11. PINEMAP + PineRefSeq = Future Forests

Konstantin Krutovsky[1,7] • *Tom Byram* [2,8] • *Ross Whetten*[3,9] • *Nick Wheeler*[4,10]
David Neale[1,11] • *Mengmeng Lu*[5,12] • *Tomasz Koralewski*[6,12] • *Carol Loopstra*[3,12]

[1]*Professor* • [2]*Assistant Professor* • [3]*Associate Professor* • [4]*Affiliate Faculty* • [5]*Ph.D. student* • [6]*Postdoctoral Research Associate* • [7]*Department of Forest Genetics and Forest Tree Breeding, George-August-University of Göttingen; Department of Ecosystem Science and Management, Texas A&M University*• [8]*Department of Ecosystem Science and Management, Texas A&M University; Texas A&M Forest Service* • [9]*Department of Forestry and Environmental Resources, North Carolina State University* • [10]*College of Forestry, Oregon State University* • [11]*Department of Plant Sciences, University of California at Davis* • [12]*Department of Ecosystem Science and Management, Texas A&M University*

The equation P = G + E + GxE describes the interactive effects of genotype (G) and environment (E) on an observed phenotype (P) and is one of the most powerful relationships in all of modern biology. Two USDA National Institute of Food and Agriculture funded projects are now working to elucidate the components of this model in unprecedented detail. The Pine Reference Sequences project (PineRefSeq) is attempting to develop the first complete genome sequence for loblolly pine (G) while PINEMAP investigates how the environment (E) interacts with individual trees and stands (GxE) to form the forest of the future (P).

The goal of the PineRefSeq project is to provide a complete genome sequence for an individual loblolly pine, which will enable future researchers to discover the nucleotide diversity in genes, promoter regions, and transcription factors. The ultimate aim is to provide an annotated list of genes describing function; regulation; and the place they occupy in biochemical pathways critical to growth, fitness, and adaptability. PINEMAP approaches the problem through the other components of the equation by examining how the environment influences phenotypes. Genetic variation is the critical point at which the two projects overlap, and the source of considerable project synergism.

Sequencing the loblolly pine (*Pinus taeda* L.) genome is far from trivial. At seven times the size of the human genome, it is one of the largest sequencing projects ever to be attempted (Figure 11.1). To further complicate matters, the pine genome is literally awash in repetitive DNA: gene families regulated in tissue specific ways, nonfunctional pseudogenes,

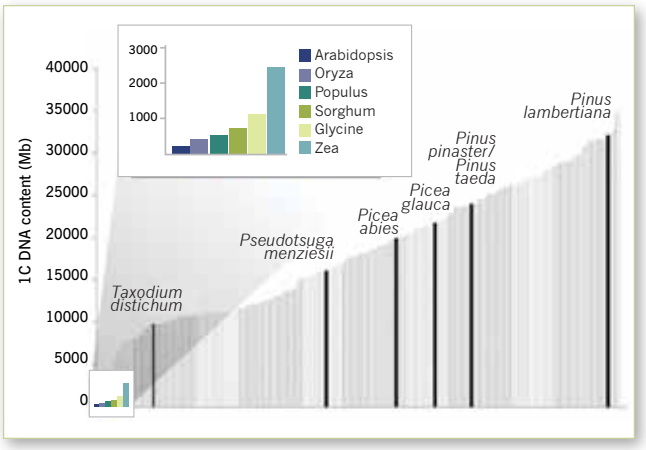

Figure 11.1. Relative genome size of several conifers compared to several species previously sequenced. Image credit: Modified from Daniel Peterson, Mississippi State University.

highly repetitive structural sequences, and possibly millions of relics of genetic reorganization in the form of mobile genetic elements, such as transposons and retrotransposons (DNA sequences that can change their position within the genome). Despite these difficulties, the project has already had considerable success and recently released a draft sequence to the research community (see http://www.nifa.usda.gov/newsroom/news/2013news/01111_loblolly_genome.html).

Meanwhile, PINEMAP is collecting measurements on phenotypes, environments, and genotypes with the goal of predicting the performance of future forests. PINEMAP researchers already are using the PineRefSeq sequence to

The way these two projects work together will be an iterative process, as PineRefSeq generates new sequences and proposes new putative functions and PINEMAP measures additional traits in novel environments.

verify and design platforms to acquire the most meaningful genotypes of the trees they are studying. Initially, this takes the form of (1) verifying genetic variation seen in existing databases from other projects, (2) designing platforms that will efficiently assay this variation, and (3) ensuring that all relevant parts of the genome are represented. The eventual goal is to identify functional differences, either in the structure of alternative forms (alleles) of important genes or in their regulation. The extensive database of interacting phenotypes and genotypes generated by the PINEMAP project will inform and validate the annotation process, including the assignment of function of the gene sequences developed by the PineRefSeq project.

Both the PineRefSeq and PINEMAP projects use the most modern genomic methods and similar next-generation sequencing (NGS) techniques. As a result, much of the sequence data generated by both projects will serve double duty. As an example, the first draft genome assembly (version 0.8 scaffolds generated by PineRefSeq) greatly helped the Texas A&M University genetics team map and cluster DNA sequence fragments consisting of several billion short nucleotide sequences generated in the PINEMAP project. These sequences were obtained from genomic DNA enriched for coding regions after hybridization with 647,634 oligonucleotide probes designed from 35,550 unigenes. By mapping the sequences to the draft genome assembly, from about 42,000 to 120,000 single nucleotide polymorphisms (SNPs) per sample were identified. These SNPs will be used further as highly informative genetic markers to study the association of genetic variation with the variation of adaptive traits and environmental variables. The PINEMAP project greatly benefits from the publicly shared Dendrome and TreeGenes databases that provide pine tree genomics data and bioinformatics tools to the forest genomics community currently maintained by the University of California Davis group of the PineRefSeq team.

In turn, data generated by the PINEMAP project can help improve and verify the PineRefSeq genome. Some of the SNPs or unigenes generated by PINEMAP have known locations on a high density linkage map and can be used as anchors for directing genome assembly in PineRefSeq project. Preliminary analysis indicates that some of the additional sequence fragments obtained by the Texas A&M University genetic team may span gaps between disjoint contigs (contiguous consensus DNA sequences assembled from shorter overlapping DNA segments) and scaffolds (ordered contigs separated by gaps where the exact DNA sequence is unknown) apparent in the early version 0.8 assembly. This will make an additional tool available to the PineRefSeq team to connect loose ends and improve the finished assembly. It may also be possible through a joint analysis of PineRefSeq and PINEMAP sequence data to infer important information on gene structure (e.g., exon and intron length and their junctions), something that neither project could easily do alone.

The way these two projects work together will be an iterative process, as PineRefSeq generates new sequences and proposes new putative functions and PINEMAP measures additional traits in novel environments. The outcome of this synergistic effort will be a better understood, more resilient, and more productive future forest.