

### Appendix A1. Key Leadership Personnel

**Richard Jorgensen** is Editor-in-Chief of the premier journal in plant biology, *The Plant Cell*, and has broad research experience in molecular genetics, population and evolutionary genetics, and epigenetics. Jorgensen's science management experience includes 4 years as Program Director in the biotech industry (\$2.4M research program, 1986-1990, and creation of a joint venture company). Jorgensen was also lead PI of a \$10.5M Plant Genome Research Program grant (10 co-PI's at 7 institutions, 1999-2004). As Editor in Chief he manages and mentors >30 Coeditors, each a leading academic active in plant biology research, as well as several journal staff members. This position also involves numerous interactions with authors, editors, staff and members of the plant biology community, including the necessity of resolving sensitive conflicts fairly and successfully.

**Gregory Andrews** does research on parallel and distributed computing. He has written a widely used book on multithreaded, parallel, and distributed programming. He has been PI on two large computing research infrastructure grants from the CISE directorate and was Co-PI on two other computing infrastructure grants. He recently served for two years as a division director at NSF; while at NSF he represented the CISE directorate on the NSF-wide Cyberinfrastructure Working Group.

**Vicki Chandler** is Director of the UA BIO5 Institute and responsible for providing space and administrative support for iPC. As a cost share she will provide salaries for the Meeting Coordinator and administrative assistants as well as the Educator. She has extensive experience in plant genetics and molecular biology, including managing large plant genomics projects that provide important community services. Dr. Chandler is responsible for the vision that created BIO5 by bringing together biologists, computer and information scientists, and statisticians to work together. The iPlant Collaborative is quintessentially a BIO5 activity and the two will be strongly synergistic.

**Lincoln Stein**, MD/PhD, is professor of bioinformatics at CSHL with over 15 years experience in bioinformatics, computational biology and biological software development. He is lead PI of Gramene ([www.gramene.org](http://www.gramene.org)), a comparative mapping database for cereals and related plants, co-PI of WormBase ([www.wormbase.org](http://www.wormbase.org)), the model organism genome database for *C. elegans*, lead PI of Reactome ([www.reactome.org](http://www.reactome.org)) a curated knowledgebase of biological pathways in human, and the lead PI for the Generic Model Organism Database project ([www.gmod.org](http://www.gmod.org)) a distributed effort to create open-source software for model organism databases. He also directs the data coordinating center for the International HapMap Project ([www.hapmap.org](http://www.hapmap.org)). Stein was director of informatics at the Whitehead Institute/MIT Center for Genome Research, now the Broad Institute. Stein will be coordinating the efforts of the CSHL component of the Information Solutions team and will participate in grand challenge research questions related to genetic diversity and genome biology. He will also supervise the development of biology "mash-up" applications.

**Sudha Ram**, McClelland Professor of MIS and Director of the Advanced Database Research Group at UA, has extensive experience with enterprise data management, large scale information systems design and developments, and semantic integration of heterogeneous biological databases using machine learning, statistical approaches, ontologies and conceptual modeling. She is the developer of the SCROL ontology and CREAM software system for integrating multiple heterogeneous databases. In the area of biological database integration, she has developed ontology based mechanisms to dynamically link multiple gene, protein, and functional databases which have been embedded into software tools. She has also developed the W7 model for data provenance which is being widely adopted by industry and implemented in software for enterprise data management. She also brings experience in digital archiving, preservation and information life cycle management and will play a lead role in bridging the gap between user requirements and technical design of the iPlant system.

**Kobus Barnard** works on integrating computer vision and machine learning research into the modeling of scientific data. He is interested in automated approaches for learning semantically viable representations from data and linking such information across multiple modalities, e.g., inferring geometric models for an organism that provides quantified morphology linked to environmental

variables or gene expression data. His interest in the PSCIC project is to enable rapid adoption of current computational thinking emerging from the computer vision and machine learning communities to help address important grand challenge problems in plant sciences. He adds experience to the oversight committee in those domains, and connections to those communities, and also brings significant experience, from the computational scientists viewpoint, of interdisciplinary thinking, having worked with a number of life scientists on multiple collaborative projects.

**Susan Brown** has conducted research in the area of adoption and diffusion of IT in a variety of contexts. In addition, she has studied the use of technology to support communication, collaboration, and knowledge exchange. Her research includes the collection and analysis of social network data with regard to these systems. Her particular contribution to this project focuses on an understanding of the issues that lie at the intersection of the technical and social systems.

**Brian Enquist** is an internationally known expert in biological scaling, ecosystems, and physiological ecology with an extensive publication record. He will act as liaison to NCEAS.

**Stephen Goff** will act as project liaison to industry. He has extensive experience in molecular and cellular biology in plants and animals. His focus for 15 years in the biotechnology industry has been gene discovery & function. He was Director of Genome Technology at the Torrey Mesa Research Institute, a subsidiary of Syngenta, where he initiated and led a large effort to build up genomics technologies to better understand model and crop plants, resulting in the publication of the genome sequence of rice (along with other groups) in early 2002. Since 2003, he has been a Senior Syngenta Fellow and Senior Technical Analyst working with Business Development at Syngenta Biotechnology. His effort is focused on new business and scientific opportunities at the intersection of plant and animal biology, with a special interest and focus on cyberinfrastructure development.

**Barbara Heath** is the managing member and lead consultant for East Main Educational Consulting, LLC. Her company offers services in evaluation, technology, and staff development in the areas of science and mathematics. Dr. Heath has a Ph.D. in Science Education (physics) from NC State University. Current and past evaluation projects include three funded by the NSF, two funded through the State of North Carolina, and two funded by other public agencies.

**Nirav Merchant** is Director of the Biotechnology Computing Facility, a campus wide core facility with primary focus on facilitating the inclusion of novel computational methods and techniques at various stages of the discovery process across all life sciences. BCF provides turnkey computational and storage infrastructure services to over 150 researchers and 200 students per year, with hands on training, special topics workshops and custom application development. His overall responsibilities include working closely with interdisciplinary teams to translate research concepts to production level tools, applications and solutions. His team is well versed in the use of high performance computing applications and resources to address research needs for various life science disciplines.

