

The Pine Reference Genome Sequence and Applied Tree Breeding

Nicholas Wheeler: MTBS, LLC
Ross Whetten: North Carolina State University



Reference Genome Sequence:

For any given organism (species), the complete and ordered “assembly” of DNA, as denoted by the nucleotides A, T, C, and G.

APCAAGTCATCCATGATT
TCCGCATAGTAGCTCATA
TCATAGTCTTCAATGCA
APCAAGTCATCCATGATC
TCATAGTAGCTCATA

Why Do We Need a Conifer Genome Sequence?

Fundamental Genetic Information

Phylogenetic Representation

Ecological Representation

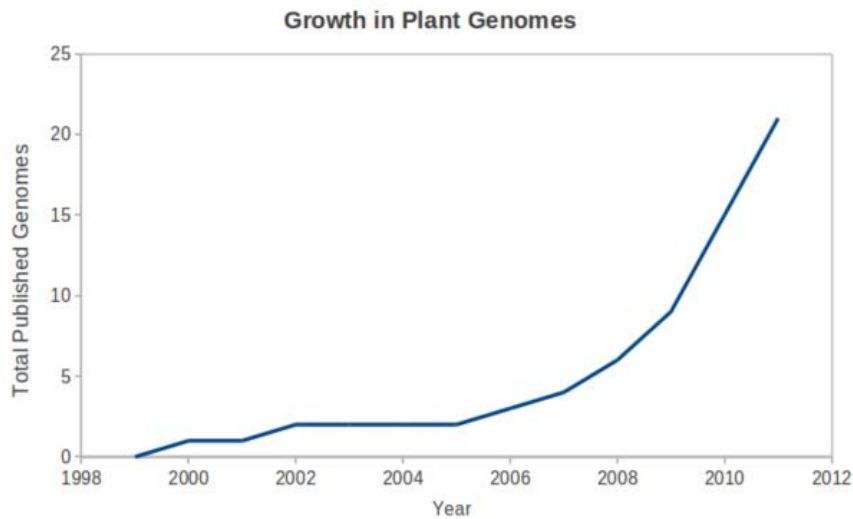
Development of Genomic Technologies

Economic Importance

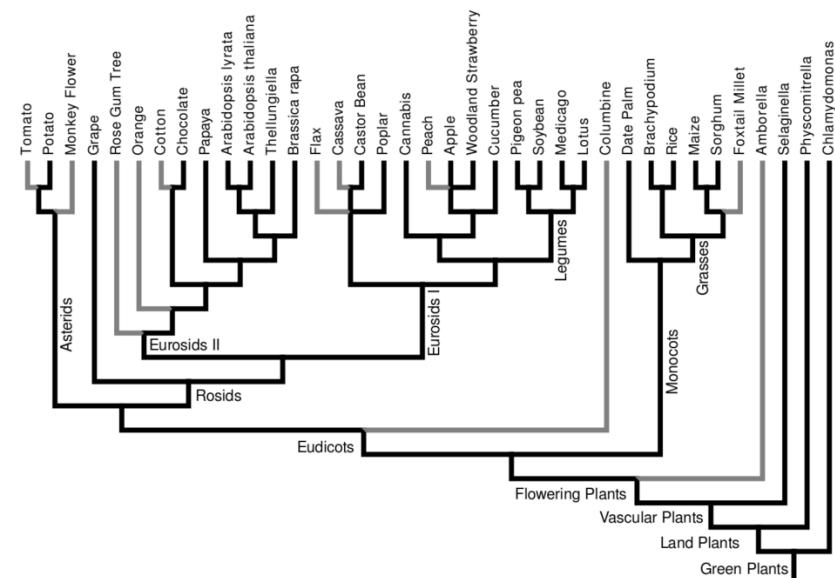


Genome Sequencing - A Short History

The rate of publication of plant genomes, updated in late 2011



A phylogenetic tree of all plants with published full genomes as of May 13, 2012



Existing & Planned Angiosperm Tree Genome Sequences

As of mid-2012

Species		Genome Size ¹ (Mbp)	# of Genes ²	Status ³
---------	--	-----------------------------------	-------------------------	---------------------

In Progress with Draft Assemblies

<i>Populus trichocarpa</i>	Black Cottonwood	500	~40,000	2.0 / 2.2
<i>Eucalyptus grandis</i>	Rose Gum	691	~36,000	1.0 / 1.1
<i>Malus domestica</i>	Apple	881	~26,000	1.0 / 1.0
<i>Prunus persica</i>	Peach	227	~28,000	1.0 / 1.0
<i>Citrus sinensis</i>	Sweet Orange	319	~25,000	1.0 / 1.0
<i>Carica papaya</i>	Papaya	372	-	
<i>Amborella trichopoda</i>	Amborella	870	-	
<i>Betula nana</i>	Dwarf Birch	450	-	1.0 / -

In Progress or Planned - No Published Assemblies

<i>Castanea mollissima</i>	Chinese Chestnut	800	-	
<i>Salix purpurea</i>	Purple Willow	327	-	
<i>Quercus robur</i>	Pedunculate Oak	740	-	
<i>Populus</i> spp. and ecotypes	Various	Various	-	
<i>Azadirachta indica</i>	Neem	384	-	

