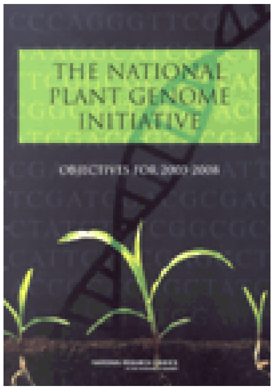


The National Plant Genome Initiative: Objectives for 2003-2008

Committee on Objectives for the National Plant Genome Initiative: 2003-2008, National Research Council

ISBN: 0-309-50317-5, 92 pages, 7 x 10, (2002)

THE NATIONAL PLANT GENOME INITIATIVE

OBJECTIVES FOR 2003-2008

COMMITTEE ON OBJECTIVES FOR THE
NATIONAL PLANT GENOME INITIATIVE: 2003–2008

BOARD ON LIFE SCIENCES

BOARD ON AGRICULTURE AND NATURAL RESOURCES

DIVISION ON EARTH AND LIFE STUDIES

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Contract/Grant No. DBI-0215024 between the National Academy of Sciences and the National Science Foundation, Contract/Grant No. 59-0790-2-089 from the Department of Agriculture, and AP02-02C001575.000 of the Department of Energy. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number 0-309-08521-7

Additional copies of this report are available from the Board on Life Sciences, 500 Fifth Street, N.W., Washington, D.C. 20001; (202) 334-2236 or the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, D.C. 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Wm. A. Wulf is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. Wm. A. Wulf are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE ON OBJECTIVES FOR THE NATIONAL PLANT GENOME INITIATIVE: 2003-2008

JEFF DANGL (Chair), University of North Carolina, Chapel Hill, North Carolina
DOUGLAS COOK, University of California, Davis, California
ROBERT HASELKORN, University of Chicago, Chicago, Illinois
ELIZABETH (TOBY) KELLOGG, University of Missouri-St. Louis, St. Louis, Missouri
ROBERT L. LAST, formerly at Cereon Genomics, Cambridge, Massachusetts
ROBERT MARTIENSSEN, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York
SUSAN MCCOUCH, Cornell University, Ithaca, New York
ERNEST F. RETZEL, University of Minnesota, Minneapolis, Minnesota
CHRIS R. SOMERVILLE, Carnegie Institution and Stanford University, Stanford, California
SUSAN WESSLER, University of Georgia, Athens, Georgia
JOHN YATES, Scripps Research Institute and Syngenta, La Jolla, California

Project Staff

ROBIN A. SCHOEN, Study Director
CLARA COHEN, Staff Officer
BRIDGET K. B. AVILA, Senior Project Assistant
NORMAN GROSSBLATT, Editor

BOARD ON LIFE SCIENCES

BOARD ON AGRICULTURE AND NATURAL RESOURCES

HARLEY W. MOON, (Chair), Iowa State University, Ames
SANDRA BARTHOLMEY, Quaker Oats Company, Barrington, Illinois
DEBORAH BLUM, University of Wisconsin, Madison
ROBERT B. FRIDLEY, University of California, Davis
BARBARA GLENN, Federation of Animal Science Societies, Bethesda, Maryland
LINDA GOLODNER, National Consumers League, Washington, D.C.
W. R. GOMES, University of California, Oakland
PERRY R. HAGENSTEIN, Institute for Forest Analysis, Planning, and Policy, Wayland, Massachusetts
CALESTOUS JUMA, Harvard University, Cambridge, Massachusetts
JANET C. KING, University of California, Davis, California
WHITNEY MACMILLAN, Cargill, Inc., Minneapolis, Minnesota (retired)
TERRY L. MEDLEY, DuPont BioSolutions Enterprise, Wilmington, Delaware
ALICE PELL, Cornell University, Ithaca, New York
SHARON QUISENBERRY, Montana State University, Bozeman, Montana
NANCY J. RACHMAN, Novigen Sciences, Inc., Washington, D.C.
SONYA SALAMON, University of Illinois, Urbana-Champaign, Urbana, Illinois
G. EDWARD SCHUH, University of Minnesota, Minneapolis
BRIAN STASKAWICZ, University of California, Berkeley
JACK WARD THOMAS, University of Montana, Missoula, Montana
JAMES TUMLINSON, Agriculture Research Service, U.S. Department of Agriculture, Gainesville, Florida
B. L. TURNER, Clark University, Worcester, Massachusetts

Staff

CHARLOTTE KIRK BAER, *Director*
STEPHANIE PADGHAM, *Administrative Assistant*

Acknowledgments

This report is the product of many people. First, we would like to thank all those who participated in the workshop on the National Plant Genome Initiative: 2003-2008 on June 6-7, 2002. Their input played an important role in the committee's deliberations.

Philip Benfey, New York University, New York
Jeffrey Bennetzen, Purdue University, West Lafayette, Indiana
Toby Bradshaw, University of Washington, Seattle, Washington
Gloria Coruzzi, New York University, New York
Rebecca Doerge, Purdue University, West Lafayette, Indiana
Michael Donoghue, Yale University, New Haven, Connecticut
Joe Ecker, The Salk Institute, La Jolla, California
Philip Hieter, University of British Columbia, Vancouver, British Columbia
Joseph Hirschberg, Alexander Silberman Institute of Life Sciences, Jerusalem, Israel
Randal Linder, University of Texas, Austin
Joseph Noel, The Salk Institute, La Jolla, California
Jim Ostell, National Center for Biotechnology Information, Bethesda, Maryland
Ron Phillips, University of Minnesota, St. Paul, Minnesota
Michael Purugganan, North Carolina State University, Raleigh, North Carolina
Marc Vidal, Harvard University, Boston, Massachusetts

Sue Rhee, Carnegie Institution of Washington, Stanford University, Stanford, California
Michael Snyder, Yale University, New Haven, Connecticut
Brian Staskawicz, University of California, Berkeley, California
James Tumlinson III, University of Florida, Gainesville, Florida
Richard Young, Massachusetts Institute of Technology, Cambridge, Massachusetts
Robert Waterston, Washington University, St. Louis, Missouri

Second, this report has been reviewed in draft form by people chosen for their diverse perspectives and technical expertise in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards of objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following for their review of this report:

Robin Buell, The Institute for Genomic Research, Rockville, Maryland
Vicki Chandler, University of Arizona, Tucson, Arizona
John Doebley, University of Wisconsin, Madison, Wisconsin
Michael Freeling, University of California, Berkeley, California
Vivek Kapur, University of Minnesota, St. Paul, Minnesota
Gill Kulvinder, University of Nebraska, Lincoln, Nebraska
Hei Leung, International Rice Research Institute, Makati City, The Philippines
Elliot Meyerowitz, California Institute of Technology, Pasadena, California
Steven Rounsley, formerly of Cereon Genomics, Cambridge, Massachusetts
David Stern, Boyce Thompson Institute, Cornell University, Ithaca, New York
Lila Vodkin, University of Illinois, Urbana, Illinois

Although the reviewers listed above have provided constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by R. James Cook of Washington State University. Appointed by the National Research Council, Dr. Cook was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring committee and the institution.

Preface

Since its inception, the National Plant Genome Initiative has become both a national and international focal point for plant biologists, creating excitement for genome scientists, crop breeders and plant physiologists alike. The diverse interests of the plant biology community, whose members work on more than 100 different plant species of economic, agricultural, or purely scientific concern, have converged on the enabling potential of plant genomics. Both the *Arabidopsis* genome project and the NPGI have leveraged important ties to the international plant biology community to further the field as a whole.

This community needs now to exploit the power of "complete" genome sequences, such as the finished sequence of *Arabidopsis thaliana,* and more recently, the draft sequence of *Oryza* spp. (rice) to make basic discoveries about plant biology, and to translate these into agricultural use. Additionally, we are poised to determine the genomic DNA sequence of several judiciously chosen additional species, as detailed in this report. There are several reasons why complete genome sequences are powerful enabling tools for discovering how genes, cells, organisms and populations function: First, because all plants have basic similarities, one can use "model" and "reference plants" with well-documented genetic features to make educated assumptions about similar features in many other species. That gives all plant biologists a head start on understanding any application in their particular crop or plant of interest. Second, the ability to specifically manipulate a plant's genetic building blocks permits us to make directed changes in its physiology. While that detailed knowledge is essential, genomics gives us a different perspective with which to obtain it—that of the big picture—because we can access and describe the activity of all the genes simultaneously. Finally, because genomic sequence is a record of a plant's evolutionary history, we can, using comparative methods, unravel what has occurred in the past, understand how the great diversity of plant form and function arose, and begin to direct that evolution to beneficial and ecologically sound uses in the future.

As the second phase of the NPGI gets under way, it will be important to keep in mind the interdependent nature of genomics research. Complete and partial genomic sequences and their attendant genomics tools are critical resources on which community members rely for individual research projects. If everyone is to make progress, funding needs to be distributed competitively in the light of stringent peer review, and the results and resources delivered need to be of the highest-quality possible, accessible without restrictions, and provided on schedule. In other words, for the sake of all of the plant biology community, the NGPI needs to fund the best science and the most qualified representatives to undertake the work. In addition, in the rush to apply genomic approaches to uncover information in many different plants, we need to remember that rapid discovery using easily manipulated model and reference species is the most efficient tool to convey knowledge to an application-oriented user community. It is thus vital that we continue to efficiently mine the model for all of plant biology—*Arabidopsis thaliana*—while building our knowledge base outward from there, through the sequencing of carefully selected reference species that we define herein, and onward to all the major crop species over time. To do otherwise is to diffuse the fundamental value of genomics as a science and as an applied tool, and does not do justice to the progress made under the first round of the NPGI.

This argument is not only supported solidly on both scientific grounds (including experience from other genome projects leading up to the Human Genome Project), it is also an economic reality. Plant genomics has made remarkable progress in a short time period with a cumulative amount of funding that simply pales in comparison to that currently available for human, mouse, invertebrate, and microbial genomics efforts. Obviously if more money were to be made available, the Initiative could accommodate a more wide-ranging set of projects to the satisfaction of many more research and commodity groups anxious to use genomics. In fact, I would argue that the competitive, peer-reviewed plant biology research funded by the NSF, DOE, USDA, and NIH is one of the best investments made by the federal government in terms of "bang for the buck," given the importance of plant biology to our society. But, given financial constraints, it makes sense to exploit the data from the model and reference species to their fullest, recruit techniques developed in non-plant genomics projects into plant biology, capitalize on falling sequencing costs when appropriate, and continue to make careful decisions on how to build the resources that the plant biology community uses collectively.

In this report, the committee on Objectives for the National Plant Genome Initiative: 2003-2008 which I chaired, makes suggestions for the next phase of the Initiative that adhere to these realities. The committee was established by the National Research Council in response to a request from the sponsors of the Initiative (OSTP, NSF, DOE, and USDA) for guidance in crafting future objectives for the program.

To assist its deliberations, the committee organized a workshop at the National Academy of Sciences on June 6-7, 2002. (The agenda is attached as Appendix A). The group that participated included plant geneticists and biochemists, evolutionary biologists, bioinformaticists, and investigators who study yeast, *C. elegans* and human pathogens. We are particularly grateful for the time they spent in both the panel sessions and working groups. Their insights added significantly to the breadth of our thinking, and are reflected in the report.

We hope that the report's recommendations, if implemented, will help the Initiative carry out its mission, launched just 5 years ago, of bringing the power of plant genomics to bear on efforts to meet the nation's needs for innovation in agriculture and energy, and in the vitally important journey towards understanding the plant world to which we, as humans, owe our very existence.

Jeff Dangl
Chairman

Contents

APPENDIXES

Executive Summary

Plant life plays important and diverse roles in our society, our economy, and our global environment. Research on plant biodiversity, development, physiology, and evolution is needed to understand plant biology in those roles and to increase our ability to use plants beneficially. The National Plant Genome Initiative (NPGI) was launched in 1998 to advance national objectives in plant biology, agricultural, and energy research. In addition to the instrumental role that NPGI played in accelerating the completion of the first plant genome sequence, that of the model plant *Arabidopsis thaliana*, it has supported the development of genomic resources, such as bacterial artificial chromosome libraries, physical and genetic maps, comparative genetic maps, and novel germplasm resources for a broad array of crop species. Functional-genomics tools enabling high-throughput screening for mutations have been developed for some crop species. And NPGI support has led to the generation of a wealth of DNA sequences deposited in public databases, mostly in the form of expressed sequence tags (ESTs), the short segments of DNA that represent gene-coding regions. That investment is creating a distributed infrastructure of experimental and database tools for plant genomics in a variety of crop species. The NPGI has, in fact, brought genomics to crop-plant biology, resulting in important new knowledge, for example, about the function of plant centromeres and kinases, and the basis of plant responses to the environment. This capability can now be leveraged to accelerate the translation of basic discovery to agriculture. The predictive manipulation of plant growth will affect agriculture at a time when food security, diminution of lands available for agricultural use, stewardship of the environment, and climate change are all issues of growing public concern.