**Carolyn Napoli** is PI of a Plant Genome Research Program grant (DBI-0421679) in support of a chromatin database (ChromDB; [www.chromdb.org](http://www.chromdb.org)). The database displays chromatin proteins for a diverse group of plants and animal species. Her primary research focus is on the fundamental mechanisms of gene regulation by chromatin and the evolutionary diversification of the chromatin proteome in plants, animals, and fungi. Her research expertise includes bioinformatics, database management, plant developmental biology, molecular genetics, genetics and epigenetics. Dr. Napoli manages a summer program for high school students to undertake independent research projects.

**Steve Rounsley** has a decade of bioinformatics and genomics experience gained at world-class genome centers and in the agricultural biotechnology industry, working with teams of biologists and software developers, and has often served the role of translator between these different fields. This diverse background will be particularly important in his role on the iPlant Oversight Committee.

**Michael Sanderson's** role is to provide expertise on computational and informatics aspects of phylogenetics and to act as liaison to NESCENT. In addition, as a systematist, he is familiar with taxonomic and nomenclatural issues arising commonly in biodiversity database activities.

**Richard Snodgrass** works primarily in temporal data management. He has written or edited six books, as well as journal and conference articles: conceptual design (4 papers), logical design (13), physical design (4), query language design (18), temporal algebras (3), implementation of temporal databases (21), temporal XML (5), and temporal databases in general (11). Recently, he has expanded more generally into scientific data management, including the representation of provenance of laboratory data and the dissemination of XML data, especially of such data that varies over time and whose schema varies over time, which are prominent features of biological data. His group is developing  $\tau$ XSchema, a compatible extension of the prominent schema language for XML that supports recovery of past versions, tracking changes, evaluating temporal queries, data versioning, schema versioning, and effective validation and dissemination of XML documents whose data and schema vary over time. Finally, he is actively studying lifecycle issues, specifically how to ensure that stable data is not accidentally or maliciously tampered with.

**Daniel Stanzione Jr.** is the founding Director of the Fulton High Performance Computing Institute (HPCI) at Arizona State Univ. The HPCI is the central hub for research computing at ASU and engages with almost 100 faculty across more than 20 disciplines dealing with large scale computational models and large volumes of data. In addition to providing supercomputing and storage facilities, Dr. Stanzione and his team work closely with researchers to aid in the scale up of computational models, as well as connecting multiple models across disciplines and connecting models to large scale, interactive visualization through ASU's Decision Theater (an immersive 3D visualization facility). Dr. Stanzione is a Co-Investigator with the Texas Advanced Computing Center in deploying the Ranger system, the first of the systems supported by NSF's Petascale Computing, and the machine that will be the largest open computing system in the world by the end of 2007. Dr. Stanzione and his team will support the national TeraGrid user community in accessing the machine and develop new training courses in petascale software development for this community. Dr. Stanzione's research includes a long-standing interest in the development of problem solving environments to support scientists in making use of high-end computing resources.

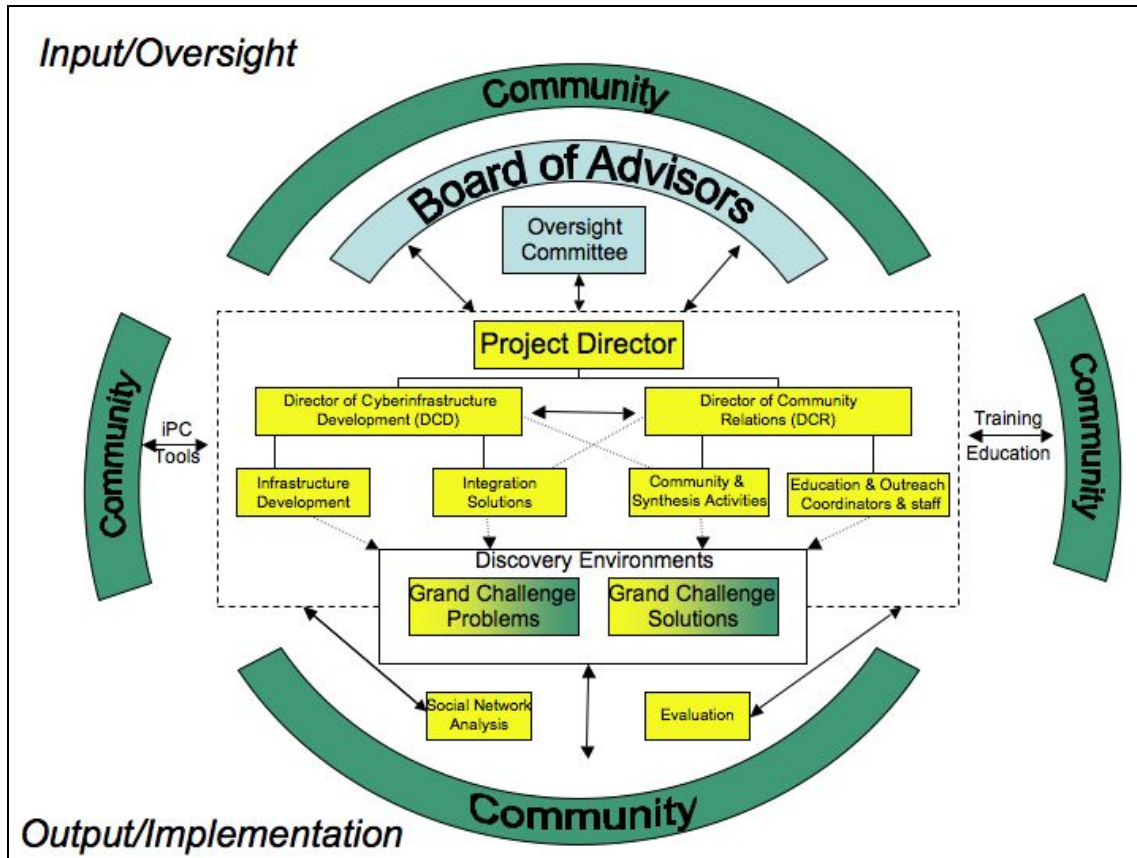
**Ann E. Stapleton** will oversee the design of policies, procedures, selection criteria and deliverables for small action teams (iPAT)--in close consultation with the PI and the project evaluation team. Dr. Stapleton will advise on policies to make sure that the community user perspective is considered at every step of the architecture design process, in coordination with the other project teams. Dr. Stapleton and project staff will organize iPATs, partnerships between plant biology, computer science, and quantitative analysis (statistics or mathematics) faculty and computer science and plant biology undergraduate and graduate students.

