege and college faculty have been instructed at training workshops conducted in 42 states and seven foreign countries. Within iPC the goal is to reach 250 high

school and college faculty per year through workshops. At these workshops examples/paths for additional involvement beyond the workshop will be provided, including examples and contact info for working with Grand Challenge teams and applications/information for iPATs. These workshops will also be excellent forums to recruit teachers and their students to the summer internships.

Statistics is a critical aspect of computational thinking, but undergraduate and graduate-level education in statistical thinking lags woefully among many biological and physical science training programs in the United States. To address this, we will connect the resources of UA's newly established Graduate Interdisciplinary Program (GIDP) in Statistics in partnership with Statistics in the Community (STATCOM), a graduate student-run consulting service that provides free statistical consulting to governmental and non-profit groups. The new GIDP in Statistics emphasizes interdisciplinary training that integrates statistical theory and methodology with pertinent, practical, subject-matter applications, while STATCOM, a program originated at Purdue University, is quickly providing a national presence in the statistical community in which graduate students become involved in consulting. Both the GIDP and STATCOM take unique approaches to engagement by integrating service learning into graduate education, and both are situated to grow within the quantitative biology portion of this Cyberinfrastructure proposal. STATCOM also includes outreach activities to K-12, as well as undergraduates, and it is anticipated that this will be our main connection to the outreach activities. Involvement of STATCOM individuals in the science fair activities (i.e., student judges, workshops for high school teachers and parents, etc.) will bring awareness of the applied nature of quantitative biology and will support our national mission in the United States to promote mathematical and statistical thinking. Anticipated outcomes of STATCOM involvement include growth of this nationally recognized student organization, education in consulting, experience in communication and outreach, and involvement of students in core activities. Dr. Rebecca Doerge will lead the STATCOM partnership, coordinating with Dr. Walter Piegorsch of the GIDP in Statistics.

8b (part 2): What mechanisms will be used to assess the impact of these efforts?

Each of the four major aspects, Dolan DNALC, iPATs, higher education, and K-12 education has set goals, products and measurable outcomes for each objective. Each will be evaluated at two levels. The first is from an external perspective where the evaluation team will seek to determine whether the education and training component has reached their goals and outcomes. The second is each internal team will conduct a longitudinal study to determine the impact of activities on their targeted groups, i.e. faculty, teachers, and students. Data will be shared between the internal and external evaluations for report preparation. For simplicity, we illustrate these mechanisms with one example from external and internal evaluation metrics.

External Evaluation Metric:

Outcome	Project Activities	Indicators of Success
Prepare the next generation: Providing diverse mechanisms for training and education to engage the community at various levels	Offer variety of mechanisms to teach the educators (1.5 day workshops, summer internships, iPATs, symposia) Develop and disseminate curricular materials (K-20) that integrate computational thinking into biology and CS Provide GC research opportunities through participation in DEs	250 high school and college faculty attend workshops per year 6 high school student interns 9 K-12 teacher interns 6 iPATs per year 90 participants in educational symposia 2 K-12 teaching modules disseminated per year 3 undergraduate teaching modules/year 10 undergraduate research fellows per year 3 faculty summer fellows/yr

Internal Longitudinal Evaluation:

DNALC workshops: Longitudinal tracking of workshop participants will measure changes in research and teaching behavior over time. Longitudinal data will be collected at three time points: pre-workshop, post-workshop, and 18 months after the workshop. All teacher surveys will be adapted from validated survey methods that have produced excellent response rates in past DNALC surveys. E-mail surveys will direct participants to an anonymous survey instrument stored at the *Survey Monkey* Internet site (<http://www.surveymonkey.com/>). This evaluation can answer a number of important questions, including: *How are iPC tools used in research? In what types of courses are iPC tools and curriculum materials used? What sort of student research projects and lab investigations have been done. How many students have been impacted by improved instruction? How many other faculty have been trained by workshop participants?*

8c. The iPC-K-12 program proposes extensive participation by teachers. How will the participation of teacher with extensive demands on their time be enabled?