The Office of Science and Technology Policy and the federal sponsors of the NPGI—the National Science Foundation, the Department of Energy, and the Department of Agriculture—approached the National Research Council for help in determining goals for the NPGI in the timeframe 2003-2008. In response to the request, the Research Council established a committee to study the future directions of plant biology and genomics and to recommend priorities for the 2003-2008 phase of the NPGI. The work of the committee was informed by a 2-day workshop held at the National Academy of Sciences on June 6-7, 2002. On the basis of discussions at the workshop and additional information, the committee developed a set of recommendations, which are summarized below and detailed in the full report. The recommendations rest, in part, on:

- Progress of the NPGI to date.
- The availability of data, software, methods, tools, biologic, and other genomics-related resources for various plant species.
- The ability of research and development user communities to absorb and rapidly exploit gene-sequence information and other genomics tools.
- The potential for international collaboration in new plant-genome activities.

The need to advance a variety of efforts in plant research and applications as rapidly as possible was balanced with the desire to proceed as economically as possible.

Recommended Goals for the National Plant Genome Initiative in 2003-2008:

1. Focus the NPGI portfolio on a small number of key plant species for in-depth development of genome-sequence data and development of functional-genomics tools.

Candidate species for detailed analysis and investment should be individually evaluated and selected by using science-based criteria that recognize: rapid and economic scientific development, the need for temporal ordering of investments to achieve economies of scale, and application to crop improvement. The criteria include genetic tractability, genome size and complexity, and the potential for translation of data and tools to agronomically important relatives. The availability of a sufficient population of researchers, including international partners, working on both a candidate species and its relatives is vital to ensure use of DNA sequence and functional-genomics tools by a coherent community (Appendix C of the report provides information on the number of publications on the 50 most-cultivated crops and a few reference species in 2000-2001).

Species for this level of large investment should be chosen from the families of Poaceae (grasses), Fabaceae (legumes), and Solanaceae (including tomato and potato) to maximize translational opportunities for the greatest number of economically important plants. Rice, maize, *Medicago truncatula,* and tomato would be examples of species that meet most criteria for immediate expanded emphasis. Because the species would be sequenced to near completion, and because their sequences will inform research and crop development for all the related crop species, we refer to them as *reference species*.

Concentrating DNA sequencing efforts on a small number of carefully selected genomes is greatly preferred over a diffuse effort on many plant species because it focuses research on identification of genes and key biologic functions in experimental contexts in which those goals can be achieved economically.

2. Enable translation of basic findings from the reference species to related crops.

To leverage the investment suggested for the reference species, essential tools for comparative genomics need to be developed, including large-insert DNA-clone libraries, physical maps, and mapping tools for their agronomically relevant relatives within the Poaceae, Fabaceae, and Solanaceae. These tools should be explicitly developed with applied goals in mind, including acceleration of plant breeding, mapping and deployment of quantitative-trait loci in breeding programs, and molecular identification of beneficial alleles of any particular gene. These tools will facilitate further genome sequencing of a variety of crops in the future as sequencing costs drop. We recognize that soybean and wheat are key crop species and thus anticipate that draft sequencing of the gene-rich regions of these genomes could commence during the 2003-2008 time frame, as the cost of sequencing declines, or funding increases.

3. Begin dissection of the evolutionary diversification of plants using genomics technologies.

The more than 250,000 species of plants have a wide variety of growth habits, adaptive responses, and useful traits. The diversity and complexity of plant genomes is increasingly recognized as a reflection of the evolutionary forces driving plant speciation. Genome comparisons across great evolutionary distances provide insights about the similarities and differences among organisms. To begin exploring and potentially harnessing that information, the NPGI should make an investment in evolutionary genomics. The investment would enable the participation of the evolutionary-biology community in the genomics revolution, thus expanding the intellectual expertise in plant biology as a whole. A broader understanding of plant evolution will also increase the palette available for crop improvement.

4. Expand investment in bioinformatics to fully leverage the wealth of plant genomics data now being generated.

The large amount of information housed in genomics databases and the expected explosion in data place huge pressures on the organization of research to effectively mine that data. The plant-research community will have to place greater emphasis on integrating bioinformatics approaches into its work. We propose a national strategy for bioinformatics that includes training, collaboration with large data centers, and bioinformatics-oriented research, such as the creation of specialized databases and new analytic tools.

5. Create new interdisciplinary training opportunities for doctoral and postdoctoral researchers.

To fully exploit genome-based data, the plant-biology community needs to expand training opportunities into disciplines that are not traditionally associated with plant biology and crop sciences, such as computer science, mathematics, chemistry, and engineering. In addition to continued support for students and postdoctoral fellows on single-grant applications, increased support for interdisciplinary training grants is needed to develop a scientific workforce capable of using the tools that multiple scientific fields offer to plant genomics.

All of those goals are important, and it is equally important that they be pursued logically and efficiently, with funding decisions determined by stringent peer review. Some will require more effort and support than others, and different technical approaches will be needed, depending on the species under consideration, the objectives of the research, and the genomic information available from other projects. The following recommendations constitute a strategy for implementing the 2003-2008 goals. The strategy assumes that the current trend of funding for the NPGI, and for the necessary complement to NPGI provided by the Arabidopsis 2010 Project to exploit the model plant species, will continue or increase.

Recommended Strategy for the National Plant Genome Initiative in 2003-2008:

(1) Support two categories of DNA sequencing:

(a) Genomic sequencing. Sequence to the level of a "deep draft" the genomes of appropriately selected and ranked reference species from Poaceae, Fabaceae, and Solanaceae. By *deep draft* we mean sequencing of the gene-rich regions of a genome so that each DNA base is represented, on average, 6 times. This is termed 6-fold coverage. *Finishing* implies filling gaps and increasing the sequence accuracy to no more than one error per 10,000 base pairs. This sequencing effort should be buttressed by sample genomic sequencing (at ~2× coverage) of related and progenitor species of these families and *Arabidopsis* (in Brassicaceae). Because having a complete sequence is fundamental to all other activities, we recommend that roughly 40% of the entire 5-year budget be used for this activity.

(b) Transcript sequencing. We recommend that roughly 12-15% of the budget be used to support sequencing of transcripts, as full length cDNAs or ESTs, as follows:

First, sequence full length cDNAs of all the *Arabidopsis* genes to generate a baseline "plant Open Reading Frame reference set"—the ORFeome—that represents the set of genes from which protein is made. Second, based on the EST data sets and unigene assemblies now available via NPGI, sequence full-length cDNAs for those genes of the other reference species that are either not found in, or are most diverged from, relatives in *Arabidopsis*. This hierarchical approach to full-length cDNA sequencing will eventually yield a plant ORFeome that incorporates many aspects of plant evolution, as well as having very high value for functional studies.

Sequence ESTs from specialized plant cell types and organs in species from which specific novelties in the expressed gene sets can be expected. We expect roughly 25 such projects of various size determined by the specific biological question.

Sequence ESTs from carefully normalized cDNA libraries of each of about 50 species chosen with reference to current programs in evolutionary genomics. This activity is aimed at facilitating an understanding of the events leading to plant diversification (see item 5, below).

(2) Translate genomics data from the model species, *Arabidopsis*, and the reference species to agricultural improvement. The public and private communities of applied plant biology should be included in efforts to translate basic discoveries to crops. In addition, assuming that prerequisite resources are developed and sequencing costs continue to drop, genomic sequencing of the gene-rich regions of soybean and wheat could be initiated. We suggest a 10% budget investment to hasten the fulfillment of translational agriculture.

(3) Support efforts to enable and deploy whole-genome functional analysis to determine the function of genes and gene networks in *Arabidopsis* and the reference species. The efforts should be ordered over time to ensure maximal leveraging of sequencing activities proposed above. These activities will require roughly 20% of the budget.

(4) Support the construction of integrated databases and analytic tools to manage plant-genomics data and make them available to the worldwide research and development community. This should include a large training component and will require about 15% of the budget.

(5) Enable plant evolutionary genomics, systematics, and population biology, through the analysis of EST sequences generated in this program as suggested in item 1, above.

(6) Support the creation of a national plant-genomics initiative steering committee. To maximize the nation's return on its investment in plant genomics, there is a need to continually strategize and coordinate research efforts on all fronts. A committee broadly representative of the plant-biology community that would take a long-term view of plant genomics could provide essential advice to the NPGI Interagency

Working Group on critical community needs, logical next steps, and specific research or technical objectives.

If implemented, the strategy proposed here will substantially affect plant biology. Furthermore, results from the NPGI will generate insights for all of biology and its applications. At the end of this decade, the plant-genome community will have made major inroads toward a more complete understanding of plants and gained new abilities to use them in productive and helpful ways.

CHAPTER ONE

Introduction

Genomics is the science and technology associated with large-scale DNA sequencing of the complete set of chromosomes of a species—its genome—and the interpretation of that sequence information. The genome is the blueprint from which an organism is built. The power derived from determining whole genome sequences is ultimately the power to understand how an organism works.

Much as a builder interprets a blueprint to construct a building, genome scientists rely on attendant technologies to maximize their interpretation of a genome's sequence. Among those technologies are DNA sequence-based methods for defining the portion of a genome that is expressed as messenger RNA (mRNA) in particular cells over the course of the organism's development; methods for determining the entire complement of proteins or small-molecule metabolites in a given cell during an organism's development; and an array of functional-genomics tools developed to assign function to single genes and to groups of genes whose protein products act in concert during a process of biologic importance. The functional-genomics tools are increasingly deployed as high-throughput technologies designed to sample an entire genome simultaneously. The robustness of the technologies varies: some are precise and lead to clear answers regarding gene function; others provide a cursory assessment of how a whole genome responds to a stimulus or perturbation—an assessment that needs refinement through classic, hypothesis-driven experimentation. The creation of very large, diverse datasets has created an enormous need for computational tools. Such tools, generically deployed under the banner of bioinformatics, are used to warehouse and make available genomics-derived data, but they are also used to analyze the data and suggest testable hypotheses. Much of this informatics-based science is critically informed by comparing many available genome sequences in the emerging discipline of comparative genomics. Genomics, then, is populated by biologists, chemists, physicists, informaticians, mathematicians, and engineers. This interdisciplinary science has captivated many scientists and driven a huge interest in the field over the last 15 years.

Research and funding priorities in genomics are associated with development of technology platforms and baseline datasets, which often takes place in large technology centers that are explicitly meant to enable a broad research community. Such communities are typically established around research involving one species or a group of related species to leverage the intrinsic power of these organisms, in contrast with the traditional method of designing sets of experiments to test a hypothesis. Therefore, large projects in genomics are often—and vitally—oriented to community service. That demands considerable community alignment toward common goals and requires that lead investigators in large genomic centers recognize that their main responsibility is to provide service to the broader research community. Rapid and open dissemination of the results of genomics research is enabling to all investigators. Without a vibrant research community to empower, the large investments required for success in genomics will be squandered. Without explicit community-service commitments from the genomics leaders of each community, those investments will not be disseminated to the community for broad discovery and application.

Genomics research, supported largely by the National Plant Genome Initiative (NPGI), has already revolutionized plant biology (OSTP 2000). The finished *Arabidopsis thaliana* genomic sequence and the completed rice draft sequences are landmarks for *all* of biology (Sanderfoot and Raikhel 2001). There are large reservoirs of sequence

Initial goals of the National Plant Genome Initiative:

- Accelerate international efforts to finish the sequencing of the genome of *Arabidopsis thaliana*.
- Sequence the genome of rice, in conjunction with an international initiative.
- Develop physical maps and expressed sequence tags for crop and other species.
- Identify functions of genes in important plant processes.
- Develop technologies and methods to advance plant genomics.
- Distribute genome data and resources.
- Conduct outreach and training projects.

information, sophisticated functional-genomics platforms, expression data, and emerging proteomics and small-molecule analysis platforms applied to genome-scale questions in plant biology. Those experimental tools have enabled a much more sophisticated analysis of plant biology than ever before possible. Comparative genomics is facilitated by the diverse genome sequences now available. The data from comparative research also aid our understanding of the evolutionary processes that diversify life forms and biologic function, of how organisms interact, and of how phenotypes are determined by genes interacting with each other and with the environment. New kinds of interrelated databases are being developed to handle the data flood. These resources are driving a renewed interest in plant biology and a greater capacity to hire new researchers, train new students at all levels, and generate new knowledge. The new knowledge will ultimately enable predictive manipulation of plant growth and will affect agriculture at a time when food security, diminution of lands for agricultural use, stewardship of the environment, and climate change are all issues of public concern.

The development of *Arabidopsis* as the “model plant” continues to enable an explosion of progress in understanding the basic processes of plant biology and has revolutionized plant biotechnology. However, it is clear that no single plant species can serve as a completely unifying experimental model for all of plant biology. For example, natural selection and domestication have resulted in many taxon-specific characteristics that are not well suited for study in *Arabidopsis*, such as nitrogen fixation in legumes, inflorescence architecture in the grasses, and the development of the fleshy fruit of tomatoes. Even homologous gene families show evidence of taxonomically restricted function and radiation, as in the example of the economically important plant disease-resistance genes. Those facts compel the simultaneous investigation of plant processes in a carefully chosen array of species.

