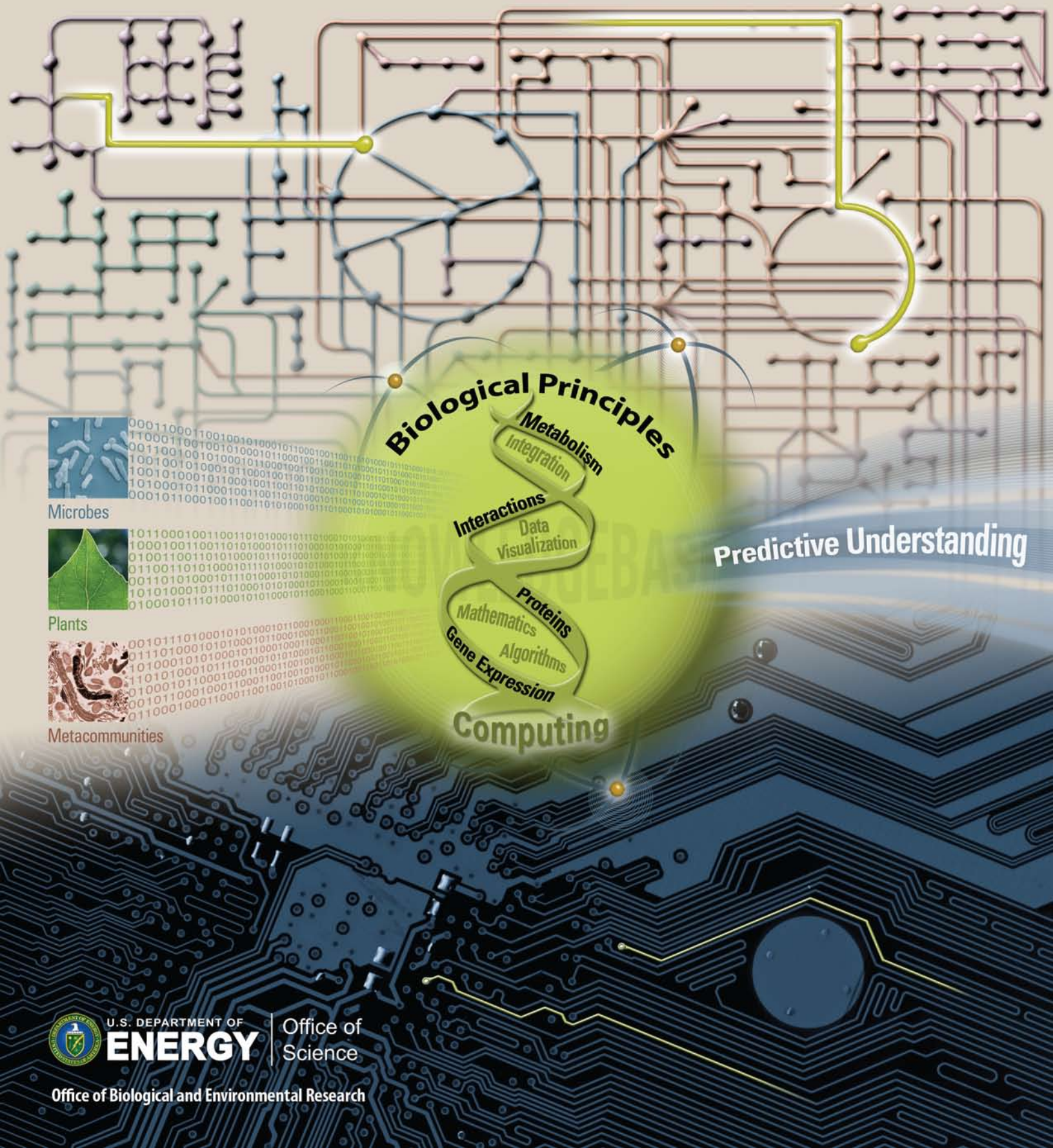
DOE/SC-0134

Executive Summary

DOE Systems Biology Knowledgebase Implementation Plan



U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

Executive Summary

DOE Systems Biology Knowledgebase Implementation Plan

As part of the U.S. Department of Energy's (DOE) Office of Science, the Office of Biological and Environmental Research (BER) supports fundamental research and technology development aimed at achieving predictive, systems-level understanding of complex biological and environmental systems to advance DOE missions in energy, climate, and environment.

DOE Contact

Susan Gregurick
301.903.7672, susan.gregurick@science.doe.gov

Office of Biological and Environmental Research
U.S. Department of Energy Office of Science

---

www.science.doe.gov/Program_Offices/BER.htm

Acknowledgements

The DOE Office of Biological and Environmental Research appreciates the vision and leadership exhibited by Bob Cottingham and Brian Davison (both from Oak Ridge National Laboratory) over the past year to conceptualize and guide the effort to create the DOE Systems Biology Knowledgebase Implementation Plan. Furthermore, we are grateful for the valuable contributions from about 300 members of the scientific community to organize, participate in, and provide the intellectual output of 5 workshops, which culminated with the implementation plan. The plan was rendered into its current form by the efforts of the Biological and Environmental Research Information System (Oak Ridge National Laboratory).

The implementation plan is available via

- **www.genomicscience.energy.gov/compbio/**
- **www.science.doe.gov/ober/BER_workshops.html**
- **www.systemsbiologyknowledgebase.org**

Suggested citation for implementation plan: U.S. DOE. 2010. *DOE Systems Biology Knowledgebase Implementation Plan*. U.S. Department of Energy Office of Science (www.genomicscience.energy.gov/compbio/).

DOE/SC-0134

Executive Summary

DOE Systems Biology Knowledgebase Implementation Plan

September 30, 2010

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

The document is available via genomicscience.energy.gov/compbio/.

Table of Contents

DOE Systems Biology Knowledgebase Workshops and Organizers

- **Using Clouds for Parallel Computations in Systems Biology. Nov. 16, 2009, at the Supercomputing conference in Portland, Oregon.**
  [Co-organizers: Folker Meyer, Argonne National Laboratory (ANL); Susan Gregurick, U.S. Department of Energy (DOE); Peg Folta, Lawrence Livermore National Laboratory; Bob Cottingham, Oak Ridge National Laboratory (ORNL); and Elizabeth Glass, ANL]
- **Plant Genomics Knowledgebase Workshop. Convened jointly by the U.S. Department of Agriculture (USDA) and the U.S. Department of Energy (DOE) on Jan. 8, 2010, at the Plant and Animal Genome conference in San Diego.**
  [Co-organizers: Catherine Ronning, DOE; Susan Gregurick, DOE; Ed Kaleikau, USDA; Gera Jochum, USDA; and Bob Cottingham, ORNL]
- **DOE Genomic Science Microbial Systems Biology Knowledgebase Workshop. Feb. 9–10, 2010, at the Genomic Science Awardee Workshop VIII and Knowledgebase Workshop in Crystal City, Virginia.**
  [Co-organizers: Susan Gregurick, DOE, and Bob Cottingham, ORNL. Co-chairs: Adam Arkin, Lawrence Berkeley National Laboratory (LBNL), and Robert Kelly, North Carolina State University]
- **DOE Systems Biology Knowledgebase Workshop at the 5th Annual DOE Joint Genome Institute (JGI) User Meeting. March 23, 2010, in Walnut Creek, California.**
  [Co-organizers: Susan Gregurick, DOE, and Bob Cottingham, ORNL. Co-chairs: Victor Markowitz, DOE JGI and LBNL, and Jill Banfield, University of California, Berkeley]
- **Knowledgebase System Development Workshop. June 1–3, 2010, in Crystal City, Virginia.**
  [Co-organizers: Susan Gregurick, DOE; Bob Cottingham, ORNL; and Brian Davison, ORNL]

These reports are available at www.systemsbiologyknowledgebase.org.

Executive Summary

A knowledgebase is a cyberinfrastructure consisting of a collection of data, organizational methods, standards, analysis tools, and interfaces representing a body of knowledge. Driven by the ever-increasing wealth of data resulting from new generations of genomics-based technologies, systems biology is demanding a computational environment for comparing and integrating large, heterogeneous datasets and using this information to develop predictive models. As a leader in systems biology research, the Genomic Science program of the Office of Biological and Environmental Research (BER), within the DOE Office of Science, supports scientific research that seeks to achieve a predictive understanding of microbial and plant systems relevant to DOE missions (genomicscience.energy.gov). By revealing the genetic blueprints and fundamental principles that control the biological functions of these systems, the Genomic Science program advances the foundational knowledge underlying biological approaches to producing biofuels, sequestering carbon in terrestrial ecosystems, and cleaning up contaminated environments. To serve the research community and address the Genomic Science program's data-intensive computing needs, this document summarizes the initial plan for creating a knowledgebase for systems biology.

