

PineRefSeq: Conifer Reference Genome Sequencing

An Adaptive Approach to the
Sequencing of the Large Genomes of
Multiple Conifer Species



U. California-Davis, CHORI, Johns Hopkins U.,
U. Maryland, Indiana U., Texas A&M U., Washington State U.

Supported by USDA NIFA AFRI



United States
Department of
Agriculture

National Institute
of Food and
Agriculture



PineRefSeq
<http://pinegenome.org/pinerefseq/>

PineRefSeq

Project Goal

To provide the benefits of conifer reference genome sequences to the research, management and policy communities.

Specific Objectives (Aims)

- Provide high-quality reference genome sequences of loblolly pine, sugar pine, and Douglas-fir
- Provide complete transcriptome sequences for gene discovery, reference building, and aids to genome assembly
- Provide annotation, data integration, and data distribution through Dendrome and TreeGenes databases.

Scientific
Advisory Committee

Tom Byram
David Jaffe
Pankaj Jaiswal
Richard McCombie

Collaborations

Sally Aitken, AdapTree, Canada,
Interior spruce and Lodgepole pine genomics
Rich Cronn, USA, Douglas-fir transcriptome
Pär Ingvarsson, Sweden, Norway spruce genome

PineRefSeq Project

Specific Aim 1

High quality reference
genome sequences of
loblolly pine and three
other conifer species

Charles Langley, Pieter de Jong,
Maxim Koriabine, Steven Salzberg,
James Yorke, Aleksey Zimin,
David Neale

Specific Aim 2

Transcriptome sequencing
for gene discovery, reference
building, and aids to genome
assembly

Keithanne Mockaitis, Carol Loopstra

Specific Aim 3

Dendrome and *TreeGenes*
databases: Annotation, data
integration, and distribution

Doreen Main, Jill Wegrzyn,
David Neale

Dendrome

Project Director
David Neale

Project Coordinator
Patrick McGuire

Training Coordinator
Nicholas Wheeler



United States
Department of
Agriculture

National Institute
of Food and
Agriculture



PineRefSeq
<http://pinegenome.org/pinerefseq/>

Why Do We Need a Conifer Genome Sequence?

- Phylogenetic Representation –
 - None currently exists. The conifers (gymnosperms) are the oldest of the major plant clades, arising some 300 million years ago. They are key to our understanding of the origins of genetic diversity in higher plants.
- Ecological Representation –
 - Conifers are of immense ecological importance, comprising the dominant life forms in most of the temperate and boreal ecosystems in the Northern Hemisphere.
- Fundamental Genetic Information –
 - Reference sequences are the fundamental data necessary to understand conifer biology and aid in guiding management of genetic resources.
- Development of Genomic Technologies –
 - The analytical and computational challenge of building a reference sequence for such large genomes will drive development of tools, strategies, and human resources throughout the genomics community.



Why Sequence Multiple Conifer Species' Genomes?

- Improve the reference sequence
 - Comparisons of related genomes improves the quality and quantity of reference sequence that can be assembled.
- Improve the Value
 - Many basic and applied research questions can be started, advanced or resolved by comparative genomics of related species.
- Improve the Efficiency
 - Return on investment is dramatically improved due to reduced costs of assembly of subsequent genomes.



The Large, Complex Conifer Genomes Present a Formidable Challenge

- **Challenges**

- The 24 Gigabase loblolly pine genome is 8 times larger than the human genome, and far exceeds any genome sequenced to date.
- Conifer genomes generally possess large gene families (duplicated and divergent copies of a gene), and abundant pseudo-genes.
- The vast majority of the genome (>95%) appears to be moderately or highly repetitive DNA of unknown function.

- **Approaches to Resolving Challenges**

- An adaptive approach that embraces current and developing “best” sequencing technologies and strategies.
- Complementary sequencing strategies that seek to simplify the process through use of actual or functional haploid genomes and reduced size of individual assemblies.



