Notes for Voice over on Sequencing Module

Slide 1

The purpose of this presentation is to describe an adaptive approach to the sequencing of very large conifer genomes. 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 genomes. The work described in these slides reflects the continuous evolution of the state-of-the-art in genome sequencing and the specific 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 draft reference sequences produced by the PineRefSeq project, formerly known as the Loblolly Pine Genome Project, will dramatically change the landscape of forest genetics.

Slide 2

The sequencing, assembly, and annotation of such large genomes requires the collaboration of many scientists and technicians with varied and complementary skills. For this groundbreaking project, faculty and staff from seven Universities and Institutes are working together.

Slide 3

As a publicly funded project, PineRefSeq is guided by a set of simple principles that will insure the genomics community benefits from the Project's output in an unrestricted and timely manner.

Slide 4

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 and will be discussed at some length during this module. The reference genome leads to the identification of all or most of the genes in an organism, and reveals features of the 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.

Slide 5

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.

Slide 6

Since the year 2000, following the announcement of the draft completion of the human genome, the pace of whole genome sequence acquisition for both plants and animals has increased rapidly. The first plant to be sequenced, *Arabidopsis thaliana,* paved the way for the sequencing of a broad spectrum of plants from across the phylogenetic landscape. The thoughtful manner in which species were selected for sequencing at labs such as the Joint Genome Institute (JGI) has provided scientists with the opportunity to conduct meaningful comparative genome analyses across the plant kingdom. As of early 2012, reference genome sequences of well over 30 plant species had been published, and dozens more are nearly completed, but not published, or in some stage of acquisition. Information on the status of species for which sequences have been completed or are in the process of being sequenced may be found at Comparative Genomics (CoGe) and JGI (Phytozome) websites, among others.

Slide 7

The first tree genome to be sequenced was that of the black cottonwood (*Populus trichocarpa*) from the Pacific Northwest USA. First published in 2006, a second version of the genome assembly and annotation has now been posted. Draft sequences of a number of other angiosperm trees have been reported, though not all have been published or are yet publicly available. This list is sure to grow rapidly.

Slide 8

Reference genome sequences for representatives of the gymnosperms, which consist principally of coniferous trees, are pending for at least 8 species, as noted here. Gymnosperms, which arose more than 250 million years ago, have substantially larger genomes than do the younger angiosperm tree species, which began to appear perhaps 130 to 150 million years ago. Most conifer genomes are 6 to 10 times larger than the human genome, for instance, and contain a great deal of repetitive DNA and large gene families. Attempts to sequence these very large genomes have had to wait for the development of next-generation sequencing technology, advanced computing capacity, and improved assembly algorithms. The first draft of a conifer genome, loblolly pine, was announced and posted at the Project's website in June of 2012, by PineRefSeq, a USDA NIFA funded Project, and sponsor of this learning module. The Project website is noted in slide two of this presentation.

Slide 9

As noted earlier, the dramatic increase in sequence acquisition was facilitated in large part by technological advances on many fronts. As next generation sequencing (NGS) technologies came on-line, the costs of acquiring a base-pair of sequence dropped precipitously. So much data were generated that labs often struggled to increase their infrastructure of data storage to accommodate it. Funding of genome sequencing clearly needed to take into account both acquisition and data storage. Furthermore, assembling the sequence into the right order requires a great deal of computing power. Short of having access to super computers, most labs have relied on server "farms" to power such memory and CPU-hogging requirements. Today (mid – 2012), around 12 to 16 million bases of sequence can be obtained for a dollar using Illumina's Hi-Seq

technology. Third-generation technology developers, such as Oxford's Nanopore, claim sequence costs may dip to 100 million bases per dollar in the very near future. These remarkable advancements have largely removed the major constraints, those being time and money, that have previously prevented sequencing the large and complex conifer genomes. In later slides we will explore how these new technologies are being employed and how the data are being used.

Slide 10

A reference sequence provides the most fundamental genetic information available – information required to understand conifer biology and aid in guiding management of genetic resources. The conifer reference sequences will identify virtually all of the genes in the genome and lay the foundation for determining their function or purpose to the species. It provides insight on how genes are regulated, the amount and distribution of repetitive elements, and the nature of how the genome evolved over time.

A reference genome for a conifer will provide phylogenetic representation in the "tree of life" where none currently exists. This will allow comparative genomic studies of gymnosperms and glowering plant that are not now possible. The conifers (gymnosperms) are the oldest of the major plant clades, arising some 250 to 300 million years ago. They provide the key to our understanding of the origins of genetic diversity in higher plants.

Conifers are of immense ecological importance, comprising the dominant life forms in most of the temperate and boreal ecosystems in the Northern Hemisphere. Many conifers, such as the lodgepole pine pictured here, have huge natural ranges. Lodgepole pine, for instance, extends from the boreal forests in Yukon to Mexico, from the Pacific coast to the Front Range of the Rockies. Reference genomes for conifers will help elucidate how they can adapt to these wide-ranging conditions, and will likely aid in assisting species to deal with anticipated global climate and environmental change. In addition, conifers represent a vast global carbon sink that ameliorates rising carbon dioxide levels.

Also, a conifer reference genome will help advance genomic technologies. The analytical and computational challenge of building a reference sequence for such large genomes will drive development of tools, strategies, and human resources throughout the genomics community.

