THE PINE REFERENCE GENOME SEQUENCE AND APPLIED TREE BREEDING

N.C. Wheeler[1] and R. Whetten

[1] Molecular Tree Breeding Services, LLC, Centralia, WA

Introduction

The PineRefSeq project was funded to produce a high-quality reference genome sequence for loblolly pine and related species. An early draft (v 0.6) of the loblolly pine genome was released at the project website in June, 2012. This draft and two subsequent improvements have been widely accessed by the scientific community. A complete reference sequence (v 1.0) should be released very shortly. Recently, draft genome sequences for white and Norway spruces were published. Long considered a task so daunting it might never be achieved, the development of next generation sequencing (NGS) technologies has opened the door to deciphering even these leviathan conifer genomes. In our talk we will discuss the continuous evolution of the state-of-the-art in genome sequencing and the adaptive approaches of a team of researchers funded by the USDA NIFA AFRI program ($14.625 million dollars over five years). It is anticipated that the reference sequences produced by the PineRefSeq project and others will dramatically change the landscape of forest genetics and pave the way for greater application of genomic resources in applied tree breeding.

The reference genome sequence

A reference genome sequence is that which results from *de novo* sequencing and assembly of a haploid complement of an organism's genome. It is the initial sequence to which all subsequent sequences are ultimately compared and therefore it must be as complete as possible given fiscal and technical constraints. The challenges to assembling the billions of bases of a conifer genome in the proper order are monumental. The reference genome leads to the identification of all or most of the genes in an organism, and reveals features of genome structure such as the amount and order of repetitive elements, the nature of regulatory elements, and so forth. Re-sequencing, or sequencing of other individuals of the same species, is vastly less time consuming and costly once a reference genome exists. Re-sequencing reveals the amount and distribution of genetic variation (mutations) within a genome on an individual or population basis.

While whole genome sequencing has become rather commonplace, and is recognized as the gold standard of genetic resource development in biological science, its history is really quite short. This is largely a function of the remarkable advancements in sequencing technology that have occurred over the last 15 or so years. The first major genome to be sequenced was the human genome. Though plans for a "Human Genome Project", or HGP, were taking shape throughout the late 1980s, the project itself did not kick off until 1990 when Congress allocated funds. Work on the publicly funded HGP was ultimately carried out by labs in 18 countries, but the bulk of the work was conducted in the USA, initially under the guidance of James Watson, and later by that of Francis Collins. A draft genome sequence was completed in 2000 and the project concluded in 2003, two years ahead of schedule, under budget, and with accomplishments far exceeding goals. The cost of sequencing declined over the course of the project from roughly

$10 per base in 1990 to around $0.09 per base at its conclusion. Today, one can obtain nearly 100,000 bases of sequence per penny. The completion of the HGP draft sequence was announced simultaneously with that of a privately funded human genome sequence project that was initiated only a few years before the announcement by Craig Venter at Celera Genomics. Since then, over 1000 individual human genome sequences have been completed and are publicly available (The Thousand Genomes Project Commission). As the quote above implies, the HGP opened the floodgates to genome sequencing of all manner of organisms.

Challenges

There are a number of challenges to obtaining a reference genome sequence for a conifer, not the least of which is the size of their genomes. Genome size varies considerably among conifer species, ranging from 6 billion to over 30 billion base pairs (Figure 1). Before next-generation sequences became available, it was estimated it would take nearly 30 years to sequence a conifer. Today, technically, it can be done in a matter of months. But size is not the only hurdle. Conifer genomes generally possess large gene families (duplicated and divergent copies of a gene), and abundant pseudo-genes. Beyond the duplicated gene (and pseudo-gene) content, it appears that the vast majority of the remaining conifer genome is composed of moderately or highly repetitive DNA of unknown or poorly understood function. Correct assembly of the genomic puzzle with so many identical or nearly identical pieces requires a more thorough and sophisticated sequencing and assembly effort than anyone has ever attempted.

