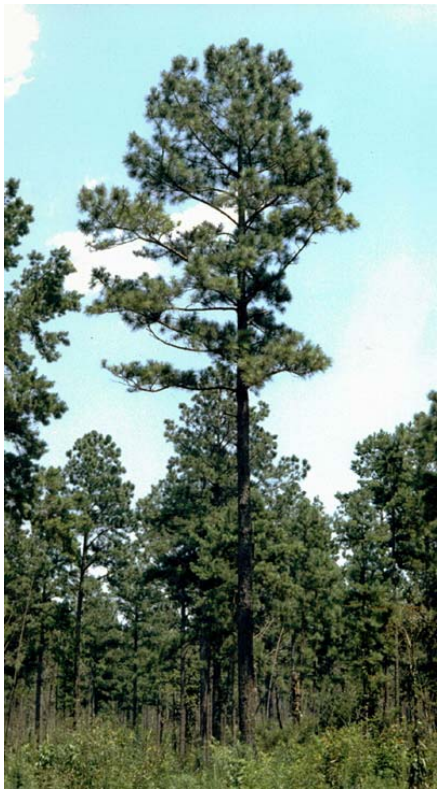
The Loblolly Pine Genome Project

A Prospectus to Guide Planning and Funding of a USDA Forest Service Led Effort to Develop an Integrated Genomics Research Program in Loblolly Pine

November 10, 2004



David B. Neale
Institute of Forest Genetics
USDA Forest Service
and
Department of Plant Sciences
University of California
Davis, California

and

Nicholas C. Wheeler
Molecular Tree Breeding Services
Centralia, Washington

Executive Summary

Loblolly pine (*Pinus taeda)* provides ca 16% of the world's annual timber supply, and grows on nearly 58 million acres of plantation and natural forest in the southeastern United States. Annually, only corn exceeds timber in farmgate economic value in America. Forest trees are the ecologically dominant life-form on ca 275 million acres in the US, and play a critically important role in carbon sequestration. In short, forests are vital to the economic and ecologic landscape of this country and loblolly pine is the single-most valuable species in those forests.

A genomics approach to describing and understanding the genetic and molecular basis of all biological processes controlling economically and ecologically relevant traits in pine is both feasible and desirable. Excellent progress has been made in many areas of genomic research in loblolly pine but the availability of a complete reference genome sequence will eventually limit progress. A Loblolly Pine Genomics Project is needed to develop a national genomics infrastructure, understand the complexity of the pine genome and develop an efficient strategy towards obtaining a complete reference genome sequence. There is a ready market for application of existing and rapidly developing genomics tools. Progress in genetic improvement of many forest tree species over the last half century has been notable (~10% increase in volume growth per generation), but traditional breeding and testing programs are expensive, time-consuming (15 to 25 years/generation), and very restrictive in the number of traits addressed in a given population. Furthermore, trees possess an abundance of natural variation that makes the potential for tree improvement large, but progress using traditional means slow. Few, if any, crops would benefit more from the development of genomic technologies that enhance our understanding of biological processes.

Support for genomics research in conifers has lagged significantly behind most major agricultural crops and model species. Initiatives are required from the pine genomics community and those who would benefit from that research (public and private land managers/owners, forest, paper, and energy industries) to enhance research support and speed progress. The purpose of this prospectus is to assist with initiative development by providing useful and complete information on the status of genomics research in loblolly pine today. Means for meeting near and long-term R&D targets are to help guide and prioritize future R&D proposals, to help scientists identify useful research collaborations, and to provide rationale and background to improve funding efforts.

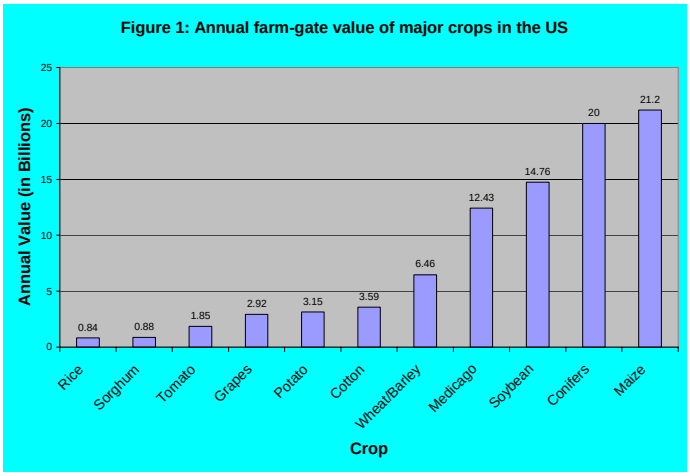
The overall benefits of the application of genomics to loblolly pine will be a vastly improved understanding of the biological and molecular basis of adaptive and economic traits of the most dominant genus of forest trees in the world[1] and significantly improved precision of genetic improvement practices that will surely reduce the time and cost of these activities.