Finally, forests and conifers in particular are of great economic worth. Globally, the gross value of the forestry sector, including manufacturing, is estimated to be US $354 billion, or about 1.2% of GDP (FAO, 2005). A third of the United States is covered by forestland and the American forest products industry contributes ~$50 billion annually to the income of ~ 1.6 million workers and their families. Southern pines provide ~58% of the timber in the US and 15.8% of the world's timber. Collectively, timber is among the most highly valued commodities in America.

In a previous slide it was shown that no fewer than eight conifer genomes are currently being sequenced. It is fair to ask why the scientific community needs more than one or a few, at most. There are legitimate reasons for seeking multiple genome sequences. Each subsequent genome sequence of related genomes will improve the quality and quantity of reference sequence that can be assembled. Conifers are old and have been evolving and diverging for 10s of millions of years and species thrive across a wide range of environments. No doubt they have developed divergent genetic approaches to dealing with these environments. Many basic and applied research questions can be started, advanced or resolved by comparative genomics of related species with varying ecological preferences. Finally, there is, in a sense, an economy of scale. Return on investment in sequencing of multiple genomes is dramatically improved due to reduced costs of assembly of subsequent genomes.

Slide 11

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. As seen in this figure, genome size varies considerably across a range of 181 gymnosperms (mostly conifers), ranging from 6 billion to over 30 billion base pairs. The enlarged inset allows for the contrast of these behemoths with the genome sizes for a number of species for which reference genomes have been published: *Arabidopsis*, rice, poplar, sorghum, soybean, and corn. Before next-gen 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 is typically the case.

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.

Slide 12

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 here, 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 figure above outlines the major elements of the genome sequencing approach that will be described in coming slides. 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 shown here as a somewhat linear flow of activities, in reality, all activities are conducted more or less simultaneously and iteratively. Each of the elements will be discussed in further detail in the coming slides.

Slide 13

Genome sequencing, like most technical efforts, has produced a specialized lexicon. Many of these terms will be introduced in coming slides, but perhaps not always in the order that helps viewers interpret earlier slides. This slide, which will be shown again and discussed at length, is intended to both introduce terms and provide an illustration of steps in sequence acquisition and assembly. Specifically, we want to introduce the terms **"paired-end reads", "contigs", "scaffolds",** and **"jumping libraries"**. Two types of paired-end reads are illustrated here. The engine behind current high-throughput sequencing methods is the paired-end reads of short stretches of sheared DNA. Sequences are obtained for 100 to 150 base pairs (bp) from each end of a DNA fragment that varies in overall length from 200 to 600 bp. In this figure, known sequence is shown in green, unknown intervening sequence is shown in orange. In many cases, the entire fragment sequence is known.

By combining overlapping known sequences, a contiguous sequence can be built. These are known as contigs. Literally millions of contigs of varying length (a few hundred to many thousands of bp) are typically created in large genome projects. Contigs can be further organized into larger groupings called scaffolds. This is done by paired-end reads of DNA fragments of much larger size than noted above,. These large fragments may be a few thousand to 40 thousand bp in length. This process is illustrated at the bottom of the slide. These vary large insert fragments are collectively known as jumping libraries.

With that, lets begin a more detailed discussion of the elements of a genome sequencing project.

Slide 14

A reference genome sequence is intended to represent a very significant proportion of the total genome sequence of an organism, and at the very least, identify and sequence through all expressed genes, their promoter regions and other regulatory elements in the genome. The task is daunting and obtaining truly complete sequences for complex genomes is simply a naïve target. However, by combining complementary strategies to obtaining sequence in sufficient quantity and quality, excellent approximations to a complete sequence are feasible. At least that is the working hypothesis being tested here. The process begins by selecting the genome to be sequenced (a single tree) and the tissues that will provide the source nucleic acids for sequencing.

In our project we have selected individual trees, one from each of three species, from breeding programs. These trees have known genetic qualities, existing genetic resources, such as mapping and QTL populations, and are unencumbered by intellectual property constraints. We have adopted two complementary approaches to acquiring sequence from DNA collected from needle and seed tissues. As previously alluded to, both approaches seek to produce actual or functionally haploid representation. The majority of sequence will be derived from shotgun sequencing of haploid DNA obtained from the megagametoyphytic tissue of a single seed! A second approach to reducing genome complexity is to divide the genome into many random haploid partitions using fosmid clonal pools, each with a genome size of about 1% of the complete genome. Assembly of these smaller pools will be done individually, and then combined in a larger assembly which can ultimately be consolidated. To help pull it all together, sequence will be obtained from an array of "jumping" or "joining" libraries. The jumping libraries consist of paired-end sequence reads of large stretches of DNA of known but variable length. The utility of the jumping library comes both from its ability to span DNA stretches between contigs, and thus pull them together in scaffolds, and to provide good estimates of the distance of undefined sequence, which assists in the final assembly.

Slide 15

The foundation of the adaptive approach to sequencing described in this module is the next-generation sequencing capability currently provided by the Illumina GAIIx, HiSeq 2000, and MiSeq platforms. Their high production capacity, flexibility, and efficiency permit the rapid and low-cost acquisition of sequence from diverse types of DNA libraries. The term library is used to indicate any collection of DNA fragments derived from all or some portion of the target genome and specially modified to facilitate sequencing using a given technology (in our case Illumina). We will describe three general types of libraries in this slide and indicate the proportional representation each will likely provide to our reference genome sequences, though the final result may ultimately vary slightly as we "adapt" to results.

