PINEMAP

Comparing Genotyping Technologies for Efficiency and Cost-effectiveness

Ross Whetten, Department of Forestry and Environmental Resources, North Carolina State University
Konstantin Krutovsky, Department of Ecosystem Science and Management, Texas A&M University
Jason Holliday, Department of Forest Resources and Environmental Conservation, Virginia Tech

Executive Summary

Several methods for genotyping using high-throughput DNA sequencing have been described using model organisms or annual crops. Application of these techniques to conifers, such as loblolly pine, is not trivial, because the pine genome is several times larger and contains more repetitive DNA than most of the species in which these techniques were developed. As part of the PINEMAP research objectives on genetic analysis of pine populations, we have compared several techniques to evaluate the relative strengths, weaknesses, costs, and benefits of each. No single technique is optimal for every purpose, so the objectives of individual experiments are important in determining which method is best suited to achieve the desired results.

Background

One of the objectives of the Genetics team in the initial phase of the PINEMAP project is to compare and contrast different methods of obtaining dense datasets of molecular marker genotypes for hundreds to thousands of individual trees. These datasets will then be analyzed, together with phenotypic data and breeding records from the cooperative breeding programs, to test the hypothesis that genetic variation for resilience to climate variation exists in southern pine breeding programs. Previous results suggest that many trees in the breeding program show good adaptability across a range of site conditions, but confirming and extending those results will be important to allow pine breeding programs in the southeastern U.S. to structure their efforts to mitigate the risk that climate change will have dramatic impacts on pine plantation productivity.

Methods

Two general methods for detecting genetic variation are being compared, and some alternative strategies are being explored for one of the methods. One method, hybrid-capture sequencing, uses synthetic “bait” sequences to capture fragments of genomic DNA that correspond to genes of interest for sequencing. This approach requires a significant investment to synthesize the bait molecules, and therefore has a higher cost per individual sample analyzed, but is expected to yield more information about genetic variation in expressed genes. The other method, restriction-enzyme-based, utilizes the tendency of expressed genes to have lower levels of DNA methylation than repetitive elements or non-expressed DNA sequences. This tendency can be used to enrich for sequences in or near expressed genes, without requiring custom synthesis of gene-specific sequences. This method has a much lower cost per individual at present, but may yield information that is less narrowly focused on the coding sequences of expressed genes. Pilot experiments have been conducted for both methods; hybrid-capture at Texas A & M University (under the supervision of Kostya Krutovsky), and restriction-enzyme-based methods at North Carolina State University (Ross Whetten) and Virginia Tech (Jason Holliday).

Results

The hybrid-capture experiment compared DNA sequences from the haploid megagametophyte and the diploid embryo from a single seed, using a pool of almost 650,000 custom-synthesized “bait” sequences to capture DNA fragments of interest, and resulted in detection of over 48,000 candidate single-nucleotide polymorphisms, or SNPs. The “bait” sequences were designed based on DNA sequences of about 35,500 putative expressed genes identified by previous projects, but the detected genetic variants are not uniformly distributed among all the putative genes that were targeted. The frequency of candidate genetic variants discovered is about 1 per 1200 bp of DNA sequence, which is consistent with previous reports of genetic variation in loblolly pine.

Photo by Gregory Powell

The NC State experiment used DNA samples from two diploid parents and 90 haploid DNAs from seeds of one of the two parents, to allow testing for genetic segregation expected of single-copy sequences. The pine genome contains an abundance of repetitive DNAs, and distinguishing allelic variation at a single genetic locus from sequence variation among multiple copies of a repeated sequence is an important challenge for any genotyping technology. Two methylation-sensitive restriction enzymes were used to fragment these 92 DNA samples, and the resulting sequences were filtered to identify fragments detected in at least 25 samples, to reduce the presence of sequencing errors in the filtered dataset. 28.4% (29,415 of 103,669) of the filtered sequence tags show a pattern of presence and absence in haploid samples that is consistent with the expected 1:1 segregation ratio for a single-copy genetic locus.

