A Proposal for New Pine Resources for Mapping and Sequence Assembly

C Echt1, R Whetten2, CD Nelson1, DG Peterson3, K Krutovsky4, C Yuceer5, JFD Dean6
1 Southern Institute of Forest Genetics, Southern Research Station, USDA Forest Service, Saucier, Mississippi, 39574, USA e-mail: cecht@sfs.fed.us
2 Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, North Carolina, USA e-mail: ross_whetten@ncsu.edu
3 Department of Plant & Soil Sciences, Mississippi State University, Mississippi State, USA e-mail: doterson@ppss.msstate.edu
4 Department of Ecosystem Science & Management, Texas A&M University, College Station, Texas, USA e-mail: k-krutovsky@tamu.edu
5 Department of Forestry, Mississippi State University, Mississippi State, USA e-mail: mcyli@ra.msstate.edu
6 Daniel B. Warnell School of Forestry and Natural Resources, University of Georgia, Athens, GA 30602, USA e-mail: jfdean@uga.edu

Haploid mapping population for genome assembly of loblolly pine, clone 7-56

*Pinus taeda* genotype 7-56:
- Used extensively in loblolly pine breeding and studies.
- 4x BAC library available (www.mgel.msstate.edu/dna_libs.htm)
- Full genome sequence is feasible and desirable
- High resolution genetic map needed for assembling physical and sequence maps.

The proposed scheme:
1) Use MDA\(^a\) to obtain >1ug DNA from each of 10,000 haploid gametic products\(^b\) of clone 7-56.
2) Genotype sufficient reference SNPs & SSRs to obtain a framework map\(^c\) of avg. 1 cM density.
3) Bin map STSs needed for assembly via selective mapping (Vision et al. 2000). Use only 384 samples.
4) Then fine map those STSs within a bin via “intra-bin mapping”. Genotype only the subpopulation of ~100 samples\(^d\) per bin that have a phase known crossover configuration for the markers that flank the bin.

- Bin map and intra-bin map populations curated by US Forest Service and distributed to genome assembly labs and others upon request.
- Web distribution of reference genotype and recombination data.

\( ^a \) MDA: multiple strand displacement amplification method of whole genome amplification using phi29 polymerase. Low-level MDA and pooling multiple aliquots per sample would give sample DNA with minimal MDA artifacts.
\( ^b \) Female haploid DNA is obtained from megagametophyte tissue of the pine seed. Seed would be obtained from open pollinations of numerous 7-56 ramets.
\( ^c \) Framework maps have a ±LOD 3 support for order. An estimated genome size of 1500M would require genotyping ~3000 SNPs + SSRs to obtain an average 1cM density. Framework ordering may be obtainable with fewer than 10,000 samples.
\( ^d \) 100 if 10,000 samples are genotyped, less if framework support is achieved with fewer than 10,000 samples.

Why new mapping populations?
The current loblolly pine mapping population is not suitable for future needs. It has limited allelic variation, limited cM resolution, variously coded sample names, privately owned genotypes and phenotype data, and is non-renewable.

We need large mapping populations that can:
- endure
- be publicly owned, maintained and distributed
- assist assembly of the pine genome sequence and BAC-based physical map
- enable studies of the interaction between sequence variation and phenotype variation.

Multi-parent population for mapping SNPs and phenotype variation in loblolly pine

The proposed scheme:
- 9 parents from distinct geographical regions and subpopulations
- 24 full-sib families each have 47 progeny
- 8 parents each have 5 families and 235 progeny
- clone 7-56 parent has 8 families and 376 progeny
- 1128 progeny = one “installation”.

Example: 7-56 topcross + partial diallel design

<table>
<thead>
<tr>
<th>Parent ID</th>
<th>Parent Provenance, State</th>
<th>Parent</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-56</td>
<td>coastal, SC</td>
<td>A</td>
<td>47</td>
<td>47</td>
<td>47</td>
<td>47</td>
<td>47</td>
<td>47</td>
<td>47</td>
<td>47</td>
</tr>
<tr>
<td>A</td>
<td>coastal, SC</td>
<td>B</td>
<td>47</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>piedmont, GA</td>
<td>C</td>
<td>47</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>north, MS</td>
<td>D</td>
<td>47</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>n central, FL</td>
<td>E</td>
<td>47</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>south, AL</td>
<td>F</td>
<td>47</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>south east, LA</td>
<td>G</td>
<td>47</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>east, TX</td>
<td>H</td>
<td>47</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>central, AR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parent colors denote range-wide population subdivisions based on SSR and isozyme data. Other diallel designs are possible and merit discussion. 1128 DNA samples + 24 controls will fit on three 384-well plates.

- Parents clonally maintained by US Forest Service
- Three installations at diverse sites, planted, managed, measured, and genotyped by US Forest Service.
- Additional installations annually archived as seed and distributed to researchers upon request.
- Web distribution of phenotype and genotype data.
- Genetic diversity of multiple parents allows mapping of most common SNPs and traits.
- Multiple families of ample size allow partitioning of genetic variance into additive and non-additive components.
- Scalable cM resolution and power of QTL detection.