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 obvious question is why so much? We will address this in the discussion that follows on sequence assembly. Suffice it to say at this point that more is almost always better than less when it comes to piecing together reference sequences.

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.

Slide 16

The majority of sequence (40 to 60X coverage) for our reference genome sequences will come from whole genome shotgun sequencing of random DNA fragments derived from haploid conifer megagametophyte. The process begins with the careful extraction of DNA from the megagametophyte tissue (1N) of a single seed. The yield of DNA from the average loblolly pine and Douglas-fir seed is about 1 to 2 micrograms (µg), and for sugar pine, perhaps 20 to 30 micrograms. Though small, this will provide more than enough DNA for whole genome shotgun sequencing libraries.

To begin library preparation a small aliquot (0.1 to 5.0 ug) of genomic DNA is shredded or sheared into random sized fragments of 1000 base pairs or less, and the broken ends of the fragments are enzymatically "repaired" to ready them for further modification. Then, adapter oligonucleotides (short DNA sequences that serve as molecular "handles" during sequencing) are attached or ligated onto the end of the fragments, which are then size-selected in include only fragments in the narrow size range (a narrowly constrained fragment size range aids sequence assembly). This mixture is then subjected to enrichment PCR amplification of about

10 cycles. The subsequent product is ready for Illumina sequencing, the technical details of which will not be discussed further here. Excellent materials explaining the process exist on the web, including a series of modules, at the Broad Institute's Illumina Bootcamp website.

The end products of the sequencing process are paired-end reads. These are paired sequences that represent the order of bases reading inward from the two ends of each DNA fragment in the submitted library. For libraries containing relatively short fragments, the entire sequence may be determined, but for longer fragments, an unspecified number of bases in the middle will remain un-sequenced. Maximum possible read lengths for each end depend on the Illumina instrument used, typically 125 bp on the HiSeq platform. Thus, for a 500 bp fragment, ~250 bp of unknown intervening sequence may reside between the obtained reads. By programming the Illumina instrument for the appropriate read lengths and choosing a library with the correct fragment size distribution, we are able to tune the length of the intervening gap, including no gap at all, or even a ready overlap of a specified size. A single sequencing run on the Illumina HiSeq 2000 machine takes ~ two weeks to complete but can produce well over a billion read pairs.

Slide 17

The sequencing of haploid genomic DNA, as described in the previous slide, largely eliminates the diploid diversity of the conifer genomes, but yields billions of short reads that must be assembled into the proper order, obviously a daunting task. Another approach to reducing genome complexity, and facilitating the assembly process, is to partition the genome into smaller bites, sequencing the individual bites, and then piecing them together, one at a time.

Historically, BACs or Bacterial Artificial Chromosomes, have been used to break the genome into bite sizes of 200 Kb or so, but BAC libraries have proven to be difficult and expensive to create in conifers. The fosmid, which uses the bacteriophage lambda particle, can be made simply, in very large numbers, and is relatively easy to manipulate. It begins with purified DNA obtained from diploid needle tissue of the target tree. The DNA is sheared and size selected (37 Kb to 40 Kb), at which point individual molecules are ligated to the cloning vector, and packaged into the lambda particles, which in turn are transfected into *E. coli* cells.

Slide 18

The *E. coli* are grown out in colonies (or clones) on media in petri dishes. Total DNA is prepared from each fosmid pool, sheared, size selected and prepared with appropriate adapters to permit sequencing on the Illumina platform, exactly like the whole genome DNA. Though the DNA used in this process came from diploid tissue, the sequence obtained from these fosmid pools is considered to be effectively haploid! This claim results from two simple features of the process: 1) each fosmid contains only a single, haploid strand of DNA and 2) the total amount of DNA included in a pool of even 4000 fosmid clones is so small that it is extremely unlikely that the same DNA segment will be represented twice.

The assembly of the sequence of this bite size genome is relatively easy, computationally, and provides for a high quality assembly of functionally haploid DNA. Combining the results of several such pools allows the sequence building process to proceed in a iterative manner. The major challenges to the fosmid pooling approach to sequencing include 1) genomic DNA preps for pine must be done with care to avoid DNA damage prior to fosmid creation, 2) the quality control steps may be onerous, and 3) care must be taken to reduce or account for E. coli and vector DNA contamination in the sequence. The total sequence acquired from this source is anticipated to vary from 1 to 5 X the species anticipated genome size.

Slide 19

For both whole genome shotgun and fosmid pool assembly a critical component of the required sequence data is the jumping library. As noted previously, the purpose of the jumping libraries is to provide a means of contracting contigs into scaffolds. Like the fragment libraries already discussed, jumping library molecules are constructed and sequenced to yield read pairs with an approximately defined physical distance in base-pairs between the two reads. The difference between these fragments and those contained in non-jumping libraries is the size of the intervening distance. This distance, as you recall, vary in size from 200 to 600 bases for non-jumping libraries (short-insert). Jumping libraries, by contrast, may have inter-read distances of 2 Kb to 40 Kb, depending on the methods used to build them. However, because Illumina sequencing can not directly utilize fragments longer than about 750 bp, construction of jumping libraries always utilizes some method of removing intervening DNA from between the two ends of each molecule to bring the overall size of the fragment below this threshold. For this project, two general methods are being used: the clone free, "mate-pair library" method used by Illumina, and the cloned "fosmid di-tag library" method, developed by scientists at JGI. It bears mentioning that the DNA input requirements of Illumina mate-pair libraries preclude the use of haploid (megagametophyte) DNA. Consequently diploid DNA is used for these libraries, but only reads that match the megagemetophyte haplotype are ultimately used in the assembly.