The multifaceted nature of the iPC allows teachers to seek a level of involvement that meets their own needs and time demands. The regional workshops are designed as a time-effective introduction to iPC tools and learning materials. A faculty member need only commit 2 days to attending a workshop that is located within commuting distance. Precollege teachers can increase their involvement and expertise by participating in intensive summer internships at UA, which allow them to develop classroom materials for themselves and other teachers. College teachers can follow-up on their regional workshop experience by forming an iPlant Action Team to pursue interdisciplinary research and curriculum development.

Because K-12 teachers have extensive demands during the school year and have little time for extracurricular activities, we have limited the longer teacher research experiences to a six week summer internship. The six-week time-frame has worked exceptionally well in our RET in Plant Genomics that has had 10 teachers each summer for the past eight years. At the onset of the six week period, teachers will be introduced to a particular topic and will work within three-member, multidisciplinary teams on strategies to integrate computational thinking into plant biology and to design teaching modules based on their work. Teachers can continue to have input into teaching module development as their time and interest permits, but iPC staff will have responsibility for final implementation of each product. CoPI Napoli (25% FTE) and a staff member with an education background (50% FTE) working closely with

DNALC and UALTC, which have extensive experience in developing educational interfaces and tools, will produce the web-based teaching aids.

9) Diversity of participation.

9a) What are specific plans for increasing the participation at all levels of women and under-represented groups?

The leadership team of iPC has an excellent track record of training women and underrepresented minorities and is committed to assuring strong participation of women and under-represented groups at all levels within iPC. We begin with ourselves and note that 40% of the iPC leadership team are women (8 total).

We will strive to have diversity in all levels of the collaborative, including the composition of the Board of Directors. This group will be composed of community leaders and be responsible for oversight of the complete project, setting priorities and approving Grand Challenge Teams; it is important to this important group be diverse. We will insure that substantial numbers of women and minorities are included and participate in iPC activities such as performing reviews of GC white papers, attending and speaking at symposia and workshops, participating in iPAT teams and internship programs. To insure diversity, we will set specific goals and continue our recruitment efforts until we reach them. Outcomes will be monitored and used to modify our recruitment approaches as necessary.

We will identify diverse pools of applicants and participants through an aggressive and broad marketing of our recruitment program to insure all the opportunities within iPC research, education and outreach activities are communicated to women and under-represented groups. An important component of recruitment is contacts and relationships with key groups and institutions with significant numbers of women and underrepresented minorities. These groups include but are not limited to: Historical Black Colleges and Universities (HBCU), Hispanic Association of Colleges and Universities (HACU), tribal colleges, and community colleges, Society for Advancement of Chicanos and Native Americans (SACNAS), American Indians in Science and Engineering Society (AISES), and college societies promoting diversity and representation by underrepresented groups. Professional societies such as the American Society of Plant Biology, Botanical Society of America, Ecological Society of America, NCSE, NESCENT, Computing Research Association (CRA), Association for Computing Machinery (ACM), and National Teacher Societies such as CSTA and NSTA will provide more avenues for recruitment.

Our approaches will be multifaceted. Select examples are outlined below:

- Web dissemination:
 - iPC website
 - DNALC (*BioMedia Group* which targets primarily at high school and beginning college students)
 - Gramene, TAIR, maizeGDB, ChromDB
 - NCSE
 - NEScent
 - Bio-Link.org
 - IEEE and ACM portals
- Direct mailing, as well as e-mail
 - Registered users of DNALC (20,000 educators including many at HACU and HBCU institutions)
 - Member of ACM's Special Interest Group on Computer Science Education (SIGCSE)
- Advertisements in society newsletters
- Recruitment at conferences targeting women and underrepresented minorities.
 - The Grace Hopper Celebration of Women in Computing
 - The Richard Tapia Celebration of Diversity in Computing Conference

- SACNAS
- MARC
- AISES
- Computer Science Teachers Association (CSTA)
- Women in Engineering Programs & Advocates Network (WEPAN)
- Seminars and talks by iPC faculty across diverse K-20 institutions with efforts targeted at increasing diversity in iPC and science and technology (mathematics and computer sciences) in general
 - HBCU, HACU and tribal colleges
 - K-12
- marketing to minorities and women through the well developed mechanisms at each of our institutions
- iPATs, small geographically dispersed collaborative teams