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

Existing and Planned Gymnosperm Tree Genome Sequences

As of mid-2012

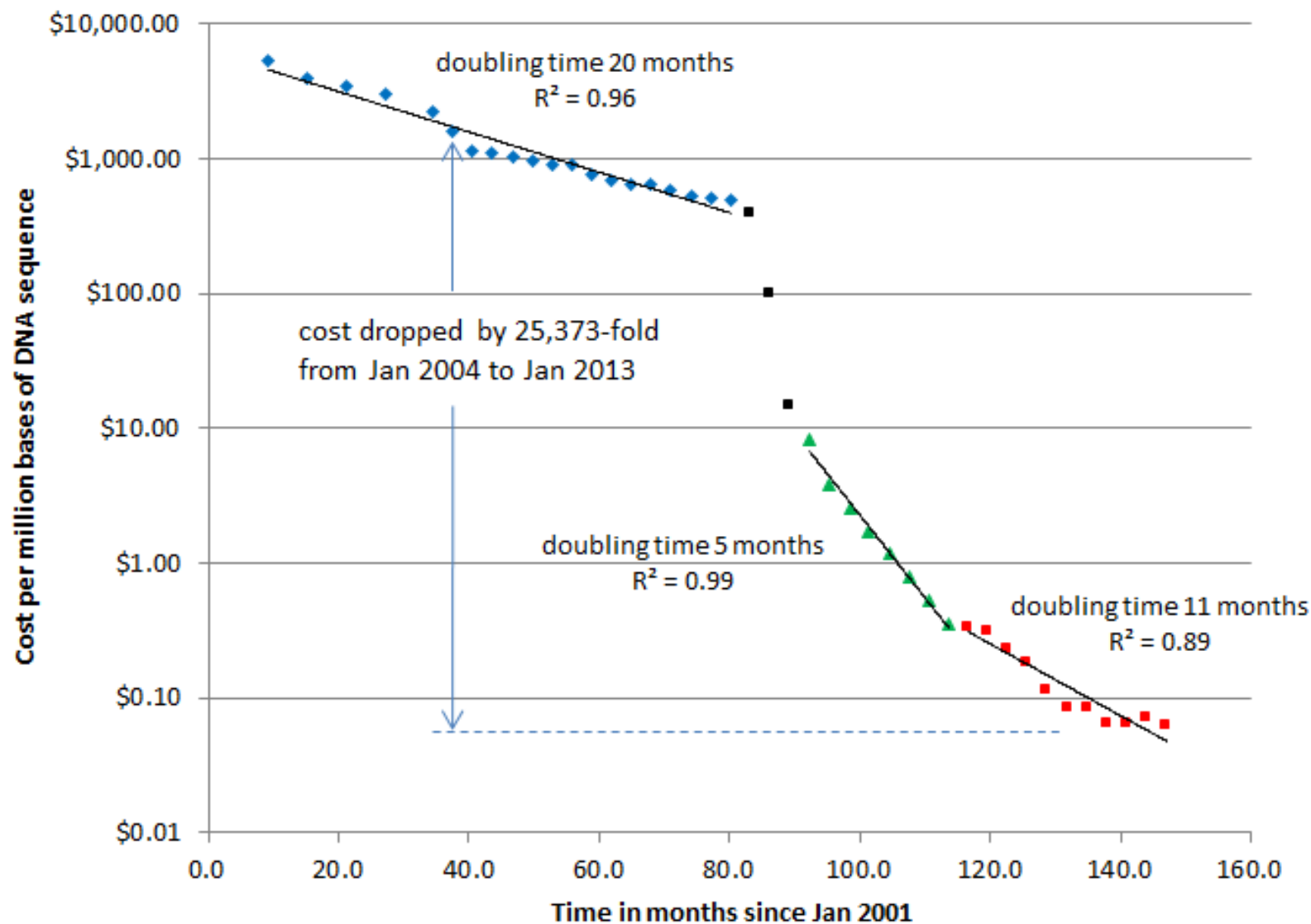
Species		Genome Size ¹ (Mbp)	# of Genes ²	Status ³
Gymnosperms				
<i>Picea abies</i>	Norway Spruce	20,000	?	Pending
<i>Picea glauca</i>	White Spruce	22,000	?	Pending
<i>Pinus taeda</i>	Loblolly Pine	24,000	?	Pending
<i>Pinus lambertiana</i>	Sugar Pine	33,500	?	Pending
<i>Pseudotsuga menziesii</i>	Douglas-fir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

