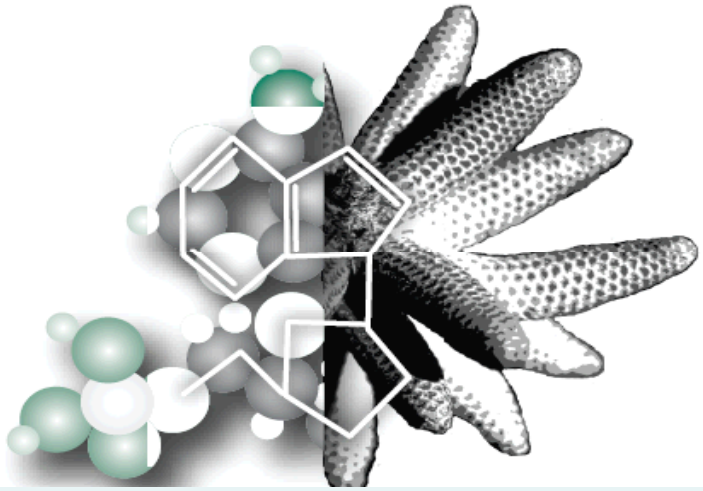
Pine Genome Initiative



Benefits of Sequencing the Pine Genome

Ron Sederoff

December 6, 2007

North Carolina State University

Perspective

In a genomic era, a plant is a machine with a finite number of known parts, its genes.

- With knowledge of functional genomics, we can understand **the molecular basis of variation and adaptation,** even though hundreds or thousands of genes may be involved.
- And that is very much what we need to know to understand development, metabolism, adaptation and evolution.

Why are pines interesting?

- There is a great deal of unique biology to be learned about gymnosperms.
- Gymnosperms are very old and successful, and one of the two groups of higher vascular plants.
- Most of our knowledge about plants is based on angiosperms.
- Pine is the best studied and most useful of the gymnosperms.

Utility of pines

- Most widely planted forest tree.
- Widely used for paper, pulp and wood products.
- Grows well on nonagricultural land.
- Substantial potential as biomass source for ethanol because of scale.
- Represents one of the nation's most abundant raw materials.

The biological problems we face are serious and complex

- Rarely single gene solutions.
- Often interacting organisms are involved. e.g. pest, pathogen and host interactions.
- Genomics makes it possible to understand the underlying biology of adaptation, and interactions of organisms in natural systems.
- Environmental effects can trigger devastating epidemics such as the mountain pine beetle crisis in the Pacific Northwest.

Mountain Pine Beetle Epidemic

- Presumably due to milder winters, the west coast of Canada has a devastating epidemic of mountain pine beetle, *Dendroctonus ponderosae killing* lodgepole pine (P. contorta).
- Trees die from the tunneling of the beetle larvae under the bark.
- An co-adapted blue stain fungus contributes to the lethality of the infestation and stains the wood.

Extent of Damage in B.C.

29 million acres affected.

- By 2007 40% of the merchantable pine wood volume is now in dead trees, representing 530 million m^3 of wood.
- 12% of all timber in British Columbia.
- Damage is projected to double to 80% of pine wood, 1 billion m^3 and 24% of all timber by 2015.

How could genomics help?

- With all genes identified, whole genome technologies can provide rigorous information on responses, and interactions of host, pest and pathogen.
- Power to detect mechanisms and to provide control solutions is optimized.
- The goal is to provide predictive models for control.

Pines for ethanol

- Pines grow well on marginal land.
- The processes for tree growth, storage, transportation, processing and purification of cellulose have been extensively developed (it is called papermaking).
- Pines can be grown on a scale commensurate with the demand, with low impact on agricultural land and water supply.
- Reduction of lignin content would reduce costs dramatically, and competitively.
- Genomic information and other modifications will be needed.

Predictive models

- The goal of systems approaches to fundamental and applied problems.
- For any organism, the first step is the full definition of its genome.
- Genome sequence is therefore, essential for a systems approach to biological problems.

Some burning questions

- What is the basis for the unique biology of gymnosperms and other trees?
- What is the basis for the large genome size?
- How different are the gymnosperms and angiosperms?
- How many genes are there in pine (and other gymnosperms)?
- How are the genes distributed?
- How do we begin to build predictive models of important processes?

Everybody is doing it !

- There is a genome sequencing objective for every major organism of major scientific or economic importance.

- Why? Because it creates information and technology that can't be obtained as readily any other way.

Can’t we get most of the information we need without sequencing?

We can get expressed genes, probably over half.