**Doreen Ware** is a research scientist at the USDA Agricultural Research Service and an adjunct assistant professor at CSHL. Ware has become the driving force behind the Gramene database and is responsible for most of the architectural and design innovations of that resource and is co-PI of that project. Ware is also co-PI on the maize genome sequencing project (in collaboration with the Washington Univ. Sequencing Center), responsible for genome annotation, coordinating the finishing activities, and data publication. She runs an active research program focused on the evolutionary biology of plant small RNAs. Prior to coming to CSHL, Ware was project manager for the Arabidopsis Biology Resource Center at Ohio State Univ. Ware will participate in grand challenge research questions related to plant development, genome biology, and the small RNA world.

**Suzanne Westbrook** is a Senior Lecturer and Associate Head in the Department of Computer Science, University of Arizona. Her research interests are in computer science education, particularly issues related to increasing the participation of women and other under-represented groups in computer science. Most recently she is pursuing ways of integrating concepts of computational thinking into education at all levels. Dr. Westbrook received her PhD from the University of Louisiana-Lafayette and has been teaching for 15 years.

## Appendix A2. Management Plan

The organizational structure, staffing and management of the *iPlant Collaborative* is designed to ensure that the organization is both open and responsive to the community, that it successfully promotes the development and use of computational thinking in plant biology, enables ground breaking research, and catalyzes community efforts to prepare a new generation of biologists. It is essential that project management be strong, central, and flexible. Strong management does not mean hierarchical, top-down management; rather, the best strong management firmly, but unselfishly guides, enables, and supports the collective activities of the organization and solicits, attracts, and ensures productive, intimate community involvement at all levels. The organizational structure of the *iPlant Collaborative* is illustrated below.



Management is the responsibility of the **Project Director (PI)**, Rich Jorgensen, who expects to spend at least 50% time on this project. He will ensure meeting the long-term project objectives through extensive interactions with the community, the Board of Advisors, the faculty Oversight Committee and NSF. He is a recognized leader with proven managerial skills leading and managing teams to achieve shared goals. His broad familiarity with much of the necessary and relevant range of science and technology is demonstrated by his role as Editor-in-Chief of the premier journal in plant biology, *The Plant Cell*, which publishes primary research in molecular and cellular biology, development, biochemistry, genetics, and evolution, as well as his broad research experience in molecular genetics, population and evolutionary genetics, and epigenetics, and his experience in the plant genetic engineering industry. Jorgensen's science management experience includes 4 years as a Program Director in the biotech industry, managing up to 15 research scientists. Jorgensen was also PI of a 5 year, \$10.5M Plant Genome Research Program grant involving 10 co-PI's at 7 institutions for the generation of research materials for the community (1999-2004). His role as Editor in Chief of *The Plant Cell* (2003-present) involves setting journal policies and managing and mentoring over

30 Coeditors, each a leading academic active in plant biology research, located in a dozen countries on 4 continents, as well as several journal staff members based in Maryland. This position involves extensive interaction with the plant biology community, including resolving sensitive conflicts fairly and successfully. Upon funding of this project, Dr. Jorgensen will develop a transition plan to pass his responsibilities as Editor in Chief of *The Plant Cell* to a new EiC prior to the normal end of his term in June, 2008. ASPB is currently interviewing EiC candidates and will select the next EiC in July, 2007. It should be possible to formally transfer responsibility by January 1, 2008.

The day-to-day activities of the project will be managed jointly by a **Director of Cyberinfrastructure Development** and a **Director of Community Relations**, both reporting to the Project Director and responding to the faculty Oversight Committee through the Project Director and the Executive Committee, which is a subset of the Oversight Committee. The Director of Cyberinfrastructure Development (DCD) is similar to a Chief Technology Officer. The DCD will be responsible for core Infrastructure Development and Integrated Solutions and will manage the core IT staff of database administrator, systems administrators, programmers, and software developers. This is a key position, requiring a national search and a competitive salary. The Director of Community Relations (DCR) will be responsible for synthesis activities, training, education and outreach at UA, UNCW, CSHL, and ASU and will manage a small UA staff (which is entirely funded by the UA cost share), including a Symposia Coordinator, an Administrative Assistant, the Outreach Educator, and the LTC teaching materials developer. The DCR will coordinate and ensure success of conferences, workshops, GCP team interactions, iPATs (working with Ann Stapleton), community wikis, etc., and will assist the faculty Education and Outreach Coordinators (Westbrook and Napoli). This too is a key position, requiring a national search and competitive salary. Together the DCD and DCR are responsible for achieving project objectives and implementing policies set by the faculty Oversight Committee. Necessarily, there will be overlapping responsibilities because the Infrastructure Development, Integrated Solutions, and Community and Synthesis Teams must coordinate and integrate; thus, the two Directors will be jointly responsible for seeing that this happens smoothly. Qualifications for both positions include a PhD in a relevant area, or equivalent experience, and project management experience.

The **Oversight Committee** will meet regularly via videoconferencing with the Project Director, the DCD, the DCR, and other staff and faculty as appropriate (monthly or more often as needed). It is comprised of Co-PIs: Andrews, Ram, Stein, Chandler; and Senior/Key Personnel: Rounsley, Barnard, Napoli, Westbrook, Stapleton, Stanzione, Merchant, Ware, and Brown. Rounsley and Barnard will have the special responsibility of ensuring communication and fostering understanding between biologists and computer and information scientists. The core Management Team will consist of the PD, DCD, DCR and the Executive Committee (PI/PD and 4 co-PIs). An administrative assistant (also funded by UA cost share) will assist the Management Team. The central management and staff of the *iPC* will be housed physically in the BIO5 Institute (including the PD) and will be virtually linked to multiple, distributed labs, centers and user communities (see Appendix 5).