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

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

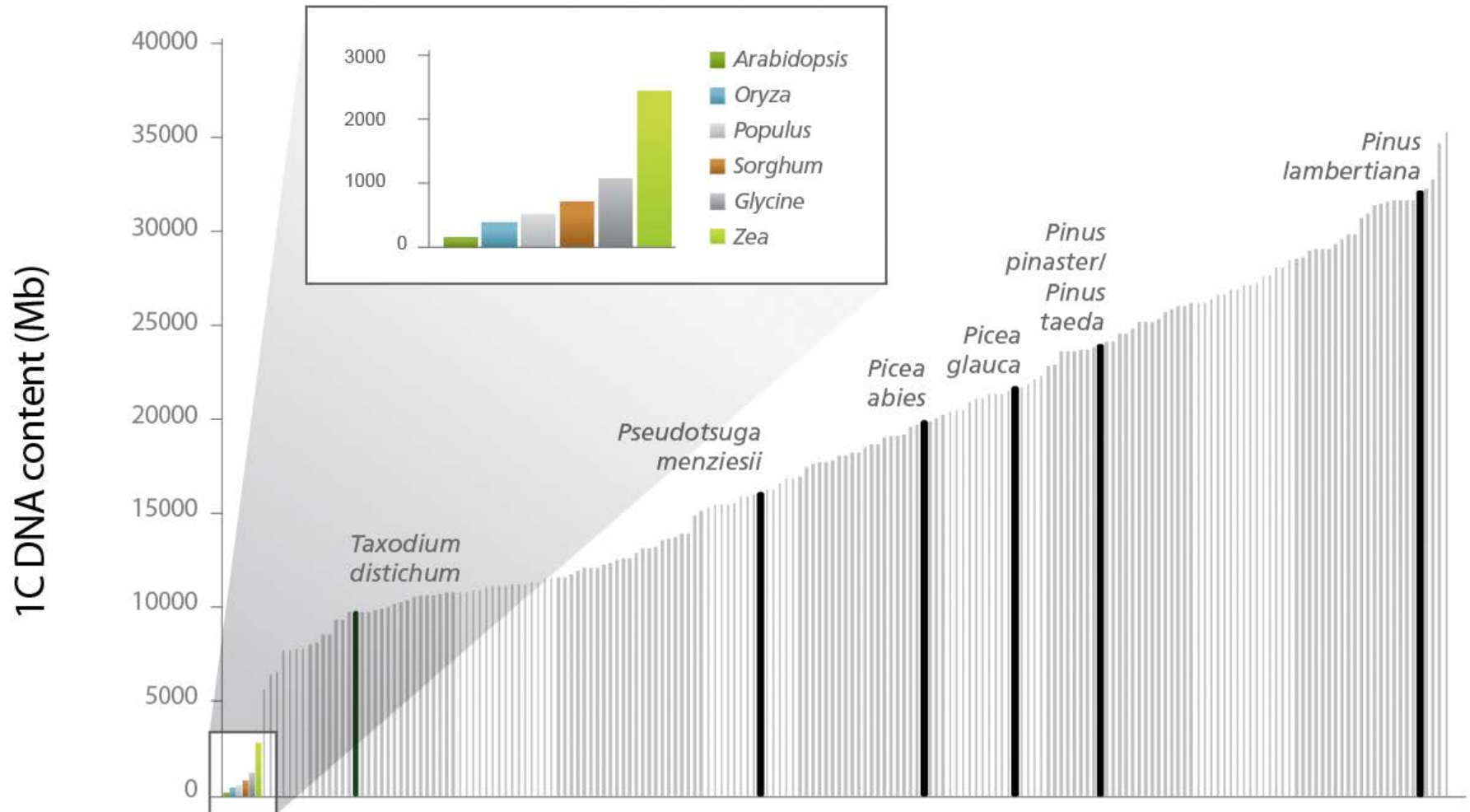
3 Status: Assembly / Annotation versions; See <http://www.phytozome.net> for all publicly released tree genomes. Conifer genomes will also be posted here as they are completed.

DNA sequencing cost at NIH genome centers



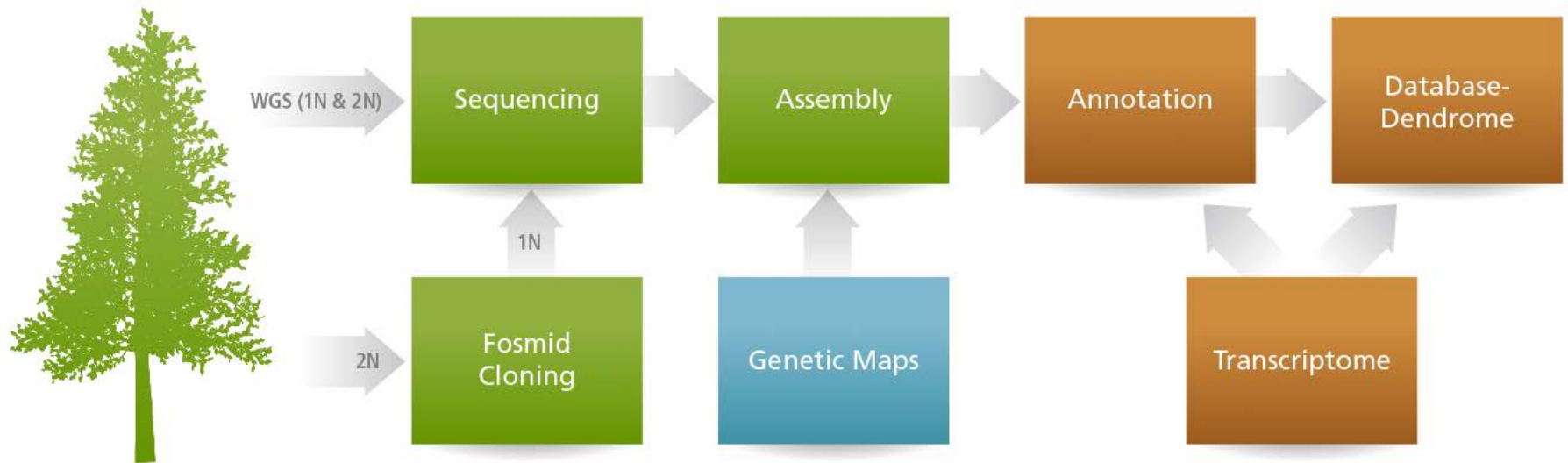
◆ Jan-01 to Jul-07 ■ Aug-07 to Apr-08 ▲ Jul-08 to Apr-10 ■ Jul-10 to Jan-13