[1] There are over 110 species of pines and comparative genomics studies show they are remarkably similar, genetically.

Introduction

Economic and Ecologic Importance

Loblolly pine is one of the most important crop species, and clearly the most important commercial timber and fiber species, in the US. The loblolly pine resource is essential to maintaining the nation's competitiveness in the global forest products markets. The native range of loblolly pine spans 14 states from southern New Jersey south to central Florida and west to Texas where it makes up more than half of the standing pine volume in the region. In 1998, 75% of the 1.6 billion seedlings planted in the United States were loblolly pine. It is the dominant tree species on 11.7 million ha of native forest and is established on over 12 million ha of plantation. The southern states provide 58% of the timber in the US and 16% of the world's timber. Collectively, timber is among the most highly valued commodities in America. Viewed as an agricultural crop, only corn currently exceeds timber in farmgate value on an annual basis (**Fig. 1**; data extracted from http://www.usda.gov/nass/ and USFS reports). The value of finished wood products exceeds 200 billion dollars a year! Furthermore, as the dominant plant species on millions of hectares, loblolly pine provides a huge and renewable/sustainable resource for carbon sequestration for a significant portion of the US, and is therefore critical in the larger picture of climate change.



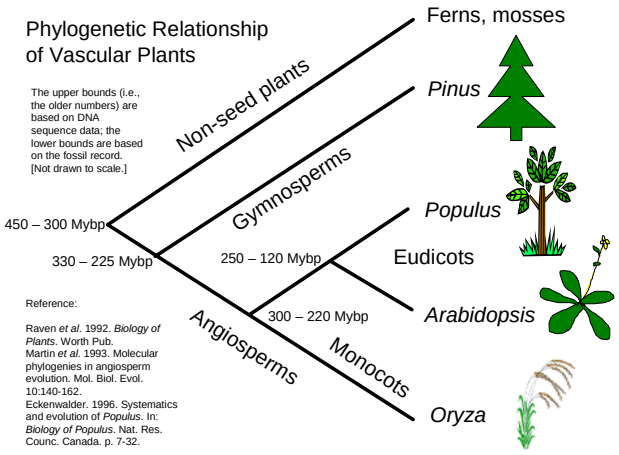
Programs to enhance loblolly pine productivity have historical roots. Loblolly pine tree improvement programs began in the southern states in the 1950s but even the most advanced programs have completed just 2-3 generations of breeding and testing. The long generation time of pine, combined with phenotypic selection of mature-tree traits, results in a very slow rate of genetic domestication. Consequently, loblolly pine would benefit tremendously from the development of any genomics technology that could accelerate breeding and improvement efforts. Realistically, pine growers in North America need to embrace genomics if they are to continue to compete globally for the fiber market. The recent completion of a whole genome sequence for poplar and rapid progress in sequencing the *Eucalyptus* genome will result in continued improvement in quantity and quality of fiber coming from these species, particularly offshore.

Finally, it is important to note that what is learned in loblolly pine will likely have utility across much of the Pinaceae, a family containing great ecological value world-wide as well as most of the commercially important species of the world. The genus *Pinus* alone contains more than 110 species, or about 20% of all known gymnosperms. Comparative genetic mapping among conifers has demonstrated high levels of genetic

similarity among Pinaceae genomes, facilitating comparative genomic analysis in this important plant family (see http://dendrome.ucdavis.edu/ccgp/).

Biological Importance of Sequencing the Loblolly Pine Genome

Gymnosperms are evolutionarily ancient, having arisen as much as 300 million year bp (see **Figure 2 below**, courtesy of http://www.ornl.gov/sci/ipgc/home.htm). The conifers, an order to which the pines belong, separated from flowering plants (Angiosperms) approximately 100 million years ago and by most measures, have evolved very slowly and conservatively. Loblolly pine therefore represents an ancestral branch in the tree of life. The pine genome is huge in base pair (bp) content (>2 x 10^{10} bp), exceeding that of Arabidopsis (1 x 10^{8} bp), poplar (5 x 10^{8} bp) and human (3 x 10^{9} bp) genomes by orders of magnitude. However, the vast majority of the pine genome appears to be repetitive (>99%), and is characterized by large gene families with many pseudogenes. Both these characteristcis serve to make sequencing of the loblolly pine genome a larger and more complex undertaking than was sequencing of the human genome, but potentially biologically more important.