As an open, computational environment for sharing and integrating diverse biological data types, accessing and developing software for data analysis, and providing resources for modeling and simulation, the DOE Systems Biology Knowledgebase (also called Kbase) will support a cultural change in biology from a focus on individual project-based efforts to open community science. The Knowledgebase would differ from current informatics efforts by bringing together the research products from many different projects and laboratories to create a comprehensive cyberinfrastructure focused on DOE scientific objectives in microbial, plant, and metacommunity (complex communities of organisms) research.

By democratizing access to data and computational resources, the Knowledgebase will enable any laboratory or project, regardless of size, to participate in a transformative community-wide effort for advancing systems biology and accelerating the pace toward predictive biology (see Fig. ES.1, below). Thus, the Knowledgebase will facilitate building a broader scientific community that will contribute to the fundamental science underlying DOE missions.

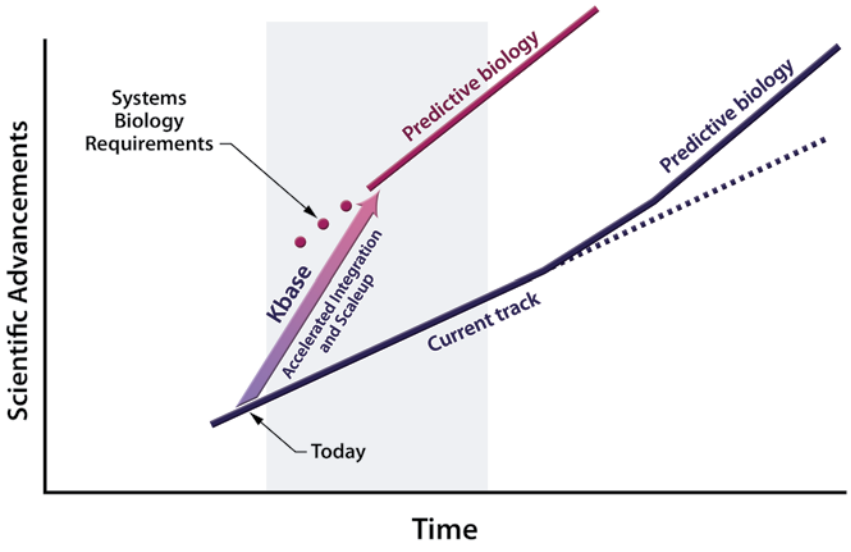


Fig. ES.1. A Faster Track to Predictive Biology. Knowledgebase-enabled integration of experimental data with models will accelerate the scientific advancements needed to improve inferences and achieve predictive biology. Building on the wealth of data being generated across many laboratories, the Knowledgebase will put biology on a new trajectory within the next decade. Scientists will hone their knowledge as they obtain answers to entirely new and more difficult generations of questions.

Systems Biology Knowledgebase Vision and Principles

The vision and justification for the Systems Biology Knowledgebase were defined in a May 2008 workshop report.[1] A key outcome for the many capabilities envisioned for the Knowledgebase (see sidebar, this page) is attaining more accurate models of dynamic cellular systems for microbes and plants. This requires a computational environment designed to support the iterative cycling of experimental design, analysis and integration of high-volume data, and modeling and simulation. As models of these cellular systems improve, they will address a progression of increasingly complex problems to help us understand and predict how these systems behave within a community of cells and organisms interacting with their environment. Ultimately, the Knowledgebase will allow users to perturb a biological system *in silico* (using "virtual experiments" on computer systems) and observe a predicted result.

By facilitating the efficient sharing of data, knowledge, best practices, and tools for rapidly developing and deploying applications for systems biology, the Knowledgebase will reduce the duplicative effort of individually establishing and maintaining similar resources for hundreds of laboratories and databases. Thus, researchers could direct more effort to scientific discovery. This open sharing and leveraging of the products from publicly funded scientific work will catalyze multidisciplinary collaborations and maximize the use and benefit of experimental results, analytical software, and modeling tools generated throughout the entire research community.

To provide the diverse capabilities envisioned for the Knowledgebase, infrastructural components will be distributed across many locations. Knowledgebase coordination, however, will be centralized and based on the following principles guiding development and operation:

- Provide open access to data, open contribution, and open-source software development—to the greatest extent possible—while simultaneously respecting a reasonable level of protection and temporary embargoes to allow publication and career development.
- Engage key stakeholders in developing the Knowledgebase, defining metrics for success, and assessing Knowledgebase performance in meeting the needs of the communities it serves.
- Support high-level policies (e.g., establishing standards for usability, interoperability, and contribution) recommended by a community-based Governance Board. This Governance Board combines features of a user advisory board and a scientific advisory board. Executive decisions (such as specifics on implementation) should be made by Project Management working closely with DOE management and the stakeholder community.

Capabilities Envisioned for the DOE Systems Biology Knowledgebase

- Curation of data, models, and representations of scientific concepts.
- Analysis (including method comparison) and inventory of results.
- Simulations and model modifications and improvements.
- Prediction-based simulation and analysis to form new hypotheses.
- Experimental design and comparison between predictions and results.

Community-Developed Plan for Knowledgebase Implementation

Building on the vision defined in the May 2008 workshop, the cumulative output of community participants in a series of five DOE-sponsored workshops[2] established the conceptual design, workflows, scope, and science to be addressed by the initial implementation of the DOE Systems Biology

[1] The May 2008 workshop report, *Systems Biology Knowledgebase for a New Era in Biology*, is available online (genomicscience.energy.gov/compbio/).

[2] Workshops are listed on p. ii.