Just as recruitment is an important part of increasing the participation of minority and underrepresented groups, so is the ease of implementing and using iPC systems and tools. Social science research has shown that attention must be paid to “ease of use” to encourage technology adoption across age and gender. It is paramount that as iPC systems and tools are introduced, we publicize their ease of use and provide training. The collaborative will monitor ease of use assessments to ensure that participants actually have a positive experience. In addition, leveraging women and underrepresented minority opinion leaders to provide social influence regarding system use will be important. The charge of the Collaborative is to ensure access, a compatible system, and well-designed training programs that address community demographics.

9b. How will the impacts of these plans for increasing the diversity of participation be monitored and assessed?

Recruitment of women and underrepresented groups and participation by people from all types of institutions will be evaluated within each component of this project, on an annual basis. Working with our advisory board we will set recruitment goals and metrics within each aspect of iPC. We will work closely with the external evaluation team who will track progress against these well defined goals and metrics.

9c. How will the assessments be used to improve the efforts of the Collaborative and inform the activities of the community at large?

Deviations from our expected recruitment goals will be communicated to the project director, oversight committee and BoD. This will result in evaluation of the recruitment methods and refinement. For example, if we expect 20% participation by group A and we achieve only 10%, we will examine the process and refine the recruitment methods. We will work closely with the external evaluation team to carryout ongoing collection and analysis of data that will inform the project team of the successes and shortcomings of each project. Data will be collected using multiple methods and in collaboration with other social scientists involved with the Collaborative. When necessary, changes to the project plan will be recommended based on the data collected. All changes will be recorded along with the rationale.

10) Partnerships

10a) How will the Collaborative's activities be coordinated with the existing large efforts on management of biological data and information, within and outside of the US?

We are very mindful that the proposed iPC will be one component of a diverse community of information management groups seeking to solve similar problems of data sharing, integration and analysis. Our challenge is to stay aware of what is going on in the greater world outside the iPC, whether it be in the biomedical sciences, physics, meteorology, the high tech world or the plant sciences, and to be alert to opportunities to exchange ideas, coordinate activities, and establish collaborations.

We have identified a series of ongoing efforts that it will be important to engage with from the beginning of the project, and have appointed liaisons to these groups based on existing relationships. A preliminary list of these efforts and their liaisons is given in response to the next question. Liaisons will report periodically to the Coordinator of Partnerships, Stephen Rounsley, who will in turn report to the Executive Committee and the BoD. As the project progresses, we will identify additional efforts to establish relationships with and designate liaisons to them. To keep abreast on relevant outside projects, we will periodically poll community members and the BoD.

Coordination between iPC and outside efforts will range from informal contacts to formalized collaborations. Possible forms of coordination efforts are illustrated by the following hypothetical scenarios:

- A large multi-center collaborative effort is developing web services protocols to share information among its centers. Since this is directly overlapping with the iPC mission it is high priority to coordinate with this group in order to avoid redundancy of effort. The liaison will engage representatives of the group to identify opportunities to share code, protocols, file formats and data types with iPC. If the opportunities for codevelopment are large, the relationship between the group and iPC could be formalized by having the iPC serve on the group's Scientific Advisory Board (or equivalent), and by having a representative of the external group serve on the iPC BoD.
- A software development group is working on a piece of open source software that would be a useful component of an iPC DE. The liaison will contact and engage the developers by participating in mailing lists, discussion forums, bug reports and feature request lists in order to keep abreast of the future development path and potentially to influence the ongoing design. If appropriate, the iPC might commit developer resources to the effort.
- Reciprocally, an iPC staff member or community member realizes that a piece of software being developed by the iPC could meet the needs of another group. We will appoint a liaison who will bring this possibility to the group's attention.
- A group of plant biologists not yet affiliated with iPC begins developing an ontology relevant to the plant sciences. As in the case of a piece of software, a liaison will engage the community via its mailing lists and discussion forums. If appropriate, the liaison will recruit interested community members from ongoing GCTs to participate as contributors to the ontology-building project.
- A bioinformatics group that works on human cancer creates a resource for storing and analyzing metabolomics data. Small extensions to this resource would be useful for studying plant physiology. The liaison will seek an agreement to extend the resource to handle plant data and to open the resource to the iPC infrastructure. As an incentive, the liaison will offer to provide "in kind" assistance to the group in the form of a visiting software developer to help implement the needed extensions.
- A group of meteorology researchers develops a distributed storage system for satellite imaging