Plant Genome Size Comparisons

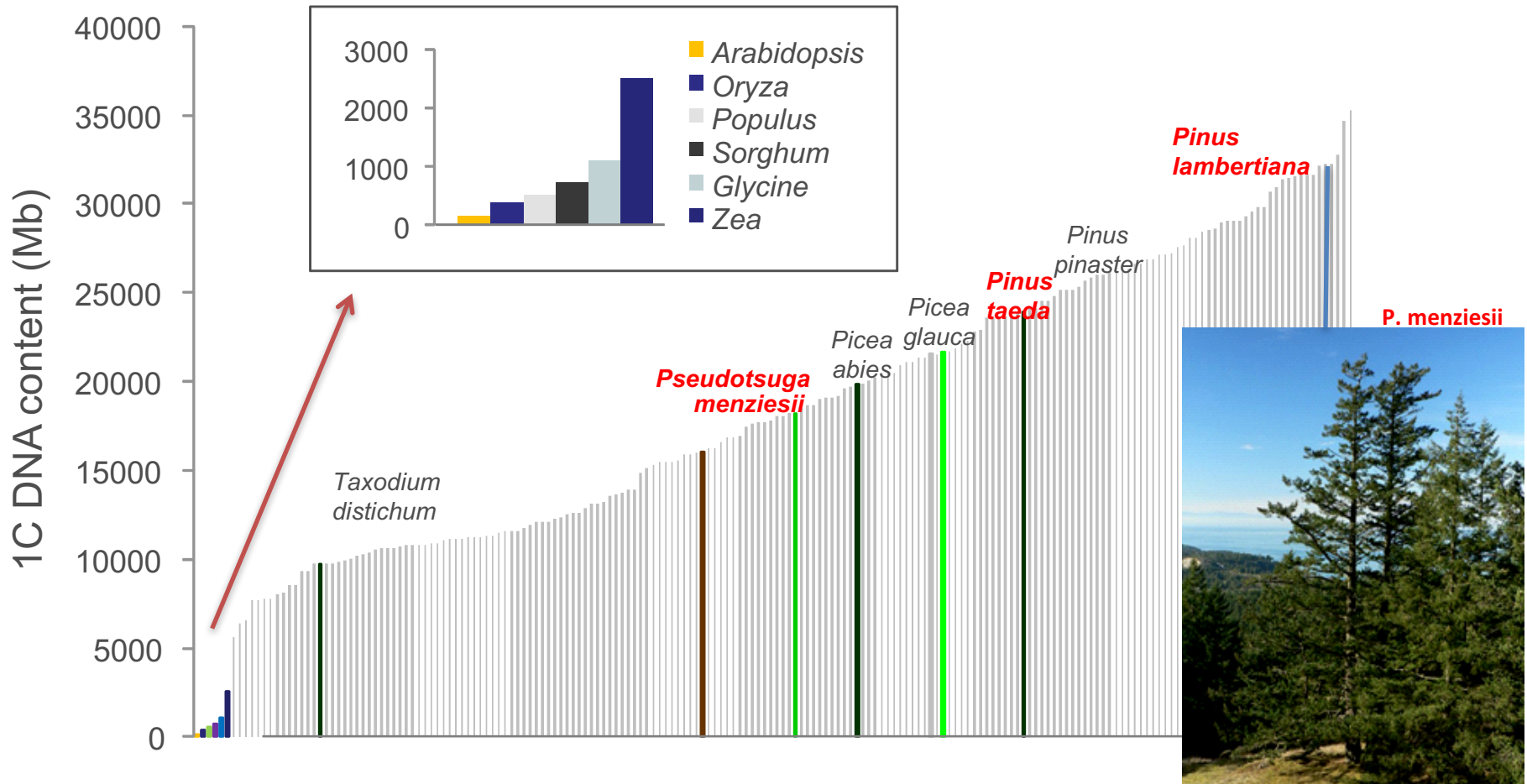


Image Credit: Modified from Daniel Peterson, Mississippi State University



United States
Department of
Agriculture

National Institute
of Food and
Agriculture


PineRefSeq
<http://pinegenome.org/pinerefseq/>

Existing and Planned Angiosperm Tree Genome Sequences

Species		Genome Size ¹	Number of Genes ²	Status ³
In Progress With Draft Assemblies				
<i>Populus trichocarpa</i>	Black Cottonwood	500 Mbp	~ 40,000	2.0 / 2.2
<i>Eucalyptus grandis</i>	Rose Gum	691 Mbp	~36,000	1.0 / 1.1
<i>Malus domestica</i>	Apple	881 Mbp	~26,000	1.0 / 1.0
<i>Prunus persica</i>	Peach	227 Mbp	~28,000	1.0 / 1.0
<i>Citrus sinensis</i>	Sweet Orange	319 Mbp	~ 25,000	1.0 / 1.0
<i>Carica papaya</i>	Papaya	372 Mbp	-	
<i>Amborella trichopoda</i>	Amborella	870 Mbp	-	
In Progress Or Planned – No Published Assemblies				
<i>Castanea mollissima</i>	Chinese Chestnut	800 Mbp	-	
<i>Salix purpurea</i>	Purple Willow	327 Mbp	-	
<i>Quercus robur</i>	Pedunculate Oak	740 Mbp	-	
<i>Populus spp and ecotypes</i>	Various	various	-	
<i>Azadirachta indica</i>	Neem	384 Mbp	-	

1) Genome size: Approximate total size, not completely assembled.