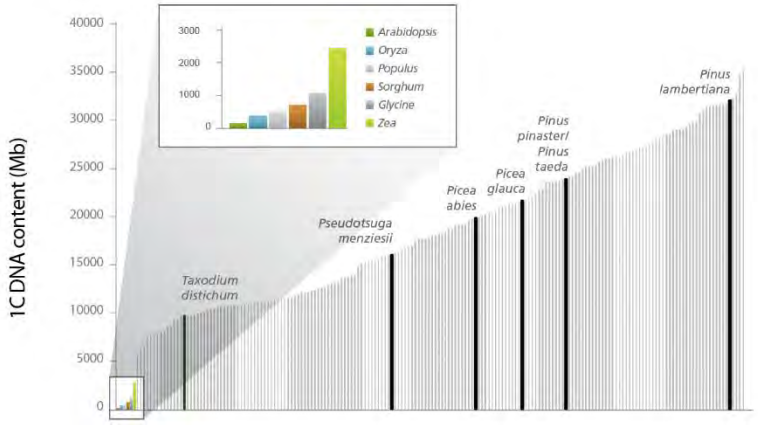

Figure 1 Range of genome sizes, in Mb (millions of bases), for an array of conifers and model plants. The image is modified from Daniel Peterson, Mississippi State University

Finally, we should note that most conifer species, and individual trees, retain an enormous reservoir of genetic diversity. Studies have shown that single nucleotide polymorphisms occur once every 50 to 100 bases throughout sampled areas of the conifer genome. Consequently, trying to sequence a diploid individual with all that variation can confound the sequence assembly function. Clearly the task of building a reference sequence for a conifer, or any other large genome organism for that matter, is a significant challenge. As scientists evaluate methods to overcome these challenges, an adaptive approach is being used. In the next slide we will describe a number of techniques that are being evaluated in our attempt to create reference genomes for three conifer species: loblolly pine, sugar pine and Douglas-fir.

Strategy for building a reference genome sequence

There is no single recipe or established strategy for sequencing large and complex genomes. Approaches for doing so are continually evolving and improving, and different organisms may require different approaches. In the project we will describe, an adaptive approach that embraces current and developing best sequencing technologies and assembly strategies will be used, carefully testing methods and techniques to ensure optimal efficiency is eventually approximated. The path chosen will be guided by approaches that will simplify assembly of the genome. These approaches can be generally described as 1) the use of complementary sequencing strategies designed to simplify the process through use of actual or functionally haploid genomes and 2) conducting assembly in iterative steps, beginning with reduced size of individual assemblies, and leading to a meta-assembly.

The process begins with a deliberate selection of an individual tree for which the genome will be sequenced and proceeds through sequencing, assembly, and annotation. These processes are facilitated by the use of large-insert and jumping libraries, genetic mapping, and transcriptome sequencing. The entire process is reliant on database creation, management, and access. Though we discuss these steps as if they were a linear flow of activities, in reality, all activities are conducted more or less simultaneously and iteratively.

Proportionally speaking, most sequence will be obtained from the whole genome DNA obtained from the haploid megagametophytic tissue of a single seed. Enough sequence will be generated from this source to represent the presumed genome size 40 to 60 fold, or 40X to 60X. As an example, consider the loblolly pine genome. It is approximately 24 Gb in size, or 24 billion base pairs, arrayed among 12 chromosomes. A 40X coverage means that 960 billion base pairs of sequence would be generated. The second major source of sequence comes from the creation and sequencing of pooled **fosmid** clone libraries, to a depth of about 5X. A fosmid clone is a unique bacteriophage lambda particle that contains a large piece of DNA from the target tree genome inserted into its own, circular DNA. The large, single-stranded inserts can be selected for size, and in this case, are typically around 37,000 to 40,000 (Kb) bases in size. Pools of fosmid clones are combined, each pool containing between 1000 and 4000 clones. A pool of 4000 fosmids would therefore contain about 160 Mb of sequence, or about 7/10ths of one percent of the total genome. Since few if any of the clones are likely to have the same fragment of target DNA as any other clone in the pool, the total sequence from that pool is effectively haploid in nature, even though it was derived from diploid tissue (needles) to begin with. 150 such pools would represent about 1X coverage of the genome. Finally, a series of jumping or joining libraries will be created from both the fosmid and whole genome DNA sources. The purpose of the jumping library is to connect or pull together sequence **contigs** into **scaffolds**. The jumping library sequences represent a very important element in assembling the reference sequence. These diverse libraries will collectively be sequenced to a depth of about 5X to 10X. The next few slides will look at each of these sources in greater detail.

