National Plant Genome Initiative

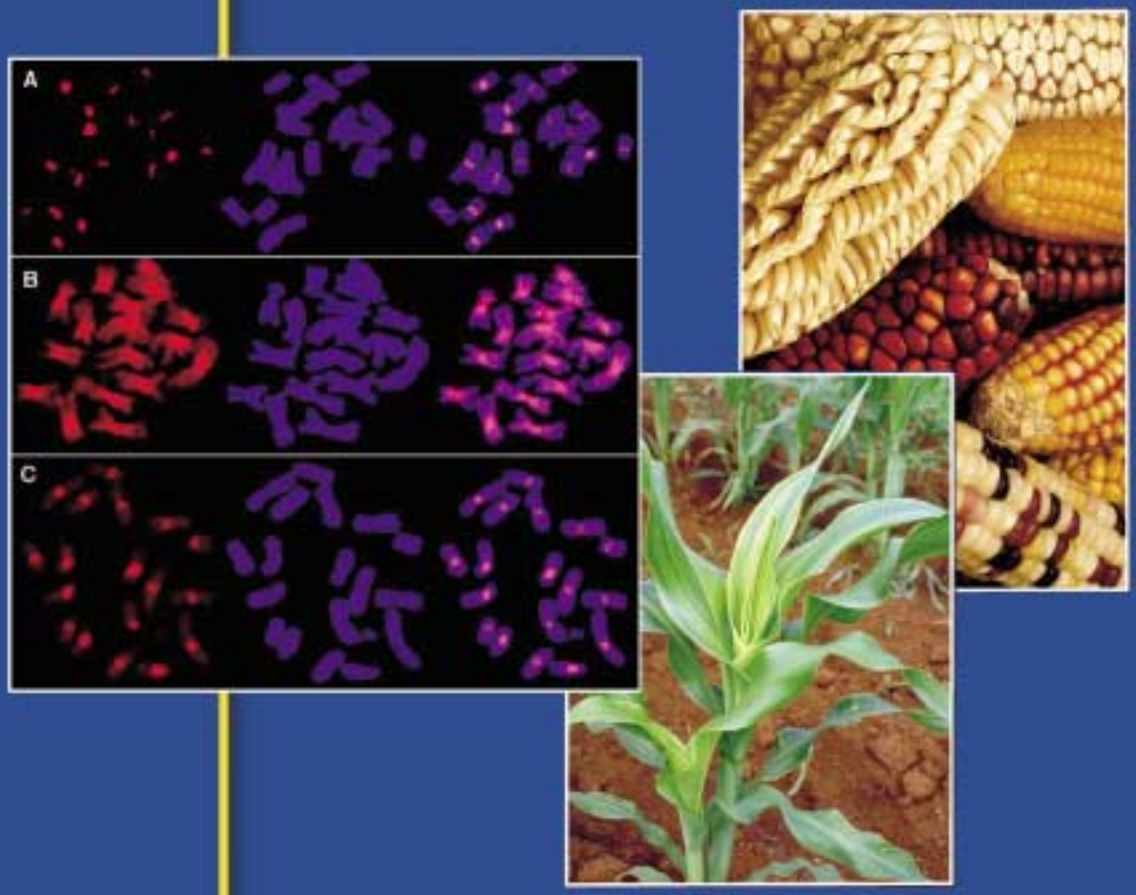
Progress Report

December 2001

National Science and Technology Council

Committee on Science

Interagency Working Group on Plant Genomes



About the National Science and Technology Council

The National Science and Technology Council (NSTC) was established by Executive Order on November 23, 1993. This cabinet-level council is the principal means for the President to coordinate science, space, and technology policies across the Federal Government. NSTC acts as a "virtual" agency for science and technology to coordinate the diverse parts of the Federal research and development enterprise.

An important objective of the NSTC is the establishment of clear national goals for Federal science and technology investments in areas ranging from information technologies and health research, to improving transportation systems and strengthening fundamental research. This council prepares research and development strategies that are coordinated across Federal agencies to form an investment package that is aimed at accomplishing multiple national goals.

To obtain additional information regarding the NSTC, contact the NSTC Executive Secretariat at (202) 456-6101.

Note: This document does not represent the final determination in an overall Administration budget decision-making process. The programs presented in this report will have to compete for resources against many other high-priority Federal programs. If these programs compete successfully, they will be reflected in future Administration budgets.

Cover Photos: The National Plant Genome Initiative has supported several major projects on maize (corn) genomics. A rich collection of mutant maize (two photos on right) provides valuable materials for maize research. An example is a project at University of Georgia on maize DNA sequences that are responsible for accurate chromosome segregation during cell division. Cover images on the left show where particular sequences are located on the chromosomes. The DNA sequences under study are shown in red or magenta, and the chromosomes are shown in blue. Some sequences are found only at centromeres (A), others are found in centromeres as well as other non-centromeric regions (B), and still others are found mostly at centromeres with minor localization elsewhere (C). [Left - courtesy of Dr. Kelly Dawe, University of Georgia; Middle - courtesy of Dr. M. G. Neuffer and MaizeDB, University of Missouri; Right - courtesy of the Agricultural Research Service, USDA]

National Plant Genome Initiative

Progress Report

December 2001

National Science and Technology Council
Committee on Science
Interagency Working Group on Plant Genomes

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

December 18, 2001

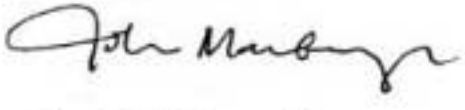
Dear Colleague:

This report provides an update from the National Science and Technology Council (NSTC) Interagency Working Group (IWG) on Plant Genomes on the progress of the National Plant Genome Initiative (NPGI). In addition to coordinating the activities of the NPGI participating agencies, the IWG monitors the progress of the NPGI and documents significant progress in annual reports. This new progress report is the third in this series.

Since its inception in 1997, significant progress has been made in all of the NPGI objective areas. Early year investments into building research infrastructure are now enabling a broad community of scientists to participate in plant genome research. New discoveries are being made that are fundamental to the understanding of the structure and function of the plant genome. The NPGI is also contributing to the advancement of the entire field of genomics by adding the unique scientific perspectives of plant biology. The impacts of the NPGI are being felt beyond the research laboratory, from the farmer's fields to the classroom and the general public.

The IWG is in the process of developing a new five-year plan. I am confident that the IWG will continue to guide the NPGI to push the frontiers of plant genomics and to respond to new challenges and scientific opportunities that will benefit all of society.

Sincerely,



John H. Marburger, III
Director

Interagency Working Group on Plant Genomes
Committee on Science
National Science and Technology Council

Co-Chairs

Mary E. Clutter
Assistant Director for Biological Sciences
National Science Foundation

Joseph Jen
Under Secretary
Research, Education, and Economics
U.S. Department of Agriculture

Eileen Kennedy
Deputy Under Secretary
Research, Education, and Economics
U.S. Department of Agriculture
(Until January 2001)

Members

Gregory L. Dilworth
Section Chief
Office of Basic Energy Sciences
U.S. Department of Energy

Noah Engelberg
Program Examiner
Office of Management and Budget

Clifford Gabriel
Deputy to the Associate Director
Office of Science and Technology Policy

Elke Jordan
Associate Director
National Human Genome Research Institute
National Institutes of Health

Sally Rockey
Deputy Administrator
Cooperative State, Research, Education and Extension Service
U.S. Department of Agriculture

Judy St. John
Associate Deputy Administrator
Agricultural Research Service
U.S. Department of Agriculture

Table of Contents

I. Executive Summary

Since its inception in 1998, the National Plant Genome Initiative (NPGI) has supported research at the frontiers of plant biology. As outlined in the long-range plan, the focus in its early years was on building the research infrastructure for plant genome research. More recently, the focus has shifted increasingly toward understanding the function of the genomes of economically important plants and investigating the genomic basis for the plant processes of economic importance. The Interagency Working Group (IWG) for Plant Genomes, under the auspices of the National Science and Technology Council (NSTC), has provided overall guidance and oversight to the NPGI. The IWG summarizes the progress of the NPGI by issuing annual progress reports; this is the third such report.

Scientific and Technical Progress:

Sequencing of the genomes of model plants

- The *Arabidopsis* genome-sequencing project has been completed, resulting in the most accurate and complete higher eukaryotic genome sequence obtained to date. Worldwide efforts to identify the function of the 25,498 *Arabidopsis* genes are underway.
- The International Rice Genome Sequencing Project (IRGSP) consortium has deposited, as of September 2001, 137 million bases in GenBank, the public database at the National Center for Biotechnology Information of the National Institutes of Health. A year ago, the total bases in GenBank was 26 million. The IRGSP has recently adopted a new strategy aimed at completion of the rice genome by late 2002, six years ahead of the original estimate of 2008. The new strategy takes advantage of the availability of rough draft sequence from Monsanto.

New research resources and discoveries

- In addition to the genome sequence of the two model plants, a large number and variety of research tools and resources for plant genome research have been accumulating and are being utilized by the entire research community. Some examples include the following: (1) a large collection of plant ESTs (Expressed Sequence Tags) in GenBank, which numbers over 1,000,000 compared to 50,000 in 1998; (2) BAC (Bacterial Artificial Chromosome) libraries for 72 plant species available to the public; and (3) a novel mapping tool for maize (corn).
- Through the use of genomic tools and genome-wide approaches, scientists are gaining new insights into the various plant processes of economic importance. Examples described in this report include the following: (1) identification of specific genes involved in plant responses to drought and salt stresses; (2) finding clues to the epigenetic phenomena, which are heritable changes without a change in DNA sequence; and (3) identification of bacterial genes that cause diseases in host plants.