Two teams will provide analysis and evaluation of how well *iPC* is meeting its goals. The **Social Networking Analysis** team, led by Sue Brown will monitor adoption of the Discovery Environments. She will work with *iPC* staff, graduate students and community participants to analyze the social networks that develop and change dynamically over time. **External Evaluation** of the *iPC* will be carried out by Barbara Heath of East Main Educational Consulting. These teams will provide reports to the PD and Oversight Committee, which will be used to assess progress and suggest adjustments as needed for approval by the BoA.

The external **Board of Advisors** will be comprised of at least eight internationally recognized members, including at least 4 plant biologists, 3 in CISE, and 1 in education and outreach. Community input will be sought for selection of members who will be chosen in consultation with NSF. The BoA will ensure broad community input into all major aspects of the project, assisted by

the Management Team and Oversight Committee. It will review and approve important decisions regarding synthesis activities, such as symposia topics and participants, selection of Grand Challenge Problem teams, and the prioritization of major iPC efforts that serve these teams. We anticipate the BoA meeting quarterly during the first year with video and telephone conferencing between the physical meetings. Thereafter, we anticipate biannual meetings, with quarterly conference calls.

#### **Management and Oversight within Specific Project Areas**

The **Integrated Solutions team** will be overseen by co-PIs Stein and Ram who have considerable experience in biology, computer science, enterprise data management, bioinformatics, genomics and software engineering. Together, they will oversee the work of IS subteams. These subteams will be assembled as needed to help address questions identified by the Grand Challenge Problem teams. These subteams will consist of members from UA and CSHL initially and later expand to include members from other institutions. Each team will be supervised by one or two research faculty chosen from UA, CSHL, and other institutions, based on their expertise in data management, workflow management, algorithms, or other areas relevant to the grand challenge problem. Each subteam will also consist of 1-2 graduate students, 1-2 agile software developers, and 1-2 postdocs, to assist in formulating a solution to the GCP. We anticipate 4-5 such subteams functioning at any given time. These subteams will also develop prototype software to address the GCP and will then interface with the ID teams to complete the transfer into production quality software and data solutions. A subset of the external BoA will help review, select, and prioritize the projects assigned to the IS subteams in consultation with the IS team leaders. All subteams will meet at least twice a month to review progress, identify commonalities across problems and solutions, and learn from each other's activities and successes/failures.

The **Infrastructure Development team** (ID) will be overseen by co-PI Andrews, who has considerable experience in computer science and cyberinfrastructure, assisted by a subset of the Oversight Committee. The infrastructure team programmers, system administrators, and user support personnel will be directly managed by the DCD. The programmers and software developers will implement production versions of IS team projects, which are selected as described above. Decisions about major upgrades/enhancements of the physical infrastructure (processors, storage, interaction devices, software, etc.) will be made by the Oversight Committee after review and prioritization by the BoA. The PD will have discretionary authority to approve small items (<\$10K each).

The **Community and Synthesis Oversight Team** will be coordinated by PD Jorgensen, with assistance from the faculty Oversight Committee. Personnel will be managed by the DCR, whereas specific faculty will mentor and advise the students and postdocs working with one or more Grand Challenge teams. Conferences, symposia and workshops will be organized by the Community & Synthesis team with extensive community input. UA faculty from a diverse variety of disciplines are available to assist with symposia, to assist specific Grand Challenge teams, and to advise students and postdocs, e.g., Rod Wing, Karen Schumaker, Ramin Yadegari, Richard Snodgrass, Sandiway Fong, Walter Piegorsch, and David Galbraith, to name a few. It is anticipated that faculty advisors and students from other institutions will contribute significantly to these teams, and we have budgeted accordingly.

The **Education Coordinator** (Westbrook) will develop training programs locally and facilitate at the national levels, while the Outreach Coordinator (Napoli; 25% FTE) will organize K-12 and community interactions. Both will interact with working groups, visitors, and synthesis activities, working together with the DCR. We will also assemble an iPC Education Advisory Board selected from national experts, e.g., individuals from NSTA, CSTA, NEA, with community input regarding membership. The myriad of challenges facing K-12 teachers, who work under tight curricular and accountability guidelines and teach diverse student bodies, dictates that our national program be managed by an experienced educator cognizant of standards, school system, etc (employed at UA; 50% FTE; full time summer commitment).

### Appendix A3. Computational and Cyberinfrastructure Capabilities

The role of the iPlant Collaborative infrastructure is to enable the solution of grand challenge problems in plant sciences. The infrastructure will be successful if (1) it is heavily used by plant scientists and their computational collaborators, and (2) it leads to research results that would not have been possible without it. Three types of infrastructure are required:

- Hardware for computation, visualization, storage, and communication.
- Software tools that facilitate discovery and experimentation.
- Staff to develop the software tools, install and administer the hardware, and provide support to users of the facility.

Below we describe these components and explain how the infrastructure will be developed and managed so that it is kept at the leading edge of technologies required to solve grand challenge problems in plant science.

#### Hardware Infrastructure

The iPlant Collaborative hardware infrastructure will augment existing facilities at the University of Arizona (UA) and Arizona State University (ASU). (See Appendix A5 for a summary of existing capabilities.) Storage will be added to support ongoing development and experimentation and to provide a persistent, reliable, and effectively unbounded repository for plant science data. The repository will ensure that key data sets are preserved beyond the lifetime of the project that produced them. It will also support reproducibility of experimental results by archiving snapshots of experimental configurations—including all software and data that was used to generate a given set of published results. UA will serve as the primary storage provider for active experiments and databases, with ASU serving as a backup for reliability and a mirror site for frequently accessed items. ASU will serve as the primary site for long-term storage,<sup>1</sup> with UA serving as a secondary site for reliability and disaster mitigation. The research storage systems at ASU, which this project would extend, provide disk-to-disk mirroring between a primary and backup site located in different campus buildings, as well as offsite tape storage.