2) Number of Genes: Approximate number of loci containing protein coding sequence.

3) Status: Assembly / Annotation versions; <http://www.phytozome.net/> ; <http://asgpb.mhpc.hawaii.edu/papaya/> ; <http://www.amborella.org> ; (purple willow – <http://www.poplar.ca/pdf/edomonton11smart.pdf> ; Neem - (<http://www.strandls.com/viewnews.php?param=5¶m1=68>



United States
Department of
Agriculture

National Institute
of Food and
Agriculture



Existing and Planned Gymnosperm Tree Genome Sequences

Species		Genome Size ¹	Number of Genes ²	Status ³
Gymnosperms				
<i>Picea abies</i>	Norway Spruce	20,000 Mbp	?	Pending
<i>Picea glauca</i>	White Spruce	22,000 Mbp	?	Pending
<i>Pinus taeda</i>	Loblolly Pine	24,000 Mbp	?	Pending
<i>Pinus lambertiana</i>	Sugar Pine	33,500 Mbp	?	Pending
<i>Pseudotsuga menziesii</i>	Douglas-fir	18,700 Mbp	?	Pending
<i>Larix sibirica</i>	Siberian Larch	12,030 Mbp	?	Pending
<i>Pinus pinaster</i>	Maritime Pine	23,810 Mbp	?	Pending
<i>Pinus sylvestris</i>	Scots Pine	~23,000 Mbp	?	Pending

1) Genome size: Approximate total size, not completely assembled.

2) Number of Genes: Approximate number of loci containing protein coding sequence.

3) Status: Assembly / Annotation versions; See <http://www.phytozome.net> for all publically released tree genomes. Conifer genomes will also be posted here as they are completed.