Technology development

- Successful application of “optical mapping” technologies to develop a whole genome map of rice and other large genomes is underway. Optical mapping is a new way to map a whole genome and was originally developed for a small microbial genome.

- A new technology that allows rapid selection of plants with mutations in any gene has been developed. It is called Targeting Induced Local Lesions in Genomes or TILLiNG, and is now being applied to animal systems.
- An elegant technology has been developed to isolate the gene-rich islands of the maize genome from the repetitive DNA which does not contain genes. The technology is based on the difference in DNA methylation between most genes and the highly repetitive components of the genome. The gene-enriched fraction is expected to represent about 10% of the whole maize genome, and should contain most of the genes. The technology is applicable to any complex genome.

Data management and informatics

- All NPGI-supported projects aim to release the results of their research to the public in a timely manner and in an accessible and useable form.
- Some groups have developed an integrated organism database where all the information as well as associated software tools are made available to the general research community. Grass and *Medicago* databases are two examples.
- Most genome projects have produced large datasets of diverse types that are often scattered in space, time, and format. This has made comparative analyses among data from different projects difficult and makes the creation of integrated databases a significant, if not often insurmountable, challenge. In recognition of the urgent need to define "an accessible and useable database" in the context of the NPGI, a workshop was convened on September 11 & 12, 2001, in Rockville, Maryland. A final report of the workshop will be discussed at the Plant and Animal Genome Conference in January 2002.

Broader Impacts:

Impact on plant science research

- One of the ultimate goals of the NPGI is to advance the field of plant sciences through the genomic revolution. The expectation is that plant genomics tools will bring new approaches to plant science research and will inspire renewed interest among beginning and early career scientists. Numerous resources, including a complete plant genome sequence, targeted gene disruptions and mutant seed stocks catalogued gene by gene, and databases containing details ranging from single sequence changes to the classification of gene families for functional genomics, provide a rich arsenal of tools for any 21st century plant biologist to undertake genetic analysis and technological manipulation of any physiological process of choice. The availability of such tools has radically changed, and will continue to change, how plant biology research is conducted.

Beyond laboratories

- Fundamental discoveries in plant genome research can often be transferred to practical applications. In order to foster efficient transfer of knowledge, some researchers have organized consortia that include public and private sectors. Examples include the following: (1) a national wheat Marker Assisted Selection (MAS) consortium including 12 public wheat-breeding and research programs across the U.S.; (2) tomato researchers working with an expanded consortium of seed companies and non-profit organizations to identify and transfer tomato genes relevant to economically important traits into other vegetable crops; (3) plant genome researchers directly addressing the problem of winter survival of alfalfa in the Midwest; and (4) an effort in Texas to increase cotton diversity and to make the information available to cotton growers.

Societal and educational impacts

- Several positive interactions between academic researchers and industry are highlighted, such as the three-way collaboration among DuPont, Incyte, and the maize mapping project. Monsanto has shared their rice genome sequence data with the international consortium, which has helped accelerate the rate of progress of the public effort. Intellectual property rights issues remain a major hurdle in forging a public-private partnership.
- Many NPGI projects integrate educational activities using various mechanisms. Some take advantage of well-established campus-wide programs, and others design new programs, such as (1) a summer program for high school students and their teachers to participate in mutant screening and characterization in maize; and (2) an opportunity for undergraduate summer research experience in both genetics and bioinformatics.
- As the focus of the projects moves toward functional genomics research and away from building research tools, there has been a significant increase in the number of projects that build and integrate graduate and postdoctoral training into project activities. These projects are providing unparalleled opportunities for cross-disciplinary training in plant biology, genomics, and bioinformatics.
- Increasing diversity in training the next generation of scientists is one of the major goals of the NPGI participating agencies. A number of partnerships have been formed between NPGI projects and various minority-serving institutions, designed to stimulate active cross-training of students and faculty as an integral part of the research project.
- Recognizing the importance of communicating plant genome research, NPGI investigators are reaching out to the public in a number of ways, such as (1) testifying at Congressional hearings; (2) giving public lectures on hot topics such as genetically modified organisms; (3) advising media on scientific content; and (4) participating in informal education activities.

In summary, research activities supported under the NPGI have resulted in many visible successes as well as potential discoveries in the making. As with any rapidly advancing field of research, plant genomics is constantly evolving. New discoveries lead to new lines of investigation. Advances in methodologies open up new ways to study long-standing questions in plant biology. New information is being put to practical use by the practitioners of plant biotechnology. The dynamic research environment is attracting young scientists to plant genomics. Disciplinary, organizational, and geographical barriers are breaking down as scientists from different backgrounds work together toward the common goal. It is clear that the NPGI has totally transformed the plant science community.

The Next Step:

The first five-year plan for the NPGI was developed in 1997 and published in early 1998. Following the roadmap outlined in the plan, the NPGI has made tremendous progress, surpassing many of the initial expectations. The initial five-year plan states that “The Initiative’s short-term goals, to be achieved over the next five years, focus on building a plant genome research infrastructure”. Now it is time to set goals for the next five years. The IWG is in the process of developing the next five-year plan for the NPGI. The target date for publication of a new plan is mid 2002.

II. Introduction

The Interagency Working Group (IWG) for Plant Genomes was appointed in May, 1997, by the National Science and Technology Council (NSTC), in response to a request from the Senate VA, HUD and Independent Agencies Appropriations Subcommittee. The IWG was composed of representatives from the National Science Foundation (NSF), the Department of Agriculture (USDA), the Department of Energy (DOE), the National Institutes of Health (NIH), the Office of Management and Budget (OMB), and the Office of Science and Technology Policy (OSTP). The charge to the IWG was to identify science-based priorities for a national plant genome initiative and to plan for a collaborative interagency approach to address these priorities. The IWG recommended establishment of the National Plant Genome Initiative (NPGI) and developed a five-year plan for the NPGI, which outlined a number of specific goals (http://www.ostp.gov/NSTC/html/npgireport.html).

The NPGI began supporting new research activities in FY1998 as a result of a significant increase in funding at NSF. In FY1999, DOE, USDA, and NSF provided additional research support for the NPGI when they began jointly supporting the U.S. rice genome-sequencing project as part of the international effort. In FY2000, increased funding for the NPGI was provided through the authorization of the Initiative for Future Agriculture and Food Systems (IFAFS). In FY2001, the DOE received an additional appropriation specifically for the NPGI. While NIH does not directly participate in the NPGI in terms of financial support of plant genome research projects, the significant NIH support for genomics research is a driving force in advancing the field of genomics in general.

Since the establishment of the NPGI in 1998, the plant genome research community has been making steady progress in all objectives outlined in the NPGI long-range plan. Progress is especially evident in building the research tools needed for the entire community to participate in plant genome research. The relatively large investment made in this area in 1998 and 1999 has begun to have visible impacts. There were an increased number of projects aiming to understand the structure, organization, and function of plant genomes by taking advantage of the research tools developed by the first NPGI projects. Genome sequencing of the model plants, *Arabidopsis* and rice, has made a huge impact in advancing the field of plant genomics. One of the most important impacts of the NPGI has been an increased level of interest and enthusiasm toward plant biology research. New tools, new discoveries, and new perspectives have opened new opportunities to study long-standing questions that lead plant biology research to the next level.

The IWG coordinates the NPGI by documenting progress and achievements of the program, as well as by reassessing priorities. The first year's progress report was issued in October 1999 (http://www.ostp.gov/html/genome/index.html); the second year's progress report was issued in November 2000 (http://www.ostp.gov/html/plantgenome/start.html). This report documents progress in the third year of the initiative, and outlines specific project highlights and previews of future developments.

III. Progress to Date

Rapid advances are being made on all fronts. Some of the highlights are described in this section in the context of the original goals outlined in the long-range plan for the National Plant Genome Initiative.

A. Scientific and Technical Progress

1. Sequencing of the model plant species; Arabidopsis and Rice

One of the goals of the NPGI is to support the complete high-resolution sequencing of the genomes of key model plant species.

The first complete sequence of a flowering plant genome

The *Arabidopsis* Genome Initiative (AGI), formed in 1996, is an international collaboration to complete the *Arabidopsis* genome sequence, the first whole genome sequence of any flowering plant species. The collaborators' efforts culminated in the most accurate and complete eukaryotic genome sequence to date. This is measured against a collection of sequences that includes the human genome, as well as the genomes of yeast, fruitfly, and worm. The *Arabidopsis* genome analysis and some accompanying major findings were presented in a landmark publication in the December 14, 2000 issue of Nature. The sequenced regions of the 125 Mb genome, including some centromeric portions but not the chromosome ends (telomeres), contain at least 25,498 genes. This constitutes the largest gene set when compared to other model organisms such as yeast (6,000 genes), *Drosophila* (14,000 genes), and *C.elegans* (20,000 genes).

A highlight of the *Arabidopsis* genome includes the existence of plant genes with counterparts in other eukaryotic organisms, including humans. These genes are necessary for life's basic machinery such as the synthesis and metabolism of proteins and biological macromolecules. It should be noted that there are some classes of shared genes that are more prevalent in plants than in animals. This is most likely because some processes, such as the ability to sense and transduce signals from the environment, while fundamentally similar in plants and animals, have evolved to a more diverse extent in plants. A varied array of such genes permits rooted plants to respond to environmental challenges from which they cannot escape. These genes allow them to redirect growth and development, as well as perform specialized plant functions. Likewise, the relatively high degree of similarity between plant genes and their bacterial counterparts can be attributed to the microbial ancestry of the chloroplast, the photosynthetic organelle in plants responsible for metabolism and energy production. There is also a significant number of genes unique to the synthesis and assembly of the cell wall, a structural element not present in animals. Cell walls are a major component of renewable biological resources like wood and cotton.