Challenges to Sequencing a Conifer Genome



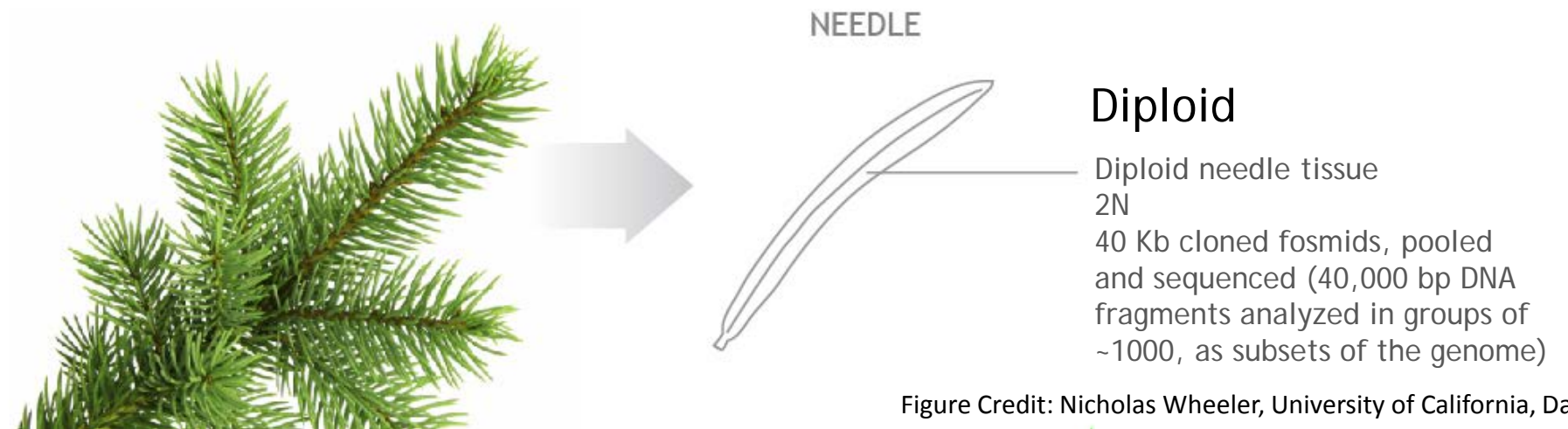
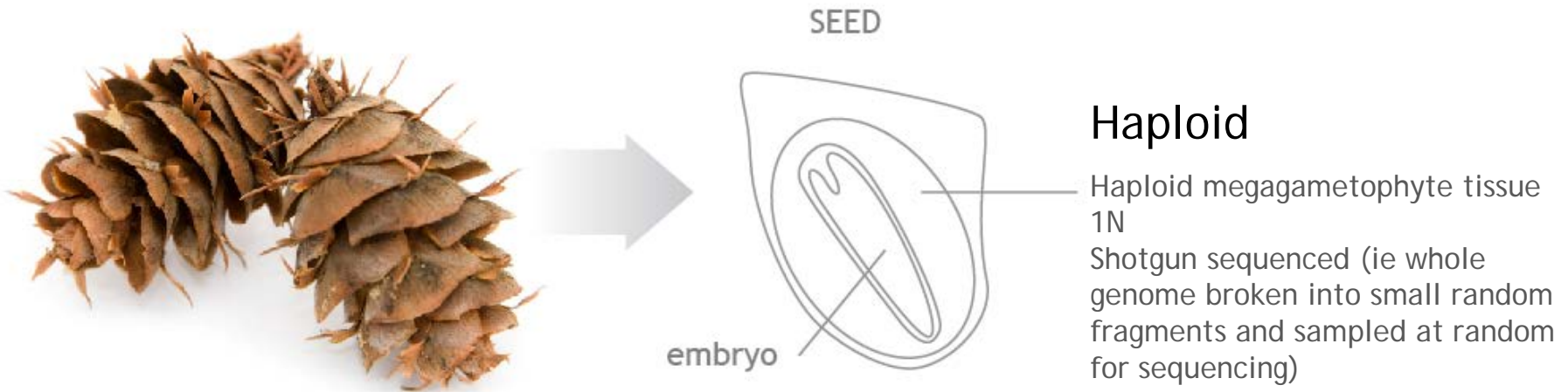
Elements of a Conifer Genome Sequencing Project

Approaches to Resolving Challenges



Acquiring the Sequence

Target Genome, Appropriate Tissues for DNA & RNA



Assembling the Reference Sequence

The Essence of Assembly

This general approach is called OLC or Overlap-Layer-Consensus.

Find pieces that fit together: Compute overlaps of reads

```

AGTGATTAGATGACTAGA
      | | | | | | | | |
GATGATAGTAGAGGATAGATTTA
  
```

Connect the pieces: Create layout of numerous overlapping reads

```

AGTGATTAGATGATAGTAGA
                GATGATACTAGAGGATAGACC
                ATAGGTAGAGGATAGACCACTCATCTAG
  
```

Create consensus sequence of contiguous nucleotides (i.e., contigs)

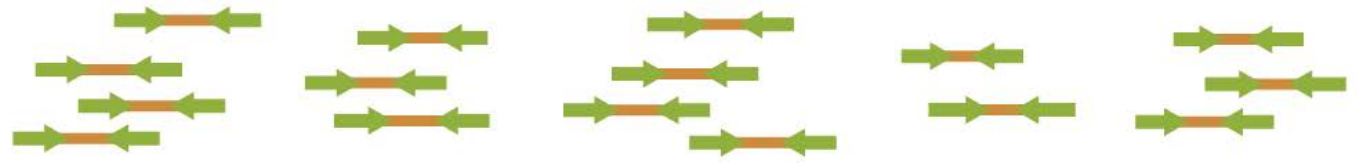
```

AGTGATTAGATGATAGTAGAGGATAGACCACTCATCTAG
  
```

Assembling the Reference Sequence

Based on Whole Genome Shotgun Sequencing

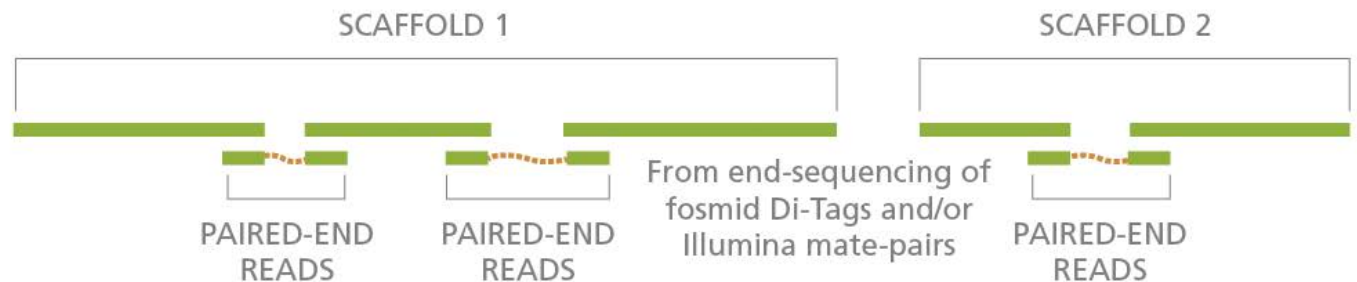
Sheared genome fragments (200 to 600 bp), prep and sequence using next-generation sequencing platform(s)