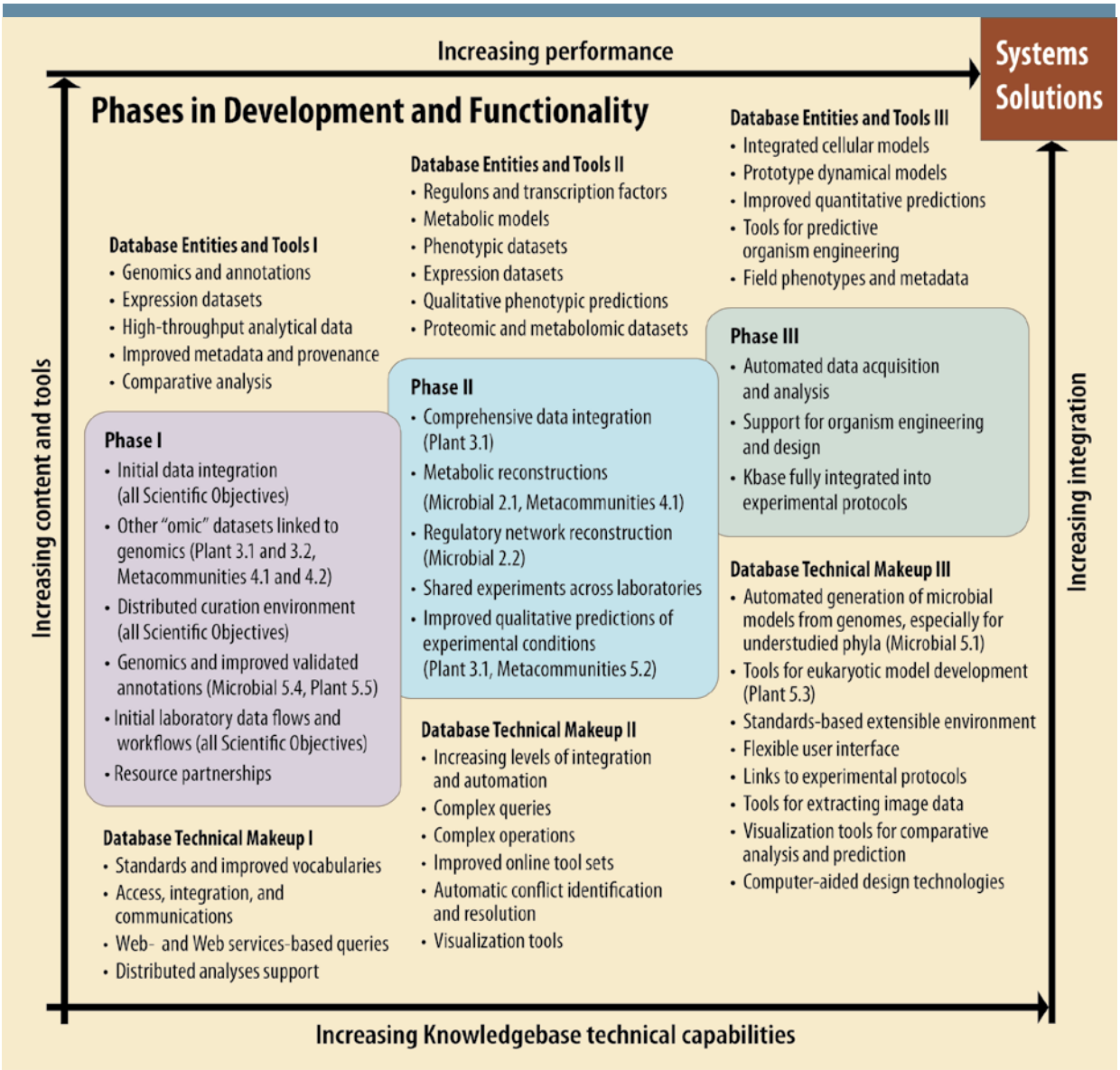


Fig. ES.2. Phases in Development and Functionality in the DOE Systems Biology Knowledgebase. This table shows three phases of technology development from less to more mature (lower left to upper right, respectively). The state of technology development for the biological systems (microbes, plants, and metacommunities) addressed in the implementation plan is in different stages of maturity. The notations in parentheses (e.g., Microbial 2.1) refer to the Science Area listed in Table ES.1, p. 6. Technologies for microbial research and analysis currently are well into Phase I. Upon implementing this plan, the Microbial Scientific Objectives will move fully into Phase II. Though technologies are less mature for plant and metacommunities, deploying the implementation plan will result in substantial progress in Phase I and Phase II. [Updated from page vii in the 2008 workshop report, *Systems Biology Knowledgebase for a New Era in Biology*.]

Knowledgebase. Technologies in computer science, bioinformatics, and data management are available now to begin the transformation of the Knowledgebase vision into reality by creating an adaptable computational environment designed for expansion and modification over the coming decade. Therefore, the Knowledgebase will be implemented in phases characterized by progressively increasing functionality (see Fig. ES.2, this page). The 3 to 5 years covered in the implementation plan will move the community from Phase I into Phase II.

One clear consensus among workshop participants is that the Knowledgebase initially should target and achieve success in specific, focused scientific objectives that were identified, developed, and prioritized

as near-, mid-, and long-term needs at the workshops. Near-term priorities were described in the greatest detail, with progressively fewer details given for the other objectives. To define the core scientific objectives, workshop participants discussed and identified the key research goals that need to be solved for three science areas relevant to DOE systems biology: microbes, plants, and metacommunities. For the six near-term scientific objectives that were identified as priorities for the Knowledgebase, representatives from the biological and computational communities worked together to translate the objectives into experiment workflows, computing system requirements, and detailed implementation plans specifying the tasks, outcomes, and integrating infrastructure needed to accomplish the objectives (see Fig. ES.3, this page). Additional objectives describe mid-term science and leveraged annotation needs that will be addressed over the coming decade.

Community involvement is critical for the success of this effort. Many consider achieving community "buy-in" for the Knowledgebase as important as overcoming the technical challenges faced in developing a community infrastructure. This powerful commitment to engaging the community is reflected in the valuable contributions from about 300 scientists who participated in the workshops culminating in the DOE Systems Biology Implementation Plan. This level of community involvement will need to continue as planning for the Knowledgebase transitions to its implementation,

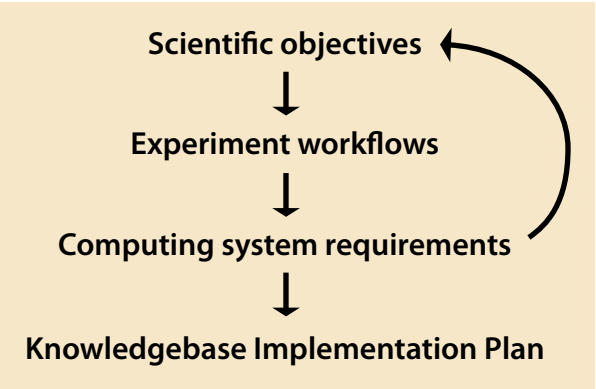


Fig. ES.3. Process for Community Development of Knowledgebase Implementation Plan.

which will require the expertise and skills from many different groups within the scientific community. Broadly, these groups represent plant and microbial researchers who design experiments and generate data; computational biologists and bioinformaticians who will develop the analysis methods and simulations that help interpret the data; and computer scientists, database developers, and software engineers who will develop the Knowledgebase infrastructure (see Fig. ES.4, p. 5).

DOE has a proven capability for linking strengths in biology and computational sciences in coordinated projects and programs. Genomic Science program collaborations involving experimental scientists, technology developers, and computational biologists have resulted in a deep understanding of specific microbes and microbial communities. In addition to advancing these ongoing efforts, the Knowledgebase will provide a unified framework for linking these different collaborations so that insights, workflows, and analytical programs resulting from these studies are more readily applied to investigations of more complex plant systems and metacommunities.

Knowledgebase Priorities and Scientific Objectives

The microbial, plant, and metacommunity science needs and objectives that will drive Knowledgebase development are listed in Table ES.1, p. 6 and described briefly in the following pages. The near- and mid-term science needs define the initial and immediate plans for the DOE Systems Biology Knowledgebase. Additional goals were identified for longer-term activities but were not further developed in detail for the implementation plan.

Microbial Sciences

In the microbial science area, the first objective is to improve the utility of metabolic network models, especially for microbes involved in biofuel production and bioremediation, so that metabolic engineering produces more predictable

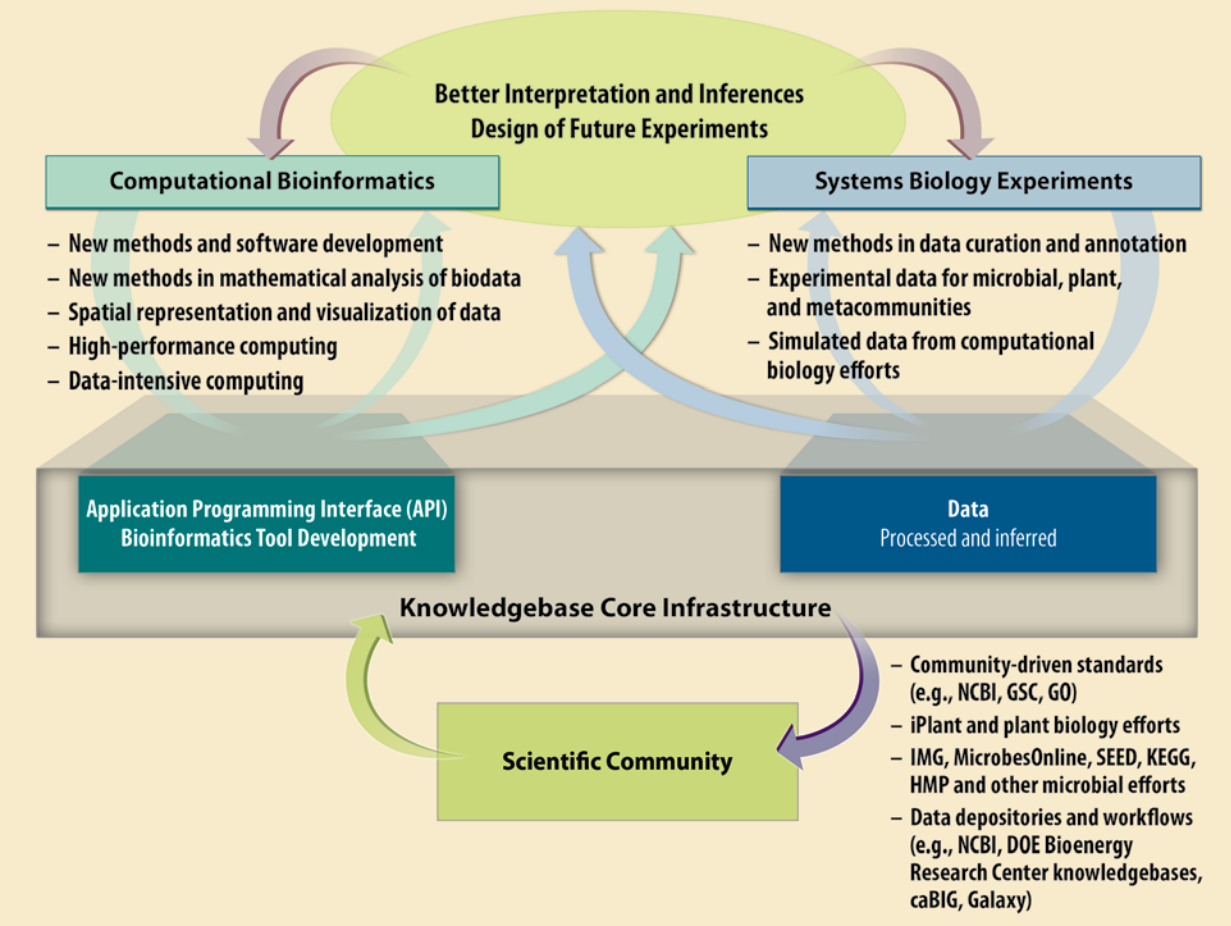


Fig. ES.4. Relationship Between the DOE Systems Biology Knowledgebase and the Larger Scientific Community.

results. The second objective is to enable automated inference of gene regulatory networks based on gene expression profiling data and then to validate inferred networks to improve prediction of cellular behavior and fitness.