The complete *Arabidopsis* genome sequence has yielded some interesting surprises. For example, there is a high degree of genetic redundancy: almost

70% of the genome is duplicated. Many genes are members of multigene families which share a high degree of sequence similarity but which may perform a range of functions. This finding may hold important clues for understanding the evolution of chromosome structure and the generation of new biological properties. Finally, almost 30% of the *Arabidopsis* genome contains genes of unknown function that have yet to be found in any other organism. This suggests that a wealth of genetic, biosynthetic, and metabolic capabilities remain to be discovered.

The international rice genome sequencing project accelerates its pace

Rice has become the model system for the grasses. This is in part because of its relatively small genome size (estimated at 430 Mb) and its common genome structure and organization that it shares with other cereal crops. Its utility as a model genetic system and its importance as a major global food source have led to an international effort to obtain the complete rice genome sequence. The International Rice Genome Sequencing Project (IRGSP) consortium has been collaborating under the leadership of a Japanese group. This group has been promoting effective and accurate sequencing by sharing materials and technologies; effective coordination has been facilitated through periodic workshops, meetings, and web-based electronic communication and exchange between all participating sequencing groups. As of September 2001, the IRGSP had generated 137 million bases of rice sequence deposited in GenBank, the public genome sequence database at the National Library of Medicine. The IRGSP now aims to complete the rice genome sequence by 2002, six years ahead of the original estimate of 2008. This is made possible due, in large measure, to technological advances in high throughput sequencing, improved strategies for dealing with difficult to assemble and link chromosome sequences, increased resources from several participating countries, and the availability of physical map and draft sequence resources provided by Monsanto.

The U.S. group is sequencing all of chromosomes 3 and 10, and half of chromosome 11, which totals approximately 20% (86 Mb) of the rice genome. The latest estimates from the U.S. group are that the completed sequence of chromosome 10, comprising approximately 5000 genes, will be available by early 2002. It is expected that the Japanese group will complete chromosome 1 around the same time. These two accomplishments will represent the first completely sequenced, publicly available, rice chromosomes and will reveal new insights into the organization of the rice genome.

The steep increase in the amount of rice sequence available to the research community has already yielded valuable information about the organization and structure of this genome. Initial gene number predictions derived from a completed genome segment indicate that the average rice gene is 2200 nucleotides in length. This is similar to the average *Arabidopsis* gene. If this average gene size holds true for the entire rice genome, the number of genes in rice may be significantly greater than that in *Arabidopsis*. As is the case for *Arabidopsis*, a significant percentage of the predicted rice genes (at least half), at present, have no predicted function. One of the challenges for future genomic efforts will be to improve computational methods for gene prediction in large, complex genomes such as that of rice.

After genome sequencing

The major rationale for investing in the complete genome sequences of *Arabidopsis* and rice is to provide the entire plant research community with a reference, or "dictionary", against which the structure, organization, and function of all other plant genomes can be compared. However, possessing primary sequence information is just the beginning. The true success of a genome sequencing project is measured by how well the research community can make use of the sequence data for their own research, and by how much the sequence information contributes to the overall advancement of the field.

Now that the *Arabidopsis* genome sequence is complete, international efforts to identify the function of all of the genes by the year 2010 have begun. At the same time, the *Arabidopsis* genome sequence information is being used by plant biologists in a wide variety of ways ranging from fundamental studies of plant processes to development of improved crops.

In anticipation of completion of the rice genome sequence, an international consortium of geneticists, molecular biologists, and information scientists has been assembled. The consortium is called the Rice Functional Genomics Consortium (RFGC); member institutions include Yale University, Brookhaven National Laboratory, Cold Spring Harbor Laboratory, The International Center for Tropical Agriculture (CIAT) – Colombia, and California State University at Fullerton. The overall objectives of the RFGC are to develop the genetic, molecular, and bioinformatics infrastructures. Such infrastructure will be necessary to eventually complete an extensive public collection of rice lines as a resource for further functional analysis of the rice genome by the entire research community.

2. Increasing our knowledge of gene structure and function of important plant processes

Perhaps the most important scientific objective of the NPGI is to elucidate the structure, organization, and function of the genomes from plants of economic importance. When the initial five-year plan was developed in 1997, it was envisioned that research in structural genomics would include the production, mapping, and sequencing of expressed sequence tags (ESTs) from a number of plant species, as well as the construction of physical maps. Functional genomics research was envisioned to include both the determination of gene functions, as well as the determination of expression patterns for pathways or networks of genes under specific environmental conditions or during development. During the past four years, research on the structure, organization, and function of the genomes of plants of economic importance has advanced considerably beyond what was proposed in the original five-year plan.

2a. Building genomics tools and research resources

For the first two years, the NPGI invested in building research infrastructure, such as research tools and biological materials for community use, in order to enable the entire research community to participate in the NPGI. The outcomes from the earlier efforts are now enriching the research community with invaluable tools to advance plant genome research.

There are now over 1 million plant ESTs in GenBank

One of the most dramatic increases in the amount of public information can be illustrated by the number of ESTs in the dbEST database (http: / /

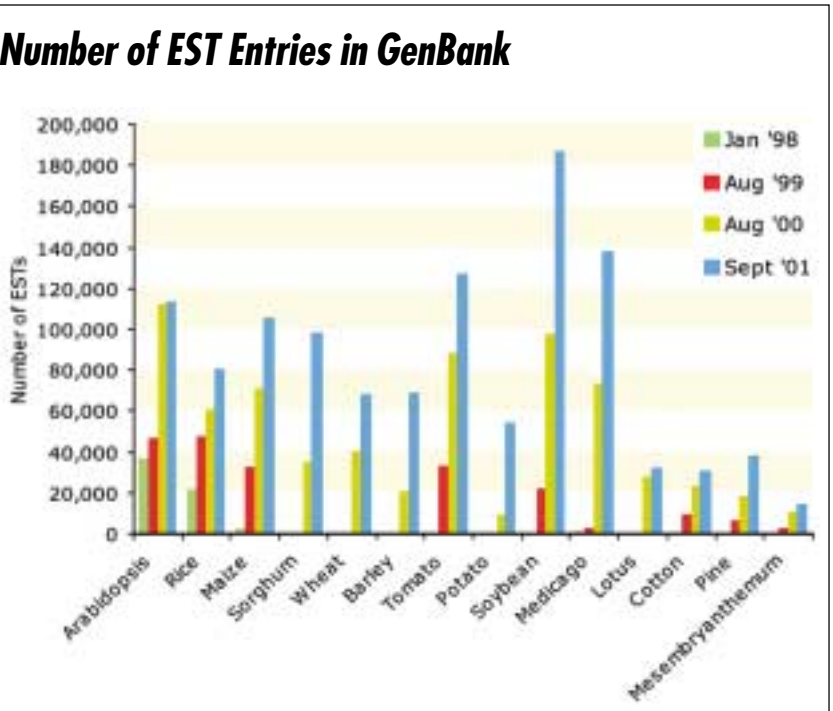


Figure 1. The number of expressed sequence tags (ESTs) for plant species available to the research community has dramatically increased in the 3 years of NPGI implementation.

www.ncbi. nlm.nih.gov / dbEST /). The dbEST now contains more than 1,000,000 plant ESTs, compared to fewer than 50,000 just three years ago (Figure 1).

Much of this increase is the direct result of support provided by the NPGI. In addition to the EST sequence information being readily accessible over the internet, the EST clones are publicly available to the research community as well. These ESTs have myriad uses for plant genome researchers, and are especially useful for identifying genes in related plants and as probes for measuring gene expression.

Table 1 lists web addresses where information about the various EST projects can be found.

Arabidopsis
http://www.arabidopsis.org

Rice
http://rgp.dna.affrc.go.jp

Maize
http://www.zmdb.iastate.edu

Sorghum
http://dogwood.botany.uga.edu/~prattlab/SorghumESTProject.html

Wheat
http://wheat.pw.usda.gov/wEST/nsf/title.html

Barley
http://www.genome.clemson.edu/projects/barley

Tomato
http://www.sgn.cornell.edu/tomato_project/progress.html

Potato
http://www.tigr.org/tdb/potato/est.html

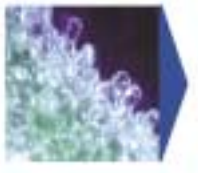
Soybean
http://www.genome.wustl.edu/est/soybean_esthmpg.html

Medicago truncatula
http://www.medicago.org

Lotus japanicus
http://www.kazusa.or.jp/en/plant/lotus/EST/

Cotton
http://cottongenomecenter.ucdavis.edu

Pine
http://pinetree.ccgb.umn.edu



Mesembryanthemum (iceplant)
http://stress-genomics.org

Table 1. Project site web addresses for various plant species. *Photos courtesy of the Agricultural Research Service, USDA and Eric Jellen, Brigham Young University.*

A complete set of oat-maize chromosome hybrids has been generated to allow precise mapping of maize genes

Most economically important plants have large, complex genomes. For example, the maize genome comprises approximately 2.5 billion base pairs and the wheat genome comprises 16 billion base pairs. Not surprisingly, the large size and complexity of these genomes have impeded integrated mapping and sequencing efforts. However, the precise mapping of the maize (corn) genome has become significantly easier due to novel mapping tools developed by a group of scientists at University of Minnesota and Oregon State University. Through systematic crosses, they have constructed a complete set of oat-maize hybrid plants, each containing the normal set of 21 oat chromosomes plus one of the 10 maize chromosomes (Figure 2). These hybrids, called "chromosome addition lines", allow rapid and precise placement of maize DNA sequences and other markers on chromosomes, including duplicated sequences or multigene families. Currently, EST sequences generated by another NPGI-supported group are being mapped onto chromosomes using these lines. With the more than 106,000 maize ESTs currently available, the resulting map will be a rich resource for both maize genome researchers and breeders. These lines and mapping data are all available to the public without restrictions at http://www.agro.agri.umn.edu/rp/genome/index.htm.