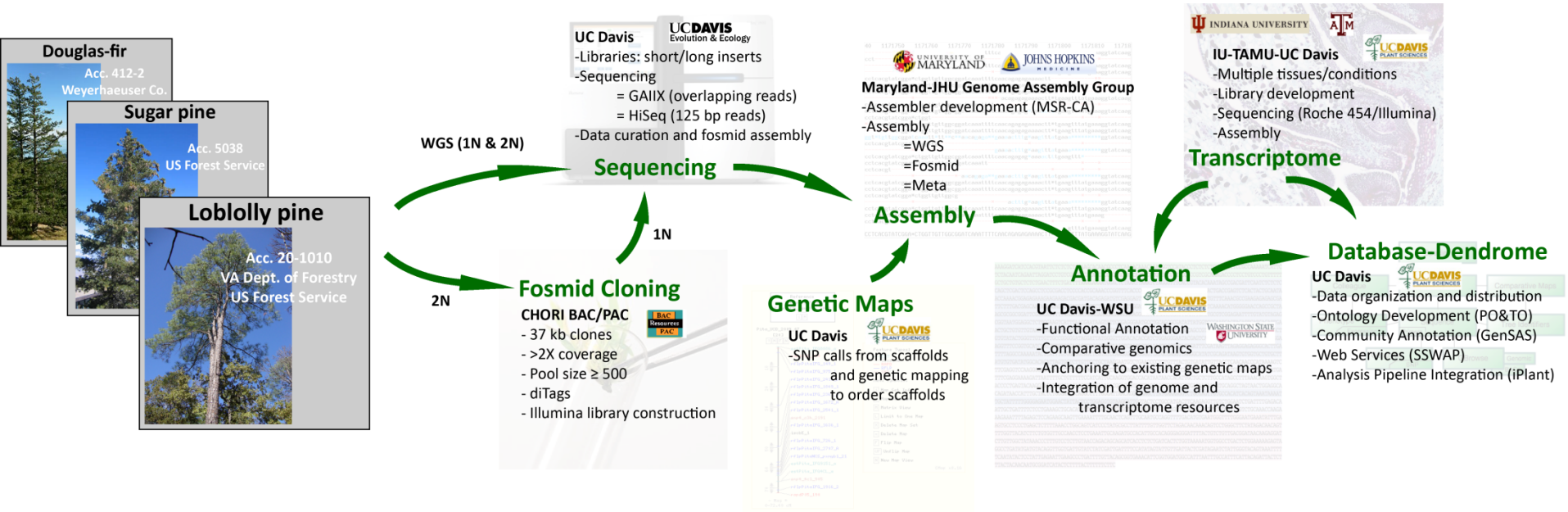


United States
Department of
Agriculture

National Institute
of Food and
Agriculture

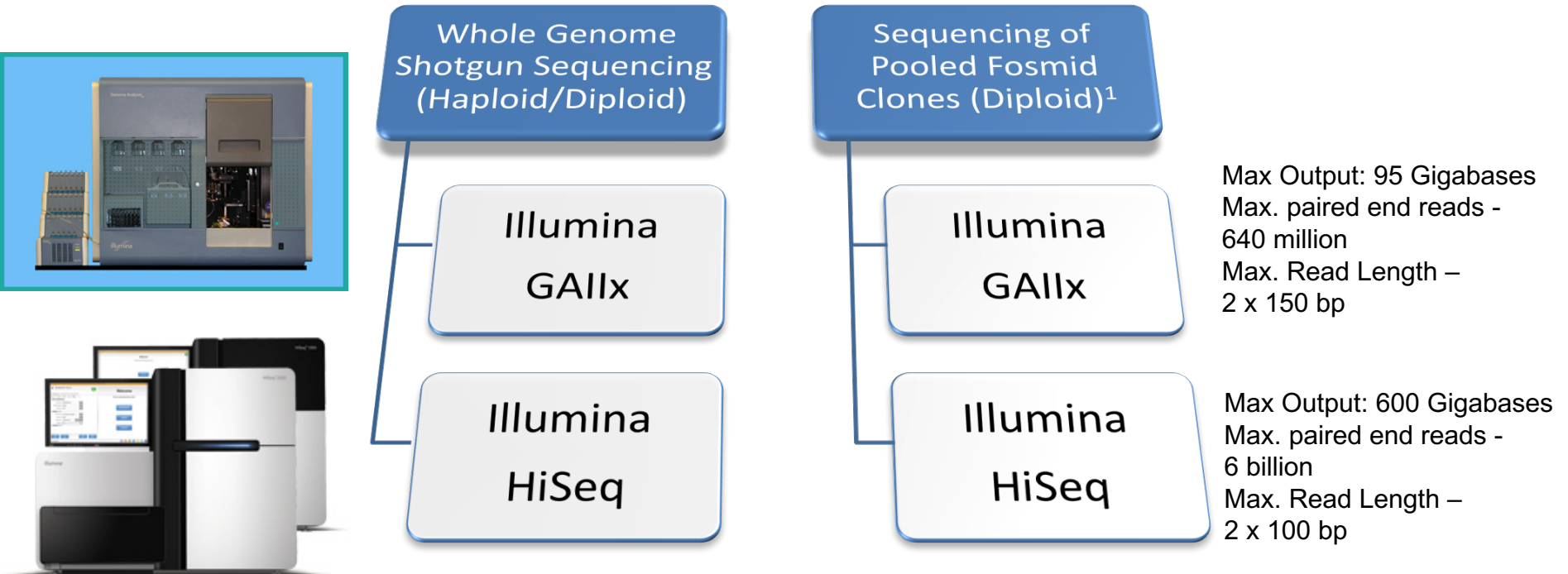


Elements of the Conifer Genome Sequencing Project



Strategy for De Novo Sequencing of the Conifer Genomes

Parallel and Complementary Approaches

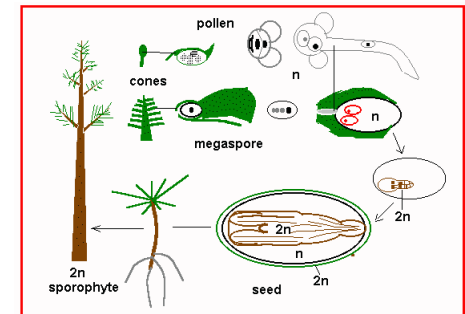


¹ Effectively haploid

Data current as of 3/2012 (Illumina)

Complementary Sequencing Approaches

- Whole Genome Shotgun (WGS) sequencing of
 - Haploid Megagametophyte: Goal - deep (>40X) representative short insert libraries from a single haploid ($1N$) segregant. Haploid genome significantly improves sequence assembly
 - Diploid Parent: Goal – deep ($\sim 8 X$) representative large-fragment paired end sequences (1 billion paired ends, 100 bp/end).