Microbial Scientific Objective 1

Reconstruct and Predict Metabolic Networks to Manipulate Microbial Function

The scientific community seeks to understand and manipulate the metabolic potential of organisms in order to understand growth and phenotypes. More specifically, this objective involves reconstructing metabolic networks, predicting the growth of organisms from their metabolic networks, understanding organisms' metabolic potential, providing scientists with software tools to interrogate and visualize metabolic networks, and enabling engineers to quickly determine the strategies necessary to remodel metabolism for specific purposes. Objective 1 will increase the speed and automation of metabolic network reconstruction and comparison and improve the accuracy of metabolic network predictions. This knowledge will lead to the informed modification of specific enzymes or the introduction of entirely new pathways, allowing researchers to determine better strategies for manipulating mass or energy flow in microorganisms. Achieving this capability will require integrating new experimental data with existing data and models of metabolic pathways, as well as developing methods to automatically create new metabolic reconstructions from newly sequenced organisms.

Current research and development in metabolic networks primarily involve two approaches:

Table ES.1. Near-Term, Mid-Term, and Leveraged Annotation Needs Supported by the DOE Systems Biology Knowledgebase

(Section numbers in first column refer to main report)

Section	Science Area	Scientific Objective
Near-Term Science Needs		
2.1	Microbial	Reconstruct and Predict Metabolic Networks to Manipulate Microbial Function
2.2	Microbial	Define Microbial Gene Expression Regulatory Networks
3.1	Plant	Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype
3.2	Plant	Assemble Regulatory Omics Data for Target Plant Species in Common Platforms to Enable Analysis, Comparisons, and Modeling
4.1	Metacommunities	Model Metabolic Processes within Microbial Communities
4.2	Metacommunities	Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function
Mid-Term Science Needs		
5.1	Microbial	Analyze Understudied Microbial Phyla
5.2	Metacommunities	Interpret Metagenomic Data to Identify Conditions Required for Growth by Key Microbial Communities Relevant to DOE Missions
5.3	Plant	Construct, Simulate, and Validate Plant Life Models
Leveraged Annotation Needs		
5.4	Microbial	Integrate Descriptions and Annotations of Microbial Genomic Features
5.5	Plant	Improve Plant Genome Annotation Datasets and Make Them More Accessible

(1) evaluating novel microbes to identify and improve desired metabolic phenotypes (e.g., recent work on *Clostridium phytofermentans*) and (2) manipulating the metabolic pathways of well-characterized microbes to enable novel functionality (e.g., initiatives to engineer cyanobacteria for photosynthetic production of alkanes and isoprenoids and recent achievements in hydrocarbon production from *Escherichia coli*). Given DOE's interest in metabolic engineering for biofuel production and bioremediation, the development of sophisticated metabolic modeling methods and experimental data for a select set of DOE-relevant organisms is a high-priority, near-term objective.

Microbial Scientific Objective 2

Define Microbial Gene Expression Regulatory Networks

In response to dynamic and competitive environments, microbes must deploy the products of diverse gene sets to survive and prosper. Expression

of the correct sets of genes at the correct levels could confer the best competitive advantage given the organism's genetic complement and the current environment. The mechanisms within cells that sense the environment and determine which gene sets should be deployed at what levels, thereby coordinating different stages of the microbe's growth and development, are collectively called the gene regulatory network. Knowledge of this network is the foundation for predicting, controlling, and designing the behaviors of microbes and their community.

The first component of this objective is to enable automated inference of gene regulatory networks, relying principally on expression profiling data. The second is to extend these inferred networks to include additional data types, both to refine network predictions and to test them. Prioritization should be given to those organisms that are key to DOE missions, with a focus on regulatory paradigms of greatest relevance to the microbe in question.

This high-priority objective can achieve near-term goals, but completion may take 2 to 10 years. The advent of genomic technology and the availability of many microbial genomes enable the development of capabilities providing data and tools from which regulatory networks and their behaviors may be inferred rather than directly measured. The regulation of networks of interactions within and among microbes defines their ability to remediate environments, improve energy crop growth, process biomass into fuels, and sequester carbon, among other things.

Plant Sciences

The first objective in the plant science area is to establish the capability to predict changes in plant biomass properties caused by genetic or environmental changes. This predictive capability is based on the mining of data that reflects the complex relationships among the physical properties of plants, their genetic makeup, and the environment in which they are growing. The second objective is to develop the capability to organize and analyze data from regulatory omics (e.g., transcriptomics, proteomics, and other large-scale molecular analyses) to improve understanding of how plants regulate gene expression in key plant species relevant to DOE missions. This capability will be critical for understanding genes, their functions, and regulation and then using this understanding to engineer plant growth and development and, in particular, biomass accumulation.

Plant Scientific Objective 1

Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype

The Knowledgebase will provide computational infrastructure to support and contextualize experimental plant phenotype data to an extent that enables researchers to predict changes in the physical properties of biomass resulting from environmental change, genetic diversity, or manipulation. Achieving this goal depends on the creation of a robust semantic infrastructure for collecting, annotating, and storing diverse phenotypic and environmental datasets. These data include measurements such as photographic images and analytical spectra that capture visible phenotypes and chemotypes related to yield, physiological performance, and sustainability.

Subsequently, the Knowledgebase will be used for data mining and analysis to understand the genetics underpinning desirable plant biomass properties relevant to DOE missions (e.g., biomass yield, conversion efficiencies to biofuels, and the ability to sequester soil carbon or contaminants). Specifically, it will serve as a basis for software applications that extract, quantify, and catalog phenotypic features from the diverse datasets and relevant metadata (data describing the primary data generated from experiments or other analyses) for data mining and further analysis. Development of a robust, semantic infrastructure for plant phenotyping research is a high-level, mid-term objective that could be carried out in 3 to 5 years and in synergy with the ongoing efforts of the National Science Foundation (NSF) iPlant Collaborative. By providing the community with a comprehensive

collection of experimental and phenotypic data for plant feedstocks important to DOE, this objective will accelerate the development and redesign of feedstocks with plant architectures, cell-wall characteristics, and other properties that improve biofuel production and carbon biosequestration.

Plant Scientific Objective 2

Assemble Regulatory Omics Data for Target Plant Species in Common Platforms to Enable Analysis, Comparisons, and Modeling

Plant regulation is known to control key aspects of plant carbon allocation and partitioning, which are critical to biomass composition and soil carbon accumulation. Regulation is also a critical distinguishing characteristic between annuals and perennials and other aspects related to sustainability. To date, we have limited understanding of how plants regulate gene expression and how this is manifested in the cell.

Assembling regulatory omics data from plant biology into common platforms is essential to DOE's systems biology mission. This objective seeks to collect several key types of regulatory omics data and associated quality metadata for six target plant species: *Brachypodium*, *Chlamydomonas*, poplar, sorghum, switchgrass, and *Miscanthus*. The assembly begins with genomic and RNA expression data (from arrays or RNA-Seq) along with small RNA and target RNA information, differential RNA processing and decay information, epigenetic markers such as DNA methylation and histone modifications, as well as available proteomic data. In the near term (1 to 3 years), classical transcriptomic data (microarrays and mRNA-Seq) as well as small RNA and basic proteomic data will be assembled. These internal and external data will be publicly accessible with user-friendly web interfaces and downloadable for power users.

Metacommunities Science

The first objective in the metacommunities science area is to determine the metabolic role of each organism residing in a community and understand which community features provide adaptive robustness to environmental change. This information will lead to improved characterizations of microbial community physiology and ecology, which are necessary for designing strategies to accelerate or ameliorate microbial activity for environmental remediation or carbon sequestration. The second objective allows scientists to study microbial communities to discover novel functions and genes within these communities. Data generated in large-scale metagenomics projects can provide the information necessary to better understand the function of poorly characterized genes. The resulting data provide actionable hypotheses about the function of many genes that have yet to be studied in detail. Additionally, scientific efforts associated with this objective will lead to the discovery of new genes that perform useful biological functions of relevance to DOE priority areas such as energy production, carbon cycling and biosequestration, and environmental remediation.