data. One of the software engineers in the IST group reads about this in a trade journal, realizes that this might solve a similar problem in the storage of biological imaging data, and brings the system to the attention of an IST faculty member. The faculty member initiates a contact with the meteorology group, and this initial contact grows into a formal collaboration.

10b) What are the main on-going efforts that require close collaboration with the Collaborative?

We have identified many of the ongoing efforts in information management and cyberinfrastructure both within and outside the plant sciences that require coordination with the iPC. We describe each of the efforts briefly, identify the primary liaison between the iPC and effort participants, and discuss the liaison's qualifications to lead in this role.

Information-Oriented Plant Science Efforts

TraitNet - *Liaison: B. J. Enquist.* This is a recently funded NSF RCN initiative that will coordinate a botanical trait-based evolutionary and ecological research. Traits are biological properties of species that influence individual fitness and govern how species interact with their biotic and abiotic environment. Its five primary goals include identify critical botanical data gaps; coordinate the standardization of collection and curation of trait data and facilitate the development of cross-disciplinary computational tools for merging, disseminating, and sharing trait data. Enquist is a senior collaborator on the Traitnet project.

Arabidopsis Polymorphism Database. *Liaison: L. Stein.* This is a newly-funded NSF project to store and analyze genotypic and phenotypic information on the natural diversity of *Arabidopsis*. Stein is a coPI on the project.

Arabidopsis Reactome. *Liaison: L. Stein (international).* This is a UK BBSRC-funded project run out of the Nottingham Stock Center to annotate signaling and developmental pathways in *Arabidopsis*. It is already coordinating with Reactome, an animal pathway database project PI'd by Stein.

CIPRes. *Liaison: M. Sanderson.* CIPRes (Cyberinfrastructure for Phylogenetic Research) is charged with developing hardware infrastructure and software tools for large-scale phylogenetic analysis and databasing, a central tool in comparative biology. Sanderson's research overlaps extensively with the project and he already collaborates with several CIPRes PIs on other phylogenetic projects.

Gramene. *Liaison: D. Ware.* This is a model organism database for rice and other monocots, which focuses on comparative genome maps. Ware is a coPI on this project.

MaizeGDB. *Liaison: D. Ware.* This is a model organism database for maize. Ware is on the SAB for this project.

NASCArrays. *Liaison: D. Galbraith.* (international) This is a microarray database of *Arabidopsis* expression data managed by the Nottingham Arabidopsis Stock Centre. Galbraith works in this area and has undertaken to act as liaison to all microarray projects.

GRIN. *Liaison: D. Ware.* This is an USDA sponsor information resource used for data storage and display of the National Plant Germplasm System (NPGS), the National Animal Germplasm System (NAGP), the National Microbial Germplasm Program (NMGP), the National Invertebrate Germplasm Program (NIGRP). Ware will undertake to act as liaison to the USDA database projects.

NESCent. *Liaison: M. Sanderson.* This project solicits broad participation from the evolutionary

biology community and has a strong informatics mission to facilitate exploitation of data sets, as well as training and mentoring in cyberinfrastructure. Sanderson serves on the NESCent informatics advisory committee and is a member of a NESCent working group on plant evolutionary genomics.

Panzea Liaison: *D. Ware*. This is a multi-center NSF-funded project to characterize genotypic and phenotypic diversity in Maize in order to identify domestication loci. Ware is a co-PI on this project.