Computational facilities will support software development and the needs of scientists who are doing computational modeling, analysis, data discovery, and other computing-intensive experiments. Existing shared-memory multiprocessors will be augmented and new ones will be installed if/when they are needed; similarly, existing high-performance compute clusters and local grids will be augmented and new ones will be installed if/when they are needed. The majority of computational resources will be located at UA, with ASU providing extra capacity to cover peak demands. We will provide a software environment similar to the one used at national supercomputing centers (e.g. Teragrid sites) and will develop tools to allow for a seamless migration to these resources. We also anticipate adding specialized computational resources such as field-programmable gate arrays (FPGAs) when the need arises. Finally, we require numerous workstations and laptops to support core and visiting staff and scientists as well as database and web servers.

Scientists on grand-challenge teams require high-end graphics workstations to visualize and potentially steer experiments that process and generate massive amounts of data. This need will

---

<sup>1</sup> We are in discussions with Google about a partnership in which they would provide archival storage.

increase as more data sets are linked together and as newer technologies such as digital microscopy generate massive amounts of data. We anticipate that eventually users will need to be able to interact with data sets and experiments using immersive, virtual reality environments. A CAVE environment already exists at UA and is available to the project; we anticipate expanding and upgrading the facility as needed to support future iPC experimental needs. In addition, iPC users will have access to the Decision Theater virtual environment at ASU.

The existing network infrastructure at and between collaborative sites is sufficient to support most anticipated uses. However, we propose to upgrade the connection between UA and ASU from 2 Gb/sec to 8 Gb/sec to support the data archiving, synchronization, and mirroring capabilities described above. This requires a one-time upgrade of interconnect hardware at the two sites; the institutions will continue to cover recurring network access fees.

Our year-by-year development plan for the hardware infrastructure is summarized below. Because technology continues to change rapidly, and because we cannot yet know the detailed requirements of the various grand-challenge teams, the development plan will have to be updated on a regular basis. The Infrastructure/Integrated Solutions Team advisory committee will provide overall direction, the iPlant Collaborative management team will oversee implementation, and the Infrastructure Director will execute infrastructure development plans in concert with his/her staff. (See Appendix A2 for details.)

*Hardware Infrastructure: Year 1*

Storage: add 50 terabytes at UA and 16 terabytes at ASU (\$3K per Tb)  
Computation: workstations/laptops at all sites; multiprocessor upgrade at UA (\$150K)  
Interaction/visualization: graphics workstations and other high-end devices (\$200K)  
Network upgrade between UA and ASU: \$50K at each site

*Hardware Infrastructure: Year 2*

Storage: add 50 terabytes at UA (\$150K) and 16 terabytes at ASU (\$48K)  
Computation: cluster/grid upgrade at UA (\$200K)  
Interaction/visualization: virtual reality (CAVE) upgrade at UA (\$200K)

*Hardware Infrastructure: Each of Years 3-5*

Storage: add 100 terabytes at UA (\$300K) and 16 terabytes at ASU (\$48K)  
Computation: upgrade or install multiprocessors, clusters/grids, workstations (\$200K)  
Interaction/visualization: upgrade or install advanced interaction facilities (\$200K)

**Software and Personnel Infrastructure**

As described in the body of the proposal, the Integrated Solutions (IS) team is critical to the success of the iPlant Collaborative. Members of this team, working in collaboration with grand-challenge problem teams, will create the software systems that make datasets useful, manage workflow, and enable discovery. Consequently, the most critical part of the iPlant Collaborative cyber-infrastructure—and the most expensive—is the people engaged in software research and development and user support.

The software creation process will follow a three-stage pipeline: (1) conception and prototyping, (2) development, and (3) deployment. Some parts of the proposed iPlant Collaborative infrastructure are well enough understood that we can immediately begin development of production tools based on

existing software, such as tools developed by Stein’s group at Cold Spring Harbor Laboratory (CSHL). Other parts of the software infrastructure—such as Discovery Environments customized for particular Grand Challenge problems and facilities for archiving specialized datasets generated by Grand Challenge Teams—are less well understood and hence require research and creation of experimental prototypes before we can know exactly what to create in the way of production systems. The conception and prototyping stage of software creation will be carried out by teams of computing faculty and graduate students working in close collaboration with plant science faculty and postdocs. The development and deployment stages will be carried out by full-time software professionals who have the expertise and employment continuity required to create and maintain production systems, as well as staff who have experience in scaling software systems to very large systems.

Although we will be able to begin prototyping some basic discovery environments (e.g. genome annotation mashups) during the first year of the project, most of the heavy computer science, algorithmic, and statistical research will occur in the second year and beyond, as the grand challenge problem teams form. Therefore, we cannot predict exactly what faculty we will need to assist GCP teams and to work with plant biologists to prepare applications to use tera- and peta-scale resources. For this reason, the personnel budget contains “pool lines” for faculty summer support, postdocs, and graduate students. The RAs and postdocs will work in pairs or small teams on prototyping projects, with faculty supervision and engagement, to facilitate cross-fertilization of ideas. The selection of projects and faculty will be determined by the iPlant Collaborative management team in consultation with the grand-challenge teams and advisory committees. Many UA faculty and students will be involved, but project teams might wholly or partly involve external collaborators. For example, Manish Parashar of Rutgers and Gregor Von Laseewski of Argonne National Labs have expressed strong interest in peer-to-peer data exchange and autonomic execution workflow, respectively.

The core programming staff will both develop production systems that result from research prototypes and evaluate/install/modify existing software packages. In addition, the core staff will maintain all installed systems and handle user support questions related to systems they maintain. We have budgeted for a full-time Director of Cyberinfrastructure Development, a database system administrator, three system administrators, six software developers at UA, two scientific programmers, and additional staff at ASU and CSHL. While these people are being recruited, we will pay for the services of some existing people in our computing support organizations to cover the transition period.