Continuous sequence – Contigs

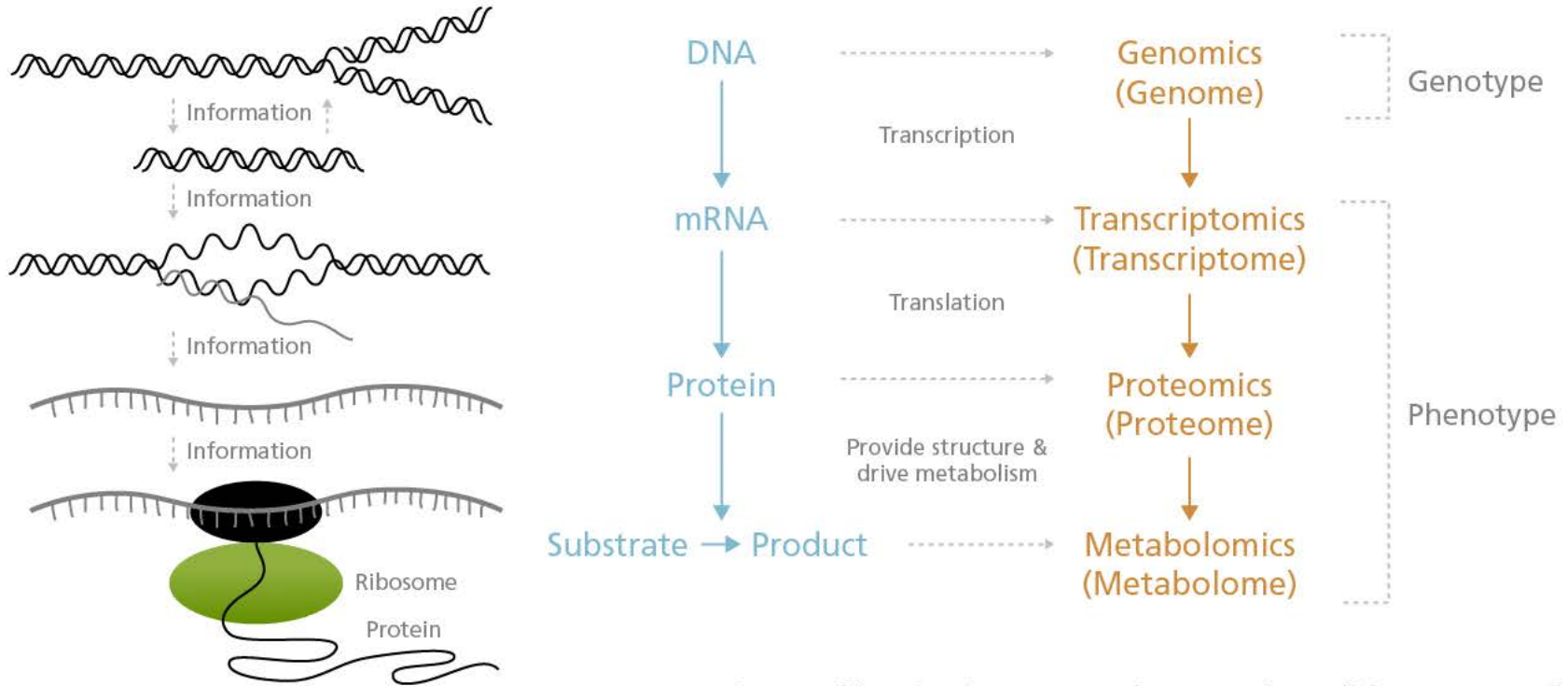


Scaffold builds facilitated by paired-end or mate-pair reads



Transcriptome (RNA) sequencing defines the genes expressed in different pine tissues

Figure Credit: Modified from Keithanne Mockaitis, Indiana University



▼ Transcriptome libraries from many tissues and conditions are needed





Preliminary Results of Transcriptome Reference Sequencing for Three Conifer Species

Updated summary of transcriptome assemblies from 454 (CCG, JGI) and RNASeq (FS) in Psme (Douglas-fir), Pila (sugar pine), and Pita (loblolly pine).



Library	N, quality filtered	Nucleotides	Transcripts Assembled	Mean Contig Length	Unique Transcripts
<i>Pila</i> Needles & Candles 454 (Newbler)	1,096,017	387,174,063	28,910	955	49,035
<i>Pila</i> Needle RNASeq (Trinity)			33,961		
<i>Psme</i> Needles and Candles 454 (Newbler)	1,216,156	419,643,998	25,041	961	92,897
<i>Psme</i> Needle RNASeq (Trinity)			99,936		
<i>Pita</i> Shoot 454 (Newbler)	874,971	205,284,775	62,342	1,124	48,842
<i>Pita</i> Callus 454 (Newbler)	882,199	344,842,307	37,322		
<i>Pita</i> Stem 454 (Newbler)	934,760	310,498,816	43,234		

Loblolly Pine: Unique complete protein coding genes: 87,602!
(Over one million alternative transcripts associated with above loci)