Now that the complete sequence of *Arabidopsis* is available, the next logical step is to continue and leverage the investment in *Arabidopsis* genomics and the investments made in the first phase of the NPGI, and to develop a small number of other *reference species* from various other botanic families, specifically chosen for using both scientific and agricultural criteria. The development of these species as references will proceed, by virtue of economy of scale and experimental rapidity, in relation to *Arabidopsis* functional genomics. Therefore, it is *vital* that the Arabidopsis 2010 Program, while not explicitly part of the NPGI, be funded at levels sufficient to complete its goals.

As the NPGI unfolds, there is a need to continually strategize and coordinate research efforts on all fronts. In the past, other national initiatives have created scientific oversight committees to perform this function. Such committees have been created for all the large genome-sequencing projects, including that for *Arabidopsis*. An oversight committee created by the NPGI could maximize progress toward national goals by serving as a focal point for discussions of critical needs (not only those related to sequencing, but all aspects of plant genomics) across the entire community and devise the most effective ways to satisfy those needs. A committee broadly representative of the plant-biology community that would take a long-term view of the genomics of economically useful plants could provide essential advice to the Interagency Working Group and sponsors of the NPGI.

ABOUT THIS REPORT

This report is the product of a National Research Council committee established at the request of the federal sponsors of the NPGI—the National Science Foundation, Department of Agriculture, and Department of Energy—to "conduct a study of the future directions of plant biology and genomics and recommend priorities for the next phase of the National Plant Genome Initiative."

The committee was asked to consider which goals are achievable in the next phase of the Initiative, and what tools will be required to meet those goals.

To inform its deliberation, the committee asked experts in the field how to build on current accomplishments in order to address major questions in plant biology and to consider whether progress toward that goal might be made by additional sequencing projects (as was emphasized in the first phase of the Initiative) or alternatively by using other strategies to "mine" the sequence data now available to elucidate biological processes and functions. A workshop held on June 6-7, 2002, in Washington, DC (see Appendix A) contributed to the committee's information-gathering process by examining key questions in plant biology, promising avenues of research, and needs for human resources and technical infrastructure. The report's recommendations encompass the issues discussed at the workshop. The following chapters describe in detail how these objectives for the next five-year phase of the NPGI will advance the use of plant genomics for a greater understanding of plant biology and its applications.

CHAPTER TWO

Sequencing: Generation of the Raw Material

The generation of new DNA sequence data will continue to be critical for the understanding of plant genomes in the near future. Although it would be desirable to sequence many plant genomes at this time, the cost of sequencing needs to drop substantially to allow immediate deep-draft or finished sequencing of more than a few genomes. By deep-draft we mean at least 6-fold coverage of the gene-rich regions, accompanied by other data, such as a physical map and sufficient sequencing information on bacterial artificial chromosome ends (BAC ends) necessary to generate a scaffold of ordered contiguous DNA sequence information. Finishing includes filling gaps and increasing the sequence accuracy to no more than one error per 10,000 base pairs. Finishing is painstaking and hence more expensive than draft sequence, and is not anticipated in every case. The current cost of finished DNA sequencing is about 9 cents per finished base, down from 50 cents per base 3 years ago, while the cost of all other high-throughput sequencing (ESTs, deep draft) is $1.50 per sequencing run. Projects proposed in NPGI should be this competitive, and should use these costs as benchmarks. The cost of sequencing will probably continue to drop during the next 5 years and beyond, and the NPGI must be positioned to take advantage of cost improvements as they occur in the next 5-10 years.

The explosive increase in understanding of biology over the last 20-30 years has been enabled by work on model genetic organisms, including *Arabidopsis*. The NPGI is best served by a structure that effectively exploits the application of detailed knowledge gained from *Arabidopsis* and judiciously selected reference species to related crop species. To that end, we recommend a detailed characterization of genomes of a small number of reference species selected on the basis of criteria detailed below, to represent key plant taxa. This should be accompanied by parallel investment in genetics and genomics tools from related crop plants that are explicitly designed to transfer knowledge gained from research in model and reference species into agronomic development. Partnership among federal agencies that span the breadth from basic to applied plant research is essential.

Species to be considered for development into reference status should be chosen according to how well they fulfill the following criteria:

1. Experimental tractability, including
   a. Forward genetics—the ability to isolate mutants and the relevant genes.
   b. Reverse genetics—the ability to target, or identify, mutations in a predefined gene.
   c. Availability of a physical genetic map.
   d. Short generation time.
   e. Ease of transformation.
   f. Ease of growth under defined conditions.
2. Low genome complexity, including
   a. Diploid genome.
   b. Small genome size.
3. Size, expertise, and ability of the research community meant to use the sequence and functional-genomics tools, including the opportunity for international collaboration.
4. Suitability for translation to agronomically valuable plants.

The few plant species that meet those criteria should be selected to encompass a range of phylogenetic diversity and to include major plant processes not present in *Arabidopsis.* Such species should already have well-developed research communities and experimental resources and should fall within key, agronomically relevant clades of the plant kingdom. In particular, we recommend that the reference species be chosen from the families Poaceae (grasses), Fabaceae (legumes), and Solanaceae (including tomato) in that order as funds allow. Those three families contain species that are both important research organisms and critical crops.

Of the roughly 250,000 plant species, humans have domesticated less than five thousand as crops, and roughly 20 species, primarily from two plant families (the grasses and legumes), provide the great majority of our food. Independent origins of agriculture were based on the domestication of cereals and legumes: rice and soybean in Asia, maize and beans in the Americas, and barley, wheat, pea, lentil, and chickpea in the Near East. The grasses and legumes are principal components of most terrestrial ecosystems. Poaceae includes sugar cane and all the major cereals: maize, wheat, rice, barley, sorghum, millet, and rye. Grass species account for about 50% of human caloric intake, and provide all cereals and most of the world's sugar, are the principal forage for animals, and occupy about 70% of the world's farmland. Fabaceae supplies nearly 33% of the human nutritional requirement for nitrogen, with a protein content that is balanced in amino acids and roughly 2/3 that of cereals. The high protein content of legumes is related to their capacity to symbiotically fix atmospheric nitrogen in ammonia; this property is a key factor in the global nitrogen cycle. Legumes are also important sources of fodder and forage for animals and of edible and industrial oils. Tomato and potato, both natives of South America and members of the family Solanaceae, have become increasingly important in the global food supply. Thus, there are compelling candidates in each of the three plant families that, on the basis of the criteria for selection outlined above, could serve as reference species. We suggest that rice and maize in Poaceae, *Medicago truncatula* in Fabaceae, and tomato in Solanaceae are strong candidates for elevation to reference-species status. They meet essentially all of the criteria proposed above. Thus, these genomes should be sequenced to at least deep-draft stage and finished if scientifically warranted. For example, we believe that the rice draft sequences now or soon available should be finished. Further, NPGI investment in functional genomics tool development, where appropriate as described in chapter 4, should be focused on these species.

The discovery of extensive conserved synteny (gene order on chromosomes) among grass genomes reflects their divergence from a common ancestral species within the last 50-60 million years. Thus, the grasses as a group constitute a unified genetic system and provide a collective model genome for the monocots. That concept holds for many plant families. Poaceae, Brassicaceae, Solanaceae, and Fabaceae are all poised to become model families in which genetic and genomic data can be extrapolated to a broad array of comparative and evolutionary studies. That potential is best exemplified in the grasses: rice and maize sequences together would permit development of a unified model system. Developing genetics and genomics tools applicable to an entire plant family requires that the advantages and idiosyncrasies of individual species in the family be delineated and explored. For example, given the goal of defining taxon-specific characters, it will be important to sample sequence the genomes of more than one species from each of the Poaceae, Fabaceae, Solanaceae and Brassicaceae. The ability to compare the gene content and structure of fully sequenced BAC clones between confamilial species is already yielding dividends. In particular, such paired-species analyses allow inference of ancestral genome structure and should reveal major changes brought about by domestication.

The combined efforts we propose for NPGI, in concert with related efforts around the world, will hasten identification of the evolutionary adaptations that characterize each plant family and will facilitate the transfer of knowledge to additional crop species in each family. In addition to *Arabidopsis* and rice, complete drafts of the genome sequences of the single-celled alga *Chlamydomonas* (CGP 2002) and poplar tree (PGP 2002) will be available by 2003, and should be exploited as scientifically appropriate. We propose that DNA sequencing consume a substantial part of the total expected funding allowance for the next 5-year phase of the NPGI (about 40% of the current total), but the cost of sequencing is a legitimate issue to consider in deciding what to sequence, how deeply, and when to begin.

Strategic and cost-effective organization of genomic sequencing projects depends on some necessary prerequisites. The approaches and costs for genomic sequencing are likely to vary, but these projects will require adherence to most of the following minimum criteria: (1) high quality physical maps and accompanying BAC-end sequences, (2) genetic maps integrated with the physical map, (3) knowledge of genome organization and complexity based on cytogenetic and pilot sequencing data, (4) existence of well characterized transcript libraries to facilitate genome annotation, and (5) public relational databases that integrate whole genome data with other data types.

Before a large-scale sequencing project for any reference species is launched, the community interested in that species should be well organized. These research communities should coordinate efforts on an international scale to make functional connections with communities of researchers working on related species. These communities should be asked to formally propose projects based on a community endorsed white paper as outlined in "Pre-project vetting" later in this chapter.

With the appropriate tools in hand, reference-species sequencing should also include low-pass random sequencing (or BAC-end sequencing, when a physical map is available) of related species, which may be other crops, as a means of gene and regulatory DNA annotation and as a mechanism for comparative genomics; and critically, establishment of a user-friendly, integrated public database.

On the basis of the experimental and agricultural criteria discussed above and the expected cost of DNA sequencing during the next 5 years, we offer the following explicit recommendations for genomic DNA sequencing for 2003-2008:

1. Finish the rice genome sequence except for heterochromatin. The current sequences available are draft sequences, and finishing is vital to establish a "gold standard" grass sequence. This project is continuing and international. In addition, sequence the gene-rich,

low-complexity genomic DNA (known as the "gene space") of *Medicago truncatula* (a legume), tomato, and maize. The unmethylated gene-rich portion of diploid higher plant genomes is typically about 200-250 Mb. This is therefore the expected size of tomato and *Medicago*, while maize, which is a partial tetraploid, is about 400 Mb.

2. Collect (~2×) sample genomic sequence from related and progenitor species of the reference species and of the model species *Arabidopsis* to use as germplasm resources and for comparative and evolutionary genomics. This will be useful for phylogenetic comparisons, single-nucleotide polymorphism (SNP) definition, and gene-model predictions; it will also provide data useful for studies of population genetics and evolution of development.

COMMUNITY STANDARDS FOR LARGE-SCALE SEQUENCING PROJECTS

Because the plant-research community will rely on the availability of the sequences of the reference species for their individual work, it is critical that sequencing be carried out effectively. In this regard, community standards for large-scale sequencing projects play an important role in moving science forward.

PRE-PROJECT VETTING

There are still serious funding constraints on DNA sequencing of large genomes, and there is a need to balance them against other goals in the NPGI. Therefore, it is vital that the research community for any species seeking NPGI funds for either genomic or deep EST sequencing prepare a community consensus white paper both to justify the project scientifically and to demonstrate community unity and community organization before the massive investment required.

ORGANIZATION

Large-scale sequencing projects should be done in a high-quality, inexpensive manner at any of the several large sequencing centers, public and private, around the world. We see no reason to build new sequencing centers as part of NPGI. Successful proposals must be competitively budgeted with respect to the per-base, per-EST, or per-cDNA cost of high-throughput sequencing. (Current estimates of sequencing costs are 9 cents per finished base, or approximately $1.50 per high-throughput sequence run.) However, assembly, finishing, and analysis require contributions from highly motivated experts. Because the most efficient finishing takes assembly and analysis into account, there is clearly a role for academic sequencing centers associated with plant biologists, even though companies and genome centers can offer competitive contracts for high-throughput sequencing. Hybrid strategies can be successful where responsibility for deliverables is held jointly by academic scientists and private-sector sequencing facilities. Whatever sequencing strategy is pursued, quality and cost metrics must be parts of the definition of deliverables to ensure maximal gene discovery per dollar invested.

DATA RELEASE

The NPGI should institute a uniform standard for sequence-data release that ensures rapid release of high-quality data (using the so-called Bermuda rules established for the human genome sequencing project). The standards may differ for different types of sequencing; they need to be carefully written and should be managed with effective oversight to serve the community in as timely a manner as possible.

METRICS OF SUCCESS

Large-scale DNA sequencing demands effective measures of success. The success of large-scale community service projects should be measured by the number of bases and clones generated and deposited into

GenBank and Stock Centers per unit of elapsed time and per dollar invested. The number of publications citing data provided by the sequencing group, and the number of new research projects initiated throughout the community reliant on a particular large-scale sequencing project are useful measures of the public value of such projects. As measures of any sequencing center's success in serving its community, those factors must be visibly advertised and regularly updated on every project Web page. The community needs to be able to understand easily what data to expect and when to expect them, so that researchers whose work requires access to the center's information can plan their experiments and activities. Progress in large-scale community-service projects should not be presented only when they are up for renewal or at the end of a funding period, and scientific advisory boards and project reviewers should be rigorous in their monitoring of this feature of large-scale service projects.