We will also employ additional students and staff to provide a virtual “help desk” staffed by people who will answer email and phone queries and create web pages and user manuals. We have budgeted for one person in year 1, two people in years 2 and 3, and three people in years 4 and 5 of the project. All iPlant Collaborative computational sites (UA, ASU, and CSHL) will cooperate in providing user support, with each focusing on the facilities and software systems that it provides.

The following table summarizes the full-time professional staff that will manage and create the iPlant Collaborative infrastructure:

Director of Cyberinfrastructure Development (1 at UA)  
 Data Base Systems Administrator (1 at UA)  
 Systems Administrators (3 at UA, 0.5 at ASU)  
 Software Developers (6 at UA, 1.5 at ASU)  
 Scientific Programmers (2 at UA, 3 at CSHL)

### User Support (1 rising to 3 at UA)

Software developers at all sites will help provide user support for the systems they create and maintain. The UA and CSHL sites also request funds for postdocs, graduate students, and faculty support for the conceptualization and prototyping of new software infrastructure; see the budget and budget justification for details.

Funds are also requested in the budget for licenses for commercial software systems that are critical to the project. The most significant is the AVS system for data visualization, which costs \$45K for a 50-user license and has an annual renewal/maintenance cost of \$9K.

### Computing Capabilities of Collaborative Personnel

Three of the PIs on this proposal have significant computing expertise: Andrews is a Professor of Computer Science at the University of Arizona, Ram is a Professor of Management Information Systems at the University of Arizona, and Stein is a Professor of Genome Informatics and Bioinformatics at Cold Spring Harbor Laboratory. All will be intimately involved with the management and development of the infrastructure. Two Senior Personnel will also play key roles in helping define and manage the iPlant Collaborative infrastructure: Dan Stanzione, Director of the Fulton HPC Center at ASU, and Nirav Merchant, Director of the Biotechnology Computing Facility at UA. See Appendix A1 for descriptions of what these people bring to the Collaborative.

Many computing research faculty and graduate students at UA and external sites will be involved with the Infrastructure, IST, and Synthesis teams to conceive and prototype new software systems. The following people have been active in working groups that prepared this proposal and will participate if the project is funded:

UA Computer Science: Kobus Barnard (computer vision), Alon Efrat (algorithms), John Hartman (virtualization/storage systems), John Kececioğlu (computational biology), Bongki Moon (data mining/bioinformatics), Richard Snodgrass (database systems), and Beichuan Zhang (computer networks)

UA Electrical and Computer Engineering: Ali Akoglu (FPGAs), Salim Hariri (autonomic workflow/distributed systems), and Ahmed Louri (high-performance computing)

UA Management Information Systems: Daniel Zeng (query optimization/data mining)

UA Linguistics: Sandiway Fong (computational linguistics and text mining)

CSHL: Richard McCombie (genome assembly/robotics), Partha Mitra (computational modeling), Doreen Ware (whole-genome analysis and annotation), Michael Zhang (biological network analysis)

We have also discussed potential collaborations with researchers at the Penn State LionShare project, Argonne National Laboratory, the USC Information Sciences Institute, and Rutgers University. As mentioned, our expectation is that as grand-challenge problems are identified and their infrastructure needs are defined, we will bring into the iPlant Collaborative those who are best suited to helping develop that infrastructure from around the nation.

**Appendix A4: Data Access, Protection and Preservation Policies**

The mission of the iPlant Collaborative is to support investigation of grand challenge questions by the plant sciences community and to foster education and training for K-12, undergraduate, and graduate students by accessing, developing, and managing digital dataset collections (both physical and virtual) and disseminating these resources as widely as possible. All the data, software tools, and other resources will be made freely and publicly available under creative commons or applicable open source terms. In fulfilling this mission we propose to develop a policy document that will outline principles, guidelines and policies for access, protection and preservation. This document will be prepared in consultation with the BoA and technology transfer experts. We will strive to ensure that the collections are:

- (1) Suitable for appropriate use based on the types of users
- (2) Accompanied by adequate documentation and metadata to enable their use
- (3) Checked and validated for quality control and include provenance where possible
- (4) Cataloged according to community developed metadata standards
- (5) Discoverable and accessible using web services and open standards protocols
- (6) Registered with appropriate authoritative repositories or information clearing houses.
- (7) Accessible to the community under policies that maximize opportunities for their use and redistribution

We do not envision collecting new data in the iPC, but rather in providing new uses of data currently existing and to be generated by the community during the funding period. The iPC will use community data to develop new analytical models, new procedures for analyzing the data, new ways to link the data, and other tools. Researchers will submit and store their data to be accessed by the community thru the iPC. We will specify guidelines for depositing data either physically (as stored copies) or as links to remote locations, and develop data sharing agreements to clarify the obligations of the data providers and users as well as the intellectual property rights.

In some circumstances we will need to obtain access to community and individual investigator-owned data resources to drive grand challenge questions forward. This must maintain the correct attribution for the data and honor the owners' IP restrictions as they carry forward in derived data sets. We will develop guidelines and policies to address these acquisitions and their resulting reciprocity. Models for these guidelines are provided by scienceforge.net, bioforge and BIOS.

The iPC through its Grand Challenge Problem teams will create Discovery Environments (DE) and license them such that users cannot prevent others from using their improvements or additions to the DE. Using the "enabling technology" that is a DE, anyone can perform analyses, make discoveries that may be patentable and that they can use to make and commercialize products, but they cannot control the DE nor prevent others from using or improving upon their improvements to the DE.

As described in the proposal, we will ensure that the data, software tools, workflows, and other resources in iPC will be made accessible to the community through a variety of mechanisms including: (1) synthesis activities through the ComSOT (2) a web-based virtual community center, extending the iPC throughout the community; (3) iPATs (iPlant Action Teams) who will train users throughout the community (4) iPETs (iPlant Education Teams) for teaching and education including undergraduate- and minority-serving institutions, and (5) by partnering and developing synergistic, integrated ties with centers, such as the ecology synthesis center (NCEAS) and the evolution synthesis center (NESCent).