PlantCyc. Liaison: *L. Stein*. This is a newly-funded NSF project to curate biological pathways in Arabidopsis, Rice and other sequenced plant genomes. Stein is a long-time collaborator with its PI, Sue Rhee.

PlexDB. Liaison: *D. Galbraith*. This is a microarray database widely used for the storage, display and analysis of gene expression data in monocots and other species (PI Roger Wise). Galbraith works in this area and has undertaken to act as liaison to all microarray projects.

The Arabidopsis Information Resource. Liaison: *D. Galbraith*. This is the model organism database for Arabidopsis. Galbraith is a long-time contributor and user of this resource, and has a good working relationship with Sue Rhee, TAIR's manager and coPI.

The Plant Ontology Consortium Liaison: *D. Ware*. This is a multi-center collaboration to develop shared ontologies to describe the anatomic structures and developmental stages of flowering plants. Ware is a co-PI on this project.

Biological Science Infrastructure Projects

NCEAS (National Center for Ecological Analysis and Synthesis). *Liaison:* *B. J. Enquist*. This project will solicit broad participation from the ecological community. NCEAS has a strong emphasis on informatics and facilitates the integration and exploration of datasets. NCEAS has been a pioneering force in ecoinformatics, training, and outreach. NCEAS has developed community informatics protocols via the SEEK project as well as developing shareware including MORPHO and KEPLER. Enquist was a post-doc at NCEAS and currently serves on several NCEAS working groups and is collaborating with the informatics staff at NCEAS.

BioSapiens. Liaison: *L. Stein* (international). This is a Europe-wide network of EC-funded projects whose mission is to share and integrate genome-scale functional information using the semantic web. The network is built around the Distributed Annotation System developed by Stein.

CaBIG. Liaison: *L. Stein*. This is a National Cancer Institute funded project to develop a grid of cancer bioinformatics services. Its mission is similar in some ways to the PSCIC, but focuses almost exclusively on software development. Stein serves on the project's Strategic Level and Data Sharing working groups and has been a software developer for three projects in the Integrated Cancer Research domain.

National Center for Biomedical Ontologies. Liaison: *L. Stein*. This is an NIH Roadmap computational biology center whose mission is to provide the infrastructure for biological ontology development. Stein is on its SAB.

Pathway Commons. Liaison: *L. Stein*. This is an NHGRI-funded project to develop a shared infrastructure for exchanging, storing and analyzing biological pathway data. Stein works closely with its PI, Chris Sander, as part of an ongoing collaboration between this project and Stein's Reactome project.

SSWAP/VPIN. *Liaison: L. Stein.* This is an NSF-funded collaboration between NCGR, CSHL and TIGR to develop technologies for identifying and exchanging biological data and computational services via the semantic web. Stein is a co-PI on this project.

The MAGnet Center. *Liaison: L. Stein.* This is an NIH Roadmap computational biology center whose mission is to develop theoretic and mathematical models to describe cell and developmental biology at multiple molecular scales. Stein is a member of the center's SAB.

The Model Organism Database Consortium. *Liaison: L. Stein.* This is a multi-institution collaboration to develop the software tools, user interfaces, and standard operating procedures needed to operating a model organism database and website. Stein is a co-PI on this project.

Information-Oriented Biomedical Sciences Projects

AHEAD. *Liaison: R. Martienssen* (international). This is an international task force to promote epigenetic research in biomedicine as well as model organisms including plants. Martienssen leads the model organisms subcommittee, and is a co-founder of the task force.

Center for Evolutionary Genomics. *Liaison: R. Martienssen.* This is an NSF-funded collaboration among four NY region institutions to investigate the origin of seed plants, especially gymnosperms using the computational techniques of phylogenomics. Martienssen is a coPI on the project.

DDBJ. *Liaison: R. Wing* (international). This is the Japanese equivalent of GenBank. Wing has high-level contacts with the DDBJ administration.

DOE Joint Genomics Institute. *Liaison: R. Wing.* This is the Department of Energy's large-scale sequencing effort. Wing is a long-term collaborator of JGI's director, Daniel Rokhsar.