CHAPTER THREE

Maximizing the Return from Reference Sequencing: Translational Agriculture

We suggest in Chapter 2 a focused sequencing effort on a small number of carefully selected genomes. This is greatly preferred over moderate investment in many plant species because it concentrates research efforts on identification of genes and key biologic functions in experimental contexts where those goals can be achieved economically. Our proposal will leverage the power of comparative and functional genomics to understand the biology of the selected crop species and their relatives. It is noteworthy here that the translation to agronomic problems will be facilitated precisely by some of the investments made in the NPGI to date. For example, physical maps of the soybean and maize genomes and extensive sets of expressed sequence tags (ESTs) are now available for many crop species. We encourage further development of mapping tools in crop species during the 2003-2008 period. It is vital that the basic discoveries anticipated from the sequencing of reference species be integrated into current efforts in crop breeding and biotechnology. The NPGI should expand to include applied-plant-biology communities, both public and private, and should involve them explicitly in translation and application of basic discovery to crop improvement. That can be achieved, for example, by developing easy-to-use molecular markers for breeding, by defining genomic intervals that carry traits of interest, and by developing informatics-based tools to hasten translation.

It is important to describe and fully exploit conserved syntenic relationships among species in Poaceae, Brassicaceae, Solanaceae, and Fabaceae so that any finely mapped qualitative or quantitative character in any member of the reference genomes can be easily bred into agronomically relevant cultivars. This effort includes the development of fine-structure genetic and physical maps of key species and the development of a comprehensive set of anchor markers. We anticipate that these tools will enable detailed characterization of the genes that contribute to specialized traits of agronomic interest, such as drought tolerance, salt tolerance, disease resistance, seed quality, and plant architecture.

To determine the genetic basis of traits of economic interest, it will be useful to map genetic variation with relatively high precision, including SNPs or simple, DNA-based, high-resolution introgression tools. Thus, it is important to have comprehensive BAC libraries for species of major agronomic interest and to use BACs for end- and low-pass draft sequencing to see which genes are in each "map bin." Coupled with identification of a polymorphism for each BAC, this approach would provide powerful tools to assign genetic variation to a small set of candidate genes quickly. Such advances will facilitate translation from reference species to other crops.

TOOLS FOR TRANSLATIONAL AGRICULTURE

The following sections outline a variety of approaches that should be undertaken to develop tools for translational agriculture.

1. *Construction of genetic maps for key, carefully chosen species.* We recommend identification of a set of several hundred conserved genes that can be used as anchor markers for comparative map construction and phylogenetic studies across relevant taxonomic distances. If possible, sets of conserved genes should be chosen on the basis of the biologic significance of the underlying genes and pathways and of genomic distribution; map positions should initially be defined in a reference genome. Comparative maps lay the foundation for high-resolution mapping of simply

inherited and quantitative traits and for gene discovery. Thus, they are an essential part of any genomics toolkit and allow information to flow between species of interest to diverse plant scientists, including breeders and evolutionary biologists.

2. *Construction of physical maps for a small number of species (a subset of 1, above) using high-quality BAC libraries.* Physical maps can be assembled from end-sequenced BACs. These are useful for finding, sequencing, and mapping genes of interest; investigating gene families; identifying transposable elements; and defining small-scale genome rearrangements.

3. *Establishment of mapping populations (preferably fixed inbreds) with genotypic segregation data.* Choice of parents for the development of mapping populations should ensure that one parent represents the genotype from which the BAC library is constructed and from which sample sequencing is performed, and that the cross segregates phenotypes of interest. The populations are useful for associating genotype with phenotype and for allele mining, and can be evaluated by different researchers interested in different phenotypes and allele combinations in different environments. The information can be readily shared via genome databases.

4. *Assuming that sequencing costs continue to drop, or funding levels increase, and that the appropriate prerequisites (see Chapter 2) are met, it will be possible to begin sequencing the gene-rich regions of additional key crop species.* Translational agriculture will eventually be simplified by the availability of genomic DNA sequence. The species identified in Chapter 2 for large-scale genomic sequencing were recommended as the top priorities for the NPGI because they fulfill essentially all of the criteria for the definition of a reference species, and because appropriate biological tools have been developed for these species in preparation for full-scale sequencing. The availability of genome sequence for other important crop species, for example, soybean and wheat, could provide valuable information for research on those crops, but they meet far fewer of the criteria described in Chapter 2 and have not yet met the prerequisites to sequencing. For example, there are not solid estimates of the size of the soybean gene space, or of the distribution of gene dense and gene poor

regions in the soybean genome; wheat is polyploid, creating sequence assembly problems, and neither wheat nor soybean is transformed in a routine manner. If the genomes of these crops were to be adequately characterized and the essential biological tools developed, draft sequencing of gene rich regions of soybean and wheat could begin in the 2003-2008 time frame. This recommendation is additionally contingent on drops in sequencing costs and fulfillment of the other, more clearly justifiable NPGI priorities outlined here.

5. *Germplasm collections with molecular genotypes.* Germplasm collections offer a larger view of genotypic and phenotypic variation than can be studied with a mapping population. Genotypic information about accessions can be derived from conserved simple-sequence repeats, SNPs, or other automated molecular marker systems. Data on such a collection can be used by breeders and by population geneticists to evaluate population structure, to determine the extent of recombination, to develop statistical models for interpreting genetic linkage patterns, for allele mining, and to perform marker-assisted selection. They can also be used to compare gene diversity in different ecotypes or subspecies, to clarify phylogenetic relationships among populations or closely related species, and to look for evidence of and evaluate the extent of allelic diversity.

6. *Transcript identification.* The genomic sequencing efforts, and the building of translational tools delineated above, are already supported by EST projects, some of which are being generated in ongoing NPGI projects from normalized libraries of different tissues, different developmental time points, or system perturbation contexts (see Table 3.1). More-extensive use of alternative gene-discovery technologies—such as serial analysis of gene expression and microbead-based and other representational methods—should be considered to complement EST projects. Comparison of ESTs or complete cDNA sequences with genomic sequence will provide critical information on gene content, family membership, and sequence diversity. Furthermore, the cDNA-based resources are crucial for accurate annotation of the genomic sequence and will provide information on allelic diversity for use in molecule-based breeding.

Table 3.1. Ten Largest Plant EST Collections by Species. (NCBI 2002)
dbEST release 080902

Species	No. of ESTs
Glycine max (soybean)	263,737
Hordeum vulgare + subsp. *vulgare* (barley)	215,714
Triticum aestivum (wheat)	175,836
Arabidopsis thaliana (thale cress)	174,624
Zea mays (maize)	165,518
Medicago truncatula (barrel medic)	162,741
Lycopersicon esculentum (tomato)	148,346
Chlamydomonas reinhardtii	112,487
Oryza sativa (rice)	104,594
Solanum tuberosum (potato)	94,257
Total plant ESTs	1,617,854
Total ESTs in GenBank	12,323,094

SOURCE: NCBI 2002.

Therefore, we advocate sequencing of full-length cDNAs from all the *Arabidopsis* genes to generate a baseline "plant Open Reading Frame reference set"—the ORFeome—that represents the set of genes from which protein is made. In addition, based on the EST data sets and unigene assemblies now available via NPGI, we advocate sequencing full-length cDNAs for those genes of the other reference species that are either not found in, or are most diverged from, relatives in *Arabidopsis*. This hierarchical approach to full-length cDNA sequencing will eventually yield a plant ORFeome that incorporates many aspects of plant evolution, as well as having very high value for functional studies.

7. *Decorating the virtual plant.* The ultimate goal of the Arabidopsis 2010 Functional Genomics Project (NSF 1999) is the development of a virtual plant whose metabolic and gene activity status can be monitored "in silico," that is, in a computer model, at any time and under any condition. Extending the concept to include information incorporated

from all the reference genomes would result in a virtual plant whose most basic functional and structural attributes would be generalizable to all plants. We envision decorating the backbone of this virtual plant with additional virtual representations of specialized cell and tissue types and, in fact, whole organs over developmental time. For example, EST sequencing of the oil-gland secretory cells of peppermint plants demonstrated a substantial enrichment, compared to leaf tissue, in expression of genes involved in oil production. In a simple analogy, the virtual plant should include the ability to make it "grow a tuber" or "develop a cotton boll." To reach the goal of incorporating important plant phenotypes (such as cotton fibers, tuber formation, apomixis, perennial habit, fleshy fruit development, nitrogen fixation, heterosis [hybrid vigor], nutrient uptake and homeostasis, and cambium development) into the framework of the reference species, it will be necessary to sample gene expression deeply in judiciously chosen, specialized cell and organ types from a variety of species. We recommend that the NPGI support approximately 25 projects to sequence ESTs from specialized plant cell types and organs in species from which specific novelties in the expressed gene sets can be expected.

PLANT INTERACTIONS WITH THEIR BIOTIC ENVIRONMENT

Up to 30% of crop yield worldwide is lost to pests and pathogens. Thus, a systematic understanding of plants should include their interactions with their environment writ large. In this regard, an expanded NPGI project portfolio including plant interactions with pathogenic, mutualist, and symbiotic organisms will have huge rewards. The 2003-2008 phase of the NPGI should include analyses of fungal genomes. The focus and criteria for selection should be related to how plants regulate their interactions with the biotic environment. Pathogenic fungi of agronomic importance that meet many of the criteria of a tractable experimental species outlined for the selection of reference species should be considered for sequencing. For example, *Magnaporthe grisea* (rice blast) is being sequenced with the support of the National Institutes of Health. Other examples might include rust (*Puccinia*), powdery mildews (*Erysiphe*), and oomycetes (*Phytophthora* or *Peronospora*). Those are all pathogens of the model, *Arabidopsis*, of the likely reference species, and of important related crop species. Thus, the strengths of concurrent analysis of both host and pathogen can be applied to understand pathogenesis in a broad array of host-parasite interactions.

Equally important is an understanding of fungi beneficial to plants, such as the mycorrhizal species and the obligate endophytes of cool season grasses. It might be premature to propose a sequencing project for such organisms, but it is nonetheless important that they be incorporated into the genomics-based plant systems biology and that experimental tools and approaches be developed for their future exploitation.

INFRASTRUCTURE FOR GENOMICS RESOURCES

Preserving high-quality specimens of genomic resources is important to empower plant-research groups worldwide. The generation of DNA sequence and translational tools will drive the need for new stock centers. Materials developed as part of federally- and internationally-funded initiatives, including collections of unique and valuable seed stocks, clone libraries, and databases have already outgrown the ability of individual labs and projects to manage and distribute them effectively to the community. Professionally managed stock centers designed to collect, organize, maintain and distribute high-quality genomic resources to the community at large are needed to facilitate genomics research.

Stock centers might be organized to manage a variety of resources developed for a specific family of plants (such as the Arabidopsis Genome stock center at Ohio State University), or they may be organized to distribute a specific type of reagent or resource for a wide range of plant families (such as BAC Resource Center currently located at Clemson University). There would be advantages to developing specialized stock centers for several species of plants in collaboration with foreign national and international institutions that house the world's germplasm repositories (along with much of the knowledge about specific plant families). Expanding the level of international collaboration and exchange would enhance access to information, germplasm and technology for scientists throughout the world and motivate the formation of partnerships that would generate novel opportunities for innovative genomics research.

CHAPTER FOUR

Functional Exploitation of Genome Sequences

FUNCTIONAL GENOMICS

Because the genome sequence of *Arabidopsis thaliana* is complete and is the most annotated plant genome sequence, this plant continues to serve as *the* model for determination of plant gene function. Research on the selected reference species—and, in fact, all other plant genomes—will be conducted with an awareness of this resource. Because it is so easy to use *Arabidopsis* experimentally for functional genomics, it should be used for applications of whole-genome, high-throughput technologies that will establish baseline knowledge and toolkits applicable to all plants. The Arabidopsis 2010 Functional Genomics Program seeks to associate every known gene in *Arabidopsis* with a protein or non-protein product so that it can be known where in the cell the product is produced, what biochemical pathway it is part of, and what possible function it has in the life of the organism (NSF 1999). Just as the finished *Arabidopsis* genome is greatly facilitating annotation of the rice genome, so will further use of *Arabidopsis* greatly simplify the functional-genomics challenges presented by the other reference species—and in fact by all of plant biology. That will greatly enhance the effectiveness of all the other plant-genomics projects and substantially lower overall cost. The committee therefore advocates that additional resources be dedicated, either as increased funding for the Arabidopsis

2010 Program or from the NPGI, to accelerate the 2010 Program goals that generate new technology platforms or plant-kingdom-wide reference toolkits aimed at similar goals in the other reference species.