Figure 2. A normal oat plant (left) with a set of 2N oat chromosomes and an oat-maize plant (right) with a set of 2N oat chromosomes plus one pair of corn chromosomes. The oat-maize plant is characterized by more upright leaves, more branched stems, and a shorter panicle. *Courtesy of Ronald Phillips and Howard Rines, University of Minnesota.*

The BAC library resource for crop genomics

The BAC Resource Center (http://www.genome.clemson.edu/) at Clemson University was one of the first plant genome resource centers funded in 1998 to maintain and distribute high quality BAC libraries to the community at a reasonable cost. A "BAC" (Bacterial Artificial Chromosome) is a package containing a long segment (usually 70,000 to 150,000 base pairs) of a genome. A "BAC library" would thus contain a collection of BACs covering an entire genome. The contribution of this resource center to the advancement of plant genome projects focused on genome sequencing and analysis has been spectacular. The BAC Resource Center maintains and distributes 72 BAC libraries including rice, maize, cotton, barley, soybean, and sorghum, serving both the public and the private sectors. The map (Figure 3) indicates the size of the community served by the Center.

Development of other genomic resources

The NPGI continues to invest in the development of research resources for plant genomics. Examples of new resources under development include the following:

Genomic Resources for the Asparagales - Asparagus, garlic, and onion are closely related members of the order *Asparagales*; they represent the most economically important non-cereal food crops produced in the U.S. The relatively close relationship between these vegetables and rice increases the likelihood that genomic resources developed for rice will be applicable to the *Asparagales*, and *vice versa*. Researchers at the

Total Orders to the BAC/EST Resource Center at Clemson University Genomics Institute (CUGI)

Figure 3. The BAC / EST resources at Clemson University Genomics Institute have been used extensively throughout the U.S.

Photos courtesy of Eric Jellen, Brigham Young University.

University of Wisconsin, California State University, Michigan State University, and The Institute for Genomics Research are coordinating their efforts to construct a cDNA library (a collection of DNA clones representing expressed genes). These cDNAs will be sequenced and used to map a common set of markers in asparagus, garlic, and onion to establish the colinearity of genetic markers (synteny) among these vegetables and rice. The outcomes of this work will help to identify *Asparagales* genes that could serve as a target for the development of improved asparagus, garlic, and onion varieties.

Development of Functional Genomics Tools for Rosaceae - An ambitious effort to develop the requisite tools to speed the process of gene discovery and characterization in *Rosaceae* species is underway by a consortium of five U.S. institutions and one European laboratory; the consortium includes Clemson University, North Carolina State University, USDA, Institute de Recerca I Tecnologia Agroalimentaries (IRTA) – Spain, and the University of California at Davis. Peach is serving as a model plant for this group of plants. The *Rosaceae* species represents a family of plants that ranks in the top five in economic importance, and includes fruit trees (i.e., apple, pear, peach, apricot, plum, cherry), berries (i.e., raspberries, blackberries, strawberries), nuts

Photo courtesy of the Agricultural Research Service, USDA.

(i.e., almond), ornamentals (i.e., roses, flowering cherry, crabapple, quince), and trees for wood (e.g., black cherry). Despite the importance of these crops, identification of genes controlling economically important characters has been limited by a dearth of genomic resources. The tools now being developed will enable rapid advances in our understanding of the genomic structure, organization, and function in this economically important family of plants.

2b. New discoveries

Taking full advantage of the publicly available tools and materials mentioned above as well as other resources, NPGI-supported projects focus increasingly on functional genomics research. Some of these projects are beginning to report exciting new findings.

Defining the plant genes involved in environmental stress responses: a common set for all plants and unique sets for individual species

One of the major differences between plants and animals is that plants are immobile. As a result, plants respond to environmental changes very differently from animals. A consortium of scientists at the University of Illinois, the University of Arizona, Purdue University, Oklahoma State University, and the University of Nevada at Reno, have set out to identify all plant genes involved in plant responses to environmental stresses such as salinity, drought, flooding, freezing, and heat. They have found that there is a set of environmental stress response genes that is conserved throughout evolution, from Cyanobacteria and algae, to fungi and higher plants. In addition, each plant species has developed its own unique set of genes involved in receiving and responding to various environmental cues. Preliminary results indicate that some of the common genes are involved in ion transport across cell

Microarray Analysis After Salt Stress

(a)

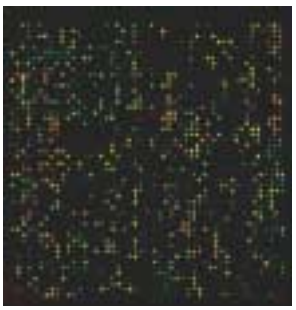
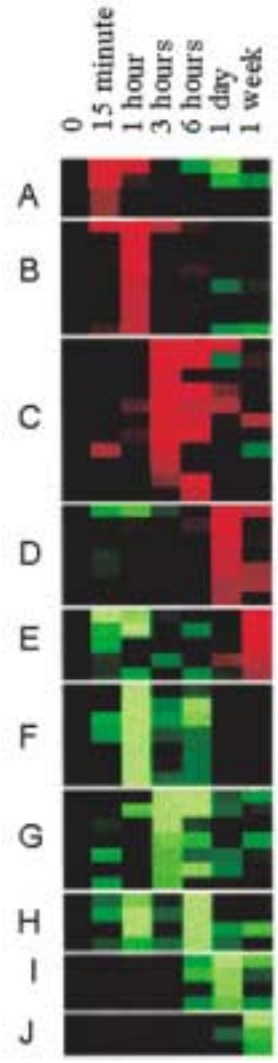


Figure 4. (a) Microarray analysis measures the gene expression of many genes at once. Each spot corresponds to a known gene sequence. In the top left microarray image, rice genes that show increased expression under salt stress are indicated in red, while those that show decreased expression are indicated in green.

(b)



(b) Results from microarray experiments are further analyzed using a computational method called cluster analysis. In this analysis, rice genes are grouped into (A) to (J) according to their relative expression levels at different time points after the plant is exposed to salt stress. The identity of genes in each cluster allows scientists to develop hypotheses about how plants perceive and respond to salt stress through changes in gene expression. *Courtesy of Dr. Hans Bohnert, University of Illinois.*

membranes. They have also found that a basic component of the environmental signal transduction pathway – how the environmental stress signal is perceived and communicated within the cell – is likely to be conserved throughout the plant kingdom. At the same time, plants have developed additional unique components. Data resulting from this project have been shared openly at http://www.stress-genomics.org/, and have stimulated new ideas and new research projects to understand the mechanisms of plant responses to stress (Figure 4). These projects are being pursued by scientists inside and outside the consortium.

Unraveling the secrets of epigenetics

"Epigenetics" refers to changes in gene function that are heritable but that do not entail a change in DNA sequence. Epigenetics is known to occur in a wide range of organisms including plants and animals. While the phenomenon has been known for over 40 years, it is only within the last few years that scientists have begun to understand its molecular basis. One emerging common mechanism appears to be the influencing of gene activity by proteins that package the DNA into chromatin; chromatin is the protein-DNA complex found in the nucleus of each cell. Scientists at University of Arizona, Johns Hopkins University, University of Missouri, University of Wisconsin, Purdue University, and Washington University at St. Louis, are conducting systematic studies to understand the mechanisms underlying epigenetic phenomena in higher plants. They have found that chromatin-modifying proteins play a key role in higher plants as well. In plants, one of the best known epigenetic phenomenon is "gene silencing," a regulatory mechanism in which expression of a particular gene is inactivated (Figure 5). Gene silencing is often observed in transgenic plants where expression of the engineered gene is preferentially inactivated. It is also observed in conventionally bred hybrids where a specific gene from one parent is inactivated. A detailed understanding of the epigenetic mechanism, how it is triggered, and what mediates this process, should lead to practical solutions and novel strategies for future plant improvement programs.

Figure 5. In this petunia flower , a gene that codes the synthesis of the purple pigment was inactivated during the flower development. Flower cells developed after the silencing event are white because purple pigments were no longer produced. *Photo courtesy of Dr. Richard Jorgensen, University of Arizona.*

How many genes does it take for a bacterial pathogen to cause disease in plants?