Software developed through the iPC will be available as open source tools unencumbered by restrictive licenses, using an open source license compliant with the principles enunciated by the Open Source Initiative (<http://opensource.osdir.com/>). Distribution of software tools will be based on the following general principles: (a) Documentation, data and other files will, wherever possible and practical, use open standards widely adopted or defined for the plant biology community. (b) All

software tools and data will have clear copyright statements indicating the ownership of the copyright on the output and the license under which it is being distributed if applicable. It is vitally important that everyone contributing to a tool is aware of the copyright for intellectual property incorporated in a tool, and the license governing its possible uses.

All tools and data will be archived in a repository with written documentation of the preservation strategy. If a software model or tool is to be patented, the development team will ensure that the existence of the patent in no way interferes with software use, modification or re-distribution under the chosen software license and that the software is licensed and publicly available, at no financial cost, for use by the general community. Software tool development teams will accept bug reports, patches, and feedback from contributors outside the tool team. We will enforce adequate testing of software modules to ensure quality assurance of the software tools. This is important to build trust in the community regarding quality and reliability of iPC tools. Where the software implements specified standards there will be testing for compliance with these standards. The testing approaches used for a software tool will vary according to the nature of the software being created, its intended uses, and the size of the software. All software tools developed in the iPC will state build-time and run-time dependencies inherent in the developed software and the licenses applicable to these dependencies: dependencies include the operating systems, compilers, development environments, web servers and similar software needed to build and/or run the software. Tool documentation will clearly state the licenses applicable to these dependencies. This allows parties interested in reusing software to determine which software they must also buy or license.

We will use version control systems to keep track of who changes files (often source code, but equally documents or data), when they change them, why they changed them and any bugs that may have been fixed by the change. These systems will also be used to determine changes between versions of the system; to assist in diagnosis of newly discovered bugs; and also to compile a list of all contributors for IP issues. We will retain intermediate versions of software by defining checkpoints for development. Projects with more than 2 institutions or 3 contributors will use a multi-user version control system. As far as possible we will retain the full change history and archive it as well. To implement these policies, we will standardize on an enterprise-wide version control system, such as CVS, Subversion or SourceSafe, and implement a standardized system for registering developers, tracking bug reports and feature requests, and allocating responsibility among teams of developers. Several possible implementations of such management systems exist, ranging from the free SourceForge framework (<http://www.sourceforge.net>), to the commercial Basecamp system (<http://everything.basecamp.com/>).

We will develop mechanisms for monitoring the storage, replication, and mirroring of data and software tools to offsite peer institutions. We will also adopt and extend current practices for information life cycle management (ILM) based on usage of data, provenance, and other factors. As the amount of data made available through the iPlant Collaborative grows over time, it will be necessary to develop mechanisms and policies to decide what data to retain, remove, or archive and when to do so. Efficient life cycle management will ensure that the right data is accessible easily when it is recovered and that obsolete data gets removed or refreshed. As a part of this effort we will constantly monitor and collect statistics on the usage of data and tools in the system, and conduct surveys to help determine the “value” of various tools and data, and other resources in iPC. These will provide input to ILM policies to be developed as a part of the overall effort. The value determined from these inputs will also help us prioritize the data, tools, models and other resources that will require long term preservation as the basis of our long term sustainability plan. We are also exploring a partnership with Google for long term archiving as discussed in Appendix A3.

We are committed to taking all the necessary precautions to ensure safety and protection of the data and software, including physical safety (e.g. from fire or other natural disasters), intruder safety, protection from theft, and digital safety (i.e., damage via electronic intrusion, viruses). We will also develop a plan for disaster recovery and remote backups to safeguard the data from unauthorized users and from damage from storage device or other hardware crashes. Disaster recovery activities

will include routine monitoring and implementing a high degree of storage redundancy based on the usage and importance rating of the data.

We will have a staff team, working with the Database Administrator, that will be responsible for maintaining, updating, and disseminating these access, protection and preservation policies on a regular basis. This team will also maintain a high level of awareness of the developments in preservation practices, technological developments, and procedures. We will also use the access, protection, and preservation policies widely adopted and encouraged by the Library of Congress. PI Ram has active research projects supported by the NSF and Library of Congress Digital Archiving and Preservation (DIGARCH) Program and will contribute her experience to this project. PI Stein's background in managing the Human International HapMap Project, Gramene and WormBase database gives him extensive relevant experience in the management and preservation of biological data sets.

In addition to following extant policies and procedures, as outlined above, we recognize that in some areas, such as provenance collection, intrusion detection and recovery, schema versioning and validation, and semantic data integration in general, the state-of-the-art is still in flux or is inadequate for the needs of such a diverse and extensive plant sciences dataset collection as proposed here. PIs Stein and Ram and senior contributor Snodgrass have interest and active research programs in these particular areas and will work with the GCP teams and other plant scientists to develop targeted solutions to support these needs. This will include the need to support reproducibility of experimental results by archiving snapshots of experimental configurations and all software and data that was used to generate a given set of published results. Key personnel Merchant's experience with full life cycle management of multi modal life sciences data sets through project ALICE (Automated Lifesciences Information Cataloging Environment) and coordination of data marshalling activities and underlying infrastructure for multiple global collaborators in National Geographic's Genographic project, provides the necessary diverse background needed to effectively and efficiently manage data sets and software for iPC.

An important concern for the iPC is long term sustainability and preservation of data, tools, services, and other resources. We expect to develop robust economic approaches to ensure survival of iPC beyond the grant period from NSF. These approaches will be based on: (a) A "Utility" model where the users of data, tools and services may be charged a fee based on their usage. This fee will be no more than that required to cover the basic costs of computing, transfer, storage and service needs (b) A "subscription" model in which participants will pay an annual access fee to use the data, tools, and other services of iPC. This model is similar to the 'academic licensing' contracts that major software and digital libraries charge educational institutions. The fee may also have different grades ranging from "bronze", "silver" or "gold" depending on the frequency of usage, types of data, tools, and services used by each participant as well as other parameters to be determined; or (c) a hybrid approach in which a user's subscription allows a certain amount of usage, after which a pay-as-you-go model kicks in. It might also be possible to make some level of service free to casual users.