Metacommunities Scientific Objective 1

Model Metabolic Processes within Microbial Communities

An overarching need for systems biology is to determine the metabolic role of each organism or key species residing in a community. This objective focuses specifically on modeling the metabolic processes within a microbial community, which requires developing metagenomics workflows and systems biology tools. In the near term, the Knowledgebase will develop workflows to analyze metagenomes and other data from microbial communities and leverage existing data and tools to create descriptive community metabolic models. The data and metadata will include the full range of current systems biology tools. Both top-down (metagenomics) and bottom-up (multispecies models) approaches were formulated for near- and mid-term goals. Eventually, these models will allow us to not only predict, but actively drive changes in the community in desired directions (e.g., accelerate environmental processes relevant to DOE missions, including environmental remediation, cellulose degradation, or carbon biosequestration).

The integration of different types of experimental measurements relating to metabolic activity is necessary for (1) generating hypotheses about the nature of interactions among community members and interactions between the community and the local environment, (2) generating hypotheses about the organisms and pathways responsible for the community's metabolic activities, and (3) predicting how the community will respond to environmental changes or to the introduction of new microorganisms. These proposed objectives will lead to improved characterizations of microbial community physiology; such characterizations are necessary for designing strategies to either accelerate biotransformation activity (e.g., uranium bioremediation) or ameliorate the outcome (e.g., acid mine drainage). Developments in microbial community understanding also will have direct benefits in understanding ecosystems and direct improvements in carbon cycling (and biosequestration in soils), as well as in biofeedstock production (via plant-associated microbial communities)—an area of immediate interest to DOE and the U.S. Department of Agriculture (USDA).

Metacommunities Scientific Objective 2

Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function

One reason to study microbial communities is to determine the novel functions and genes of organisms within these communities. Data generated in large-scale metagenomics projects can provide the information necessary to better understand the function of poorly characterized genes. As metagenomic (along with metatranscriptomic and metaproteomic) data are rapidly coming online, a critical scientific objective is the development of approaches for mining the data to identify previously unknown genes and for leveraging the wealth of metadata associated with metagenomic datasets. Information about gene-organism co-occurrence can help identify testable hypotheses about the function of newly identified or poorly characterized genes. Additionally, scientific efforts associated with this objective will lead to the discovery of new genes that perform useful biological functions of relevance to DOE priority research areas. Improvements in identifying unknown genes and their function will help to reduce potential error propagation in gene-calling databases.

Reliable functional annotations are critical prerequisites of a successful research program in systems biology. This objective will accelerate efforts aimed at characterizing the function of currently understudied genes. Additionally, the tools developed as part of this project will be valuable assets to scientists generating new datasets by allowing them to leverage Knowledgebase-associated datasets in the analysis process and to generate actionable hypotheses. A key element is to handle diverse types of associated metadata.

Further Scientific Goals: Mid-Term Science and Leveraged Annotation Needs

The community identified other desirable and achievable scientific goals to improve functionality of the Systems Biology Knowledgebase. Several feasible medium- and high-priority needs were considered important for the Knowledgebase. One of three mid-term scientific needs is to analyze understudied microbial phyla. The goal of this scientific objective is to understand the role of unclassifiable members of a microbial community in terms of genetic and phenotypic comparison. To achieve this objective, physiologic and metabolic datasets must be linked to metagenomic annotations to provide context and evidence. Another mid-term objective is metagenomic interpretation to identify conditions required for growth by key microbial communities relevant to DOE missions. This would improve our ability to cultivate (and isolate) target species from these communities. The third science need is to construct, simulate, and validate plant life models to enable semiautomated inference, construction, simulation, validation, and query of complex, multilevel (i.e., gene, protein, metabolite, small RNA, organelle, cell, and tissue) datasets. These plant life models would be used to

integrate and explore experimental data types collected during studies of plant feedstocks that impact bioenergy production and carbon cycling.

Two of the identified near-term science needs are for improving annotation of both microbes and plants—high-priority objectives that would be immediately leveraged by the Knowledgebase project. In addition, the increasing number of large and complex metagenomic sequence data (hundreds of gigabases for soil, for example) requires advances in algorithms for assembly. The DOE Joint Genome Institute (JGI) is the lead organization in primary sequencing and annotation for organisms of DOE and community interest. These organisms include microbes, plants, fungi, and microbial communities. The DOE JGI is pursuing and developing plans to improve its approaches for incorporating ongoing technology advancements. Programmatically, the DOE JGI would have the primary mission to develop and carry out implementation of improved annotation pipelines. The DOE JGI and the Knowledgebase will closely collaborate to reach these mutual goals.

Infrastructure and Architecture

The DOE Systems Biology Knowledgebase will be a large-scale system that:

- Makes massive amounts of biological data freely available to the scientific community, through hosted services and as links to external resources.
- Provides high-performance and scalable computational resources.
- Supports a large user community with tools and services to enable researchers to use the Knowledgebase.

To meet these requirements, the Knowledgebase must be designed with a highly elastic architecture that enables computer scalability on demand to meet the ever-changing computational requirements of scientific users. This elastic architecture must be supported by continual expansion and scaling to accommodate new data, computational platforms, and software innovations. The overall goal for the architecture is to support the creation of a broad-based, scalable Knowledgebase that provides a set of services to underlying data and computational resources. Decisions about the design and implementation of the architecture are critically important to the efficient and low-cost sustainability of the Knowledgebase. These decisions will be based on the following core set of architectural principles defined in the plan:

- **Open.** Provide the community with a published set of open-source application programming interfaces (APIs) to access Knowledgebase resources in an automatic fashion using software.
- **Extensible**. Enable the community to use the APIs to extend the capabilities of core Knowledgebase resources.
- **Federated.** Provide users with transparent access to a federation of physically distributed heterogeneous computational and data resources.
- **Integrated.** Create mechanisms to integrate existing databases and tools essential for the DOE systems biology community.
- **Exploit data locality.** Implement mechanisms for transparently moving requested analyses to execution sites that can best exploit data locality and provide maximum performance.
- **Modular.** Promote modular, component-based design for codes that can be readily connected to build pipelines for executing complex, multi-step analyses.
- **Scalable.** Expand Knowledgebase system architecture to accommodate increased use and functionality by transparently incorporating additional computational and storage resources.

The Knowledgebase infrastructure must be a rich collection of services and hardware. The problems faced by scientists require a variety of computing and data platforms and applications that do not fit nicely into a system based on a single hardware or software platform. Knowledgebase hardware and services include data repositories, data storage or data warehouses, data centers at multiple

locations, virtualization, data parallel processing on commodity hardware, cluster computing, and high-performance computing (HPC). Data and metadata representation and registries are other key aspects. This collection also will enable semantics-based searches of metadata (such as ancillary experimental data, ontologies, controlled vocabularies, and data models). With the inclusion of ESnet and Internet2 as the underlying network backbone, the Knowledgebase infrastructure is a cloud-based system providing a unique and valuable resource for biologists and offering these capabilities:

Platform as a Service. The Knowledgebase will provide a software platform for users to share, use, develop, and deploy bioinformatics applications that serve DOE systems biology. The platform will support users in exploiting the computational and data resources in the Knowledgebase cloud. By providing facilities that support the complete life cycle, from building to using Knowledgebase-enabled software, users can receive the full benefits of the Knowledgebase infrastructure.

Infrastructure as a Service. This will allow users to leverage the Knowledgebase hardware, thereby reducing local operational costs associated with purchasing, installing, and maintaining hardware as well as reducing the burden on local facilities to house the hardware. Advances in hardware virtualization now make it possible for users to create images of their local system that can be shared via the Knowledgebase with other users, enabling the replication of scientific results and the sharing of analysis environments.

Data as a Service. This will provide community data curation services and allow users to store, access, share, and curate heterogeneous data in the Knowledgebase, reducing the need to buy additional storage and to scale their existing infrastructure. Providing data services to the biological research community in a time when data accumulation rates are increasing exponentially will enable research scientists to spend more resources focusing on biological problems.

The primary architecture recommendation is for a layered architecture blueprint (see Fig. ES.5, p. 12). The four layers will consist of a user access layer, an infrastructure layer, a federation layer, and a layer of federated hardware resources. The foremost task for the Knowledgebase platform is to provide the scientific user with access to the underlying Knowledgebase-associated data, shielding the user from how that access is achieved (e.g., federated versus centralized, cloud-based versus central server). It also should provide the user with elementary analysis and visualization tools to apply to that data, mechanisms for storing intermediate results, standards for exchanging data between tools, and ways to connect analysis tools for creating *ad hoc* workflows. In addition, the platform should provide a low-threshold infrastructure for tool development, reuse, and dissemination.