Given the huge sequencing capacity currently available in the United States, and the creation of appropriate genetic resources, acquiring a DNA sequence of the loblolly pine genome is now a realistic goal for the pine community. However, a series of ordered research objectives must be accomplished first to better understand the complexity of the loblolly pine genome and develop a national genomics infrastructure for this species. DNA sequencing costs are lowering rapidly and it is anticipated that the background research, infrastructure development and affordable sequencing will converge in less than 5 years, at which time a complete genome sequence for loblolly pine can be determined. Such a resource would clearly place the United States in a competitively favorable environment.

Status of Funding for Loblolly Pine Genomic Research

Funding for genomic studies of pines (trees in general) have lagged significantly behind agricultural crops relative to their overall value and contribution to this country's economy. Total funding for pine genomics research from all sources since 1988 is slightly over $18.0 million dollars. From 1998 to 2003, NSF funding for pine genomics research was ~ $7.5 million. During that same period, maize received $128 million, rice $60 million, and *Medicago* $24 million. In short, funding levels and continuity have not been conducive to making rapid progress in genomics research of pines in the US. Members of the pine genomics community have therefore determined that a collective, integrated approach to conducting research and acquiring federal funding is required to

improve the rate of scientific discovery and progress in this important crop species. They have proposed the creation of a virtual organization called: ***The Loblolly Pine Genome Project.***

The Loblolly Pine Genome Project (LPGP): Purpose and Goals

The LPGP has grown out of a pair of workshops held in Davis, CA in May of 2003 and the Jekyll Is, GA in June, 2004. Comprehensive reports from these meetings may be viewed at **http://dendrome.ucdavis.edu/lpgp/.**

Purpose of LPGP
The purpose of the LPGP is 3-fold:

- *To help guide and prioritize R&D efforts in loblolly pine genomics*
- *To assist with the development of genomics infrastructure and research collaborations among members of the genomics community*
- *To provide rationale and background for improving the funding status of genomics research in loblolly pine in the United States.*

In short, the LPGP seeks to elevate loblolly pine to the status of model conifer species for genomics studies, and the pine genomics community in the US, to world leader's in tree genetic discovery.

Project Goals
Ultimately, the scientific goal of the LPGP is

> *to describe and understand the genetic and molecular basis of all biological processes controlling economically and ecologically relevant traits in pine by applying and advancing state-of-the-art technologies and, through open scientific collaborations and exchanges of information.*

The technology transfer and outreach goal is

> *to reduce the time and expense, and increase the genetic gain potential, of tree improvement programs by providing requisite tools and molecular reagents to the industrial community in a timely manner.*

The Strategic Research Plan

To become the world's leaders in pine genomics research, institutions participating in the LPGP must undertake an aggressive, directed research plan targeting technology and infrastructure development over the next 5 years. *There is an urgent need to find and identify all the genes and regulatory elements in the pine genome, characterize the genetic variation in them through DNA sequence analysis, and, finally, determine the relationship between that genotypic variation and expressed plant phenotypic variation.* This will require substantial and dedicated funding from US government sources, such as