We will develop and nurture partnerships with large companies such as Google, IBM, and EDS to support the basic maintenance, upgrade and replacement of obsolete hardware and software, and will explore the possibility of technology transfer to one or more non-profit organizations to support iPC resources over the long term. An alternative approach is to hand off all or a part of the services offered by the iPC to an existing commercial or non-profit organization at the end of the project period. To assist in developing and assessing these strategies, we will enlist the help of the top ranked entrepreneurship program in the Eller School of management at the University of Arizona. This program assists in developing business plans, getting venture capital and helping to establish startup companies using students and faculty teams. A special focus of this program has been the establishment of software services companies and non profit organizations. We will also involve the technology transfer offices of the participating institutions, as well as the Board of Advisors in these sustainability efforts.

## **Appendix A5. Institutional Capabilities**

Organizational leadership is described in the Management Plan (A2), and technical expertise in Key Personnel (A1) and Infrastructure (A3); thus, this section describes space, infrastructure and technologies, for synthesis, analysis and communication. iPC will leverage existing, well integrated and complimentary institutional resources at the University of Arizona (UA) and other collaborating institutions.

The interdisciplinary BIO5 institute is committing faculty offices and open office modules to house project management, educational and outreach management, conference management, the core infrastructure team and visiting researchers working on synthesis activities within its new building designed to promote interdisciplinary interactions (see letter of support). The Thomas W. Keating Bioresearch building houses 370 researchers from 13 departments and 6 colleges, who work within open laboratory wings, adjacent to faculty offices and open office modules organized into research neighborhoods. BIO5 has conference rooms equipped with video and audio conferencing along with gigabit speed connections to all desktops and wireless internet access. For multi site remote conferencing iPC will utilize the mobile Access Grid; smaller online group meetings, training and application sharing sessions will be conducted using cross platform web meeting systems (IBM Sametime & Adobe Connect), these tools allow web casting, recording and annotation of meetings for future reviews with eventual transformation into training material.

The BIO5 building also houses the Biotechnology Computing Facility (BCF) that has a well equipped data center; providing customized and virtualized storage and CPU resources to large scale bioinformatics projects on campus on a cost recovery basis. BCF infrastructure is tailored to promote remote team science, all tools and compute intensive applications are accessible remotely. Participating team members and applications are linked using presence awareness tools; allowing instant web based meetings and collaborations; complementing the data center is a data processing and analysis laboratory that houses desktop based tools and a molecular visualization setup. BCF staff provide assistance and training with analysis applications and domain specific topics; assisting with migration of applications and data to compute intensive environments, such as campus wide Condor/Condor-G grid, clusters and shared memory machines. With the Triesman Center BCF offers rapid 3D printing services to create hard copy models from complex 3D visualizations for easier interpretation. All of these resources and staff expertise will be available to iPC.

BCF works cooperatively with the centralized Research Computing (RC) group and utilizes their data center and facilities to avoid duplication of resources; these include shared memory machines (SGI Altix 4700); “interactive supercomputing” connectors that allows popular desktop based applications like MATLAB, R etc. to transparently utilize a cluster for backend processing for rapid and interactive analysis; CAVE visualization environment and experienced staff for building complex visualizations. The “computing condo” arrangement at BCF and RC will allow iPC to add dedicated compute and storage nodes to existing infrastructure without expensive upfront renovations or configuration and integration delays. Data centers at BCF and RC have adequate capacity to handle iPC growth. Network Technology Solutions (NTS) provides the campus wide networking and connections to the external Internet, Internet 2 and National Lambda Rail. NTS will assist iPC to ensure high level of performance and reliability for connections to participating institutions.

Learning Technology Center (LTC) provides access to tools and resources to develop instructional material, assessment tools and frameworks. LTC staff will provide necessary customization and improvement of tools used by iPC outreach and education teams. Synthesis meetings will be

streamed, recorded and distributed using expertise from LTC. UA is home to the internationally renowned Learning Games Initiative (LGI), a transdisciplinary, inter-institutional research group that studies, teaches with, and builds computer games for educational contexts and will work closely with iPC and LTC staff to develop innovative learning material.

The Fulton High Performance Computing Institute (HPCI) at Arizona State University (ASU) has state-of-the art machine rooms and over 2000 node linux cluster and large scale storage capabilities; through a partnership with the Texas Advanced Computing Center, the HPCI participates in the support of the NSF TeraGrid, and aids users in making the transition to the use of national-scale resources. The Decision Theater at ASU is an immersive environment setup to aid groups with visualization and collaborative decision making and will complement the CAVE setup at UA. The compute infrastructure resources at CSHL and University of North Carolina Wilmington are not described herein as neither site will participate in computing infrastructure.

CSHL has three meeting venues. The main venue is Grace Auditorium, a 365-seat facility equipped with state of the art audio and visual technologies including the ability to capture and stream meeting proceedings campus-wide and to the internet. Smaller meetings (10-25 attendees) are held at the Banbury Center, located approximately five miles away on a fifty acre estate in the adjacent Village of Lloyd Harbor. Meetings of an intermediate size are held at the Woodbury Genome Research Center; its 86-seat auditorium is specifically set up to allow for computer-oriented meetings and training sessions. In addition, the Dolan DNA Learning Center features a state of the art computer training lab with seating for 30 simultaneous participants

The University of Arizona is expected to assume management of Biosphere 2 facility and campus, summer of 2007, which is located 35 miles north of campus in an isolated and beautiful setting. Current plans involve the establishment of a new institute to initiate and carry out interdisciplinary programs addressing grand challenges. The Director of the Institute, Dr. Pierre Meystre is enthusiastic about iPC using its conference facilities. The campus includes three conference rooms that can seat from 40 to 120 participants, a suite of 36 dual occupancy offices, and modern housing facilities in a “village” of 28 furnished two- to four-bedroom casitas with fully equipped kitchens. The campus is fully networked, with fast (cap I) Internet access to UA libraries. It provides ideal facilities for short-term retreats, summer schools and conferences, as well as months-long programs.