Genome Assembly Statistics for Recently Sequenced Species

Year	Common Name	Scientific Name	Assembly Size (GB)	Predicted Size (GB)	N50 Contig (KB)	N50 Scaffold (KB)
2011	Potato	<i>Solanum tuberosum</i> L.	0.7	0.8	31.4	1320.0
2011	Orangutan	<i>Pongo abelii/pygmaeus</i>	3.1	3.1	15.5	740.0
2011	Nake Mole Rat	<i>Heterocephalus glaber</i>	2.7		19.3	1590.0
2011	Atlantic Cod	<i>Gadus morhua</i>	0.8		2.8	690.0
2011	Coral Reef	<i>Acropora digitifera</i>	0.4	0.4	10.7	190.0
2012	Gorilla	<i>Gorilla gorilla gorilla</i>	2.9		11.9	914.0
2012	Oyster	<i>Crassostrea gigas</i>	0.6	0.6	19.4	400.0
2013	Radish	<i>Raphanus sativus</i> L.	0.4	0.5	25.0	
2012	Wheat	<i>Triticum aestivum</i>	5.5	17.0	0.6	0.6
2013	Loblolly Pine	<i>Pinus taeda</i>	20.1	22.0	8.2	30.7



**What use is a reference genome
sequence to applied tree
breeding ?**

CAAPCAAGTCATCCATGATT
TCCGCATAGTAGCTCATA
TCATAGTCTTCAATGCA
APCAAGTCATCCATGATC
CATAGTAGCTCATA

Applied tree breeding

- **Primary goal:** Produce improved genetic material for deployment as planting stock, while maintaining sufficient genetic diversity to manage risk.
 - ✓ Understanding biological mechanisms is not a goal, but it can be a tool.
- **Primary tool:** Modeling the genetic basis of phenotypic variation in breeding populations.
 - ✓ Phenotypes measured in field tests of progeny from structured mating designs.
 - ✓ Genetic information primarily based on pedigree records
 - ✓ BLUP (best linear unbiased predictor) relies heavily on kinship information!

- For 20 years, tree breeders have been fascinated by the prospects of employing genetic markers to make tree improvement faster, less expensive and more efficient (MAS or marker assisted selection).
 - QTL: quantitative trait locus
 - Association genetics
 - Genomic selection
- Though markers have found many routine applications today (parental exclusion, clonal identity, PMX/WPA, etc), MAS per se has largely not proven tractable until recently.
- Recent advances in sequencing technology and analytical approaches spur current interests.
 - Use of markers in kinship matrices for BLUP
 - Modeling of genetic merit (genomic selection)

Where are we today with respect to a strategy for using genomic information in applied tree breeding?

- Approaches that rely on statistical associations between marker loci (allelic variants) and phenotypic traits, such as linkage based QTL or LD based Association Genetics, present obstacles to applied tree breeding.
- A good, reliable and relatively modest sized marker set could be used now to dramatically improve estimates of kinship matrices in traditional BLUP analyses.
- Whole-genome modeling of genetic merit has potential for MAS based largely on kinship (identity by descent).

Predictive modeling of genetic merit

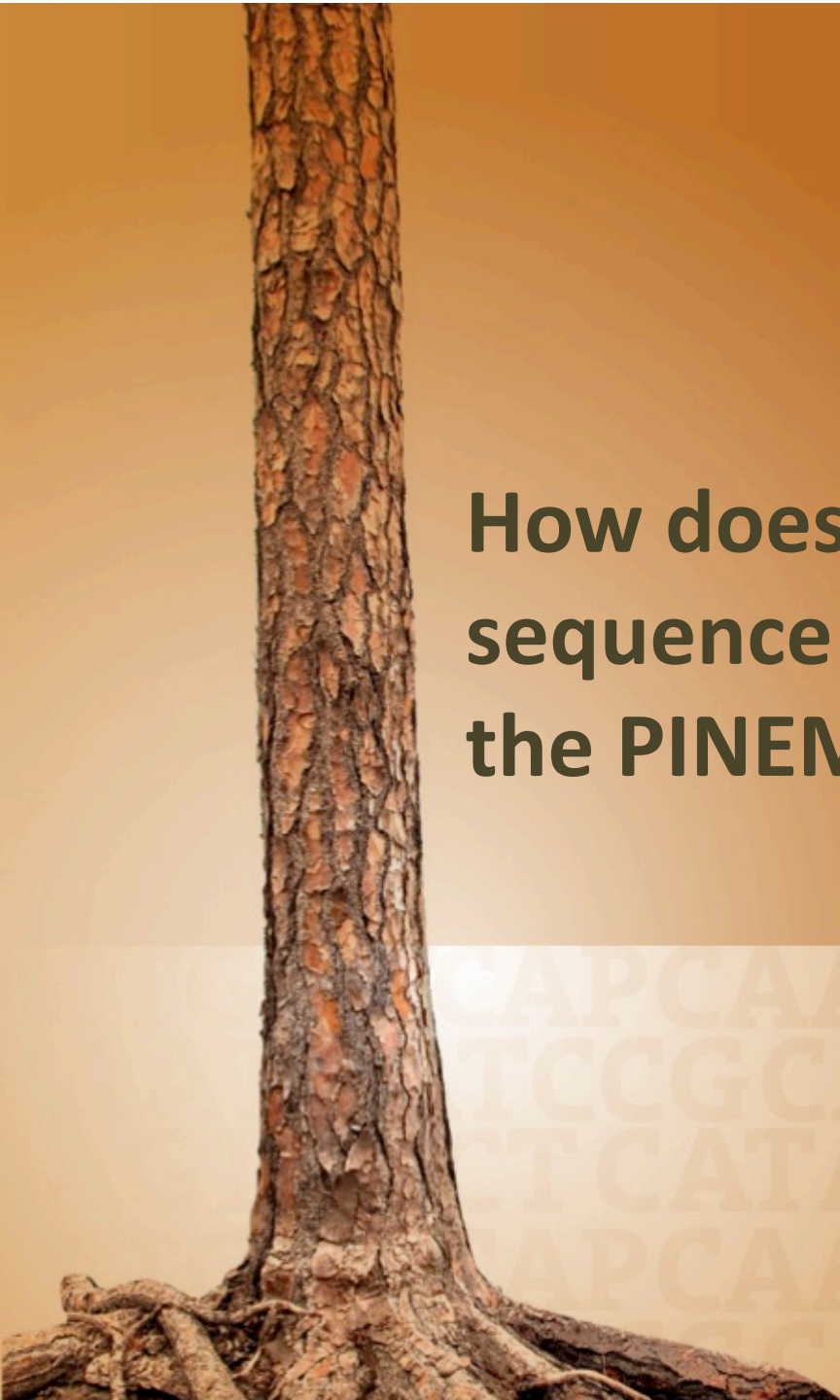
- A population derived from a limited number of parents mated in a structured design will be the initial test case
- Identity-by-descent methods using parental haplotypes and progeny phenotypes will allow modeling contributions of chromosome segments to phenotypic variation.
- No model selection or statistical significance testing required.
- Crosses among progeny will be used to test the predictive power of the models in the next generation
- **Simply stated, the greater the proportion of a progeny's genome that comes from parents proven to be superior, the better. Markers can identify parental chromosomal segments precisely.**

