

Pine Gene Discovery Project

Final Report -- 08/31/1997 -- 02/28/2001

**Ross W. Whetten
Ronald R. Sederoff
Claire Kinlaw
Ernest Retzel**

April 2001

Work Performed Under Contract No. DE-FC07-97ID13550

**For
U.S. Department of Energy
Office of Energy Efficiency and Renewable Energy
Washington D.C.**

By North Carolina State University

PINE GENE DISCOVERY PROJECT

FINAL REPORT
08/31/1997 -- 02/28/2001

Ross W. Whetten
Ronald R. Sederoff
Claire Kinlaw
Ernest Retzel

April 2001

Work Performed Under Contract No. DE-FC07-97ID13550

Prepared for the
U.S. Department of Energy
Office of Energy Efficiency and Renewable Energy
Washington D.C.

Prepared by
North Carolina State University
Raleigh, N.C.

FINAL TECHNICAL REPORT

IDENTIFICATION NUMBER: DE-FC07-97ID13550

PROJECT TITLE: PINE GENE DISCOVERY PROJECT

RECIPIENT: DEPARTMENT OF FORESTRY, NORTH CAROLINA STATE
UNIVERSITY

SUBCONTRACTORS:

CALIFORNIA STATE UNIVERSITY, HAYWARD
UNIVERSITY OF MINNESOTA, MINNEAPOLIS

CONTACT INFORMATION:

RONALD SEDEROFF
FOREST BIOTECHNOLOGY GROUP
DEPARTMENT OF FORESTRY
2500 PARTNERS II BUILDING
840 MAIN CAMPUS DRIVE
RALEIGH NORTH CAROLINA 27695-7247
(919) 515-7800
(919) 515-7801 (FAX)

FOREWORD

This final technical report presents results from the DOE sponsored "Pine Gene Discovery Project", carried out by North Carolina State University, California State University at Hayward, and University of Minnesota at Minneapolis.

Libraries of genes expressed during pine xylem differentiation, shoot tip elongation, and pollen cone maturation were made at North Carolina State University by Isabel Allona and Gisele Gurgel. DNA sequencing reactions were carried out at North Carolina State University by Gisele Gurgel, with additional technical support from Reenah Schaffer, Sabrina Piercy and other members of the Forest Biotechnology Group. Sequencing reactions at California State University at Hayward were carried out by Tan Ho and Claire Kinlaw. Analysis at University of Minnesota was carried out by Sheila St. Cyr, Sopheak Sim, Elizabeth Shoop, and Rod Staggs, with support from members of the Computational Biology Centers. The contributions of all of these individuals is gratefully acknowledged.

This project was supported by a research grant from the Department of Energy Agenda 2020 program (number DE-FC07-97ID13550) and by contributions from Westvaco, Rayonier, Georgia-Pacific, and Union Camp.

Abstract

Integration of pines into the large scope of plant biology research depends on study of pines in parallel with study of annual plants, and on availability of research materials from pine to plant biologists interested in comparing pine with annual plant systems. The objectives of the Pine Gene Discovery Project were to obtain 10,000 partial DNA sequences of genes expressed in loblolly pine, to determine which of those pine genes were similar to known genes from other organisms, and to make the DNA sequences and isolated pine genes available to plant researchers to stimulate integration of pines into the wider scope of plant biology research. Those objectives have been completed, and the results are available to the public. Requests for pine genes have been received from a number of laboratories that would otherwise not have included pine in their research, indicating that progress is being made toward the goal of integrating pine research into the larger molecular biology research community.

TABLE OF CONTENTS

	Page
FOREWORD	i
ABSTRACT	ii
TABLE OF CONTENTS	iii
INTRODUCTION	1
OBJECTIVES	5
MATERIALS AND METHODS	5
RESULTS	7
DISCUSSION	8

Introduction

The growing world population places increasing pressures on all natural resources, including forests, as a larger population tries to meet its needs from a constant land base. Demands for forest products such as paper and timber typically increase as nations undergo industrial development, and at the same time, growing populations require more land for housing and agriculture. Forests serve important environmental and ecological roles, as well as providing forest products, and a diminishing area of forested lands is available to meet those needs as well. Forest lands are increasingly divided into two categories; intensively managed plantations for production of industrial raw materials, and lightly-managed natural reserves for protection of natural forest ecosystems. Application of the most advanced methods to forest regeneration and silviculture will be essential to maximize the sustainability and productivity of plantation forests, so that reserve lands can be managed primarily for environmental benefits.