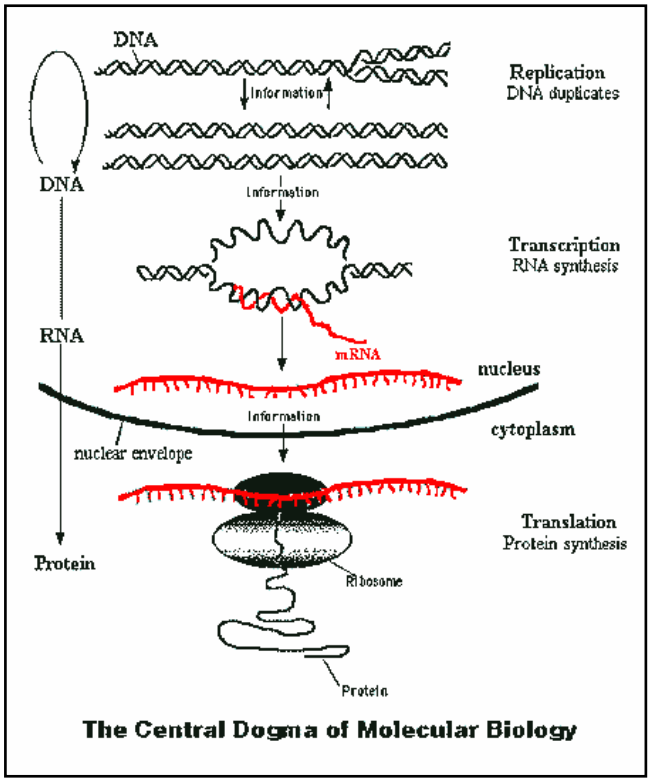
a USDA/NRI program in pine genomics and/or the US Forest Service Agenda 2020 program.

The following sections describe a series of specific research objectives, the current status and needs of pine genomics in each of those areas, and a description of how that information can be applied by the scientific and industrial communities. It concludes with a brief section on how the forest products industry can assist in advancing a sponsored genome program.

Finding, Identifying and Characterizing Genes

The long-term objective of the LPGP is to discover, identify and characterize all the genes in the pine genome. By so doing, we lay the foundation for describing and understanding the genetic and molecular basis of all biological processes controlling economically and ecologically relevant traits in pine. The technologies for accomplishing the objective are largely in place, and are undergoing continuous improvement, becoming faster and less expensive.

Finding the Genes: Finding a gene means identifying all, or a portion, of the genetic code of that gene. Genes are located along an organism's chromosomes, which are made of DNA, a double-stranded, helical macromolecule comprised of building blocks (bases or nucleotides). There are 4 kinds of bases, represented by the letters A,T,G, and C. These letters are read in groups of 3 (called codons). Genes typically start and stop with unique codons, and generally have far fewer than 5,000 total bases (1660 codons). If, as we are led to believe by studies with humans, *Arabidopsis* plants, and rice, most organisms are likely to possess between 25,000 and 50,000 genes, then fewer than 250,000,000 bases are needed to encode all biological functions. Unfortunately, most organisms have far more DNA than needed. Humans have nearly 3 billion bases in each cell with a nucleus, and loblolly pine has around 20 billion. Needless to say, finding genes is not a trivial task. Those working with pine have initially chosen a gene discovery method based on the Central Dogma of Molecular Biology (see Fig. 3). That is, the genetic message in (DNA) is transcribed by another nucleic acid molecule called messenger RNA (mRNA) which in turn moves from the nucleus of a cell to the cytoplasm where it is translated, one codon at a time, into a protein (each codon codes for an amino acid building block). Using



The Central Dogma of Molecular Biology

lab techniques, the mRNA in a tissue can be separated and the original DNA message that encoded it can be determined. The result is an accumulation of short DNA sequences (300 to 900 bases in length) that represent expressed genes (ESTs). Such "libraries" consist of many types of genes, some of which are represented hundreds or thousands of times, others only once. Finding expressed genes that occur in low copy number, likely to be the most important genes biologically, is, naturally, the most difficult and time-consuming task.

In loblolly pine, as many as 19,000 unique genes, perhaps half of the anticipated total, have been discovered. Much more work is needed to find the remaining expressed genes. This will require the construction of more expressed gene libraries. However, some important genes are expressed under such exceedingly rare conditions that they would never be detected in this way, even with an unlimited supply of expressed gene libraries. Consequently, a complete genome sequence is required if we are to identify all the genes.

Gene Identification: Finding the genes (knowing the base sequence) must be followed with identifying the function of the gene. That means answering the question, "what biological processes do the genes influence?" For model species such as *Arabidopsis*, this step is typically done by creating genetically transformed plants with mutations in the gene of interest and looking for phenotypic aberrations. As a result, vast libraries of gene sequences with known function exist in *Arabidopsis* databases. Tree gene sequences can be checked against these databases, and tentative functions can be assigned based on similarity. Alternatively, one useful experimental approach to determining gene function is the microarray. Using nano-technology, short gene sequences can be attached to glass-slides by the thousands, the slides are hybridized with gene transcripts from plants that have been experimentally treated or challenged, and a record of which genes are being expressed in response to the treatment are revealed. An objective of the pine genomics community is to develop an array with all or most of the pine genes (> 35K sequences). To date, 3 different arrays are being constructed that will include as many as 19K unique sequences (among them). Much additional work will be required to standardize and improve this technology.