For the clone-free mate-pair approach, whole genome DNA is fragmented and size sorted to fragments of desired length, typically with a mean size somewhere between 2 Kb and 5 Kb, with a variance of +/- 10%. The ends of the fragments are labeled with biotin, a small, covalently bound molecule commonly used to tag DNA molecules for later retrieval due to its high binding affinity to another molecule called streptavidin. Each DNA fragment is then circularized by joining the ends together with ligase. These large, circular molecules are then re-fragmented into random pieces, with one piece from each circle containing the labeled, joined ends. The biotin-labeled "junction fragments" are subsequently bound to streptavidin coated magnetic beads, and non-labeled (non-junction) fragments are washed away. As with the non-jumping libraries previously discussed, adapters are ligated onto the free ends to make the molecules "sequencer ready" and the library is amplified by 18 cycles of enrichment PCR. A final size selection procedure isolates molecules with a size between 350 bp and 650 bp, ideal for Illumina sequencing. An astute observer may notice that the original fragment ends are now oriented "backwards", pointing towards the center of the molecule. Assembler software is designed to anticipate this when using paired-end reads from this type of jumping library, and processes them appropriately.

Slide 20

The fosmid di-tag approach relies on specialized vectors that ultimately allow for the creation of short paired-end fragments that are ready for sequencing directly on the Illumina platform. One common approach to deleting the internal sequence is the process of nick translation and cleavage. These libraries typically consist of much larger pieces of DNA being spliced out, typically in the range of the fosmid insert size of 35 to 40 Kb. Since these libraries produce paired-end reads representing only the ends of the fosmid inserts, they are not applicable for scaffolding fosmid assemblies themselves. However, a di-tag library constructed from a sufficient number of unique fosmid clones offers great utility in assembling the WGS by joining contigs across gaps too large to be spanned by the 2 to 5 kb libraries described in the previous slide. Collectively, these two approaches give a range of insert sizes that allows for building scaffolds of considerable length. The anticipated amount of sequencing required to develop the di-tag resources is also likely to be about 5X in depth.

In the following slide we will illustrate how the process of assembling the sequenced pieces together should work, hypothetically.

Slide 21

So far we have described the types of DNA libraries that have been created for sequencing and the platform used to generate the sequence. Our outline of activities has been brief, largely void of technical details, and lacking in description of the various experiments due diligence requires before investing fully in any specific sequencing strategy or chemistry. Still, the general approaches have been addressed.

At this point a great deal of data has been generated and electronically stored in the form of millions of "**reads**", each read consisting of relatively short sequences (~100 to 160 bases) from each of two paired-ends of a DNA fragment that varies in total size from roughly 200 to 600 bases (short-insert libraries) or up to 40 Kb in the case of large-insert libraries, with interior unknown sequence of variable length depending on the library origin. With billions of sequence reads in the bank, the grueling task of assembling the sequence in its proper order begins. The current slide is intended to illustrate the general concept of how a genome assembly is created. We will address specifics of the process in following slides.

The **sheer** abundance of sequence obtained (40 to 60 fold the genome size) insures that most, though typically not all, stretches of genomic DNA sequence will have been sampled and sequenced multiple times. By aligning stretches of sequence that overlap with each other, it is possible to extend the total length of sequence that is known. This contiguous sequence is called a **contig**. Contigs may be built using as few as two reads or literally millions of reads, stretching from 100s to 10s of thousands of bases in length. A complete reference genome would theoretically consist of one contig for every chromosome, or linkage group, in an organism. This would imply that virtually all nucleotides in the sequence were known. Such sequences exist for only the smallest of genomes (bacterial or organelle genomes for instance). In reality, the assembly of large genomes begins with the building of 10s or 100s of thousands of contigs and similarly large numbers of sequences that do not overlap with any others, called singletons. In the case of the short paired-end reads obtained from the Illumina sequencing platform, the end sequences are aligned until sufficient depth of coverage is found to identify the entirety of sequence in one stretch.

The next step in the assembly consists of pulling **contigs** together into **scaffolds.** If the contigs are reasonably close enough to one another it is often the case that two reads from the same pair will span the distance between them, as noted in the illustration shown here. Again, a scaffold may consist of a highly variable number of contigs and the goal of the reference sequence is to have as few scaffolds as possible, ideally one per linkage group or chromosome. For any two scaffolds, the ordered relationship between them will remain unknown without additional sequencing reads or traditional genetic approaches such as genetic mapping.

Before further discussing how mapping and other complementary activities enhance the assembly of a genome, we will discuss in greater depth how the steps just described are accomplished.