It has been known for some time that the ability of the bacterial pathogen, *Pseudomonas syringae*, to cause diseases such as bacterial speck disease in tomato and other plants is dependent upon genes activated by a virulence factor called the "HrpL alternative sigma factor" (Figure 6). Virulence factors play an important role in the relative severity of plant diseases. In order to identify genes involved in the steps leading up to visible disease symptoms, a consortium of scientists at Cornell University, The Institute for Genomic Research,

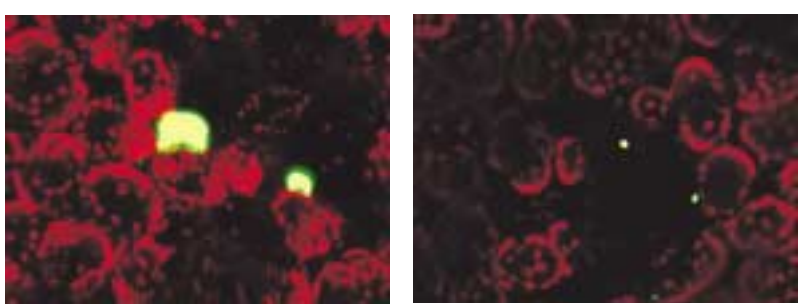


Figure 6. Pseudomonas syringae infects plants by injecting virulence factors. In these confocal microscope images, plant cells appear red and the bacterial colony appears fluorecent green. Left: Large wild-type bacterial colony grows in plants. Right: Mutant bacteria lacking genes for virulence actors fail to grow in plants. *Photo courtesy of Dr. Alan Collmer, Cornell University*.

Boyce Thompson Institute, the University of Missouri, Kansas State University, and the University of Nebraska has developed a draft genome sequence of this bacterial pathogen. The scientists used a combination of computational and gene expression techniques to identify virulence-implicated genes under the control of the HrpL alternative sigma factor. They have identified more than 20 genes in the bacterium that appear to be involved in the establishment of bacterial speck disease in tomato and are now working to determine specific roles for each of these genes. Once scientists understand how these bacterial genes interact with the host plant genes in causing a disease, they will be able to develop rational strategies for the development of disease-resistant plants and for disarming disease-causing bacteria. Since *Pseudomonas* is one of the most prevalent bacterial groups in the environment affecting the health of plants and animals, this study will contribute not only to the understanding of specific pathogen-tomato interactions, but also to the broader understanding of microbe-host interactions.

On the horizon

Some projects that have recently been funded address long-standing questions about fundamental plant processes. The questions can now be answered directly with genomic tools that have become available within the last few years. A few examples of these new projects are given below to provide a glimpse of exciting discoveries expected in the near future.

Understanding the genomic basis of apomixis - Apomixis is a naturally occurring mode of asexual reproduction in flowering plants that results in embryo formation without the involvement of fertilization of the egg. Seed-derived progeny of an apomictic plant are genetically identical or exact clones of the maternal plant. Among economically important plants, this phenomenon is known to occur in grasses, citrus, and cotton. Groups at the University of Georgia and Texas A&M University are focusing their collaborative study on a common form of apomixis that occurs in grass species. They hope to uncover the genes involved in this process and the mechanisms by which their action and interaction are regulated. These findings could have significant economic implications because they can be used to fix superior gene combinations in crop plants within two generations. At the same time, the outcomes will provide fundamental information about a mechanism of reproduction unique to plants.

Genomic approaches to physiological ecology - Water use efficiency (WUE) is among the most fundamental constraints shaping the evolution of almost every aspect of plant structure from cells to whole plants. Water utilization and the water use efficiency of growth also remain the most important determinants of terrestrial productivity in both natural and managed ecosystems. A team of scientists at the Boyce Thompson Institute, Oklahoma State University, the University of North Carolina, and Cornell University, has embarked on an ambitious project to identify and characterize genes that act singly and in concert to regulate the development and plasticity of essential plant traits controlling WUE. This project not only addresses one of the long-standing fundamental questions in plant biology, but also represents a novel use of genomics approach in environmental biology research.

Dissection of the genetic basis of polyploidy: the organization of extremely complex genomes - Normal human cells, except for egg and sperm cells, contain two sets of chromosomes. Each set of chromosomes is termed "N", so cells containing two sets of chromosomes are referred to as 2N or diploid. In contrast to humans, many plants have multiple sets of chromosomes (e.g. 4N, 8N), a phenomenon referred to as "polyploidy". Polyploidy occurs in many crop species. For example, alfalfa is a 4N plant, wheat is 6N, cotton is 4N, and sugarcane is 8N. Stonecrop, an ornamental plant, is known to have 80N. Increased N number appears to provide advantages such as increased variability, larger cell / tissue / plant size, and increased photosynthetic capacity. At the same time, the large size and the complexity of polyploid plant genomes have made it difficult to study the phenomenon. The recent advent of genomics has opened up new ways of studying polyploidy and the NPGI supports a number of projects aimed at understanding the structure, organization, and function of polyploid plant genomes. These projects include a systematic study in Brassica and maize by a group of scientists at the University of Wisconsin, the University of

Washington, Cold Spring Harbor Laboratory, Texas A&M University and the University of Missouri; a study focused on the wheat genome by a group of scientists at the University of California at Davis; and a study aimed at understanding the dynamics of polyploidy in cotton by a team of scientists at Iowa State University. Results from these studies will have broad implications in plant biology. They will provide a fundamental understanding of speciation in plants and the potential for an entirely new breeding strategy based on polyploidy. They may also impact biomedical research since polyploidy in animals is usually associated with abnormal cell growth such as cancer.

3. Technology development

Rapid advances in genomics are, in part, a result of continued development of new techniques and methods, especially those that allow high throughput processing of data and information. The plant genome research community has benefited from general advances in genomic technologies for many other organisms. At the same time, the plant genome research community has developed its own tools, either by adopting and modifying the available technologies for plant systems, or by inventing techniques that are uniquely suited for plant genome research.

Optical Mapping: a new way to map a large genome

A researcher at the University of Wisconsin, Madison has been successful in mapping large portions of the rice genome using a microscope to visualize the positions at which specific enzymes cut along its length. The technology is called "optical mapping" and it has been applied successfully to the construction of a whole chromosome map for *Plasmodium*, the human malaria parasite. The application of this technology to the larger and more complex rice genome has posed significant technical challenges; however, recent progress indicates that the researcher should be able to construct a whole genome map of rice (Figure 7). The optically mapped whole genome can, and is already beginning to, facilitate and accelerate the assembly of the rice genome sequence information being generated by the International Rice Genome Sequencing Project. In addition, the whole genome map can be used to decipher and identify the regions containing repetitive elements. Such regions are common in a large genome such as rice and are difficult to identify and resolve by other means. The whole rice genome map will also precisely anchor BACs and other clones with respect to chromosome location, thereby facilitating the assignment of gene positions along the 12 rice chromosomes. The rice optical map thus has tremendous potential in helping to identify and characterize sequence gaps (both the location and length) and to provide an independent means for global verification of completed sequence.

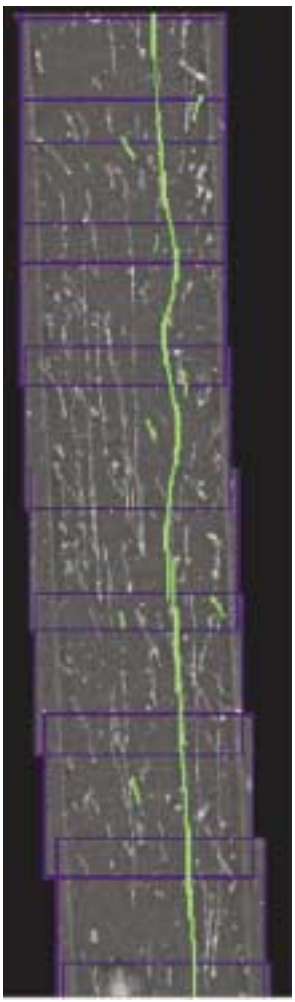


Figure7. An optical map of a portion of a rice chromosome, constructed from multiple images (each framed in blue) of a spread preparation of digested rice DNA. The digested genomic DNA and standard DNA molecular markers are labeled in green. *Photo courtesy of David Schwartz, University of Wisconsin.*

TILLiNG: a novel technology for rapid selection of a mutation in any gene from mutant plant

A new technology that allows rapid selection of mutants from any gene in any plant has come out of a proof-of-concept project being conducted at the Fred Hutchinson Cancer Research Center in collaboration with the University of Washington and the Institute for Systems Biology. All that is required is a collection of plants treated with a chemical that induces

single base changes (or "point mutations") and some sequence information about the gene of interest. This rapid screening method is called **T**argeting **I**nduced **L**ocal **L**esions **i**n **G**enomes or TILLiNG. The initial development and testing of the methodology was done using *Arabidopsis* plants; however, it has since been shown that TILLiNG can also be applied to other organisms. Currently, the investigators are helping scientists set up TILLiNG for rice, barley, maize, mouse, and worm mutants. As with any high throughput technique, informatics tools are an integral part of TILLiNG. Various software tools that keep track of samples, catalogue mutants, screen specific genes, and visualize the search results have been developed and updated during the course of the project. A TILLiNG service, along with the associated software tools, is available to the community through the project website at http: / / blocks.fhcrc.org / ~steveh / Welcome_to_ATP.html.

Concentrating the gene-rich regions of a large genome

All large-scale genome projects conducted to date have employed either a "minimum tiling path" or a "whole genome shotgun" sequencing approach. Both of these methods produce pieces of sequence representing the whole genome that are then assembled using computational methods. However, these approaches are not readily applicable to many economically important plant species such as maize, soybean, or wheat, due to the huge size and complexity of their genomes. For example, the maize genome is as large as the human genome, but also contains huge arrays of a small number of sequences repeated over and over again. Most of these highly repetitive sequence regions do not contain genes but are instead interspersed with gene-rich "islands". A group of scientists at the Cold Spring Harbor Laboratory have developed an elegant technology to isolate the gene-rich islands of the maize genome from the repetitive DNA (Figure 8). The technology is based on the difference in DNA methylation between most genes and the highly repetitive components of the genome.