Gene Characterization: Gene discovery and identification is succeeded by gene characterization, a multi-dimensional category with several long-term objectives. First, we seek to describe genetic variation in all identified genes by re-sequencing genes in populations of individuals. Typically, this variation is described by single base polymorphisms, or SNPs. Secondly, it is desirable to place all genes on a reference genetic linkage map. Finally, we are working toward the completion of a physical map of loblolly pine, and ultimately the complete integration of the physical and genetic maps.

To date, relatively few genes have been characterized on a population basis in loblolly pine (< 100 genes), and there is considerable need for building a high-throughput pipeline to accelerate this effort. Also, the LPGP seeks to place all expressed genes on a genetic linkage map (currently there are fewer than 350). Work is just now beginning on the creation of a physical map of loblolly pine, a process that will likely require 5 years and substantial financial support to complete.

Applications for Tree Genetics and Improvement: Finding, identifying and characterizing all the genes that make up a tree are a daunting, but technically feasible tasks. As progress toward these goals is made, we dramatically increase our fundamental understanding of the genetic basis for all biological functions and add substantially to our ability to domesticate or manipulate select tree populations to meet human needs. We learn how the genome is organized, how it has evolved, and the nature of gene families, with functional and non-functional members. The creation of microarrays reveals how specific genes are expressed under varying environmental conditions, the first step in identifying candidate genes for specific traits, and tailoring environmental treatments to obtain desirable phenotypes. Candidate genes may be used to select superior individuals (see association studies in the next section) or as reagents for genetic engineering experiments. Creation of informative molecular markers provides for critical quality control in orchard and cloning programs, allows for robust mapping of quantitative trait loci (QTL) and facilitates breeding and testing methods. Indeed, the simplest application of molecular markers, genetic fingerprinting, can literally save a company millions of dollars by insuring that appropriate clones are represented in seed orchards or, are the source of starter materials for somatic embryogenesis multiplication.

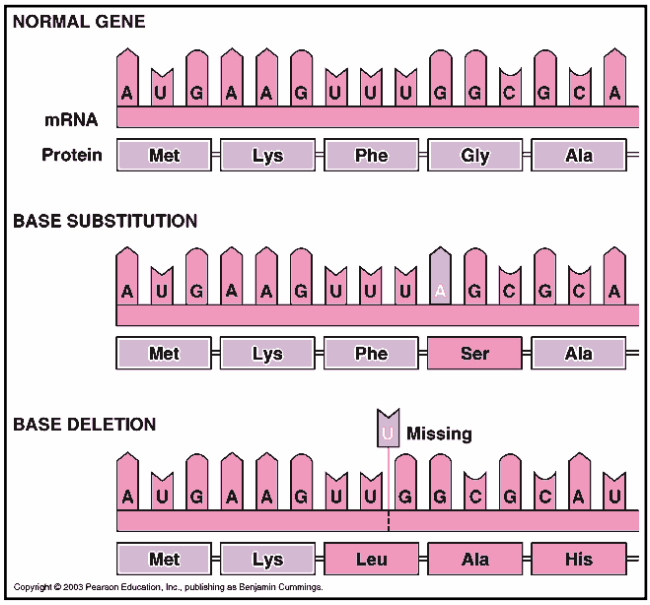
Associating Genotypes and Phenotypes

Identifying and characterizing variation in the DNA of expressed genes, as noted previously, provides the foundation for associating genotypes with phenotypes. Selection for a desirable trait, such as high wood specific gravity, based on the presence of a single nucleotide out of billions in the pine genome is potentially the most fundamental and profound application of genomics. Developing the infrastructure to identify associations between genotypes and phenotypes is the focus of this section.

Most traits of interest to forest geneticists are quantitatively inherited (i.e. controlled by the collective action of many genes with small effects, resulting in continuous or quantitative variation). Historically, we have relied on direct phenotypic assessments and quantitative genetic analytical methods to gradually improve tree populations for these traits (i.e. shift allele frequencies), a slow and inefficient process. By the early 1990's, developments in DNA marker and genome mapping technology resulted in the construction of genetic linkage maps and the molecular dissection of quantitative traits (what came to be known as Quantitative Trait Locus mapping). QTL reveal the number and genome location of genes affecting a trait as well as the magnitude of effect of each gene, greatly expanding our knowledge of forest genetics. For the tree breeder, QTL offer the possibility of indirectly selecting for phenotypic traits using molecular markers (Marker Aided Selection, MAS) but the method is restrictive (typically of use only in the pedigree within which the QTL were located) and tells us nothing about the actual genes controlling the trait. Ideally, we would like to know which genes affect a trait, and how variations (mutations) in those genes influence phenotypes. An approach called 'association genetics' provides a means for doing this.