However, it is precisely because *Arabidopsis* does not do many things of importance in plant biology that we envision the development of functional-genomics toolkits in the reference species. This pertains most obviously to those genes not present in *Arabidopsis* or that are highly diverged from the closest *Arabidopsis* homologue. Nevertheless, it is also vital to develop testable hypotheses about gene function among closely related species. Conserved gene function is the key to construction of valid comparative maps and to manipulation of germplasm via breeding (introgression of traits) but can be difficult to assign by sequence alone. This is particularly true when minor amino acid changes can lead to altered function, as in the enzymes of secondary metabolism and transcriptional regulators. Hence, it is vital to develop large collections of sequence-tagged mutants, comprehensive large-insert libraries and physical maps of a variety of important species radiating in an evolutionary sense, from the references. Conserved function can be hypothesized on the basis of synteny (genes flanked by the same genes in two species may be related by descent). This information will drive testable hypotheses about gene function in other organisms. One can test those hypotheses readily by accessing mutant lines in *Arabidopsis* or the reference species from public stock centers.

EXPANDING THE FUNCTIONAL-GENOMICS TOOLKIT

Thus far, the plant-biology communities have been technology users, not creators. We endorse expenditure of funds for technology development and infrastructure that address critical questions specific to plant genomics. For example, the lack of high-throughput, robust transformation systems in many plant species and the lack of gene-replacement techniques are impediments to rapid advancement. Equally important, and equally elusive, is the development of cell cultures that maintain a differentiated state.

To achieve economies of scale, it is important to place a high priority on reaching the genome-sequencing goals for the reference species before large-scale investment in some functional-genomics tools. However, other functional-genomics tools require little or no genomic sequence. For example, forward genetics and characterization of insertion mutants or chemically induced mutants can be accomplished in the absence of genomic sequence. A very deep and robustly annotated unigene set of ESTs can be used to make informative microarrays. However, other functional-genomics tools, such as protein chips and high-throughput proteomics require substantial cDNA or genome sequence before they can be appropriately designed and deployed in a cost-effective manner. What should be avoided are costly forays into functional genomics technologies and projects that yield partial or ambiguous results, due to incomplete sequence information, and that will need to be repeated when the full genome sequence becomes available.

Our specific recommendations for tool development in the model and reference species over the 2003-2008 timeframe are based on having complete or nearly complete genome sequences and large EST collections before the beginning of large-scale investment in functional genomics. In the short term, that means that such investment may be limited to *Arabidopsis*, *Oryza*, *Chlamydomonas*, and *Populus*. Development of some functional-genomics tools in the reference species can begin (and some have) based on, for example, deep EST projects ongoing in NPGI. Other tools will require staged development as genome or EST sequences (full unigene sets) become available. For example, high-throughput proteomics as used to identify proteins in complex mixtures is only effective when a sequenced and annotated genome is available, and thus is limited currently to only *Arabidopsis* and rice. The technologies might come on line for each of the other proposed reference species at different times, depending on the progress of sequencing and gene annotation. Alternatively, it might not be necessary to develop each technology for each species if the biologic question is best addressed with functional-genomics tools in the model or reference species. Thus, it is critical to scientifically justify investments in functional-genomics tools.

It is important to distinguish between pilot projects in functional genomics aimed at establishing a technology and full-genome, high-throughput use of mature technology. The distinction might in some instances lead to delay in deployment of a given technology until the sequence is available. A strong case can be made that existing infrastructures in the yeast, *Caenorhaliditis elegans,* and *Arabidopsis* communities should be expanded to make these tools available efficiently to the crop-plant communities.

Eventual efforts (10-year goals for all the reference species) should encompass the following:

- *Development of the essential genetic toolkits.* These will include comprehensive sets of sequence-indexed mutants, accessible via database search and immediately available as seed stock; robust polymerase chain reaction or chip-based mapping tools; and robust conditional expression systems for sensitized and saturating genetic screens for rare alleles.
- *High-throughput methods for predicting and experimentally validating gene models.* Validated species-specific gene models enable accurate identification of genes from genomic sequence and cross-genome comparisons. Validated models also help to identify conserved *cis*-acting elements. Furthermore, full-length cDNA sequencing of diverse mRNA populations enables the eventual construction of high-resolution whole-genome arrays to use in gene-expression studies, and the materials to generate protein chips.
- *Technologies for measuring gene expression.* Robust, high-density arrays or chips hybridized with mRNA populations from a variety of organs and developmental stages will generate a genomewide database that contains snapshots of all the transcriptional changes during a plant's growth (the transcriptome). We suggest further development of rapid and inexpensive ways to assess cell-specific gene expression in multiple species, preferably at the single-cell level. Spatial and temporal expression of genes at several stages of development requires high resolution, high-throughput in situ hybridization methods; single-cell mRNA population analysis; and preparation of specialized tissues and cell types.

- *Technologies for profiling protein dynamics.* It is critical to define the temporal and spatial regulation of protein synthesis and destruction throughout a plant's life cycle. Technologies are needed that are more comprehensive (that display more proteins with greater dynamic range and better quantification) and that detect and measure the factors that regulate these events in plants. It is important to know both cell type and subcellular localization of all proteins. To this end, subcellular fractionation methods that are robust and clean must be developed.
- *Technologies for building protein networks.* Defining genome-scale protein-protein interaction networks—including spatial, temporal, and quantitative measurements—will require development of a variety of tools whose infrastructural basis has been set in yeast and *C. elegans* research. These include simple purification of protein complexes with affinity tags or immunoaffinity isolations and mass spectrometry for identification of the components; protein arrays; and high-throughput protein-protein interaction screens. We will also require new methods to detect dynamic interactions in vivo that are not ready for general use in any species.
- *Biochemical genomics.* Many of the above aims are part of a global approach to the biochemical activities and function of each gene product. Pharmacologic approaches, such as identification of small-molecule inhibitors or activators of gene function, have not been a traditional strength of plant biology. The committee endorses development of platforms to define small-molecule ligands for known proteins, with the ultimate goal of defining an inhibitor for every protein function. Also, small-molecule substrates or inhibitors tethered to affinity probes should be used to measure and identify enzymatic activities in cells. Small-molecule analytic methods continue to evolve rapidly and become more accessible to biologists, although state-of-the-art technologies are expensive to buy and run. These technologies should be made available to plant researchers in their own laboratories, through service facilities, and in multi-investigator funded projects.
- *Systematic manipulation of gene-product expression and activity.* The overriding goal of functional genomics is to elucidate the physiologic

function of each gene product by systematic alteration of its concentration in the cell. Rapid, genome-scale systems to silence gene expression are a desirable goal, as is the capacity to target mutations, insertions, and deletions to specific genomic regions and genes via allele replacement. Those techniques will serve as research tools and enable allele replacement in crop improvement.

- *Natural variation as a source of functional information.* The development of quantitative-trait loci (QTLs) and linkage-disequilibrium analytic tools in the model and reference species is vital for assigning function to genes. High-throughput mutant and allele detection systems are also vital, as are reliable systems to detect single-base mismatches. They require appropriate mapping populations, particularly as related to exploitation of natural variation in crop and noncrop species in which the focus is on the identification of valuable *alleles*, not only the elucidation of gene functions at particular loci. The study of multitrait quantitative genetics is an important way to assign function to some genes (such as those whose mutations result in lethal phenotypes) and to discover proteins that are rate-limiting for important traits. Plant biology's historical exploitation of natural variation as the raw material of breeding provides a wealth of extractable information, as does the availability of wild accessions of many species.

CHAPTER FIVE

Genomics and the Major Transitions in Plant Evolution

There are more than 250,000 species of plants. They represent a wide variety of growth habits, adaptive responses, and useful traits. Such diversity and complexity are increasingly recognized as a reflection of the evolutionary forces driving plant speciation. Plant genomics is increasingly capable of providing DNA-based analytic tools for comparing genomes across great evolutionary distances and for providing insights about the similarities and differences among organisms, the basis of ecologic adaptations, and their origins and persistence. There is untapped value in natural variation as a source of functional information because natural variation has led to variation in function that cannot be uncovered in typical forward or reverse mutant screens. In fact, many important issues in evolutionary and ecologic genomics can be addressed with existing or proposed fully sequenced models, such as *Arabidopsis*, rice, maize, and *Medicago*, taking advantage of their large collections of cultivars and wild relatives with diverse life forms (perennial and annual), mating systems, and ploidy levels. Yet, although those species are useful for addressing some questions in ecologic genomics, they do not reflect the widely divergent templates for evolutionary genomics provided by the total breadth of plant or other species.

Current sequencing costs are still too high to invest in the broad set of species currently used in evolutionary biology, so a selection of additional species must be targeted to explore plant biodiversity. We advocate a modest investment in the period 2003-2008 for development of genomics resources in species outside the model, references, and their crop relatives. As an initial step, we recommend the survey of roughly 50 species with EST sequencing. ESTs remain the most rapid and cost-effective way to sample gene content and discover new genes. This effort will go a long way toward making gene discovery and comparative genomics possible.

To ensure that genomics investment in additional species builds effectively on existing resources and on the EST sequencing suggested above, we recommend that the evolutionary-genomics community pursue further the selection of a small number of key species (5-10) spanning critical evolutionary nodes in preparation for communitywide genomic investigation over the next 3-10 years. The species should be selected to provide a broad view of the evolutionary potential of genomes and a deep understanding of gene diversity and adaptation. Other federally funded projects—such as Deep Green (2002) and its successor, Deep Gene (2002)—have identified a phylogenetically diverse array of species for application of genomics. The NPGI should give priority to developing tools for species from this set or from among species closely related to them. In the context of evolutionary studies, the genomes of cyanobacteria, from which up to 20% of the genes in contemporary plants originated, and the eukaryotic algae also are legitimate objects of study in the NPGI. The specific approach taken to achieve the goals of evolutionary studies broadly is not immediately obvious and will require consensus building. Our key concern is that, for any given species or evolutionary question, there needs to be at least some minimum concentration of scientists ready to exploit genomics data.

Beyond that concern, criteria for choosing any evolutionary-genomics focal species should include:

- Distributed position in the phylogeny, with emphasis on early branches of the green plant phylogeny.
- Genome size, with emphasis on small genomes and simple ploidy.

- Genetic tractability:
  Ease of crossing and population development.
  Ease of growth.
  Short generation time.
- Availability of existing tools:
  Germplasm collections.
  Mapping populations.
  Genetic and physical maps.
  BAC or P1-derived artificial chromosome clone collections.
  EST collections.
  Mutant collections.
- Size of the research community vs required investment.
- Importance of the focal species or close relatives:
  In terms of agriculture.
  In ecology and conservation.

Tools for genomic studies of diverse focal species should be explicitly comparative and should be developed with nonspecialists in mind, with the aim of broadening the community of researchers who have access to and can effectively use plant-genomics data in the future. By expanding the essential toolkit available to evolutionary biologists interested in diverse taxa in the next 5 years, and by urging the relevant community to coalesce around a set of common goals, the stage can be set for the expansion of plant genomics into evolutionary questions. We hope that this modest investment will prepare the evolutionary genomics community for a much larger investment in genome sequencing in the next 5-10 years.

CHAPTER SIX

Development of a National Strategy for Plant Bioinformatics

When the NPGI was launched, it was recognized that the long-term success of plant biology depended on researchers' obtaining seamless access to the disparate and massive datasets arising from genomics research and to the tools needed to examine and analyze the data. There is now a flood of sequence and other plant data, and with it has come the need to expand access to the collective data being generated, so that biologists working on a wide array of plants can find answers to a diverse set of research questions. Making the data that are representative of the entire Kingdom of plant life available and usable to the scientific community is a major undertaking—one that requires a national strategy for plant bioinformatics.

Bioinformatics is a broad discipline that exploits the richness of large datasets to generate research findings. More than a set of tools, bioinformatics is a research approach that includes the engineering of information systems (such as the creation of databases), the development of analytic methods (such as data-mining tools to extract biologically significant patterns in sequence or other data), and the creation of computation-based, predictive models that use multiple types of data to understand how plant systems operate. As a framework that enables investigators to access, integrate, analyze, and compare large datasets, bioinformatics is central to genomics research.

In the short term, a national strategy for bioinformatics requires the plant-research community to place greater emphasis on integrating bioinformatics approaches into its work. That includes training, collaboration with large data centers, and bioinformatics-oriented research itself, such as the creation of specialized databases or new views into genomic data that lead to novel insights. General databases will be needed to provide community services for the reference species, and they should be developed with community participation. The stewards of data and the creators of databases and tools should not act independently but should communicate and coordinate with each other and with public genome repositories to develop common platforms, standards, and interfaces.

In the long term, the common platforms and specifications will become the foundation of a "genomics grid" that will allow appropriately trained investigators to harness the power of a broad network of distributed databases, tools, and computing power from their desktops. That vision of the future requires investment in a computational infrastructure (hardware and software) needed not only for plant biology but for all of genomics research nationally. The NPGI should be a leader on the path to that goal.

To lay the groundwork for this vision, we offer the following specific recommendations for the next 5-year phase of the NPGI.

1. Support the development of community databases as tools to generate knowledge.

Scope and participation: In the context of the NPGI, bioinformatics must serve the unique information needs of diverse research groups focused on different plants and different research goals. The relevant research groups, nationally and internationally, must be active participants in the development of dynamic, interoperable, specialized databases.

Databases should provide an intellectual focus for the integration and interpretation of a wide spectra of biologic data. If properly conceived and constructed, a dynamic, distributed database interrelating everything from nucleotide sequences to ecologic data will provide a research tool that will potentiate new kinds of discoveries in biology.