Slide 22

So we know how and from where the conifer sequence will be derived and a general idea of how it will be put together. Historically, simply obtaining the massive quantity of sequence for large genomes was a show-stopping stumbling block. However, next-generation sequencing technology has made sequence acquisition fast and affordable – it is simply no longer an obstacle. The tallest hurdle to be cleared remains – assembling the massive quantity of sequence into its proper order. The task of genome assembly is akin to assembling a giant jigsaw puzzle. The loblolly pine tree genome is about 7 times larger than the human genome, and the first draft assembly is nearly 5 times larger than any other plant genome previously sequenced and assembled (wheat estimated genome size is 17 gigabases but the draft assembly is only 5.5 Gb).

Assembling the conifer genome requires putting together a jigsaw puzzle of 15 to 20 billion pieces or reads, each read being about 100 to 150 bp in length. The process illustrated in this slide is called the OLC or Overlap-Layout-Consensus algorithm. It is the approach taken for all of the early genome assemblies that were based largely on Sanger sequencing methods**,** which yield sequence reads of 600 to 1000 bp rather than the 100 to 250 bp reads typical of Illumina technology. Well-known software packages using this approach are the Celera Assembler, PCAP, Phusion, and Arachne. The best known of these is probably the Celera Assembler which, with many enhancements added in recent years, can assemble data sets of up to 1 billion reads. An alternative approach to assembly is known as the de Bruijn graph assembly method. Assembly programs based upon the de Bruijn approach include SOAPdenovo, Allpaths-LG, Velvet, and Abyss. These packages are commonly used on Illumina-generated sequences. In the next few slides we will discuss these alternative approaches and the overall strategy for assembling the massive conifer genomes that the PineRefSeq project has adopted.

Slide 23

The De Bruijn graph approach builds contigs by first building a special graph. In this graph, each node represents a sequence of length k, called a k-mer, that was found in a read. To create the graph, we draw an edge between two nodes X and Y when the last k-1 letters of X are equal to the first k-1 letters of Y. The graph is built by processing the reads one at a time, and adding nodes whenever we see a k-mer that we haven't seen before.

We create contigs for the assembly by finding paths through the graph that visit each edge at least once. The method can only roughly estimate the graph of the genome from reads due to sequencing errors and lack of coverage. Repetitive sequences create cycles (loops) in the graph which make reconstruction more difficult.

Slide 24

A comparison of the benefits and drawbacks of the two principal approaches to assembling sequences suggests that OLC is generally desirable, but earlier OLC systems were simply too computationally intensive to handle the large conifer genomes. Some implementations of the graph method are computationally efficient, but they have a number of drawbacks that may significantly reduce the quality of the assembly. The adaptive approach used in the PineRefSeq project is to combine elements of both methods. This is done with a custom software package created by collaborators at the University of Maryland. It is called MaSuRCA (pronounced mazurka), which stands for Maryland SuperReads – Celera Assembler.

The hybrid approach begins by using the graph method. Specifically, Illumina reads are error corrected and then used to create a de Bruijn graph where each k-mer is unique in the graph. Many reads will extend to the same "Super-read" or contig, effectively reducing the amount of data by a large factor. In the initial loblolly pine assembly, an average super-read consisted of 30 to 50 Illumina reads. The process reduces the 20 billion reads to fewer than 1 billion super-reads. The super-reads are subsequently assembled with the modules from the Celera OLC assembler (CABOG) using additional long mate pairs for linking the contigs.

The MASURCA assembler has been demonstrated to work on conifer genomes. It is currently designed to handle data sets of up to 30 billion reads. Using efficient multi-threaded code, it can handle a WGS assembly of the 22 Gb loblolly pine genome on a computer with 48 cores and 1 Terabyte of RAM in 1 to 2 months.

Slide 25

We conclude the section on genome assembly with a slightly different schematic view of the overall assembly strategy. Note that in this slide we reference use of the transcriptome and genetic mapping as aids in the final

assembly process. In the slides to follow we will address how these aids are used to help complete the first full draft of a whole genome sequence.

(Note: MaSuRCA - Maryland SuperRead Celera Assembler)

Slide 26

Genetic maps provide a basis for anchoring and orienting the genome sequences – they help validate the integrity of the existing genome assembly. In this figure, four of the 12 chromosomes or linkage groups (LG) of loblolly pine are portrayed with arrays of markers genetically mapped using a large (~500) full-sib progeny derived from a controlled cross between two parent trees. Most of the markers used in this map are SNPs or single nucleotide polymorphisms discovered through re-sequencing of expressed genes in an array of unrelated loblolly pine trees. In the blow-up feature, it can be seen that two or more markers are often binned at the same location on the LG implying the markers reside next to one another. In fact, they may be quite some distance apart (thousands of bp for instance), but they can not be separated genetically because too few meiotic events exist to capture rare cross-over events. Greatly increasing progeny size, coupled with much larger marker arrays, can significantly enhance the utility of genetic maps in pulling whole genome sequence scaffolds together. Fortunately, conifers lend themselves well to both progeny and marker expansion.