Identity by descent and imputation (filling in the blanks)



What use is a reference genome sequence . . . in applied tree breeding?

- As a reference for re-sequencing elite individuals to identify functional alleles or haplotypes, and consequently, to provide superior estimates of kinship.
- As a physical map of marker locations, to guide imputation of missing genotype data
 - Essential for matrix-based methods of analysis
 - Allows accurate imputation of progeny from structured mating design based on known parental haplotypes
- As the fundamental framework for knowledge of conifer genes and regulatory elements, to enable future advances in MAS strategies as technology develops.



**How does the reference genome
sequence project intersect with
the PINEMAP project?**

CAAPCAAGTCATCCATGATT
TCCGCATAGTAGCTCATA
TCATAGTCTTCAATGCA
APCAAGTCATCCATGATC
CATAGTAGCTCATA

PINEMAP Objectives

The PINEMAP project is comprised of extension, education, and research elements, the latter of which includes a set of genetics objectives.

The genetics team is investigating the genetic basis of pine productivity and adaptive traits by conducting linkage and association mapping to identify alleles that can be screened in populations, helping to accelerate improvement of productivity and adaptive traits.

The conifer reference sequence project will directly affect this effort by identifying all, or nearly all, of the genes in the loblolly pine genome, and inform scientists on the nature of gene regulation.



Landscape genomics attempts to explain which genetic and environmental factors play a role in how organisms adapt to their surroundings.





Approaches and their relationships

$$\text{Phenotype} = \text{Genotype} + \text{Environment}$$

Provenance or Common Garden Trials

Phenotype X **Environmental** Associations

Marker Assisted Tree Breeding

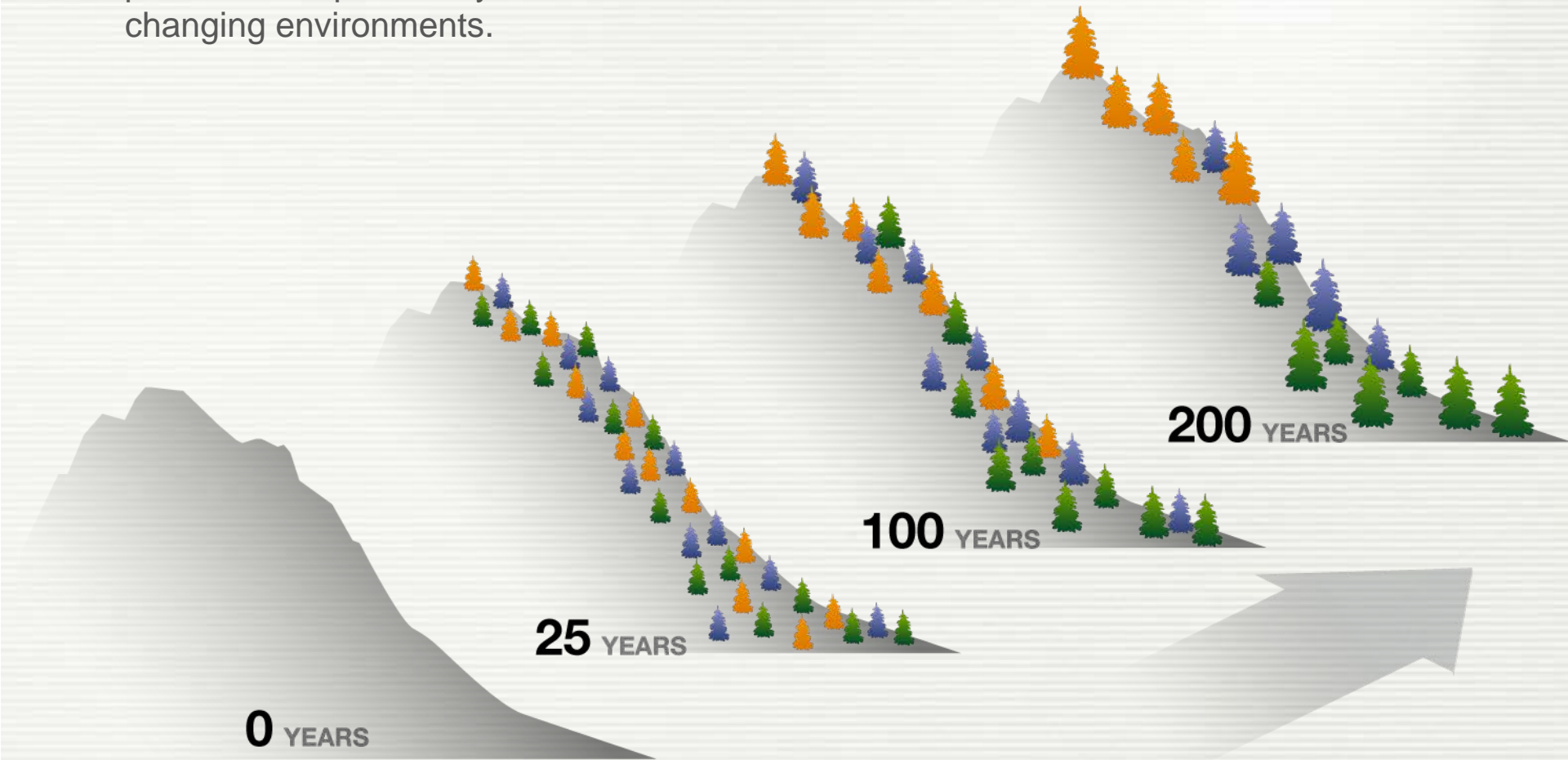
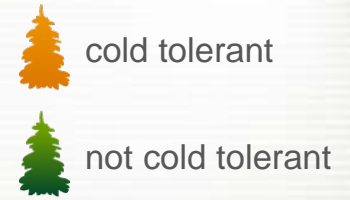
Genotype X **Phenotype** Associations