ENCODE and modENCODE. *Liaison: L. Stein* (international). These are NIH-funded projects to identify and characterize all functional elements in the genomes of human, fruitfly and *C. elegans*. Stein is coordinator of the modENCODE DCC, and a member of ENCODE metadata working groups.

ENFIN. *Liaison: L. Stein* (international). This is a large multi-center EC-funded project to develop computational models of fundamental biological processes in metazoa. Stein is on the ENFIN SAB.

Ensembl. *Liaison: L. Stein* (international). This is a European Bioinformatics Institute effort jointly funded by the Wellcome Trust and EMBL to annotate and visualize the genomes of all species, currently focused on vertebrates, but soon to be extended to other clades, including plants. Stein is a former Ensembl scientific advisory board member and a close collaborator of Ewan Birney, the head of EBI's sequence database operations.

Microarray databases (non-plant). *Liaison: D. Galbraith* (international). These general microarray databases include SMD, EMBL-EBI ArrayExpress, NCGI-GEO, CATMA and RARGE-RIKEN. Galbraith works in this area and has undertaken to act as liaison to all microarray projects.

National Center for Biomedical Information. *Liaison: D. Galbraith.* This is a National Library of Medicine-funded center that supports a variety of databases spanning the biomedical literature, protein and nucleotide sequences, diseases and genomes.

The Epigenome. *Liaison:* R. Martienssen (international). This is a European Union Network of Excellence project headed by T. Jenuwein, consisting of more than 50 research groups dedicated to epigenetic mechanisms and networks in plants, animal models and humans. Martienssen is on the scientific advisory board of this project.

UCSC genome browser. *Liaison:* L. Stein: This is a genomics platform used for the analysis of multiple metazoan genomes, but primarily the vertebrates. It is likely that plants will be added to the genome browser in the future. Stein is a collaborator with Jim Kent, head of the genome browser effort.

Information Technology Projects

STATCOM. *Liaison:* R. Doerge. Statistics in the Community (STATCOM) is a graduate student-run consulting service that provides free statistical consulting to governmental and non-profit groups. STATCOM originated at Purdue University, but has rapidly grown into a national infrastructure for statistical education and consulting. Doerge was a founder of STATCOM, and will serve as iPC's liaison to that project.

Gene Ontology. *Liaison:* L. Stein (international). This project is a collaboration of model organism and protein databases and biological research communities actively involved in development and application of the Gene Ontology.

EC-US working group on Plant Biotechnology. *Liaison:* D. Ware (international). The working group is part of the larger EC-US Task Force on Biotechnology Research and is a means of increasing the mutual understanding of US and European Community activities and programs related to biotechnology research.

Wikipedia. *Liaison:* V. Chandler (international). Wikipedia is a community-driven encyclopedia of human knowledge. Dr. Chandler will liaise with this project by scanning for opportunities for iPC faculty, staff, grand challenge team members, as well as members of the education, outreach and social science teams to author and edit relevant Wikipedia articles that contain strategically-placed references to iPC resources. This will help build the iPC community and increase the project's visibility in the broad community of researchers, students, journalists and lay people.

Commercial Endeavours

Dole Foods. *Liaison:* S. Goff. Dole Foods is investing \$1B in nutritional biotech in Kannapolis NC and have announced an agreement with RedHat to provide open source bioinformatics support at the North Carolina Research Campus. Goff has contacts with David Murdock, owner of Dole Foods.

Syngenta. *Liaison:* S. Goff (national and international). Syngenta is a multinational agricultural biotech company with offices in both the United States and Europe. It is currently investing tens of millions of dollars into upgrading its informatics infrastructure, aimed at enhancing computational support for product research & development. Dr. Goff is a Senior Fellow at Syngenta and is currently the Science Advisor for the informatics infrastructure project.

This is an early list of collaborations and is not considered exhaustive. It is anticipated as the consortium progresses additional and have established a mechanism whereby liaisons are assigned. In addition to individual organization professional societies such as ASPB, IEED, and National Science Teachers associated will be handled through faculty participation in their respective societies.