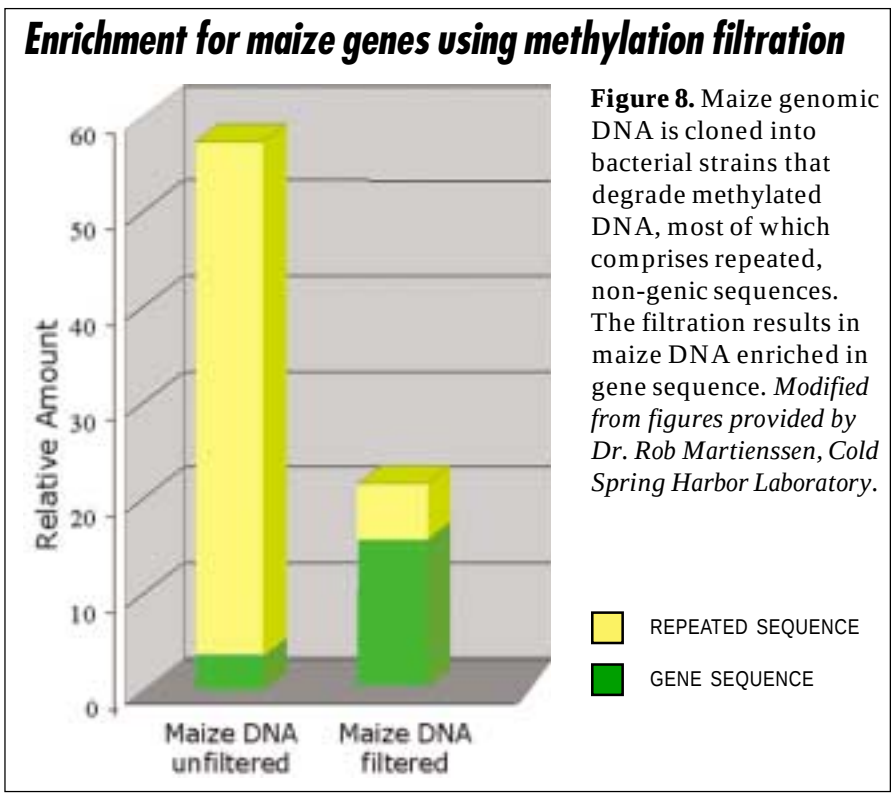


Figure 8. Maize genomic DNA is cloned into bacterial strains that degrade methylated DNA, most of which comprises repeated, non-genic sequences. The filtration results in maize DNA enriched in gene sequence. *Modified from figures provided by Dr. Rob Martienssen, Cold Spring Harbor Laboratory.*

The gene-enriched fraction is expected to represent about 10% of the whole maize genome, and should contain most of the genes and few repetitive sequences. In principle, this technology is applicable to any complex genome, and is in fact already being applied to other species such as sorghum. More importantly, this technology makes it possible to realistically consider large-scale sequencing of the maize genome in order to identify, sequence, and assemble all the maize genes in an efficient and cost-effective manner.

Under construction

Advances in plant genome research continue to enable research in new directions, which in turn motivates scientists to develop novel technologies to reach new goals. Examples include the following:

Developing robust maize transformation systems - After surveying the maize research community, the Maize Executive Committee identified an urgent need for a straightforward and reproducible transformation system for maize. A team of plant transformation specialists assembled from Iowa State University, Purdue University, the University of Wisconsin, and North Carolina State University is using cutting-edge genomic resources and technologies, some developed through prior NPGI-funded projects, to develop tools, as well as provide training and a transformation service to the broader community.

Gene targeting in plants - Genome sequencing projects are providing a wealth of knowledge about the individual genes carried in plant genomes. The ultimate goal of such projects is to understand the function of each of these genes. One of the most powerful approaches to characterizing gene function involves the analysis of organisms that contain specific alterations in genes of interest. It would be useful to specifically remove one gene and replace it with another gene. Currently, there is no method to target specific genes for mutation or replacement in plants. There is a pressing need for the development of methods that make directed gene modifications of plant genomes rather than the non-targeted methods currently available. A project has been initiated at the University of Utah to develop a gene targeting methodology that utilizes DNA sequence information resulting from the various public genome sequencing projects. The methodology is designed to inactivate a target gene or to introduce modifications into a target gene. A similar approach developed by a member of this group has already yielded spectacular results in the fruitfly. If successful in plants, the gene targeting methodology will allow researchers to produce an almost unlimited variety of genetic changes in plants. This tool will be useful both for basic research and for future efforts to improve crops.

4. Informatics and data management

A hallmark of the NPGI is the open sharing of data and information generated by each project. Primary data such as genome sequences and ESTs are routinely deposited in GenBank at the National Center for Biotechnology Information of the National Institutes of Health. Some organism-specific databases are integrated and maintained by the Agricultural Research Service of the U.S. Department of Agriculture. Many of the projects are generating new integrated databases that also incorporate tools for genomic analysis. The data are analyzed and released to the database, and tools for analysis are made available at the same web site. These tools can be used to perform additional analyses on the site database or can be used for external data.

GRAMENE: A resource for comparative mapping among the grasses

GRAMENE is a curated, open-source, Web-accessible data resource for comparative genome analysis of the grasses, to which all cereal crops belong. The database is relational and the foundation datasets are derived from the various rice genome projects. GRAMENE currently contains: (1) the publicly-available rice genomic sequence (finished and unfinished); (2) all available annotations on the genomic sequence; (3) integrations of several rice genetic maps with each other, and an integration of the genetic maps with the physical map; (4) mappings of rice BAC (Bacterial Artificial Chromosome) end sequences, microsatellites, and other sequence-based markers onto the rice genome; and (5) functional and sequence information on all published rice proteins. On this foundation, sequenced-based integration between rice, sorghum, and maize maps will be built, which will facilitate comparative genomics studies among the grasses. GRAMENE is a collaboration of scientists at Cold Spring Harbor Laboratory, Cornell University, and the USDA Agricultural Research Service. The sequence annotations, maps, and other information are available at http://www.gramene.org.

The Medicago truncatula project: A model species for legumes

An international consortium of scientists has developed *Medicago truncatula* as a powerful model system to study the genomes of legumes. The U.S. participants in this international effort are located at the University of California at Davis, the Samuel Roberts Noble Foundation, the University of Minnesota, Texas A&M University, and The Institute for Genomic Research. The web site at http://www.medicago.org links to the project database and provides access to all raw data, sequence annotated for predicted genes, and reports of comparisons made with genes from other organisms. The genetic map section of the web site also contains a graphic display of sequence homologies between *Medicago* and *Arabidopsis*, and their map position in the *Arabidopsis* genome. A series of clickable elements takes the user through a series of increasingly detailed views that culminate in the underlying sequence relationships. The database and associated informatics tools are providing invaluable resources to legume researchers working on soybean, other beans, alfalfa, and tree legumes, as well as to the broader plant genome community.

Establishing a minimum standard for management of plant genome research data

Most genome projects have produced large datasets of diverse types that are often scattered in space, time, and format. This has made comparative analyses of data from different projects difficult and makes the creation of integrated databases a significant, if not often insurmountable, challenge. Because of the urgent need to define "an accessible and useable database" in the context of the NPGI, a workshop was convened on September 11 & 12, 2001, in Rockville, Maryland.

The workshop participants concluded that there is a need for a common set of capabilities present in all databases that will allow the broadest access. Further, they recommended that the community begin to develop a standard format for data definition and exchange in order to facilitate machine and human parsing of the data. As a follow-up activity, seven working groups will recommend the minimal functionality needed to be incorporated into any database so that the data will be readily accessible to the wider community. The plan is to widely circulate draft recommendations for community input, and then publish a report by early 2002 that will eventually be used as a standard by all future NPGI projects.

In the works

Development of statistical methodology for agricultural genomics - A recent trend in the statistical analysis of microarray data has been the application of traditional experimental design and linear model techniques to gene expression data sets. An obvious advantage of this approach is that it provides researchers with an established method for analyzing microarray data, and it requires only standard statistical packages rather than custom software. Researchers at the Purdue University Computational Genomics Facility are investigating various proposed linear models-based analysis techniques and their application to data sets from gene expression experiments.

Web resources for the computation and display of physical mapping data - The long-term goal of this project by investigators at Clemson University is to provide flexible web-based resources that will make it easier for the user to explore relationships between pieces of physical data. One area in which this will be useful is in analyzing detailed genomic maps used to assemble sequences. Flexible tools will be created to search sequences and associated annotation files in order to show the distribution of markers and genes along a chromosome and within smaller sequence assemblies. As part of this effort, a computational genomics course for graduate computer scientists will be designed and taught, and an on-line tutorial will be developed to maximize the utility of these resources within the broader community.

B. Impact on Plant Science Research

One of the ultimate goals of the NPGI is to advance the field of plant sciences through the genomic revolution. The expectation is that plant genomic tools will bring new approaches to plant science research and will inspire renewed interest among beginning and early career scientists.

Numerous resources, including a complete plant genome sequence, targeted gene disruptions and mutant seed stocks catalogued gene by gene, and databases containing details ranging from single sequence changes to the classification of gene families for functional genomics, provide a rich arsenal of tools for any 21st century plant biologist to undertake genetic analysis and technological manipulation of any physiological process of choice. The wide-ranging utility of these resources is apparent in the 153,000 web site references retrieved by a search using the phrase "plant genome" in a popular web search engine. Following the links can lead a scientist to experimental protocols, discussion groups, germplasm resources, seed stock centers, as well as to genetic and physical maps, sequences, and markers.