An investment in databases for reference species must be supported by an investment in interoperable species-specific databases. The databases may incorporate information from related species (comparative-genome databases) and should include core information for cross-species referencing. Thus, for instance, a rice database might provide a basic data model that could meet the database needs of all cereals if funds were available to curate nonrice data into a parallel version of the rice database. Such a model is being pursued by the Gramene database. In general, it is neither desirable nor economically feasible to support separate databases for all species; there must be other mechanisms, such as data warehousing for smaller projects in related community databases.

In order for community databases to succeed, data maintenance needs to be recognized as a valid activity, and supported accordingly. This is especially true as a database grows and additional dedicated support personnel are needed. The Arabidopsis Information Resource (TAIR) constitutes a model for some aspects of the scope and level of research and service desirable for all the other reference-species databases (TAIR 2002). Each of the reference species will need financial support at least comparable with that received by TAIR. Note that TAIR is underfunded (in budget and staff) relative to central databases dedicated to *Drosophila* and *C. elegans* (personal communication, Chris Somerville), a reality that gives an estimation of the support required for success, in as much as those model animal genomes are comparable in size with the *Arabidopsis* genome.

Database design: The long-term vision of a bioinformatics strategy is to create a decentralized collection of independent and specialized databases that are developed and maintained by different groups and communities but that operate as one large, distributed information resource with common controlled vocabularies, related user-interfaces, and curation practices. An example of a collective effort to develop a common vocabulary is the Gene Ontology Consortium (2001). Other standards for interoperability are evolving in semantics and syntax, and these standards-developing activities can be enhanced by their adoption in the community databases and in cooperation with the national data repositories. Databases might also be designed to incorporate information from related species; they would be comparative-genomics databases that would include core information for cross-species referencing. Examples of this cross-referencing mechanism are the distributed annotation system (DAS) and the developing distributed service registration environment, bioMoby (bioMoby 2002).

Standards for the exchange of data and derived information between databases must be developed not only within the plant community but also in the international genomics community. Therefore, cooperation with national data resources, such as the National Center for Biotechnology Information (NCBI), is critically important. It is essential that the databases be available for participation in the international scientific community.

The current organization and operation of many of the community research databases for plant species will need to change dramatically if they are to take on this role and successfully accommodate the full sequence of a species' complete genome. As a data resource, these databases should be prepared to handle huge volumes of incoming data, annotate them automatically, present them to the research community in a timely fashion, work with the national data resources, and develop or adopt a curation model for the data. In addition, the databases must become a platform for comparative studies with data from related species, and their managers must recognize their responsibility as members of the larger genomics community. In this environment, even the technical details of managing the computer system will be more demanding because not only the species community but the global genomics community will depend on its availability 24 hours per day, 7 days a week. Hardware, software, and data redundancy capabilities will become major design considerations.

The Journal Concept for Community Database Curation

The annotation of genomics data maintained in support of biological research activity provides much of the value and success of the community database. This has been demonstrated in the model organism databases for *Drosophila* and *C. elegans*, where there is substantial support for curation activities. These model organisms have the advantage of small genomes and hence finite and limited data sets. In the plant community, where comparative genomics will become an essential tool to leverage related information, new models for annotation must be explored to accommodate the exponential growth of integrated comparative information. The real annotation of genomic information is in the published literature, and a new paradigm is needed to foster, as a curational activity, the incorporation of information derived from the literature into the database.

Community databases might also develop the analytical tools to enable launching, accomplishing, and even publishing primary research results. The implications of this direction are profound, allowing the community database to become a dynamic mechanism to lead, respond to, and integrate genomics-research efforts. When a database environment is capable of providing analytical services, the database also has the potential to become a vehicle for publication of those results.

Four types of curation activities could therefore be envisioned within a community database: 1. The algorithmic annotation of data; 2. The inclusion of literature related to genome information; 3. The publication of new methods and derived results; and 4. The potential publication of negative results. The latter three areas fall into categories best supported by peer review and publication.

To accomplish those goals, the concept of structuring a database in concert with scientific journals is attractive; for example, databases could have editors and reviewers. Some of the information in the databases in fact, will require peer review, and new mechanisms to support such publication can be developed in concert with the traditional means of publication. This curation-publication model builds on the strengths of both systems: the immediacy and community ties of the database, the need for timely and effective curation, and the peer review and recognition of the journal.

The development of a model for the inclusion of the published literature as a scholarly activity alters the view of that activity, and provides a check for the accuracy of interpretation. The publication of new insights developed from database services provides a closed loop for the database activities, and again provides a direct mechanism for peer review. Finally, the publication of negative results gives added value to the database as a source of information not traditionally having an outlet, but essential to the progress of genomic activities.

The databases must be robust, extensible, scalable, and maintainable. When possible, plant databases should use off-the-shelf software for their infrastructure and for the development of major data-mining tools. All data models for the databases will need to be published in an electronic format, kept up to date, and documented in detail. Database-associated software (such as parsers and loaders) will need to be made available to the community. The methods used in the preparation of derived information (methods and standard operating procedures) must also be published and available for review and replication. Those strategies will encourage the bioinformatics and computational research efforts essential to address the challenges awaiting us in the next decade by minimizing the duplication of effort in database development and deployment.

Relationship with national data repositories: Currently, community databases often incorporate data that are not validated, because including them can provide additional insights for users. However, these data often contain errors and are frequently asynchronous with data in the national public repositories (such as GenBank). In developing the long view of bioinformatics, we must address the need to develop a gold standard for data quality in our national repositories. If national repositories can certify the correctness of the data they contain, then the essential role of community-oriented databases will be to present integrated and alternative views into the data. A clear understanding of this relationship and greater collaboration with the national repositories might result in more effective curation of plant data. As a matter of efficiency and for the archival maintainence of reference genomic datasets, community-oriented databases should contribute to ensuring the quality of the data at the national repositories but not duplicate the services available at NCBI, which is charged and qualified to certify, update, and maintain plant-genome data and to augment the fundamental tools available to large genomics projects. Such tools may include services that coordinate identifiers across multiple databases and provide a critical link between the database-as-publication and the publications tracked by the National Library of

Medicine and the National Agricultural Library. Increased interactions between the bioinformatics community and NCBI will potentiate an entirely new view of what can be derived from genomics data.

Oversight: It is imperative that plant databases be implemented and managed in such a way as to ensure responsiveness to present and future community needs. An essential component of any database-management structure will be advisory committees that can provide critical periodic evaluations of the success of the databases in meeting the needs of the plant-biology community and work in concert with representatives of NCBI. Because of the convergence of research in plant biology around common sets of goals and of reference and model organisms, the management and advisory committees for databases should include members from outside the immediate community served by the databases.

2. Support research on new algorithms and technologies.

Beyond the development of integrated information resources appropriate to the plant community, sophisticated analytical tools must be developed to handle the flow of large, multidimensional datasets and to allow biologists to analyze and interpret the data in an interactive fashion. Examples of this kind of specialized application are statistical analysis of microarrays, comparative sequence alignment and QTLs, and data mining.

Computational resources need to be developed that apply the most advanced techniques in the domains of computer and computational science and that are only now being conceptualized in those fields. New research must be funded in database-management systems designed for native genomic information, algorithms for data mining, supervised and unsupervised machine learning, statistical analysis of multiple views of nontraditional data, and data visualization. That kind of research needs to be conducted on a computational infrastructure appropriate to the scale of the problems. Generalized infrastructures capable of supporting such research are envisioned as a set of technologies that include globally distributed datasets, distributed and interoperable databases, and interconnected clusters of computers that could be used to solve computationally intense problems. The high-performance, distributed computational architecture can be provided by technologies such as those being developed for grid computing. In the future, the development and maintenance of a genomics grid will allow many more investigators to participate in exploiting genomics data by making a vast array of data resources and computational tools generally available, thus leveling the playing field for biologic researchers.

Like the databases themselves, algorithms, software, substantive scripts, and analytical methods developed and applied with support from the NPGI should be made freely available. A large community of computer-science and bioinformatics developers have embraced the open-source model of software development, which provides an environment for availability and cooperative development of tools. Just as the immediate release of genomic sequence data was considered an essential component at the initiation of the genomic sequencing efforts, so will the availability of high-quality software affect the development of bioinformatics. The impact of such source-sharing has already been dramatic in the furtherance of bioinformatics goals with such tools as BLAST and Ensembl. Such broad community efforts should be strongly encouraged.

3. Ensure that NPGI-funded community databases contain a substantive informatics training component.

There is a shortage of researchers with interdisciplinary training that spans biology and bioinformatics. Over the long term, the shortage will be addressed by undergraduate and graduate programs being developed at many universities. In the meantime, there is a role for community databases in increasing these skills in their respective user communities. The databases established for the reference species should, as an element of their mission, develop and organize short courses and encourage exchange visits between investigators associated with the database and user sites.

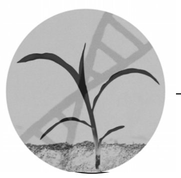
Training can also be integrated with database research and development. Bioinformaticists who are responsible for community databases must be able to meet the projected demands of the community to incorporate increasingly diverse information into databases. There should be some support for database-design brainstorming sessions and for short- and long-term visits at a database or computing center (for example, to examine critical needs in new database construction or develop strategies for migration to improved hardware and software platforms). Through training efforts, therefore, community databases can foster a collaborative interface between biologists and computational scientists.

CHAPTER SEVEN

Achieving Interdisciplinary Training

A glaring bottleneck in plant genome sciences is the paucity of biologists adequately trained in quantitative disciplines, such as mathematics, statistics, physics, and computer science. As plant biology moves toward becoming a predictive science, biologists trained in such fields will be in even greater demand. In the short term, biologists with training in computer science, rather than computer scientists with training in biology, may bring the most benefits, although the converse should be encouraged. Ideally, with the progression of the science, teams of biologists, bioinformaticists, and computer scientists will build on the strengths of the individual disciplines, with the bioinformaticists performing a facilitating role in the translation of data to research findings and in helping the community to develop the needed skills.

Therefore, we strongly urge the founding of training programs designed specifically to recruit students and postdoctoral scientists with degrees in the above disciplines into biology as a whole and plant biology in particular. These training programs need to be embedded in existing bioinformatics and genomics training environments that are demonstrably interdisciplinary and successful. Both individual fellowships and formal training grants need to be supported by the NPGI, as do in-depth training courses modeled on the Cold Spring Harbor courses. Training grants should be complemented by encouraging the development of both semester-long and short courses in all aspects of bioinformatics, statistical genomics, evolutionary biology and computational biology. There is a long history of timely and focused training support in emerging disciplines among the agencies that support the NPGI, and we encourage the resurrection of past models designed to recruit people with new intellectual outlooks into plant biology. The need for different types of expertise in plant biology is particularly striking in database creation and maintenance and in the statistical analysis of complex data, such as data from mRNA-expression and protein-profiling experiments. Therefore, in combination with the overall bioinformatics goals outlined in the previous section, the NPGI should support interdisciplinary efforts to bridge the widening gap between biologists and scientists trained in quantitative disciplines.



References

BioMoby 2002. The BioMoby Project. http://www.biomoby.org as if August 1, 2002.

CGP 2002. Chlamydomonas Genome Project home page. http://bahama.jgi-psf.org/prod/bin/chlamy/home.chlamy.cgi as of August 1, 2002.

Deep Gene. 2002. Deep Gene home page http://ucjeps.herb.berkeley.edu/bryolab/deepgene/ as of August 1, 2002.

Deep Green. 2002. Green Plant Phylogeny Research Coordination Group. 2002. Deep Green: Understanding the Diversity of Plants. http://ucjeps.berkeley.edu/bryolab/greenplantpage.html as of August 2, 2002.

Gene Ontology Corsortium 2001. Creating the Gene Ontology Resource: Design and Implementation. Genome Research 2001. Aug11 (8):1425-1433.

NCBI 2002. National Center for Biotechnology Information. dbEST database. Release date: August 9, 2002.

National Science Foundation 1999. Realizing the Potential of Plant Genomics: From Model Systems to the Understanding of Diversity. Report of a Workshop. http://www.nsf.gov/pubsys/ods/getpub.cfm?bio011 as of August 1, 2002.

Office of Science and Technology Policy 2001. National Science and Technology Council, Committee on Science, Interagency Working Group on Plant Genomics. National Plant Genome Initiative, Progress Report, December 2001.

PGP 2002. Populus Genome Project home page. http://bahama.jgi-psf.org/prod/bin/populus/home.populus.cgi as of August 1, 2002.

Sanderfoot AA and Raikhel NV. 2001. Arabidopsis could shed light on human genome. Nature 2001 Mar 15; 410 (6826): 299.

TAIR 2002. The Arabidopsis Information Resource. http://www.arabidopsis.org/ as of August 2, 2002.