The implementation plan recommends that the Knowledgebase initially consist of up to seven data centers on ESnet, upgraded to interconnections at 100 gigabytes (GB), each with petabyte (PB) storage. The storage is expected to double every 2 years. Each scientific data center would be associated with one of the six scientific objectives, and one data center would focus on coordinating the Knowledgebase core infrastructure development. Coordination of infrastructure development would be replicated as required to the other locations. Co-locating and sharing computational resources at some data centers, if possible, also are recommended. Compute clusters for virtualization could be co-located with compute clusters that support data parallel applications. For example, the cluster to support virtualization is expected to have 1000 to 3000 nodes running standard Linux with 1 to 2 PB of scratch storage. It is expected that HPC resources will be provided by existing DOE Office of Advanced Scientific Computing Research (ASCR) facilities. Facilities must have space and infrastructure to expand.

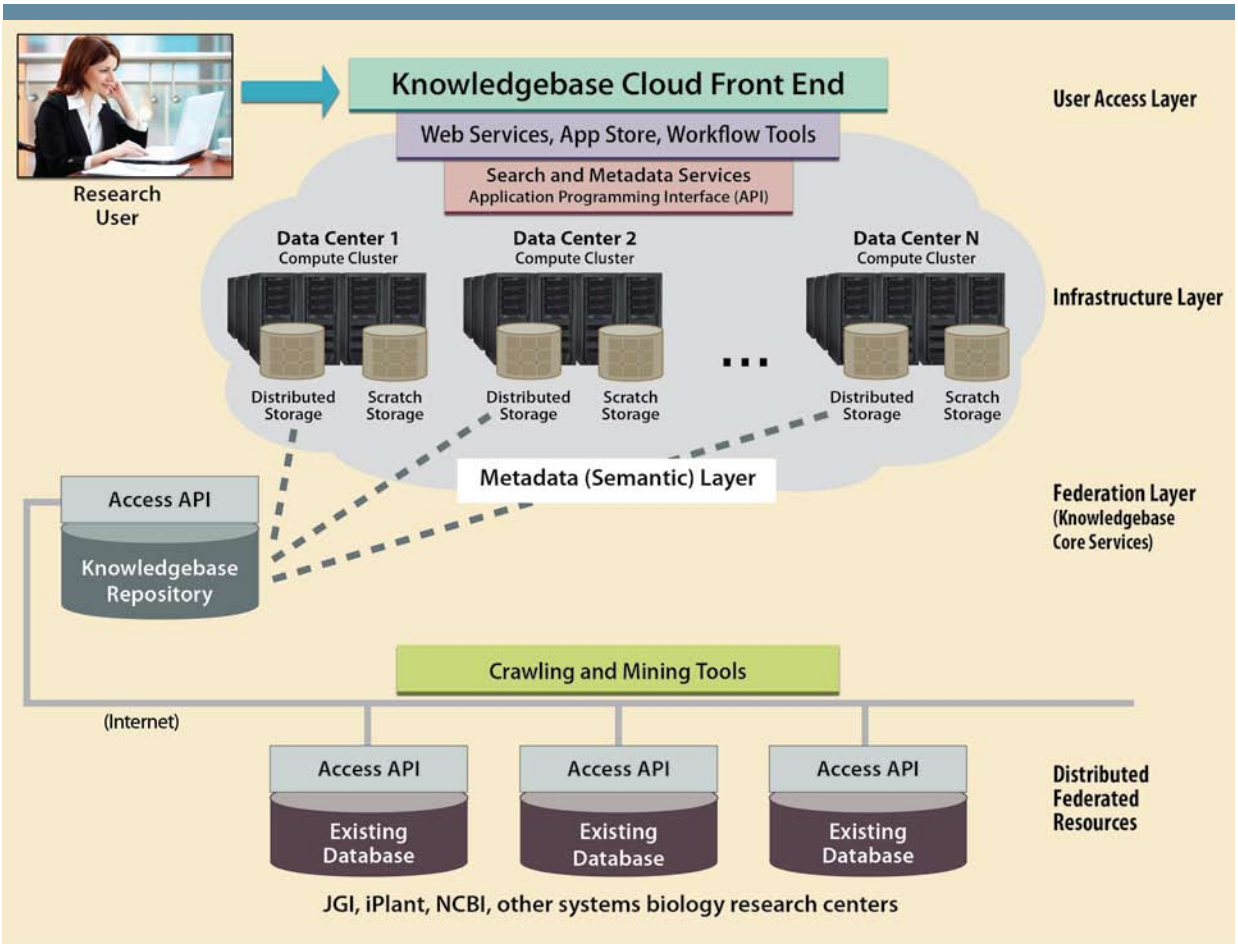


Fig. ES.5. Knowledgebase Architecture Overview. The architecture comprises four layers. The higher layers will span all systems within the Knowledgebase. The federated layers will reside within each of the specific data centers. Not every federated resource may be in every data center. The purpose of each layer and component is described in Chapter 7 of the DOE Systems Biology Knowledgebase Implementation Plan.

Key Systems Biology Knowledgebase Partnerships

The Knowledgebase is providing a unique impetus toward support and acceleration of the DOE systems biology research community. However, this effort is not operating in isolation of other synergistic efforts. Critical partnerships that will be leveraged are described in the implementation plan. These include the DOE JGI, DOE ASCR, National Center for Biotechnology Information (NCBI), and the NSF iPlant Collaborative.

The Knowledgebase will work with the DOE JGI to ensure that analysis tools developed are cross-compatible and that sequencing data and experimental data are shared to support a robust annotation system.

The Knowledgebase effort may employ several ways to leverage the exascale computing capability being developed in ASCR. Most of the scientific targets for the Knowledgebase are data driven and at a scale that, for the foreseeable future, likely could be met by a more moderate sized community cluster. However, opportunities exist for "co-design" partnerships between ASCR and the Knowledgebase to address unique problems such as the combinatorial analysis of biological networks. These partnerships may lead to better solutions for the potential "all versus all" comparisons in data-rich biological problems.

Another key partnership would be with NCBI, the major repository of primary sequence data. In addition to archiving bibliographic information (e.g., PubMed), NCBI has begun collecting more comprehensive biological information other than sequence data. NCBI recognizes the value of the Knowledgebase as an infrastructure that can fill the gap in the analysis and understanding of biological systems by providing users with a single portal to a variety of tools, resources, and multiple data types. The Knowledgebase will work with NCBI to share experimental data, cross-reference analysis resources, develop community-supported standards for new types of data, and develop tools that are cross-compatible for data analysis and data visualization. A Knowledgebase-NCBI working group will be formed and will meet on a regular basis to facilitate this collaboration.

Finally, the NSF iPlant Collaborative is another key partner that focuses on connecting plant biologists with plant breeders and supporting computational and analysis resources for these user communities. iPlant also is developing hardware and software tools for phenotyping plants in the field. In collaboration with the Integrated Breeding Platform, iPlant will support seed storage, phenotyping databases, pedigree support, and portable software and hardware tools useful for field biologists, and these resources can be leveraged for the Knowledgebase plant objectives. The Knowledgebase also will work with the iPlant community to establish common data standards and cross-compatible analysis tools.

Knowledgebase Development Timeline

The DOE Systems Biology Knowledgebase Implementation Plan describes the tasks needed to provide the research community with a comprehensive cyberinfrastructure to advance systems biology over the next several years. The basic timeline of the project is shown in Fig. ES.6, p. 14. Details of the specific tasks and needed expertise are presented within the report.

The implementation plan outlines additional work to continue and expand initial Knowledgebase efforts over the next decade. By providing open access to data and tools that can address biological problems in various application areas, the Knowledgebase will have impacts beyond the scope of the specific targeted objectives, and it will directly impact the pace of biological research throughout the broader scientific community. Ultimately, the Knowledgebase will provide access to data, simulations, and tools to continue to move biology from a descriptive to a predictive science. The ability to make inferences based on broad community-derived datasets will help answer current research questions and will allow new (currently unanswerable) questions to be posed and tested.

The success of the Knowledgebase will be determined not only by demonstrating clear progress toward accomplishing the focused scientific objectives outlined in the implementation plan, but also by how effectively the research community can use and benefit from Knowledgebase resources and services. To be effective, the Knowledgebase must identify and address the needs of stakeholder communities and coordinate with other synergistic efforts. As a resource that is accessible to all, the Knowledgebase will catalyze new collaborations across disciplines and provide the community with a computational environment for testing hypotheses and investigating biological systems at a scale and scope not possible today. The Knowledgebase has the potential to open a new paradigm of biological science, truly engaging a systems approach.