 **Landscape Genomics**

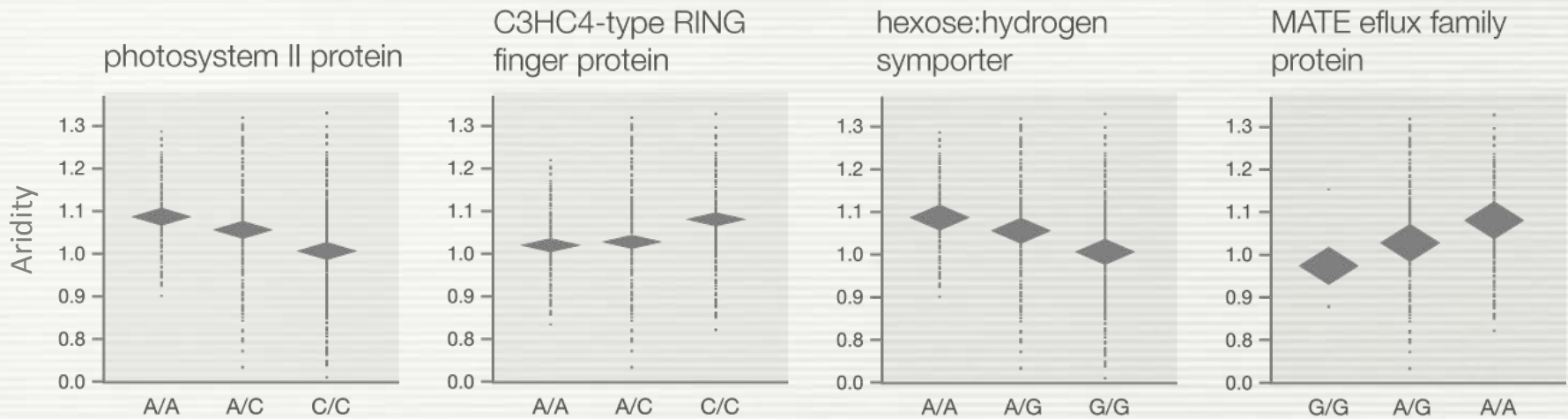
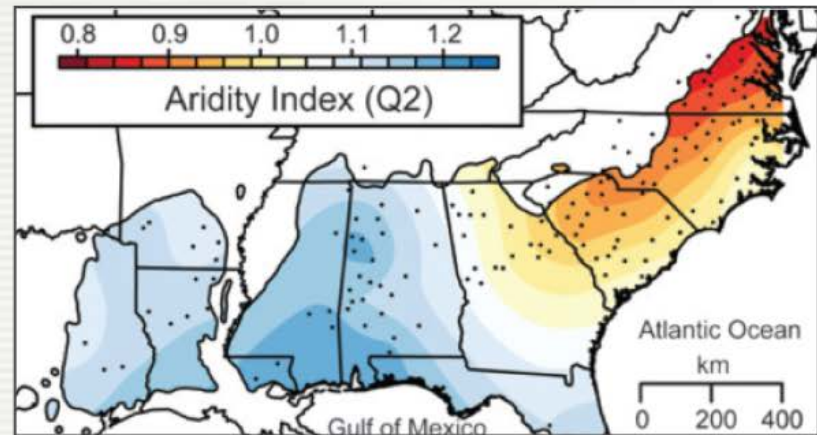
Genotype X **Environmental** Associations

What is the scientific basis?

An organism's genetic makeup determines its adaptive potential and probability of survival in diverse and changing environments.



Genotype by environment associations



Acknowledgements

PineRefSeq: David Neale, Charles Langley, Jill Wegrzyn and all the project investigators and associates at University of California, Davis; Children's Hospital Oakland Research Institute; Johns Hopkins University; University of Maryland; Indiana University; Texas A&M University; and Washington State University

PINEMAP: Tim Martin (Project Director, UF-Gainesville); Tom Byram (TFS and TAMU) and Ross Whetten (NC State) (co-leaders of PINEMAP Genetics team); Gary Peter (UF-Gainesville), Jason Holliday (VT), Kostya Krutovsky (TAMU), Dana Nelson (USFS-SIFG), Steve McKeand and Fikret Isik (NC State) (PINEMAP Genetics team investigators); and the staff and students of the three loblolly pine breeding cooperatives (NCSU CTIP, UF CFGRP, and TFS-TAMU WGFTIP)



United States
Department of
Agriculture

National Institute
of Food and
Agriculture



www.pinemap.org



www.pinegenome.org/pinerefseq

Tutorial modules on conifer genomics available online

<http://www.extension.org/pages/60370/conifer-translational-genomics-network-online-modules>

Module 16: Landscape Genomics

Module 17: Conifer Reference Sequencing – PineRefSeq