APPENDIX A

Workshop on the National Plant Genome Initiative: 2003-2008

National Academy of Sciences
Washington, D.C.
Auditorium
June 6-7, 2002

AGENDA

Thursday, June 6

11:30 am Informal lunch for meeting attendees and speakers,
NAS Great Hall

1:00 pm Welcome and overview of workshop goals.
Jeff Dangl, Chairman

1:05 pm Origin of the Plant Genome Initiative.
Ron Phillips, University of Minnesota

1:15 pm SESSION I - NEW GENOMICS TECHNOLOGY
Moderator, John Yates, Scripps Research Institute, Syngenta
Panelists:
Marc Vidal, Harvard University
Michael Snyder, Yale University
Richard Young, Massachusetts Institute of Technology
John Yates, Scripps/Syngenta

2:35 pm Discussion.

3:00 pm Session II - Advances in Sequencing
Moderator, Ron Phillips, University of Minnesota
Panelists:
Robert Waterston, Washington University
Rob Martienssen, Cold Spring Harbor Laboratory

3:40 pm Discussion.

3:55 pm Snack Break

4:10 pm Session III – Structural, Functional, and Comparative Genomics
Moderator, Sue Wessler
Panelists:
Joe Ecker, The Salk Institute
Jeff Bennetzen, Purdue University
Toby Bradshaw, University of Washington
Philip Hieter, University of British Columbia

5:30 pm Discussion.

5:55 pm Break

6:10 pm Session IV – Biotic Interactions
Moderator, Doug Cook
Panelists:
Brian Staskawicz, UC Berkeley
Jim Tumlinson, U Florida

6:50 pm Discussion.

7:15 pm Informal buffet dinner for workshop presenters, committee and workshop attendees, NRC Refectory.

8:00 pm Dinner discussions, NRC Refectory.

9:00 pm Adjourn.

Friday, June 7

8:30 am Objectives for the day. Jeff Dangl

8:40 am Session V – Biochemistry and Gene Expression
Moderator, Joseph Hirschberg
Panelists:
Philip Benfey, NYU
Gloria Coruzzi, NYU
Chris Somerville, Carnegie
Joseph Noel, Salk

10:00 am Discussion.

10:25 am Snack Break

10:45 am Session VI – Evolution and Natural Variation
Moderator, Elizabeth Kellogg
Panelists:
Michael Purugganan, NCSU
Michael Donoghue, Yale
Randal Linder, UT Austin

11:45 am Discussion.

12:10 pm	SESSION VII - CHALLENGES IN INFORMATICS Moderator, Ernie Retzel Panelists: Jim Ostell, NCBI Sue Rhee, Carnegie Institute Rebecca Doerge, Purdue
1:10 pm	Discussion.
1:35 pm	Break for Lunch.
2:00 pm	Working group discussions.
4:15 pm	Reconvene in plenary for summaries of discussions.
5:00 pm	Adjourn.

APPENDIX B

Committee Biographies

Jeff Dangl (*Chair*) is the John N. Couch Professor of Biology at the University of North Carolina at Chapel Hill. His research involves pathogen recognition by plants and the evolutionary processes of disease resistance in plants. He is the recipient of the John L. Sanders Award for Distinguished Undergraduate Teaching/Service, UNC-CH, 1998, and the Prize for Young Researchers, State of Nord-Rhein-Westfalen, Germany, 1991. He also serves on the editorial boards of *Cell, The Plant Journal, Molecular Plant-Microbe Interaction, Trends in Plant Sciences* and *Current Opinion in Plant Biology*. Dr. Dangl is a past member of the North American *Arabidopsis* Steering Committee and the NSF Eukaryotic Genetics panel. He is a current member of the NIH CDF-1 Study Section. Dr. Dangl received his BAS in Biological Sciences and English from Stanford University in 1981 and a PhD in Genetics from Stanford University in 1986. He was an NSF Plant Biology postdoctoral fellow from 1986-1989 and a founding Research Group leader of the Max Delbrueck Laboratory of the Max Planck Society in Cologne, Germany from 1989-1995.

Douglas Cook is Professor of Plant Pathology at University of California, Davis, and a Fellow of International Graduate School in Bioinformatics and Genome Research at the Universitat Bielefeld. He also serves as Director of the UC Davis College of Agricultural and Environmental Sciences

Genomics Facility. He has been active in establishing the legume, *Medicago truncatula*, as a model system for biological and genomics studies. His research interests on *M. truncatula* include symbiotic nitrogen fixation and the translation of genomic information to crop legume species. His research group is also contributing to an international effort to characterize the transcriptome of *Vitis vinifera* (grape). He is a member of the International Steering Committee for Grape Genomics and of the US Legume Genomics Initiative. He received his PhD at the University of Wisconsin-Madison and postdoctoral training at the Carnegie Institution of Washington's Department of Embryology.

Robert Haselkorn is a Distinguished Service Professor of Molecular Genetics and Cell Biology at The University of Chicago. He has been a leader in demonstrating how the filamentous, heterocystous cyanobacteria accomplish biological nitrogen fixation and photosynthesis simultaneously. Recently, he has studied acetyl-CoA carboxylase genes in wheat and in parasites. He was a Guggenheim Fellow at the Institut Pasteur, and a recipient of the Darbaker Prize of the Botanical Society of America and the Gregor Mendel Medal in Biological Sciences from the Academy of Sciences of the Czech Republic. He is a member of the National Academy of Sciences and a fellow of the American Academy of Arts & Sciences. He received his PhD in biochemistry from Harvard University and studied plant viruses in Cambridge, England as a postdoc. Dr. Haselkorn is chairman of the Board of Directors and a co-founder of Integrated Genomics, Inc., a genome sequencing and bioinformatics company.

Elizabeth "Toby" Kellogg is the E. Desmond Lee and Family Professor of Botanical Studies at the University of Missouri-St. Louis. Her current research focuses on the evolution of development, identifying genetic changes that correlate with differences among species, genera and families, working specifically with the grass family, which includes the cereal grasses and their numerous wild relatives. She has studied evolution of C4 photosynthesis and evolution of sex expression, but most of her current work involves the architecture of inflorescences—characteristics that have been used for hundreds of years in grass classification. Incorporating an evolutionary tree in this research identifies specific branches of the tree where changes have occurred that have led to modern grass phenotypes. Her lab has been involved in collaborative work producing evolutionary trees for the grass family and many of its 10,000 species, and has helped establish a well-supported phylogeny of the family with the goal of identifying the candidate genes that may be responsible for changes in inflorescence morphology. Dr. Kellogg is the recipient of the Engler Medal of the International Association of Plant Taxonomists and the Hoopes Prize for excellence in supervising undergraduate research. She has served as the President of the Society of Systematic Biologists and on the editorial boards of Australian Systematic Biology, Molecular Biology and Evolution, and the International Journal of Plant Sciences. She received her PhD in biology from Harvard University.

Robert L. Last is a Visiting Scientist at the Max Planck Institute for Chemical Ecology in Jena, Germany. He was Director of Discovery Genomics at Cereon Genomics, LLC in Cambridge, Massachusetts from 1998-2002. From 1989 to 1998, Dr. Last was on the staff of the Boyce Thompson Institute for Plant Research at Cornell University, and he was an adjunct Professor of Genetics and Development at Cornell. He was awarded an NSF Presidential Young Investigator Award in 1991, and named a Monsanto Fellow in 2002. His research interests included plant stress adaptation, secondary metabolism, and amino acid biosynthesis. Dr. Last did postdoctoral research in plant genetics at Whitehead Institute from 1986-1989, and received his PhD in Biological Science from Carnegie-Mellon University in 1986. Dr. Last's service to the scientific community has included chairing the Plant Molecular Biology Gordon Conference, organizing the Cold Spring Harbor *Arabidopsis* Genetics Course from 1995-1997, and serving as a member of the NIH Biological Sciences 1 Postdoctoral Fellowship Study Section. He is currently an Associate Editor of the journal Plant Physiology and a member of the Keystone Conferences Scientific Advisory Board.

Robert Martienssen is a Professor at Cold Spring Harbor Laboratory, Cold Spring Harbor, New York and leads the plant biology group there. Dr. Martienssen is a plant geneticist, working on transposons, genome biology, and developmental genetics of maize and the model plant *Arabidopsis thaliana*. He has developed reverse genetics strategies using transposons that have become powerful and widely used tools in plant genetics research and led to the formation of gene function databases for *Arabidopsis* and maize. In addition to his full-time academic position, Dr. Martienssen is a co-founder and member of the Board of Directors of Orion Genomics, an agricultural genomics company based in St Louis, MO. Dr. Martienssen received his BA in Natural Sciences (Genetics) from Cambridge University, England, in 1982, and his PhD from the Plant Breeding Institute and Cambridge University in 1986. He held an EMBO postdoctoral fellowship at the University of California at Berkeley from 1986-1988 and has been on the faculty at Cold Spring Harbor Laboratory since 1989. Dr. Martienssen was a co-recipient of the Kumho International Science Award in Plant Molecular Biology and Biotechnology, 2001.

Susan McCouch is associate professor of plant breeding and plant biology at Cornell University. The focus of her research program is to develop and apply molecular tools for rice improvement. She has served as a Plant Genome Panel member for the USDA National Research Initiative, on the tri-agency (NSF/USDA/DOE) Panel on Plant Genome Initiative, and the National Academy of Sciences (NSF) colloquium “Protecting Our Food Supply.” She received her PhD in plant breeding, genetics, and entomology from Cornell University.

Ernest Retzel is director of the Center for Computational Genomics and Bioinformatics at the University of Minnesota. His research interests include high-performance distributed computing, genomic databases and data mining, visualization, and extending databases for genomic information and automated analysis. He has served on the Scientific Advisory Board of the National Center for Genome Resources, the

University of Nevada Genome Center, and several computer-industry advisory boards. He received his PhD in microbiology from the University of Minnesota.

Chris R. Somerville is director of the Department of Plant Biology at the Carnegie Institution of Washington and professor of biological sciences, Stanford University. Dr. Somerville received his BSc in mathematics and a PhD in genetics from the University of Alberta. Dr. Somerville has pioneered the use of the small mustard plant, *Arabidopsis thaliana*, as a model species for plant molecular genetics. The subjects of his research contributions include plant genomics, embryo development, and the synthesis of structural and storage components of plant cells. Dr. Somerville is a member of the US National Academy of Sciences and a fellow of the Royal Society of London and the Royal Society of Canada. He has received the Alexander von Humboldt US Senior Scientist Award, a Presidential Young Investigator Award from the National Science Foundation (NSF) and the Charles F. Schull Award and the Gibbs Medal from the American Society of Plant Physiology. He has been awarded honorary degrees from Queens University, Wageningen University, and the University of Alberta. Dr. Somerville is on numerous editorial boards, and has served on various advisory panels for NSF, the National Institutes of Health, the US Department of Agriculture, and other agencies and institutions. Dr. Somerville is chairman of the Board of Directors of Mendel Biotechnology, Inc., a company that uses functional genomics to study plant genes. He has also served as a consultant to many companies, including Unilever, DuPont, Monsanto, Eli Lilly, Pioneer, and Dow.

Susan Wessler is Distinguished Research Professor of Plant Biology and Genetics at the University of Georgia at Athens where she has been since 1983. Her research involves transposable elements and their impact on the evolution of genes and genomes. She has been involved in the isolation of many plant elements including Activator/Dissociation (Ac/Ds). Most recently, her lab discovered miniature inverted repeat transposable elements (MITEs), the most predominant element associated with plant genes. Currently she is heading up a collaborative project that is using computational and experimental approaches to identify and characterize most of the transposable elements in the two sequenced rice genomes. She received her Ph.D. in Biochemistry from Cornell University and began her studies on plant transposable elements while a Postdoctoral Fellow of the American Cancer Society at the Carnegie Institution of Washington in Baltimore. She is a member of the U.S. National Academy of Sciences. She is currently Associate Editor for Plant Physiology and is on the Editorial Boards of the Proceedings of the National Academy of Sciences and Current Opinions in Plant Biology.

John Yates is professor of cell biology at the Scripps Research Institute and director of protein and metabolite dynamics at the Torrey Mesa Research Institute. His laboratory uses tandem mass spectrometry as a technique for characterizing a proteome, using detailed information yielded by the mass spectrometer to identify proteins from complex mixtures. His research draws on biology, chemistry, and computer science to increase the scope, sensitivity, and throughput of technologies for practical proteomics. He is a recipient of the Pehr Edman Award in Protein Chemistry, and serves on the editorial advisory boards of several journals, including the Journal of Proteome Research and is an associate editor of Analytical Chemistry. He received his PhD in chemistry from the University of Virginia.

APPENDIX C

Number of Publications in 2000-01 of 50 Most Cultivated Species

The 50 most-cultivated species were found by searching FAO's FAOSTAT agricultural database of primary crops by area harvested. The number of publications was obtained by searching ISI's Web of Science, a data base of 5,700 journals, for key words (species or genus names) in titles of articles. Number of research groups was estimated by eliminating research groups bearing the same institutional address and at least one common author (to avoid double counting). Model species are provided as references. Frequently cited papers are defined as being among the top 5 most-commonly cited papers in 2000 and 2001.