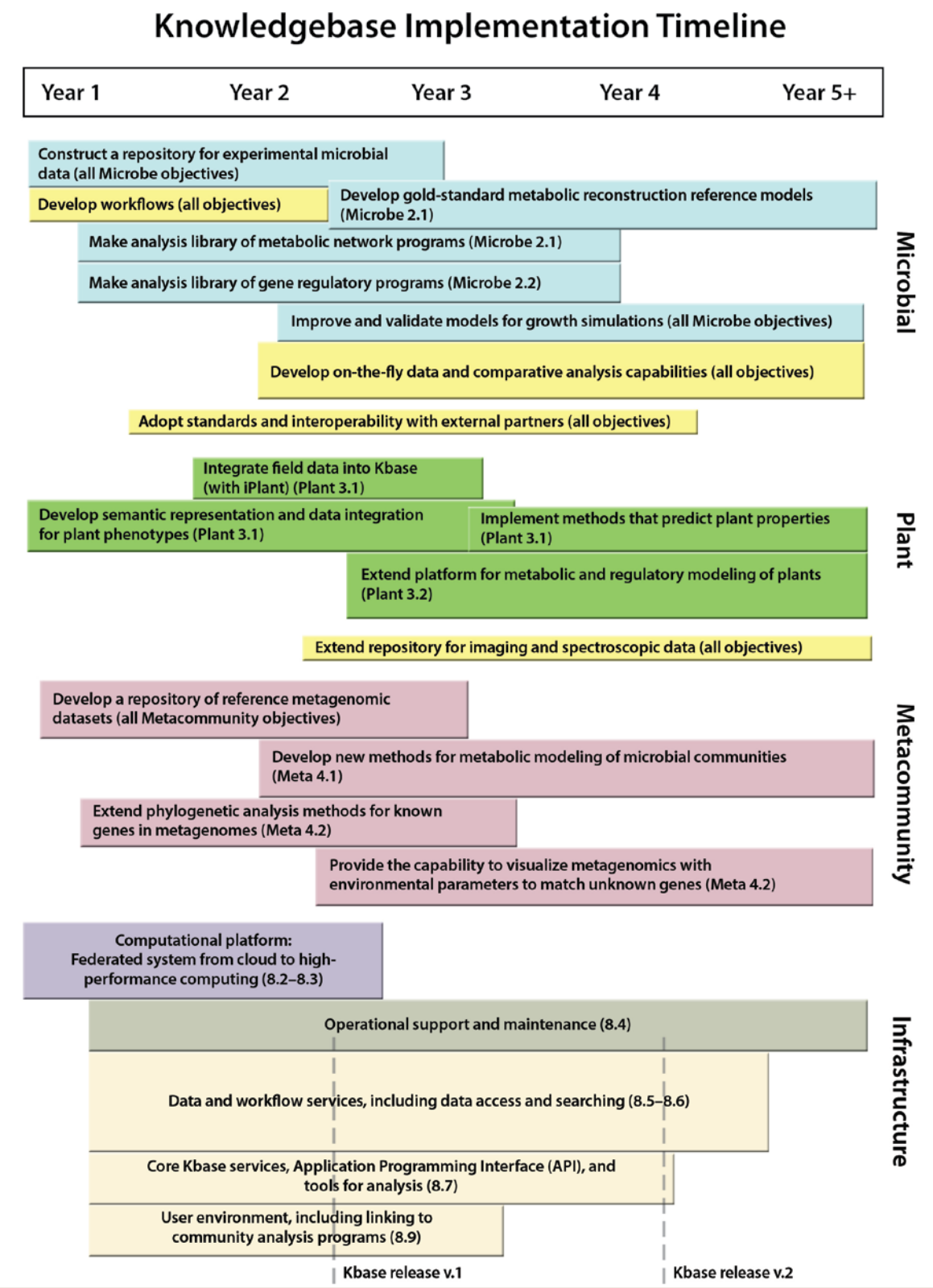


Fig. ES.6. High-Level Timeline for the Knowledgebase Implementation Plan. Light blue boxes represent tasks in the Microbial Science Area. Green boxes represent tasks in the Plant Science Area. Pink boxes represent tasks in the Metacommunity Science Area. Yellow boxes are tasks that apply to all objectives across all science areas. Infrastructure includes hardware (purple), operations (gray), and software (beige). Box height indicates relative level of effort for each task.

Contributors and Observers at DOE Systems Biology Knowledgebase Workshops[1]

Paul Adams
Lawrence Berkeley National Laboratory

Eduard Akhunov
Kansas State University

Eric Allen
University of California, San Diego

Martin Allgaier
Lawrence Berkeley National Laboratory

Gordon Anderson
Pacific Northwest National Laboratory

Adam Arkin
Lawrence Berkeley National Laboratory

Steve Baenziger
University of Nebraska

Scott Baker
Pacific Northwest National Laboratory

Nitin Baliga
Institute for Systems Biology

Jill Banfield
University of California, Berkeley

Ali Barakat
Pennsylvania State University

William Barbazuk
University Florida

Chris Bare
Institute for Systems Biology

Eric Beers
Virginia Tech University

Alex Beliaev
Pacific Northwest National Laboratory

Jeffrey Bennetzen
University of Georgia

David Benton
University of Wisconsin

Rex Bernardo
University of Minnesota

William Berzonsky
South Dakota State University

Devaki Bhaya
Stanford University

Paul Blum
University of Nebraska, Lincoln

Ben Bowen
Lawrence Berkeley National Laboratory

Jim Bradeen
University of Minnesota

Mya Breitbart
University of South Florida

Tom Brettin
Oak Ridge National Laboratory

Jim Bristow
Lawrence Berkeley National Laboratory

Jeff Broughton
Lawrence Berkeley National Laboratory

Charles Brummer
University of Georgia

Marcia Buanafina
Pennsylvania State University

Robin Buell
Michigan State University

John Burke
University of Georgia

Victor Busov
Michigan Technological University

Patrick Byrne
Colorado State University

William Cannon
Pacific Northwest National Laboratory

Shane Canon
Lawrence Berkeley National Laboratory

Brian Cantwell
University of Tennessee, Knoxville

John Carlson
Pennsylvania State University

John-Marc Chandonia
Lawrence Berkeley National Laboratory

Christopher Chang
National Renewable Energy Laboratory

Amy Chen
Lawrence Berkeley National Laboratory

Dylan Chivian
Lawrence Berkeley National Laboratory

Tim Close
University of California, Riverside

James Cole
Michigan State University

Frank Collart
Argonne National Laboratory

Luca Comai
University of California, Davis

Robert Cottingham
Oak Ridge National Laboratory

Carlos Crisosto
University of California, Kearney

Richard Cronn
U.S. Department of Agriculture

Thomas Davis
University of New Hampshire

Brian Davison
Oak Ridge National Laboratory

Paramvir Dehal
Lawrence Berkeley National Laboratory

Narayan Desai
Argonne National Laboratory

Adam Deutschbauer
Lawrence Berkeley National Laboratory

Katrien Devos
University Georgia

Patrik D'haeseleer
Lawrence Livermore National Laboratory

Amit Dhingra
Washington State University

Mitch Doktycz
Oak Ridge National Laboratory

David Douches
Michigan State University

Andrew Doust
Oklahoma State University

Jorge Dubcovsky
University California, Davis

Ismail Dweikat
University of Nebraska

Ronan Fleming
University of Iceland

David Francis
Ohio State University

[1] This list does not include participants from the *Using Clouds for Parallel Computations in Systems Biology* workshop held at the 2009 Supercomputing meeting because it was a large open meeting without a formal participants list.

Many people attended multiple workshops.