<http://hcs.osu.edu/hcs300/gymno.htm>

- Direct Sequencing of Pooled Fosmid libraries
 - *P. taeda* fosmid pool size is 500 clones
 - Complexity of individual pools well within available assembler's specs.
 - Effectively haploid, facilitating assembly ($2N$ fosmid library but $1N$ pool).

Fosmid Pooling:

Genome Partitioning to create reduced genome complexity

- The immense and complex diploid pine genome can be chiseled into bite-sized, functionally haploid, pieces using variable insert-sized fosmid clones.
- Fosmid pools (~500 clones) with combined insert sizes far less (~100 Mbp) than a haploid genome size assures a haploid genome representation.
- Pools will consist of short (250 to 750 bp), long (3 – 5 kbp) and paired-end fosmid (37-40 kbp) inserts.
- The final dataset from sequencing the fosmid pool will be ~180 X depth, and have pool labels to facilitate assembly.



Sequence Assembly

- Independent assemblies will be constructed for each of the complementary sequence databases.
 - Paired haploid sequences should be possible to assemble from the haploid / diploid DNA templates.
 - This can be facilitated by using fosmid overlap and genetic mapping of literally 10s of thousands of SNP markers selected from scaffolds.
- Assembly will be iterative, using a combination of the De Bruijn Graph and Overlap Layout Consensus (OLC) strategies.
- Multiple assemblers (Celera, SOAPdenovo, Allpaths-LG), including our own (MSR-CA) will be used and compared.

Transcriptome Assembly Summary

Full transcript assemblies developed from previous sequencing efforts as well as additional sequencing from Mockaitis (CGB-IU)/Loopstra (TAMU) will be used in functional genomics studies, WGS assembly, and profiling of gene expression differences in response to biotic and abiotic stresses.



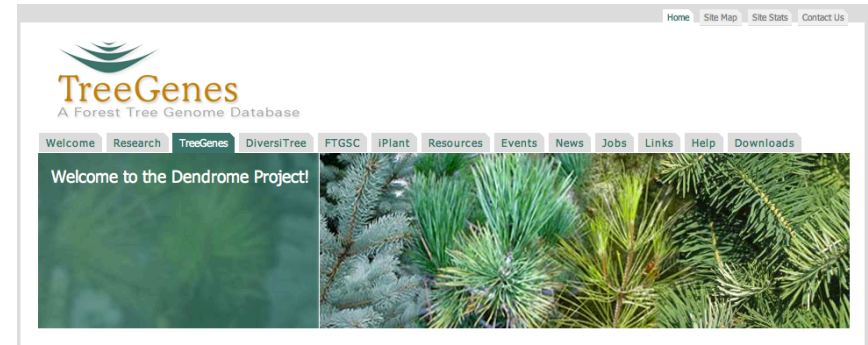
Douglas-fir RNASeq (FS) & 454 (JGI); Sugar pine RNASeq (FS) & 454 (JGI); Loblolly pine 454 (JGI & CGB-IU) & RNASeq (UCD);

Annotation and Database Management

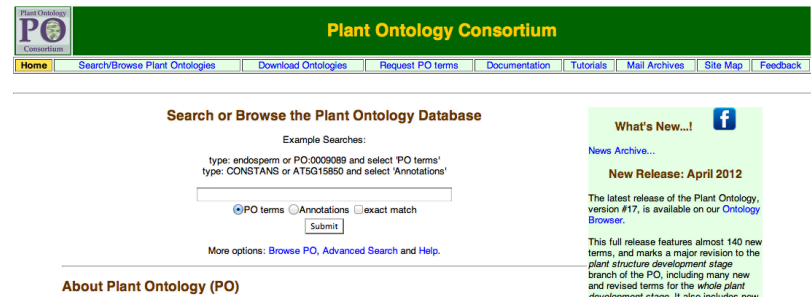
- Genome and transcriptome sequences will be delivered via the *TreeGenes* (Dendrome) database.
- Comprehensive SNP resources will be distributed through *Dendrome's DiversiTree* interface.