- We can get their approximate locations.
- And their alleleic diversity.
- It is very useful. Isn’t that enough?
- No.

What do we get from full sequencing?

- Regulatory sequences.
- Introns.
- Missing genes, rarely expressed genes.
- Pseudogenes
- Hidden duplications
- Full genome tools
- Full and precise locations
- Epigenetic tools
- Genome structure and evolution
- The beginning of a “systems” approach.

Pine is a MEGAGENOME (~20,000 Mb)

- It is 7X the human genome.
- It is 150X the Arabidopsis genome.
- It is 50X the poplar genome.
- Sequencing costs are coming down.
- May be quite difficult. A somewhat different strategy is needed.

Some problems

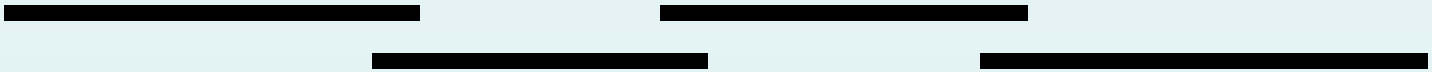
- New high throughput methods are designed for “resequencing”, not sequencing from scratch.
- The difference is an order of magnitude, or maybe more.
- Most of the genome is repeated. Large numbers of repeated sequences are expected. Pseudogene frequency may be high.
- There is almost no information of distribution of repeats and structure of heterochromatin.

One way to do it: A two step plan.

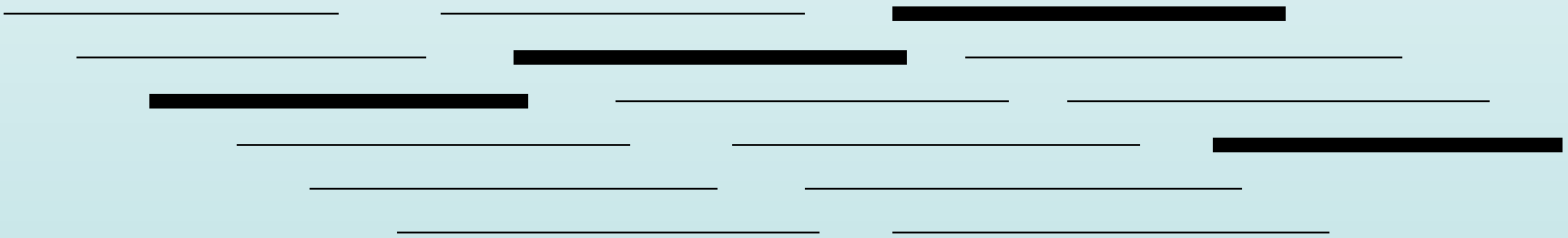
- 1. **Build the map.** Build the complete physical map. Some new tools are needed for high throughput genetic mapping and long range physical mapping.
- 2. **Then do a 1.3X sequence.**

The key is a minimum tiling path.

Minimum path (1.3X)



Shotgun sequencing (30X or more)



Some “new” techniques would help.

- Genome reduction methods.
  - Chromosomes or chromosome fragments
  - Selective amplification (e.g. AFLP methods)
  - Methyl filtration
  - Cot analysis
  - Repeated sequence sites and borders.
- High resolution mapping populations (Echt et al).
- High throughput genetic mapping (arrays e.g. DArT).
- Some cytogenetics, e.g. FISH mapping.

Building THE MAP

- Some deeper EST sequencing
- Put ESTs and other markers on the map.
- 20X (or more) BAC coverage, fingerprinting and end sequencing.
- Long range fragment or chromosome segment mapping.
- Building the physical map by integrating the genetic and physical maps.
- Determining the minimum tiling path.

What about without a map?

- Expect at least 20X more sequence needed.
- Currently limited by the number of machines in existence.
- And it might not work because of the difficulty of assembling such a large genome with all the repeats.

What will it cost? How long would it take?

- About $22 million has already been funded, or committed to pine genomics projects (1999-2007).
- **The MAP.** About the same level of effort could create an integrated physical map.
- **The SEQUENCE.** The cost, time and effort of sequencing pine today would then be in line with many other plant genome projects being done today.

What will happen with cost and throughput of sequencing in the near future?

- "It's like the wild west out there"
  - Jeffrey Krummel (Solexa Rep).

  Costs and throughput will likely change by a factor of two in the next year.

  Dozens of companies developing new platforms.

  Sequencing itself may soon be less expensive than making the map and the assembly.

  The first pine sequence then provides a framework for many more.