The transcriptome

A complementary element of building a reference genome sequence is the characterization of the organism's **transcriptome**. The transcriptome is the entire set of RNA transcripts in the cell, tissue, or organ from which the RNA was collected. It is the product of all the genes that are

being expressed at that time and place. Since the transcriptome is tissue and time specific, any given attempt to sample it will surely under-represent the total complement of functional genes in the genome. It is simply a snapshot in time of what genes are being expressed, and how they are being expressed. Consequently, it is desirable to sample transcripts (mRNA) from many tissues, collected under different environmental conditions, and at different times in the development of the plant if a "complete" characterization of the plant's transcriptome is desired. In the conifer reference genome project, two dozen or more RNA/cDNA libraries have been used to characterize the transcriptome of each selected genotype or individual. In many cases, libraries are collected from plants that have been subjected to experimental treatments or stresses. cDNA, or complementary DNA, is the product of reverse transcription of mRNA. The cDNA's have only DNA sequence that is used to code for a gene product (does not contain intron sequences for instance). Collectively, cDNA libraries produce an array of EST's or Expressed Sequence Tags.

The transcriptome is substantially more complex than the genome. While the DNA content of an individual is virtually constant throughout every cell, the transcript of a gene may vary considerably. Transcripts may be modified, alternately spliced, edited, and degraded on the way to being translated into protein. The transcriptome can help us understand how cells differentiate and respond to changes in their environment.

While the most direct way to identify a gene is to document the transcription of a fragment of the genome, such as is done with the sequencing of ESTs, protein coding sequences may also be identified by a process known as ab initio gene discovery using software that recognizes features common to protein coding transcripts. This is done by analyzing the genome sequence directly. Generally, both approaches are used, though for the latter all putative genes must be confirmed by a second line of evidence before they may be elevated to gene status.

Use of the reference genome sequence in applied tree breeding

The primary goal of applied tree breeding is to produce genetically improved planting stock while maintain sufficient genetic diversity to manage risk. Tree breeders seldom know which genes they are selecting for or the biological mechanisms they are influencing. Genomic resources offer the promise of providing such insights. Today, the primary tools of the tree breeder is the genetic test. Phenotypic data from tests are, by-in-large, analyzed using BLUP, or best linear unbiased predictor methods, which rely heavily on kinship relationships between trees in the breeding program. While breeders have long been fascinated by the prospects of employing genomic resources to make tree improvement faster, less expensive and more efficient, attempts to do so have generally come up short, until now.

Development of the pine reference genome sequence will provide many genomic resources, the most important of which in the near future being genetic markers. A virtually unlimited number of markers will be readily available and could find immediate use in two applications: 1) the improvement of kinship matrices in BLUP analyses, and 2) the modeling of genetic merit using genomic selection.

In the long run, the identification of virtually all of the genes in the genome, an understanding of their function, and how those genes are regulated, should enable future advances in marker assisted selection strategies that could transcend and complement current breeding practices.

Resources

Wheeler, N., and Neale, D.B. Reference Genome Sequencing [Online Learning Module]. Pine Reference Sequence. eXtension Foundation. Available at: http://www.extension.org/pages/67931 (verified April 29, 2013)

One of 17 learning modules with a focus on genomics in tree breeding, this module expands on material presented in this talk/abstract with a voice-over Camtasia video presentation. http://www.extension.org/pages/60370/conifer-genomics-learning-modules