As can be seen in figure 5 (below), changes (mutations) in the DNA code at even a single base can have profound effects on the protein product of that gene. For instance, the

substitution of one base for another can change one amino acid while the loss of a base all together results in a shifting of the codon reading frame and likely a complete loss of function for that protein. Such a mutation in the CAD gene of a loblolly pine plus tree has resulted in the loss of function of that gene and an associated change in the lignin chemistry in that tree. Studies have subsequently shown a relationship between that simple mutation, decreased cost of pulping of wood from that tree and increased growth of progeny with the mutation. Using carefully controlled studies with appropriate populations, statistical associations between such mutations and phenotypic traits can be, and have been, found for other genes in loblolly pine.



The long-term objectives of the LPGP are to identify mutations in all expressed genes and to seek associations between those mutations and economically/ecologically important phenotypic traits. Initially, genes will be prioritized based on their known or suspected function (candidate genes) as determined in the gene discovery and identification components of this project. To meet this objective, 4 research areas need to be addressed: 1) development of appropriate populations for association testing, 2) development of methods to rapidly and cost effectively genotype test trees for mutations in candidate genes, 3) development of new, faster, and less-expensive methods to phenotype large numbers of plants for heritable traits, and 4) development of improved computational approaches for integrating genotypes and phenotypes.

Applications

The large-scale identification of associations between specific genotypes (characterized by single base mutations) and phenotypes of important traits will profoundly change the way tree improvement is practiced in the United States. Selecting for molecular markers at the earliest of ages will dramatically increase gain per unit of time at a fraction of the cost of field testing. The technology will have utility for both population improvement and selection of clones. Most notably, selection based on allelic effects in candidate genes will have utility for all loblolly pine families. Significantly, simultaneous selection for multiple traits is feasible.

Information and Material Management

A world-class genomics project is ultimately dependent upon a strong, integrated, and well-supported information resource database such as the Dendrome Project (http://dendrome.ucdavis.edu) that allows for the capture, organization, querying and

archiving of the vast amount of information being generated by multiple research institutions. In addition, the biological resources must be archived, curated and distributed through a Genetic Stock Center, which for pine and other forest trees does not yet exist. Secured funding from non-traditional (non-grant supported) is absolutely necessary for development of these pivotal resources.

Funding of a Loblolly Pine Genome Project

It is recommended that a 5 year program at a significant dollar amount be funded to support a Loblolly Pine Genome Project. The 5 year goals will be to 1) develop a national genomics infrastructure and 2) develop a basic understanding of the complexity of the pine genome and the function of all expressed genes. The goal of a subsequent 5 year program will be to obtain a complete genome sequence of loblolly pine. The Loblolly Pine Genome Program should be administered by the CSREES under the National Research Initiative and/or by the USDA Forest Service Agenda 2020 Program. Specific objectives for the first 5 year program would include:

Gene Discovery

1. Complete a 500,000 EST database.
2. Obtain a complete full-length gene coding sequence resource.
3. Complete BAC library construction.
4. Construct a complete physical map.
5. Complete dense expressed gene genetic maps.

Gene Function

1. Complete construction of high-density gene expression arrays.
2. Negotiate access to high-efficiency DNA transformation systems held in the private sector.
3. Develop better tools for mutation analysis and the study of gene function.

Genotype to Phenotype

1. Develop multiple association mapping populations.
2. Develop allele discovery and high-throughput SNP genotyping platforms.
3. Develop high-throughput phenotyping technologies.
4. Develop computational methods for association genetics and marker breeding technologies.

Comparative Genomics

1. Develop a comparative map and sequence genomics infrastructure so that loblolly pine genomic information can be compared and leveraged against similar efforts in other pines and conifers worldwide (e.g. Monterey pine, Maritime pine, Scots pine, Douglas-fir and Norway spruce)

Genome Database and Genetic Stock Center

1. Support development of a public genome database and genetic stock center for archiving, curating and distribution of genome resources.