George Garrity
Michigan State University

Jack Gilbert
Plymouth Marine Laboratory

Bikram Gill
Kansas State University

Paul Gilna
Oak Ridge National Laboratory

Jim Giovannoni
Cornell University

Adam Godzik
Sanford-Burnham Medical Research Institute

Steve Goff
iPLANT

Jose Gonzalez
South Dakota State University

Ian Gorton
Pacific Northwest National Laboratory

Pam Green
University of Delaware

Jeff Grethe
University of California, San Diego

Arthur Grossman
Stanford University

Masood Hadi
Sandia National Laboratories

Daniel Haft
J. Craig Venter Institute

Steven Hallam
University of British Columbia

Maria Harrison
Cornell University

Caroline Harwood
University of Washington, Seattle

Loren Hauser
Oak Ridge National Laboratory

Patrick Hayes
Oregon State University

Sam Hazen
University Massachusetts

Terry Hazen
Lawrence Berkeley National Laboratory

Karla Heidelberg
University of Southern California

Alyssa Henning
Cornell University

Matthias Hess
Joint Genome Institute

Eva Huala
The Arabidopsis Information Resource

Phil Hugenholtz
Joint Genome Institute

Amy Iezzoni
Michigan State University

Eric Jackson
U.S. Department of Agriculture

Keith Jackson
Lawrence Berkeley National Laboratory

Scott Jackson
Purdue University

Janet Jansson
Lawrence Berkeley National Laboratory

Jerry Jenkins
Hudson Alpha Institute for Biotechnology

Nicholas Justice
University of California, Berkeley

Udaya Kalluri
Oak Ridge National Laboratory

Peter Karp
SRI International

Kimberly Keller
University of Missouri

Bob Kelly
University of North Carolina

James Kelly
Michigan State University

Robert Kelly
North Carolina State University

Joonhoon Kim
University of Wisconsin, Madison

Matias Kirst
University of Florida

Kerstin Kleese van Dam
Pacific Northwest National Laboratory

Steve Knapp
University of Georgia

Robin Kodner
University of Washington

Mavrommatis Konstantinos
Joint Genome Institute

Anthony Kosky
Joint Genome Institute

Julia Krushkal
University of Tennessee, Memphis

Cheryl Kuske
Argonne National Laboratory

Nikos Kyrpides
Joint Genome Institute

Miriam Land
Oak Ridge National Laboratory

Bob Landick
University of Wisconsin

Carina Lansing
Pacific Northwest National Laboratory

Jan Leach
Colorado State University

Jared Leadbetter
California Institute of Technology

Jim Liao
University of California, Los Angeles

Libbie Linton
Utah State University

Konstantinos Liolios
Joint Genome Institute

Jenny Yan Liu
Pacific Northwest National Laboratory

Miron Livny
University of Wisconsin

Thomas Lubberstedt
Iowa State University

Thauos Lykidis
Lawrence Berkeley National Laboratory

Yukari Maezato
University of Nebraska

Krishna Mahadevan
University of Toronto

Laura Marek
Iowa State University

Victor Markowitz
Lawrence Berkeley National Laboratory

Hector Garcia Martin
Lawrence Berkeley National Laboratory

Sergei Maslov
Brookhaven National Laboratory

Xavier Mayali
Lawrence Livermore National Laboratory

Michael Mazourek
Cornell University

Maureen McCann
Purdue University

Phil McClean
North Dakota State University

Susan McCouch
Cornell University

Lee Ann McCue
Pacific Northwest National Laboratory

David Mead
Lucigen Corporation

Barbara Methe
J. Craig Venter Institute

Folker Meyer
Argonne National Laboratory

Richard Michelmore
University of California, Davis

Jonathan Millen
University of Rochester

Amit Mitra
University of Nebraska

Todd Mockler
Oregon State University

Gary Muehlbauer
University of Minnesota

Lukas Mueller
Cornell University

Aindrila Mukhopadhyay
Lawrence Berkeley National Laboratory

Kristen Munch
National Renewable Energy Laboratory

Seth Murray
Texas A & M University

Gerard Muyzer
Delft University of Technology

Ambarish Nag
National Renewable Energy Laboratory

David Neale
University California, Davis

Joseph Onyilagha
University of Arkansas, Pine Bluff

Andrei Osterman
Burnham

Elizabeth Ottesen
Massachusetts Institute of Technology

Krishna Palaniappan
Lawrence Berkeley National Laboratory

Bernhard Palsson
University of California, San Diego

Jiwan Palta
University of Wisconsin

Chongle Pan
Oak Ridge National Laboratory

Nicolai Panikov
Northeastern University

Morey Parang
Oak Ridge National Laboratory

Charles Parker
Names for Life, LLC

Bahram Parvin
University of California

Cameron Peace
Washington State University

Zhaohua Peng
Mississippi State University

Andy Pereira
Virginia Tech University

Amanda Petrus
University of Connecticut

Madeleine Pincu
University of California, Irvine

David Pletcher
Lawrence Berkeley National Laboratory

Mihai Pop
University of Maryland

Iris Porat
Oak Ridge National Laboratory

Jason Raymond
University of California, Merced

Jenny Reed
University of Wisconsin

David Reiss
Institute for Systems Biology

Susanna Repo
University of California, Berkeley

Kathryn Richmond
University of Wisconsin

Errol Robinson
Pacific Northwest National Laboratory

Dmitry Rodionov
Burnham Institute

Dan Rokshar
Joint Genome Institute

Margie Romine
Pacific Northwest National Laboratory

Pam Ronald
University California, Davis

Jeffrey Ross-Ibarra
University of California, Davis

Steve Rounsley
University of Arizona

Nagiza Samatove
North Carolina State University

Herbert Sauro
University of Washington

Gary Sayler
University of Tennessee, Knoxville

John Warner Scott
University of Florida

Alexander Sczyrba
Joint Genome Institute

Joao Setubal
Virginia Bioinformatics Institute

Adrian Sharma
Massachusetts Institute of Technology

Judy Silber
Sound Vision Production

Blake Simmons
Lawrence Berkeley National Laboratory

Steve Singer
Lawrence Berkeley National Laboratory

Steve Slater
University of Wisconsin

Kevin Smith
University of Minnesota

Carol Soderlund
University of Arizona

David Spooner
University of Wisconsin

Dina St. Clair
University of California, Davis

Dan Stanzione
University of Texas

Ramunas Stepanauskas
Bigelow Laboratory for Ocean Sciences

Rick Stevens
Argonne National Laboratory

Steve Strauss
Oregon State University

Leonid Sukharnikov
University of Tennessee, Knoxville

Wesley Swingley
University of California, Merced

Ernest Szeto
Lawrence Berkeley National Laboratory

Tatiana Tatusova
National Institutes of Health

Ines Thiele
University of Iceland

Brian Thomas
University of California, Berkeley

Christian Tobias
U.S. Department of Agriculture

Susannah Tringe
Joint Genome Institute

Jerry Tuskan
Oak Ridge National Laboratory

Edward Uberbacher
Oak Ridge National Laboratory

Allen Van Deynze
University of California, Davis

Richard Veilleux
Virginia Tech University

Wim Vermaas
Arizona State University

Wilfred Vermerris
University of Florida

John Vogel
U.S. Department of Agriculture

Judy Wall
University of Missouri

Dong Wang
University of Nebraska

Zhong Wang
Lawrence Berkeley National Laboratory

Derrick White
University of Nebraska

Owen White
University of Maryland

Steven Wiley
Pacific Northwest National Laboratory

Tanja Woyke
Joint Genome Institute

Cathy Wu
University of Delaware

Shizhong Xu
University of California, Riverside

Koon-Kiu Yan
Yale University

Jian Yin
Pacific Northwest National Laboratory

Will York
University of Georgia

Janice Zale
University of Tennessee

Karsten Zengler
University of California, San Diego

Hongyan Zhu
University of Kentucky

Observers

Paul Bayer
Department of Energy

Jennifer Bownas
Oak Ridge National Laboratory

Peter Bretting
U.S. Department of Agriculture

Kris Christen
University of Tennessee

Daniel Drell
Department of Energy

Cheri Foust
Oak Ridge National Laboratory

Joe Graber
Department of Energy

Susan Gregurick
Department of Energy

Holly Haun
University of Tennessee

John Houghton
Department of Energy

Stephen Howell
National Science Foundation

Randy Johnson
U.S. Forest Service

Ed Kaleikau
U.S. Department of Agriculture

Arthur Katz
Department of Energy

William Klimke
National Center for Biotechnology Information

Shing Kwok
U.S. Department of Agriculture

Neocles Leontis
National Science Foundation

Liang-Shiou Lin
U.S. Department of Agriculture

Betty Mansfield
Oak Ridge National Laboratory

Gail McLean
U.S. Department of Energy

Larry Nagahara
National Cancer Institute

Jack Okamura
U.S. Department of Agriculture

Frank Olken
National Science Foundation

Cathy Ronning
Department of Energy

Susan Schexnayder
University of Tennessee, Knoxville

Jane Silverthorne
National Science Foundation

Marvin Stodolsky
U.S. Department of Energy

Sharlene Weatherwax
Department of Energy

Shireen Yousef
Department of Energy

Table of Contents from
DOE Systems Biology Knowledgebase Implementation Plan

Printed with soy ink on recycled paper 

U.S. Department of Energy Office of Science 2010