- Recent collaborations resulted in the inclusion of wood-related terms in Plant Ontology (#17).
- Continued integration of conifer-related terms into Plant Ontology and Gene Ontology is in progress.
<http://www.plantontology.org/amigo/go.cgi>



- Primary annotation provided through collaboration with GDR (Genome Database for Rosaceae).
- Community-level annotation provided by integrated web-based tools (GDR GenSAS and *Dendrome* Gbrowse).



Schedule of Activities

Aim/Task	Year 1		Year 2		Year 3		Year 4		Year 5	
Methods Development										
Lob Pine Primary Seq										
Lob V1.0 Assembly										
Lob Seq Polishing										
Sugar Pine Sequence										
Douglas-fir Sequence										
Transcriptome Reference										
Transcriptome Functional Analysis										
Data Integration And Distribute										
Genome Annotation										

Practical Applications of a Conifer Reference Sequence

Through a series of Federally funded (USDA, NSF) research projects, our knowledge of the conifer genome has expanded tremendously over the last 15 years. Specifically, the forest genetics community has:

- Identified large libraries of putative genes (ESTs – sequences of genes that are expressed in plant tissues).
- Developed very large inventories of genetic markers such as SNPs (Single Nucleotide Polymorphisms) and SSRs (Simple Sequence Repeats)
- Vastly improved the density of genetic maps.
- Identified statistical associations between allelic variation in known genes and variation in traits of commercial or ecological importance.
- Developed techniques for using markers to improve tree breeding.

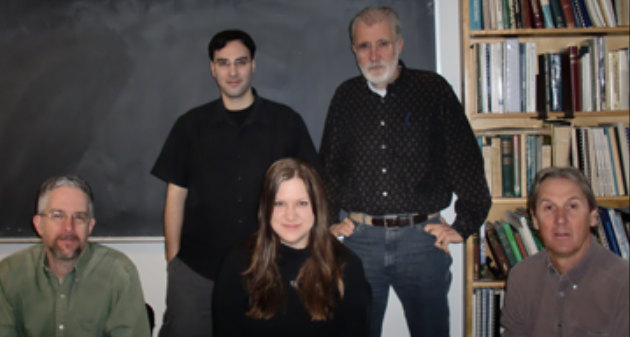
Despite this progress, practical applications such as Marker Assisted Selection (MAS) remain illusive because our efforts have only revealed a fraction of the genome's information.



Practical Applications of a Conifer Reference Sequence

A complete conifer reference sequence and knowledge of essentially all the genes in the genome will substantially change the nature of tree improvement in domesticated tree populations and land management in natural populations by:

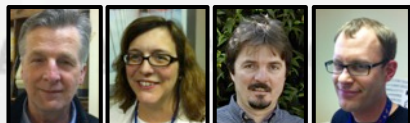
- **Fully informed MAS approaches such as Association Genetics and Genomic Selection**
- **Improved tree breeding methods**
- **Identifying populations resilient or susceptible to environmental stress and climate change**
- **Informing decisions on assisted migration**



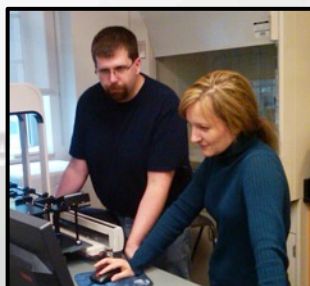
PD David Neale (r), co-PD Jill Wegrzyn (c), and (l to r) John Liechty, Ben Figueroa, and Patrick McGuire UC Davis



The Maryland Genome Assembly Group featuring co-PD Steven Salzberg and Daniela Puiu (Johns Hopkins U) and co-PD Jim Yorke and Aleksey Zimin (U of Maryland)



(l to r) Co-PI Pieter de Jong, Ann Holtz-Morris, Maxim Koriabine, Boudewijn ten Hallers CHORI BAC/PAC



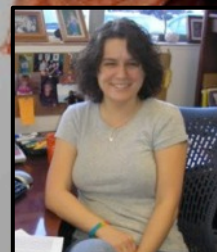
Co-PI Keithanne Mockaitis and Zach Smith Indiana U



Co-PI Chuck Langley (r) and (l to r) Marc Crepeau, Kristian Stevens, and Charis Cardeno UC Davis



Co-PI Carol Loopstra and Jeff Puryear TAMU



Co-PI Dorrie Main WSU



United States Department of Agriculture

National Institute of Food and Agriculture



<http://pinegenome.org/pinerefseq/>