The most important tree species for plantation forestry in the U.S. is loblolly pine (*Pinus taeda* L.); over 1 billion loblolly pine seedlings are planted every year in the southeastern U.S. Much progress has been made in the past few decades in understanding the basic genetic and physiological mechanisms that underlie plant growth and productivity, and these findings have been widely applied to improvement of annual crops. Perennial crops such as loblolly pine have presented several barriers to parallel investigations, including their large size and long generation times. Determination of the complete genome sequence of the model plant *Arabidopsis*, and the pending completion of the rice genome

sequence, will provide the basis for moving ongoing research into plant growth into a integrative phase in which the molecular genetic information from DNA sequences can be placed into the context of field research on physiological and agroecological constraints on productivity. One important research need for inclusion of loblolly pine in this phase of integrative plant biology is DNA sequences from genes expressed during pine growth.

Pines are an ancient lineage of land plants, separated from annual crop species such as rice and soybean by at least 250 million years. This divergence seems extreme when comparing these species at the level of whole plants, but is less daunting when approached at the molecular level through comparative analysis of gene sequences. Such cross-species comparisons of gene sequences can provide important links between discoveries in intensively-studied model organisms and less-intensively studied but more economically and ecologically important organisms such as loblolly pine.

Determination of DNA sequences has become, thanks to recent improvements in automated laboratory instrumentation, a powerful means of gaining understanding of relationships between organisms, and of providing experimental reagents for further detailed analysis of the genetic and biochemical mechanisms that underlie normal growth and development. Comparisons of DNA sequences from many different organisms, ranging from bacteria and fungi through animals and plants, have shown that many genes are shared in all living organisms. Similarities in the basic biochemical processes of life can be used to explain the similarities in the genes that carry out those processes, and key differences between organisms in physiology or biochemistry are frequently accompanied by differences in the genes present in those organisms. These observations have led to a

growing appreciation for the value of comparative analysis of DNA sequences as a means of integrating biological observations on an enormous variety of organisms. The occurrence of particular genes or sets of genes is a characteristic that can be shared between widely divergent forms of life, and this pattern of similarities and differences in gene content provides an important counterpoint to observations of the life histories and physiology of whole organisms.

There are two ways of viewing the comparative analysis of DNA sequences. The first is as a means of pulling together, and finding relationships between, apparently unrelated observations from a variety of species. The second is in extending the observations made on intensively-studied species by generating hypotheses about species that are less well-known. DNA sequences are required in both cases, and this has given rise to a widespread movement in biology to obtain DNA sequences from many different organisms. It has been said that the determination of DNA sequence of expressed genes, as a first step, and of the complete genome, in the final phase, is an important and worthwhile goal for every organism of significant ecological, scientific, or economic value. This is equally true for organisms that are not convenient experimental systems as for those that are - for loblolly pine as well as for *Arabidopsis*, dairy cattle as well as mice. The integrative function of bringing together physiological observations on pine, cattle, *Arabidopsis*, and mice is followed by the extensive function of generating hypotheses based on the detailed analysis of model systems and testing them in the larger context of non-experimental organisms such as pines and cattle.

The Pine Gene Discovery Project was initiated in 1997 with the objective of providing an initial glimpse of the genes involved in pine growth and development. The exact number of genes present in pines is not known, but a random sample of genes can be made by sequencing complementary DNA, or cDNA, copies of messenger RNA molecules present in tissues and organs during growth. Different tissues and organs express overlapping subsets of the total complement of genes present in an organism, and analysis of cDNAs from different sources provides information about the degree of similarity between tissues. The tissues and organs used for the Pine Gene Discovery Project were differentiating secondary xylem, elongating shoot tips, and maturing male strobili, or pollen cones. Secondary xylem, or wood, is one of the dominant features of pine biology - the dominant role of trees in forest ecosystems is due to the woody stems that allow them to compete for light by growing to greater heights than herbaceous annual plants. Wood is also an important industrial raw material, leading to interest in changing wood properties in trees harvested for industrial processing. Tree improvement in the past forty years has focused largely on increasing productivity and form, rather than changing wood properties. The rate of tree growth and overall tree form are heavily influenced by the growth rate and form of elongating shoot tips, because the shoot tip gives rise to all the aerial portions of the plant. Vegetative growth, such as occurs in shoot tips, is distinct from reproductive growth and differentiation. Reproductive tissues, such as the male strobili or pollen cones, may express a very different subset of genes from vegetative tissues, including genes that may be involved in controlling the onset of reproductive maturity and the degree of fertility.

Objectives

The objectives of the Pine Gene Discovery Project were to:

1. Obtain 10,000 partial DNA sequences of genes expressed in loblolly pine,
2. Compare those sequences to previously-described gene and protein sequences in the public sequence databases, and
3. Make the sequences, the results of analysis, and the cloned genes available to plant biology researchers.

The long-term goal of the Pine Gene Discovery Project was to stimulate molecular biology and genetic research on loblolly pine, by providing essential reagents and information to industry and academic scientists.