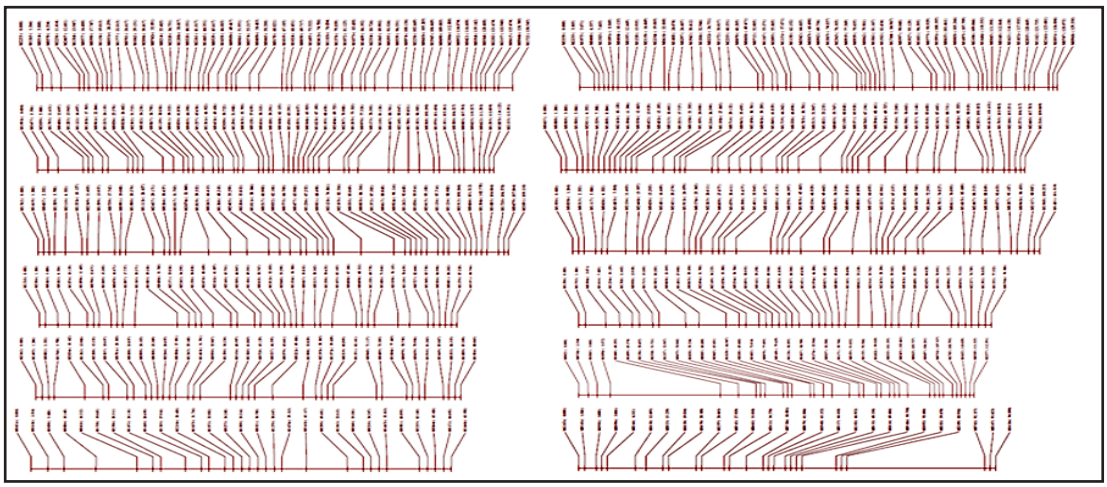

Figure 1. This image shows a preliminary linkage map of markers, with 12 groups that presumably correspond to the 12 pine chromosomes.

The Virginia Tech experiment used a single restriction enzyme, in conjunction with random shearing of DNA, to test a variation of the restriction-enzyme-based approach on samples from a diploid parent and 7 haploid offspring. Analysis of data from that experiment show that the pine DNA sequences are contaminated with high levels of sequence from synthetic adapters used in preparing samples for sequencing. Consequently, the decision was made to drop the shearing-based protocol from further consideration, and focus attention on the first two alternative procedures. Data analysis is continuing for these preliminary datasets, and additional experiments are also planned to extend these preliminary results and address new questions.

Implications

One key question for application of these genotyping technologies will be whether a single technology is suited for all the experiments envisioned as part of the PINEMAP genetics component, or if different technologies will be better for different experiments. The cost difference between the hybrid-capture and the restriction-enzyme-based methods is significant at this point, although further optimization of the methods used may reduce that difference in the next year. A second key question is how best to utilize the reference genome sequence for loblolly pine in future analyses. Preliminary experiments have tested the utility of comparing the restriction-enzyme-based sequences to the draft (version 0.6) assembly with some success, and as more refined versions of the reference genome sequence are released, additional value will be gained from those comparisons.

In summary, preliminary results comparing genotyping technologies show that cost-effective high-throughput genotyping of pine breeding populations is feasible at a cost of about $30 per individual tree for the restriction-enzyme-based methods. Hybrid-capture sequencing may provide more information about coding sequences of expressed genes, which could be an important asset for detecting genetic variation that underlies phenotypic variation in adaptive traits in loblolly pine. The PINEMAP project includes several different genetics experiments based on different populations, and both the restriction-enzyme-based and hybrid-capture sequencing methods are likely to find application in achieving the project objectives.

For additional information on PINEMAP genotyping research, contact Ross Whetten (ross_whetten@ncsu.edu).

USDA United States Department of Agriculture National Institute of Food and Agriculture

The Pine Integrated Network: Education, Mitigation, and Adaptation project (PINEMAP) is a Coordinated Agricultural Project funded by the USDA National Institute of Food and Agriculture, Award #2011-68002-30185.