Rank	Most Cultivated Crops Common Name	Area Cultivated (ha)	Key Word Searched	Average number of publications 2000/01	Average number of research groups 2000/01	Average times that a frequently cited paper is cited 2000/01
1	Wheat	213,816,865	Triticum durum	11	10	2
	Wheat	213,816,865	Triticum aestivum	88	78	8
2	Rice, Paddy	151,541,091	Oryza sativa	113	90	9
	Rice, Paddy	151,541,091	Oryza	124	101	10
3	Maize	137,596,759	Zea mays	88	81	7
4	Soybeans	75,539,904	Glycine max	54	46	3
5	Barley	54,268,528	Hordeum vulgare	42	40	6

continued

Rank	Most Cultivated Crops Common Name	Area Cultivated (ha)	Key Word Searched	Average number of publications 2000/01	Average number of research groups 2000/01	Average times that a frequently cited paper is cited 2000/01
6	Sorghum	42,634,739	Sorghum bicolor	24	22	4
7	Millet	37,401,852	Pennisetum	19	17	1
	Millet	37,401,852	Echinochloa	9	8	1
	Millet	37,401,852	Panicum	8.5	7	2
	Millet	37,401,852	Setaria	8	6	1
	Millet	37,401,852	Paspalum	16	14	2
8	Seed Cotton	33,978,922	Gossypium	30	28	3
9	Groundnuts in Shell	25,538,439	Arachis hypogaea	28	27	3
10	Rapeseed	23,961,310	Brassica campestris	11	9	9
	Rapeseed	23,961,310	Brassica napus	76	70	7
	Rapeseed	23,961,310	Brassica rapa	21	19	4
11	Beans, Dry	23,240,843	Phaseolus vulgaris	73	67	5
	Beans, Dry	23,240,843	Phaseolus	89	82	5
12	Potatoes	19,301,006	Solanum tuberosum	47	44	4
13	Sugar Cane	19,245,069	Saccharum officinarum	9	9	0
14	Sunflower Seed	18,397,973	Helianthus annuus	33	31	3
15	Cassava	17,019,975	Manihot esculenta	19	14	3
16	Alfalfa for Forage+Silage	15,728,690	Medicago sativa	26	24	4
17	Oats	12,863,933	Avena	26	24	3
18	Coconuts	10,833,538	Cocos nucifera	8	7	1
19	Coffee, Green	10,766,384	Coffea	21	16	3
20	Cow Peas, Dry	9,862,865	Vigna	52	49	3
	Cow Peas, Dry	9,862,865	Vigna unguiculata	23	23	3
21	Oil Palm Fruit	9,712,150	Elaeis guineensis	11	11	2
22	Rye	9,672,849	Secale cereale	16	14	3
23	Sweet Potatoes	9,391,910	Ipomoea batatas	12	13	2
24	Chick-Peas	8,582,102	Cicer arietinum	25	23	2
25	Olives	8,076,533	Olea europeaea	14	13	5
26	Natural Rubber	7,903,421	Hevea	14	13	2

Rank	Most Cultivated Crops Common Name	Area Cultivated (ha)	Key Word Searched	Average number of publications 2000/01	Average number of research groups 2000/01	Average times that a frequently cited paper is cited 2000/01
27	Sesame Seed	7,783,611	Sesamum indicum	10	9	1
28	Grapes	7,305,355	Vitis vinifera	40	37	5
29	Citrus Fruit,Total	7,201,853	Citrus	185	152	5
30	Cocoa Beans	7,186,666	Theobroma	13	12	1
31	Peas, Dry	6,227,031	Pisum sativum	56	49	9
32	Sugar Beets	5,979,044	Beta vulgaris	25	23	6
33	Apples	5,594,443	Malus	24	23	3
33	Apples	5,594,443	Malus domestica	9	8	1
34	Plantains	4,671,010	Musa	25	24	3
34	Bananas	4,201,809	Musa	25	24	3
35	Tobacco Leaves	4,182,338	Nicotiana tabacum	40	36	4
36	Yams	4,050,872	Dioscorea	14	14	0
37	Pigeon Peas	3,830,215	Cajanus cajan	14	14	1
38	Lentils	3,765,463	Lens culinaris	10	10	2
39	Tomatoes	3,745,229	Lycopersicon	66	60	4
39	Tomatoes	3,745,229	Lycopersicon esculentum	43.5	41	4
40	Oranges	3,598,625	Citrus sinensis	8	7	1
41	Watermelons	3,133,412	Citrullus	7	6	1
42	Triticale	3,059,216	Triticale	30	26	2
43	Mangoes	3,036,343	Mangifera indica	18	16	1
44	Linseed	3,032,743	Linum usitatissimum	13	13	1
45	Buckwheat	2,787,780	Fagopyrum esculentum	5	5	1
46	Cashew Nuts	2,767,486	Anacardium occidentale	5	5	0
47	Onions, Dry	2,738,699	Allium cepa	27	23	5
48	Cabbages	2,623,010	Brassica oleraceae	36	31	7
49	Clover for Forage+Silage	2,362,691	Trifolium	37	34	5
50	Plums	2,295,961	Prunus domestica	2	2	0

continued

Rank	Most Cultivated Crops Common Name	Area Cultivated (ha)	Key Word Searched	Average number of publications 2000/01	Average number of research groups 2000/01	Average times that a frequently cited paper is cited 2000/01
	Other Model & Reference Species		Arabidopsis thaliana	277	235	31
			Brachypodium	2	2	1
			Chlamydomonas	100	79	16
			Lotus	24	21	7
			Medicago truncculata	20	17	11
			Populus	64	55	11

SOURCE: FAOSTAT, 2001, http://apps.fao.org/page/collections?subset=agriculture; ISI Web of Science, 2000-2001.

Glossary

Anchor markers - Genes conserved across genomes that can be used in comparative mapping and phylogenetic analysis.

Annotation - Adding pertinent information such as start/stop codons, intron/exon boundaries, 5' and 3' untranslated regions, gene coded for, amino acid sequence, or other commentary to the database entry of raw nucleotide sequence information.

Apomixis - Asexual plant reproduction in which meiosis and fertilization are altered so that only one parent contributes genes to the offspring.

Bacterial artificial chromosomes (BAC) - Vectors used to clone large (100-300 kb) inserts of genomic DNA that can be stably maintained in *Escherichia coli* cell cultures.

BAC end sequencing - Sequencing of the ends of BAC clones (roughly 600 nucleotides), a process that is useful for creating a tiling pathway of BACs across a chromosome.

Bermuda rules -The data release policy of the International Sequencing Consortium, deriving from the principle that sequence will be of greatest public benefit if freely available. Under the rules, assemblies of 1-2 kilobases are deposited in public data banks every 24 hours, and no patents are filed.

Bioinformatics - The science of managing and analyzing biological data, including genomic research data, using advanced computing techniques.

Brassicaceae - The mustard family. Members include Arabidopsis, canola, broccoli, rape, cabbage, kale, cauliflower.

Cis-acting elements - DNA sequences in the vicinity of the structural portion of a gene that regulate gene expression.

cDNA (complementary DNA) libraries - A collection of DNA clones representing a population of messenger RNA from which all non-coding, intron sequences have been removed.

Comparative Genomics - The comparison of gene and genome structure, function and evolution across taxa.

Coverage - representation of the accuracy of sequencing. For 5× coverage, a given base has been examined, or "covered," 5 times.

Diploid - A state in which each type of chromosome is present as a pair of homologous chromosomes.

DNA (deoxyribonucleic acid) - The fundamental molecule encoding genetic information. DNA is a double-stranded molecule held together by weak bonds between base pairs of nucleotides. The four nucleotides in DNA contain the bases adenine (A), guanine (G), cytosine (C), and thymine (T).

DNA sequence - The relative order of the nucleotide bases making up the DNA along the chromosomes.

Draft sequence - The determined order of base pairs of a chromosomal area at a level of 4 to 5× coverage.

Evolutionary nodes - Points of evolutionary divergence, representing ancient speciation events.

Expressed Sequence Tags (ESTs) - The result of large-scale partial sequencing of randomly selected cDNA clones. ESTs are a useful tool for gene identification, localization, and mapping.

Fabaceae - The legume family. Members include soybean, beans, cowpeas, peas, and alfalfa as well as numerous tropical trees.

Finished sequencing - Additional sequencing needed to fill gaps, reduce ambiguities, and increasing sequence accuracy to no more than 1 error per 10,000 base pairs. The finished version will provide an estimated 8× to 9x coverage of each chromosome.

Fixed inbred lines - Plant lines in which all loci are homozygous and, if selfed, breed true. Populations of fixed inbred lines may be developed from biparental crosses, i.e., Recombinant Inbred, (RI) lines or Doubled Haploid (DH) lines. These are considered immortal populations because each line retains its genetic integrity when selfed.

Forward genetics - Identification of mutants followed by genetic crosses to locate the genes in which the mutations occurred.

Functional genomics - The analysis of genes, their resulting proteins, and the role played by the proteins in an organism's biochemical processes.

Gene-rich regions - DNA sequences that contain a high percentage of coding sequences, and less than average amounts of non-coding DNA.

Genetic linkage - The residing of genes closely together on the same chromosome arm. Linked genes tend to recombine less frequently than unlinked genes.

Genetic map - A linear designation of the relative positions of genes on chromosomes and the distance between them, in linkage units, based on frequency of intergenic recombination.

Genome - The entire complement of genetic material in a chromosome set.

Genomics - The science and technology associated with the large-scale DNA sequencing of the complete set of chromosomes from a species and the interpretation of that sequence in terms of its organization, function and evolution.

Genomic DNA - DNA containing both coding (exon) and noncoding (intron) sequences.

Genotype - Genetic composition of an individual.

Heterochromatin - Highly compacted DNA containing very few genes.

Heterosis - The observation that in some circumstances, a hybrid offspring exhibits higher fitness (or productivity) than either of its parents.

Homeostasis - The tendency of an organism or population to reach equilibrium and resist change.

Homologue - A gene related to a second gene by descent from a common ancestral DNA sequence. The term may apply to the relationship between genes separated by the event of speciation or to the relationship between genes separated by the event of genetic duplication.

High-throughput sequencing - A fast, bulk method of determining the order of bases in DNA.

Immortal mapping populations - See Fixed inbred lines (above).

Library - An unordered collection of related pieces of DNA (for example, cloned DNA from a particular organism) whose relationship to each other can be established by physical mapping.

Linkage disequilibrium - A state in which alleles are inherited together more often than can be accounted for by chance, indicating that the two alleles are physically close on the DNA strand or that they are simultaneously the product of selection.

Loci (plural of locus) - Positions of genes or other markers on a chromosome.

Messenger RNA (mRNA) - Nucleotide segments carrying information from DNA and serving as a template for protein synthesis.

Microarrays - Sets of miniaturized chemical reaction areas that may also be used to test DNA fragments, antibodies, or proteins. Microarrays can be used to monitor genetic diversity based on DNA-DNA hybridization, or to measure changes in gene expression, based on hybridization of a given mRNA population to cDNA embedded in a silica chip.

Model species - A plant species that can serve as a unifying experimental model. The committee refers to *Arabidopsis thaliana* as the model plant species.

P1-derived artificial chromosome (PAC) - Vectors based on a bacteriophage (a virus) PI genome used to clone large DNA fragments that can be stably maintained in *Escherichia coli* cell cultures.

Phenotype - Detectable, outward manifestations of a given genotype.

Phylogeny - Evolutionary relationships among organisms; the developmental history among organisms.

Physical map - A map representing physical distances between genes and other markers (for example, restriction enzyme cutting sites), as measured in nucleotide base pairs.

Ploidy - The number of chromosome sets in a given organism.

Poaceae - The grass family. Members include rice, maize, sorghum, sugarcane, wheat, barley, oat, fescue.

Polymerase chain reaction (PCR) - A technique used for amplification of specific DNA segments.

Polyploid - An organism with more than two sets of a basic, or monoploid number of chromosomes, such as triploid, pentaploid, or hexaploid. Many plant genomes are polyploid.

Protein chip - Microarray technology for protein profiling.

Quantitative trait loci (QTL) - Genetic loci that affect a quantitatively inherited trait.

Reference species - Plant species that serve as references for the species in major agronomically relevant plant taxa.

Reverse genetics - Isolation and sequencing of a desired gene, and subsequent creation of mutations in it.

Serial analysis of gene expression (SAGE) - A tool allowing analysis of overall gene expression.

Shotgun sequencing - Sequencing method that involves randomly sequenced cloned pieces of the genome, in contrast to “directed” sequencing methods, in which pieces of DNA from known chromosomal locations are sequenced.

Single nucleotide polymorphisms (SNPs) - DNA sequence variation that occurs when a single nucleotide (A, T, G, or C) is altered. SNPs can be useful in detecting genetic variation among individuals in a given population.

Solanaceae - The nightshade family. Members include tomato, potato, tobacco, eggplant and petunia.

Synteny - Linkage of genes along a chromosome. Conserved synteny refers to the conservation of gene order on chromosomes of different species.

Transcription - The synthesis of an RNA copy from a sequence of DNA; the first step in gene expression.

Transcriptome - The full complement of activated genes, transcripts, and mRNA in a given tissue at any particular time.

Transformation - The process of integrating exogenous DNA into the genetic material of another organism.

Transposable elements - Genetic elements that may "jump" to new locations, often disrupting the function of the genes into which they are inserted. These elements sometimes encode enzymes that synthesize an identical copy of the insertion into a new site.