Remote distance and financial considerations are no longer a major limiting factor for plant research, as evidenced by the inception of the first virtual conference on genomics. Investigators or interested community members can use the World Wide Web from anywhere in the world to easily access sites for exchanging ideas and reading new developments on a range of topics like functional genomics, structural genomics, computational approaches for expression profiling, metabolic profiling, data standardization and management, implications of genomics research, and proteomics.

The genomic revolution has irrevocably changed plant biology.

C. Impact of NPGI Research beyond Laboratories

Photo courtesy of the Agricultural Research Service, USDA.

The initial NPGI five-year plan states that "The ultimate goal of genomics is to understand the structure and function of every gene in an organism. With the intent of exploiting that knowledge for the betterment of society, the NPGI will pursue this goal by focusing on plant species important to agriculture, environment, energy and health." The spectrum of research supported by the NPGI is indeed broad; discoveries of new genes of economic importance and the understanding of the function of such genes function are quickly being applied to develop improved plant-based products and practices.

Bringing genomics to the wheat fields

With the overall goal of transferring new developments in genomics to wheat breeding and production, investigators at the University of California, Colorado State University, Cornell University, Kansas State University, Montana State University, North Dakota State University, Purdue University, University of Idaho, University of Minnesota, University of Nebraska, USDA and Washington State University lead the organization of a national wheat **M**arker **A**ssisted **S**election (MAS) consortium that includes 12 public wheat-breeding and research programs across the U.S. In the MAS program, molecular markers are used as chromosome landmarks to facilitate the precise introgression of small chromosome segments carrying the genes of interest. Available molecular markers will be used to transfer genes for resistance to fungi, viruses, and insects, as well as gene variants related to improved bread, pasta, and noodle quality. These genes will be incorporated into a minimum of 240 adapted cultivars or breeding lines belonging to all major market classes of U.S. wheat, and since they are transferred by normal recombination, the resulting lines will not be classified as transgenic. These improved cultivars will transfer the value of genomic research to the wheat growers' fields.

Harnessing investments in genomics of model species for vegetable improvement

A critical question in determining strategies for investment in crop research is the extent to which investments in model species benefit other crops. A team of scientists at Cornell University, Volcani Center – Israel, USDA, California State University, New Mexico State University, Northeastern Organic Farming Association, and J. Haapala Farmers' Cooperative, is developing tools for genomic analysis of peppers to assess the degree to which resources already available in tomato and other Solanaceous plants will be relevant. They will focus on the resources applicable to disease resistance and fruit quality, two important objectives for pepper in particular and for vegetable improvement in general. Specific genes identified as relevant to economically important traits in pepper will be transferred through an expanded consortium of seed companies and non-profit organizations. This will maximize the impact of the work in the form of new varieties and novel traits, which will benefit consumers and the producer community.

Genomics of biomass production and its relation to winter survival in alfalfa

Alfalfa is widely grown throughout the U.S. and the world for its nutritious forage. Over the past 20 years, there has been no improvement in alfalfa yield in the Midwestern U.S. One way to improve yield is to increase production in late summer and autumn when plants typically become dormant; however, increased autumn forage production has been associated with winter injury. A second way to increase yield may be through a marker-assisted selection approach. To tackle these problems, researchers at Iowa State University are using a genomic approach to identify genomic regions associated with biomass production, both throughout the year and at individual harvests within the year, plant height during the autumn, and winter injury during the following spring. They have developed map tools to locate Quantitative Trait Loci (QTLs), chromosomal regions important for yield, plant height, and winter injury. Results obtained to date indicate that a different QTL often exists for each trait, suggesting that improved yield and winter survival can be realized concurrently.

Reducing the genetic vulnerability of cotton

Photo courtesy of Texas A&M University, Soil and Crop Sciences Department.

The U.S. cotton production system exemplifies the complex challenges that must be met in order to reduce the genetic vulnerability of a major crop. Genetic vulnerability results from a combination of a crop's evolutionary history, trends in breeding and biotechnology practices, and grower decisions based on inadequate available information, all in response to the inevitable pressures imposed by processor and consumer requirements. Researchers at the University of Georgia, Georgia Agricultural Extension Service, Texas A&M University, and Texas Agricultural Extension Service are engaged in the development of the following: (1) a Web-accessible resource that will enable producers to reduce short-term field genetic vulnerability through better-informed decisions about use of existing germplasm; and (2) 'user-friendly' germplasm containing new gene combinations useful for short-term cotton improvement, plus new genetic stocks useful for long-term research. These resources will help increase cotton germplasm diversity, which, in turn, will help decrease genetic vulnerability of cotton.

D. Societal and Educational Impacts

From the beginning, the NPGI has encouraged more than the simple pushing forward of the frontiers of knowledge in plant genomics. It has strongly encouraged networking among the various genome projects, interactions across disciplinary and geographic boundaries, international research collaborations, partnerships with industries and growers, education and training of the next generation of scientists,

increased participation of under-represented groups, and outreach activities to communicate the outcomes of plant genome research to the general public. In fact, NPGI project investigators are playing a significant role in addressing each of these issues.

Interaction with industry

A number of projects funded through the NPGI include collaborations with industry. For example, a project funded at the Boyce Thompson Institute is making a collection of maize plants containing the mobile DNA tag, Ac, distributed at regular intervals. This project is being carried out in collaboration with an industrial partner, DeKalb Genetics (now part of Pharmacia). DeKalb has provided the project with about half of their required summer nursery field space and all the winter nursery space and support. They also provide salary and support for a research technician to work specifically on this project. Both the public and private sectors will benefit from the collection of lines being generated. The seeds are being deposited in a public stock center and there are no intellectual property issues associated with these lines that will impede their use.

The goal of one of the largest NPGI projects is to provide integrated maps of the maize genome. This project is led by a senior scientist at the University of Missouri with a number of collaborators at Clemson University and the University of Georgia. The outcomes of the project have been enhanced by a collaboration with DuPont and Incyte to anchor BACs with 10,000 unique EST assemblies. More than 166,000 BACs from two of the project libraries have been delivered to Incyte for quality control and BAC filter production. DuPont has already provided information for more than 3400 ESTs and completion of the project goal is anticipated by January 2002. As data are received by the project, they are integrated into the displayed BAC data at Clemson and are accessible to the public with no restrictions.

The International Rice Genome Sequencing project has continued to enjoy the benefits of the rough draft rice genome sequence being provided by Monsanto. As predicted when Monsanto first offered to share their data with the public efforts to sequence the rice genome, their data have indeed helped to accelerate the rate of sequence data release into the public databases by the International Consortium.

The degree to which industry is making use of data and information resulting from the NPGI projects is difficult to gauge. However, it is clear that young scientists being trained by the NPGI-supported projects are in high demand by major industrial laboratories as researchers, and as research and development managers.

Education

Investigators at the Boyce Thompson Institute participating in various NPGI projects have established the Emerson Summer Program, which allows high school teachers and students to participate in screening Ac-tagged maize populations containing many interesting mutants. Because maize mutants are easy to spot and many of them have interesting visual mutations, they naturally pique the curiosity of young students. The demand for places in this program has been so great that it has opened up additional summer slots and now accepts students during the school year.

Many NPGI projects take advantage of existing education and training programs on participating campuses. For example, a project at the University of Wisconsin at Madison participated in three such programs: (1) the 'Summer Science Institute for High School Students', a program designed mainly to allow minority students to gain hands-on research experience and basic academic skills, and to explore career opportunities in science; (2) the 'Undergraduate Research Scholars Program', a new campus-wide initiative designed to give first- and second-year undergraduates their first exposure to original and innovative research; and (3) the 'College Access Program' which is mainly designed to encourage and support middle and high school students from under-represented groups to pursue higher education in science.

The Maize Gene Discovery project supported nine undergraduate students during Summer 2001. Five students worked in bioinformatics and four worked in plant genetics and phenotypic analysis. Students were trained jointly at Stanford University, Iowa State University, the University of California at San Diego, the University of California at Berkeley, and the University of Illinois project sites. An exciting aspect of this training program was the opportunity afforded to students to first receive training in bioinformatics at Iowa State University, and to then apply this training to their specific research projects upon returning home.

Training a new generation of plant genomics researchers

In the past two years, there has been a significant increase in the number of projects that integrate graduate and postdoctoral training. These projects are providing unparalleled opportunities for cross-disciplinary training in plant physiology, genomics, and bioinformatics. NPGI projects have provided increased training opportunities in cutting-edge plant biology since 1998. For example, the Plant Stress project, mentioned earlier in this document, alone has trained 26 postdoctoral researchers, 33 graduate students, and 79 undergraduates during this time period.

A group of NPGI investigators at University of Arizona collaborate in recruiting Native American and Hispanic undergraduate students and providing them with opportunities to participate in hands-on research experience in various aspects of plant genomics research. In addition, the same group of scientists works with high school science teachers from predominantly Hispanic or Native American high schools during the summer. The teachers receive hands-on experience in plant genomics research and develop science lessons / experiments that can be carried back to their students during the school year. The ties forged between the teachers and the project personnel continue throughout the school year.

The cellulose genome project, led by a group of scientists at the University of California at Davis and Texas Tech University, has established a formal exchange program with Alabama A&M University, a historically black university. An African American Master's student spent three weeks in the UC Davis laboratory and then continued the research at his home institution. He has now completed his degree and recently accepted a position with a major agribusiness. A project studying the *Agrobacterium* transformation in plants, based at Purdue University and State University of New York (SUNY) Binghamton, has two students from Tuskegee University working on the project. The students spent the summer at the project site. They have taken their research projects back to Tuskegee University and continue to work on them during the school year, maintaining links with the Purdue / SUNY groups.