Slide 27

In this figure, we illustrate how genetic mapping can enhance the genome assembly process. Mapping seeks to place one or more genetic markers on every scaffold. If two scaffolds exist on the same linkage group, and markers on different scaffolds are genetically linked, the scaffolds can be combined into “super scaffolds” so to speak. The distance between two linked scaffolds will remain defined by a sequence gap of unknown length unless further sequence is obtained. Mapping provides an efficient means of anchoring and orienting scaffolds. That is, scaffolds can be arranged properly with respect to one another, and the sequence orientation (5’ to 3’) can be revealed. In this illustration, for instance, scaffolds one and two are shown to map in the order in which they were originally portrayed. Genetic linkage could have just as easily shown that scaffold two belonged before scaffold one. While the PineRefSeq project seeks to pull the assembly together with genetic maps, a thorough mapping project could require a great many markers (say, hundreds of thousands). Once considered economically beyond reach, technology advances now make this objective tractable. The PineRefSeq project will strive to map literally millions of SNP markers using the rapidly developing approach of Genotyping by Sequencing, or GBS.

Slide 28

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.

The term transcriptomics refers to the global analysis of gene expression in an organism, or expression profiling. Methods and approaches to characterizing the transcriptome and gene expression have evolved rapidly over the last decade and will be discussed further in the following slides.

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.

Slide 29

The illustration shown here is a schematic representation of the structure of a eukarotic gene. It will be discussed momentarily but first we reflect on the utility of the transcriptome to the building of the reference genome.

The transcriptome, though representing only a small fraction of the total conifer genome, is arguably the center of attention for both basic and applied genetic and evolutionary interests. Characterizing expressed genes and the mechanisms that regulate them is central to our characterization of the genome as a whole. A clear understanding of the transcriptome will help identify gene location, boundaries, and boundaries of their introns and exons, alternative splicing sites, upstream and downstream promoter and regulatory sequences, gene families and pseudo-genes, and much more. By identifying all expressed genes in an organism, the transcriptome can assist in defining when the reference genome sequence build is nearing completion.

In the PineRefSeq project a priority is given to the early development of useful transcriptome reference sequences. These are compiled by mixing existing EST datasets, derived from cloned cDNA libaries (Sanger and Roche 454 sequencing), with new data sets obtained from the Illumina RNA-SEQ sequencing of cDNA fragments obtained from the multitude of new libraries.

Let's use the illustration above to explain how comparing the transcriptome reference sequence to the genome sequence can facilitate both de novo genome assembly and gene characterization. Before the RNA transcript (mRNA) serves as a template for protein biosynthesis, non-coding sequences (introns) are eliminated, coding sequences (exons) are fused (referred to as 'splicing') and the 5' and 3' untranslated regions (UTRs) are post-transcriptionally modified. What typically gets sequenced from mRNA is the open reading frame. Open reading frames (ORFs) that are translated into a protein always start with the initiator codon AUG and end with one of the terminator codons UGA, UAA or UAG. Basic promoter sequence motifs such as TATA and CAAT, additional promoter elements such as ERE (ethylene response element, in some plant genes), and up- or downstream regulatory regions on the same strand as the coding region are called *cis*-elements. Cis-elements are not typically captured in the transcriptome sequence, nor are the intron sequences. Ultimately, these sequences can be provided by the reference genome. Conversely, RNA-seq

provides such abundant sequence that full-length or near full-length cDNA sequence reads are obtained more frequently and this may more fully inform the genome sequence build. The transcriptome provides a good quality check on the whole genome assembly.

Slide 30

Beyond building a reference genome sequence, the study of the transcriptome and gene expression provides countless insights into how organisms develop, genes interact, and so forth. We name but a few lessons here:

Transcriptomics contributes to

1. An understanding of genes and pathways involved in biological processes. This is a sort of "guilt by association": genes with similar expression may be functionally related and under the same genetic control mechanisms.

2. Elucidating the function of unknown genes based on their spatial and temporal expression.

3. Identifying marker genes as diagnostic tools for disease or stress susceptibility or tolerance.

4. Our understanding of gene expression, which can be viewed as a proxy for cis- and trans- regulation, thus allowing us to make indirect inferences about genetic differences among individuals or species.

5. Our understanding of the relationship between gene expression and changes in the proteome and metabolome.

Slide 31

An example of the transcriptome sequencing efforts from the conifer genome sequencing project outlined in this presentation shows preliminary results of combined sequencing efforts based on Roche 454 and Illumina RNASeq sequencing platforms for three conifer species: sugar pine, Douglas-fir, and loblolly pine, each represented by multiple tissue cDNA libraries. The data in the table represents the total number of sequencing reads (N) used to assemble the transcript sequence, the total number of nucleotides sequenced, the number of transcripts assembled for each library, the average length of the resulting transcripts (in base pairs), and the total number of unique transcripts for each species. The column titled library reflects the species, tissue from which RNA was derived and the software program used to assemble the transcript contigs.

At this point in the assembly of the reference transcriptome the number of unique transcripts is presumed to be a significant over-estimate of the true number of expressed genes in the organism. Two or more transcripts may represent the same expressed gene, either as alternatively spliced products, or simply two different sections of the full-length transcript that have not been combined due to lack of sufficient sequence to pull them together.

Slide 32

Genome annotation is the process of attaching biological information to sequences. It is the element of a genome project that analyzes the linear combination of DNA bases and provides an interpretation of their structure, function, and relevance. Without annotation, the sequence itself is largely void of useful information for biologists. The annotation process is multi-faceted, relies heavily on analytical software, but cannot be done well without manual curation by knowledgeable scientists.