Materials and Methods

Libraries of pine genes expressed in different tissues and organs were prepared using two different methods. A library of genes expressed in differentiating pine secondary xylem was prepared in the Lambda ZAP cloning vector (Stratagene), according to the manufacturer's instructions. Material for this library was collected from a 35-year old grafted loblolly pine of genotype 7-56, through the cooperation of Westvaco Corporation. Libraries of genes expressed in pollen cones and in shoot tips were prepared from materials collected from loblolly pine trees growing in the North Carolina State University Schenk Experimental Forest. These libraries were prepared in the Lambda TriplEx cloning

vector (Clontech) according to the manufacturer's instructions, using the SMART-PCR method to prepare cDNA from one-microgram samples of total RNA.

Plasmid clones were obtained from the xylem library by superinfection of an *E. coli* culture, already inoculated with the lambda library, with M13 helper phage that provides the functions essential for excision and packaging of the pBluescript phagemid in an M13 protein coat. The phagemid supernatant was then used to inoculate JM109 host cells, and ampicillin-resistant colonies were picked for use in plasmid purification.

Plasmid clones were obtained from the shoot tip and pollen cone libraries by inoculation of an *E. coli* strain bearing a plasmid that provides Cre recombinase functions. The plasmid sequences in the lambda TriplEx vector are flanked by Lox sites, which are acted upon by Cre recombinase to excise the plasmid as a circular molecule. Ampicillin-resistant colonies were picked from the excision plates for use in plasmid purification.

Plasmids were purified from 1.2ml *E. coli* cultures grown in 96-well format, using a modified alkaline lysis system (REALprep, Qiagen) according to the manufacturer's instructions. DNA sequencing reactions were carried out with dRhodamine dye-terminator reaction mix (ABI), purified by gel filtration on Sephadex G-50, and analyzed on an ABI 377 sequencer. Some DNA sequences were also determined using a Licor 4000 model sequencer and dye-labeled primers. DNA sequence analysis was carried out at the University of Minnesota Computational Biology Center, using software developed there.

Results

The results of this project were:

1. A total of 10,309 partial DNA sequences were obtained from pine genes expressed in three different tissues and organs. The genes were obtained from immature pollen cones (1,745 sequences), differentiating secondary xylem (1,969 sequences) and elongating shoot tips (6,595 sequences).
2. These sequences were compared to gene and protein sequences in the public databases, and the results of the comparisons are available on the World Wide Web at <http://web.ahc.umn.edu/biodata/doepine/>. These comparisons showed that about 10% of the pine genes discovered during this project are not similar to any previously known gene, including all the genes predicted to occur in the complete genome sequence of the model plant *Arabidopsis*. This finding suggests that pines have some genes that do not occur in herbaceous model plants, and that gene discovery projects in pine will continue to identify new genes until the sequences of all pine genes are known.
3. The sequence analysis results are available through WWW, and cloned genes from the project are available to researchers by request. The names, addresses, and other contact information for the Principal Investigators of the project are linked to every sequence derived from this project, and any researcher interested in obtaining one of these clones has been provided with it.

The long-term goal of the Pine Gene Discovery Project has already been achieved, in an important sense, through the increased number of plant biologists studying pine. The

availability of reagents and information useful in comparing pine genes to those in other plants has greatly increased the interest of many plant biologists in including pine in their studies. Pines are a member of a very ancient and successful lineage of land plants, and there is considerable value to be gained from including pines in comparative studies with other plants. The ultimate objective is to gain an understanding of the mechanisms that have allowed these trees to endure for over 250 million years, coping with environmental changes as well as competition from other kinds of plants.

Discussion

The original research plan specified that the cloned pine genes would be submitted to the American Type Culture Collection for long-term storage and dissemination to the research community. This was not successful. ATCC stopped responding to inquiries regarding submission of clones from the Pine Gene Discovery Project shortly after the project period began, and has never acknowledged any further attempts at communication since. This was taken to mean that ATCC was not willing to undertake the acquisition of a large number of cloned genes which would require substantial investment for storage and dissemination, in the absence of a strong demand for the genes and a large community of users willing to contribute to the support of the ATCC. As a result, the Principal Investigators have taken on the responsibility of distributing cloned genes from this project. There have been several dozen requests for cloned genes over the course of the project, and requests are still filled periodically.

The original project plan also specified that every sequence obtained, and the results of comparison to the public databases, would be placed into Genbank, the U.S. public sequence database. After consultation with Dr. Ernest Retzel (the subcontractor responsible for analyzing the sequences at University of Minnesota), it was decided that the interests of the research community would be best served by submitting only the highest-quality DNA sequences to Genbank, and making the remaining sequences available through the University of Minnesota Web site. This approach allows the research community access to all the DNA sequences, without degrading the utility of Genbank as a public resource by deposition of lower-quality data.