NPGI-supported projects have involved a diverse student population.

Photos courtesy of Dr. David Bird, North Carolina State University.

The plant-nematode interaction project at North Carolina State University has been invited to participate in a "Village of Science and Technology" project at St. Augustine's College, a historically black college in Raleigh. One of the planned activities includes a regular "Genomics Club" which involves faculty as well as graduate students and postdoctoral researchers from the North Carolina State University's project. The project personnel believe that this activity will inspire more students from under-represented groups to consider scientific research as a career choice.

Public outreach

Some NPGI investigators make special personal efforts to reach out to the public, while others integrate public outreach activities within their research activities. It is important that scientists themselves understand that it is their responsibility to communicate their research results to the public at large.

Outreach to farmers - The soybean functional genomics project is unique in that the national soybean growers groups contribute significantly to the efforts of a consortium of academic researchers including the University of Illinois, the University of Missouri, the University of Minnesota, Northern Arizona University, and Iowa State University. Because of these close ties, the project director, Dr. Lila Vodkin, serves as a liaison between the research community and the end-users. In the past year, the project investigators have given talks to soybean producers at the Midwest Soybean Conference, the Advanced Judicial Academy for judges, farmers, and high school students. Dr. Vodkin was also interviewed for an article published in the ISF Farm Manager Newsletter entitled 'New tools to speed gene mapping'.

Individual efforts to reach general public on topical issues - Dr. Maureen Hanson of Cornell University, a co-PI on the maize chloroplast genome project led by a scientist at University of Oregon, participated in the PBS / Frontline program entitled "Harvest of Fear". Her comments were an important contribution to the balanced discussion of the scientific basis of genetic modification of plants for enhanced agronomic qualities. Dr. Hanson is active in outreach activities in her local area. She has hosted a visit from a small Vermont College biology / philosophy class. In a similar manner, Dr. Nina Fedoroff, at Pennsylvania State University, has given presentations on issues in plant biotechnology to various fora, including a National Academy of Sciences conference on "The Role of New Technologies in Poverty Alleviation and Sustainable Development", an AAAS Science and Technology Policy Forum on 'Biotechnology and agriculture: promise and peril', and at a Science Writer's Workshop entitled ' The future of GMOs'.

Communicating to lawmakers - During the last 12 months, there have been several congressional hearings related to plant genome research. At a hearing of the House Science Committee two of the NPGI awardees, Dr. Vicki Chandler at University of Arizona and Dr. Daphne Preuss at University of Chicago, provided invited testimonies about the importance of plant genome research for the future of science in our nation.

Traveling exhibit - The Missouri integrated maize mapping project has received funding from the NSF Informal Science Education program to construct a traveling exhibit entitled, "Superior Food Quality: Do the Tools Used to Achieve the Goal Make a Difference?" The goal of this project is to inform the general public on issues surrounding genetically modified organisms and to provide them with basic scientific information about crop improvement using maize as an example. This project focuses on maize because it has been an important food crop for thousands of years and continues to be bred for superior agronomic traits. The traveling exhibit will present the history of maize domestication, including the key races of maize and the techniques used in maize breeding and production. How genes are manipulated using breeding and molecular approaches is described. The exhibit concludes with a science-based analysis of the pros and cons for genetically modified organisms (GMOs) and the reasons for wanting to modify maize are reviewed. The exhibit will be made available to field days, state fairs and other public forums over the next two years.

Family math night - An Iowa State University mathematician, currently involved in several plant genome research projects, has been interacting with teachers, pupils, and parents from a local elementary school. At informal gatherings, he introduces them to non-standard mathematical topics such as algorithms, modeling, biomathematics, and combinatorics. Through games and novel instructional materials that use familiar objects, participants unknowingly learn the basic concepts of applied mathematics. Based on the success of this pilot activity, Family Math Night has expanded to other elementary schools in the area. Investigators are presently converting existing materials into a web-accessible form. The Family Math Night is now integrated into the plant genome research activities at Iowa State University, which provides students and postdoctoral researchers the opportunity to interact with the general public.

IV. New Goals for the National Plant Genome Initiative

Many technical and scientific breakthroughs have been made during the past year of the NPGI. These breakthroughs have opened up new opportunities and at the same time they have posed significant challenges for plant researchers. The Interagency Working Group for Plant Genome will consider both the opportunities and challenges in developing new goals for the NPGI.

As scientists begin to use genomic information from model species to compare both gene structure and chromosome organization in their selected plant system, there are some technological barriers to be overcome. For example, large genomic segments (1 million bases and larger) that contain significant incidents of tandem sequence repeats are unstable when propagated in bacterial hosts for DNA amplification and storage. Clones of this type exhibit deletions and rearrangements that are not easily avoided by conventional molecular biology protocols. These and other similar difficulties in maintaining and propagating large genome segments from cereals and other crops will limit progress in interspecies gene organization comparisons and the development of models for genome evolution. New methods need to be developed to overcome this limitation.

When the International Rice Genome Sequencing Project completes its efforts, we will know 99.99% of the genes in the rice genome. This will be more than sufficient for researchers to advance the genomics of grasses. Nevertheless, there will remain small gaps in sequence because of a lack of technology to completely close all the gaps of a genome that is three times larger than the other completed model genomes of *Arabidopsis*, *Drosophila*, and *C.elegans*. Human and mouse genome sequencing projects are encountering the same technical limitations. Since the rice genome is so much smaller than the human or mouse genome, it provides a nice system to explore ways to close these sequencing gaps. This is an opportunity for the plant genomics research community to take a lead in developing new methods to overcome this limitation.

The first generation of maize (corn) genomic tools (maps, ESTs, cDNAs, tagged genes, filtered libraries) has set the stage for a concerted effort to selectively identify the genes contained within large complex genomes such as maize or wheat. It is not cost effective or feasible to sequence and assemble complete genomes of this type because of the large amount of highly repetitive DNA. However, methods have been developed to concentrate regions of the maize genome that are enriched in genes. This provides a new opportunity to determine if it is possible to sequence only the enriched regions and obtain sequences of all the genes on the genome. A large plant genome will likely represent the test system for this kind of effort; it will likely be the proof of principle for this approach for other large genomes including plants and animals.

Creating data management and informatics tools to access and make use of the data will be a major challenge in the coming year. For example, the above mentioned sequencing strategy requires a whole new set of informatics tools to track clones, assemble raw sequence data, finish sequence, annotate genes, and transfer the information into the public database. Also, functional genomics is clearly the next phase of the plant genomics revolution, and the development of coherent resources for data acquisition, management, and storage remains the highest priority and a significant challenge for the NPGI. Significant informatics infrastructure must be developed if scientists are to progress towards an integrated view of plant biology that combines information about variations in metabolic information, protein expression and gene activity in response to environmental cues and temporal and developmental programs.

The Interagency Working Group (IWG) for Plant Genomes is in the process of developing the next five-year plan for the National Plant Genome Initiative. The informatics workshop mentioned above is part of that activity. There will be several stakeholders meetings to seek input from the community to identify scientific opportunities and technical challenges that will bring plant genome research to the next level. Budgetary needs for the next phase of the NPGI will be considered as an integral part of the planning process. The next five-year plan is expected to be published in late spring or early summer of 2002.

Acknowledgement

The Interagency Working Group on Plant Genomes acknowledges the assistance of Sharlene Weatherwax (Department of Energy), Jane Silverthorne (National Science Foundation), Ed Kaleikau (U.S. Department of Agriculture), Leland Ellis (U.S. Department of Agriculture), Machi Dilworth (National Science Foundation), and Christopher Cullis (National Science Foundation) in preparation of this progress report. Sarah Zielinski (National Science Foundation) helped with illustrations. Cynthia Lohmann (National Science Foundation) helped edit the report.

Abstract

The Interagency Working Group (IWG) for Plant Genomes was appointed in May, 1997, by the National Science and Technology Council (NSTC). The charge was to identify science-based priorities for a national plant genome initiative and to plan for a collaborative interagency approach to address these priorities. The IWG recommended establishment of the National Plant Genome Initiative (NPGI) with the long-term objective to understand the structure and function of genes in plants important to agriculture, environmental management, energy, and health. In addition to coordinating the activities of the NPGI participating agencies, the IWG monitors the progress of the NPGI and documents significant progress in annual reports. This new progress report is the third in this series.

Since its inception in FY1998, the NPGI has supported research at the frontiers of plant genomics. Progress has been made in all areas including new scientific discoveries, development of research tools and resources that allow the entire scientific community to participate in the NPGI activities, and increased training opportunities for students. As with any rapidly advancing field of research, plant genomics is constantly evolving. New discoveries lead to new lines of investigations. New methodologies open up new opportunities to study long-standing questions in plant biology. The dynamic research environment is attracting young scientist to plant genomics. NPGI project participants have been reaching out to the K-12 teachers and students as well as the general public in an effort to communicate their science to the society at large. In the next 12 months, the IWG will develop the next five-year plan for the NPGI that builds on the advances made to date and expands the frontiers of plant genomics.

For further information, contact:

National Science and Technology Council Executive Secretariat at (202)456-6101 (voice) or (202)456-6027 (fax).

Also available on the NSTC Home Page at http://ostp.gov/NSTC/html/NSTC_Home.html.

EXECUTIVE OFFICE OF THE PRESIDENT
Office of Science and Technology Policy
Washington, D.C. 20502