Annotation may be roughly divided into two categories: functional annotation of the transcriptome, regulatory sequences, and non-protein coding genes, and annotation of the structural features of the genome

like repetitive sequences, duplications, and SNPs. The next few slides will describe elements of these annotation processes, beginning with functional annotation.

Slide 33

A high priority for any genome sequencing project is the identification and classification of “all” genes in the genome into families of known or putative function, a process generally referred to as functional annotation, or functional genomics. Functional genomics focuses on dynamic activities such as gene transcription, translation, and protein-protein interactions. As noted in Wikipedia, functional genomics attempts to answer questions about the function of DNA at the levels of genes, RNA transcripts, and protein products. A key characteristic of functional genomics studies is their genome-wide approach to these questions, generally involving high-throughput methods rather than a more traditional “gene-by-gene” approach. Ultimately, the goal of functional genomics is to understand the relationship between an organism's genome and its phenotype, through gene expression and proteomic studies. Fully describing these relationships is part of the annotation process.

Functional annotation is a multi-step process that relies heavily on analytical software tools, the most used of which are the family of BLAST programs. BLAST, or the Basic Local Alignment Search Tool, finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to existing sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. BLAST products are available on line at the National Center for Biological Information (NCBI) or may be downloaded to local servers.

A number of other programs or steps are useful for functional annotation of a putative gene including GO, or gene ontology, which will be discussed in the following few slides, and a series of applications to better understand the nature and function of the putative protein products. **SignalP** predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms. A **signal peptide** is a short peptide chain that directs the transport of a protein. **TMHMM**, orTied Mixture Hidden Markov Model, is a program that predicts the structure of trans-membrane helices in proteins. **Pfam** is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models. **PROSITE** is a protein database. It consists of entries describing protein families, domains and functional sites, as well as amino acid patterns, signatures, and profiles in them.

As DNA and protein databases continue to improve, the functional annotation process that depends solely on alignments and comparison of known sequences will improve, but it will always be incomplete and prone to error. As Gibson and Muse note, **identifying** the true function of many genes will ultimately require classical genetic, biochemical and cell biological methods and experiments.

Slide 34

Gene Ontology, or **GO**, is a process that seeks to unify the representation of gene and gene product attributes across all species. More specifically, the project aims to: 1) Maintain and develop its controlled vocabulary of gene and gene product attributes; 2) annotate genes and gene products, and assimilate and disseminate annotation data; and 3) Provide tools for easy access to all aspects of the data provided by the project. Gene Ontology consists of 3 areas that are briefly described using examples in the illustration on this slide. Those areas are defined as biological process, molecular function, and cellular component.

The ontology concept is now being widely expanded to include other areas of biological relevance such as defining terms that describe phenotypic attributes.

Slide 35

An example of GO is presented here as a comparison of transcriptome assignments between two species of forest tree, the American chestnut, which has been virtually extirpated from the forest landscape due to the introduction of an exotic disease, and the Chinese chestnut, which possesses resistance or tolerance to the same disease organism. Geneticists use such ontologies to help direct their search for the genes that confer resistance, so they may be used in introgression breeding or genetic engineering experiments.

Slide 36

In addition to identifying all genes in a reference genome, the functional annotation process includes identification of conserved regions that possess regulatory control of gene expression and non-protein coding genes that produce functional RNA molecules.

Regulatory sequences are typically identified as conserved regions or motifs that act as protein binding sites for transcription factors. These are generally identified by comparative genomic approaches such as phylogenetic foot-printing or shadowing.

The number of genes encoding functional, but non-protein coding, RNA molecules may exceed 1000 in some organisms. As described clearly by Gibson and Muse, the array of genes controlling RNA molecules that in turn influence all manner of cellular functions is large and complex in nature. One such class of genes are those that code for microRNAs which are “short (~22 nucleotides) hairpin RNAs that bind to the 3’ UTRs of mRNAs and function in gene regulation either by repressing translation or promoting mRNA degradation”.

Slide 37

An ideal annotation would be inclusive of many bits of information. This compendium of data would include the complete DNA sequence and reads that were used to build the sequence. It would include all single nucleotide polymorphisms known or detected in the individual from which the sequence was derived, or from re-sequencing of one or more subsequent individuals. It would include knowledge of promoter and regulatory regions controlling gene expression, to the extent they are known. It would include all known messenger RNA molecules, as predicted by multiple splicing sites, all potential protein products and their modifications, and an array of information on gene expression and protein interactions.

Obviously, a complete annotation will exist for relatively few genes in most newly sequenced genomes, and even approximating a partial list of these traits for most genes may take years, or more expense than the initial sequence build. As the process evolves, the classification of gene sequences may change. The annotation nomenclature is typically defined as the following: A predicted gene sequence that matches the entire length of a known gene sequence is called a **Known Gene.** A **Putative Gene** is one that contains a region or set of regions conserved with a known gene. A putative gene may also be referred to as “like” or “similar to”. A predicted gene sequence that matches that of another gene or EST with unknown function is called an **Unknown Gene**. Finally, a predicted gene that does not contain significant similarity to any known gene or EST is called a **Hypothetical Gene.** For all definitions given here, the term protein is frequently used interchangeably with gene.

Slide 38

Though functional annotation of the expressed portion of the genome is glamorous, it represents just a tiny fraction of the whole genome in conifers. Complete annotation requires characterization of other features of the genome, collectively referred to as structural annotation.

The most common feature of higher plant genomes including conifers is repetitive DNA. Ninety percent or more of most plant genomes probably consist of repetitive DNA which comes in two general types, tandem repeats and interspersed repeats. Tandem repeat elements or motifs include micro- and mini-satellites while interspersed repeats are dominated by transposons (moving or jumping genes) and retro-transposons of varying length, from 100 bp up to 10 kb or larger. These elements may occur over a million times in the genome. Another class of repetitive DNA that affects genome structure is the gene family. Large gene families may possess many inactive genes or pseudogenes. In some organisms large segments of the genome may be duplicated and located, either on the same chromosome or on other chromosomes. Of course, the ultimate duplication event is of entire genomes (polyploidy).

The content of the nucleotides guanine and cytosine show wide variation across the typical genome which can affect the accuracy of sequencing and may play a role in chromatin assembly mechanisms. Finally, there remains keen interest in the structure and function of centromeric and telomeric regions of the chromosomes, which are comprised largely of heterochromatin – or large sections of long, highly repetitive stretches of DNA including transposable elements and inter-chromosomal duplications.

Slide 39

A database is an *indexed collection* of information. The explosion of genomic information in recent years has led to an equally rapid expansion in database resources to capture and manage the information. There is a tremendous amount of information about biomolecules in publicly available databases but individual projects invariably must develop their own database resources, created and maintained by increasingly knowledgeable and sophisticated bioinformatics experts. We begin our discussion of database resources by noting the key functions that should be served by a project level database, or “genome browser”. For the PineRefSeq project the primary genome database and browser is maintained at TreeGenes, an element of the Dendrome Project housed at the University of California at Davis. The project will collaborate closely with the Genome Database for the Rosaceae, or GDR, to provide a comparative genomics platform.

The principle functions or resources that a project related database should provide include the following:

Capture of DNA sequences and related annotations, check all data for accuracy, and post data to internal databases.

As soon as possible, distribute data to relevant external databases.

Maintain and track all original information developed and all actions taken on the datasets.

Provide computational resources for retrieval and analysis of searchable databases.

Provide visualization tools such as Genome Browsers to enable users to effectively mine the available sequences.

Provide links to all relevant databanks external to project server resources.

Provide appropriate documentation and tutorial support for project resources.

Provide thorough and timely responses to requests for data or assistance.

Slide 40

Public database resources are abundant, variable in quality, ease of use, level of curation, and so forth but there are key sites that provide most of what is needed for a comprehensive study of any genomic resource. A complete listing of current databases is provided annually in the January issue of the publication “Nucleic Acids Research”. In the 2011 issue over 1300 editorially selected databases were listed. As noted by Arthur Lesk in his book entitled Introduction to Genomics, resource databases may be grouped in a number of categories including nucleic acid sequences, amino acid sequences of proteins, protein and nucleic acid structures, small-molecule crystal structures, protein functions, gene expression patterns, metabolic pathways, and networks of interaction and control. On this page and the ones that follow we discuss but a few of the most used database resources available today.

Nucleic acid and protein sequence databases are probably the most frequently visited. The EMBL, DDBJ and NCBI GenBank sites are core resources for nucleic acid sequence queries. We will discuss NCBI in greater detail in a bit.

Slide 41

One of the most useful and comprehensive database collections is the NCBI, part of the National Library of Medicine. It is home to a multitude of useful databases including GenBank, Protein, PubMed, and many others, as noted in this graphic. NCBI provides interesting and useful summaries, browsers, and search tools. Entrez is their database search interface for NCBI. Here you can search on gene names, chromosomal location, diseases, articles, keywords, etc. Sequence resources are organized by BioProjects which can present relevant descriptions, protocols, resulting sequences (both raw and assembled). NCBI has tutorials to assist with learning the ropes of how to use the site.

Slide 42

As the second anniversary of the PineRefSeq Project nears (February, 2013), significant strides toward Project goals have been made. The first loblolly pine genome sequence release of data, version 0.6, made in June of 2012, garnered instant interest from around the globe. In the six months following that release, over 3,000 FTP downloads and 1000 NCBI blasts against the draft reference were received from scientists in over 30 countries, but the majority of hits came from colleagues in the United States and Canada. Though we can not say how the data is being used, in most cases, we do know that it is being used extensively by some of our tree improvement cooperatives to improve their tree breeding and selection capabilities.

At the North Carolina State University Industry Cooperative Tree Improvement Program, the reference genome will serve two key functions: 1) As a reference for re-sequencing elite individuals to identify alleles in all genes for which markers exist, and 2) As a physical map of marker locations, to guide imputation of missing data. This capability is essential for all matrix-based methods of analysis. It allows accurate imputation of progeny from structured mating designs based on known parental haplotypes. The combination of these two capabilities will allow analysis of the genetic basis of phenotypic variation and incorporation of additional genomic data. It will not only improve traditional quantitative genetic methods, but will facilitate the future of genomic selection methods.

The first full release of the loblolly pine sequence will be made in January 2013. We anticipate an equally enthusiastic response from users to the new, vastly improved assembly. As our team works to finish that reference sequence, efforts are already underway to build sequences for additional species, Douglas-fir and sugar pine.

Slide 43

The sequencing and assembly of large conifer genomes was a distant vision as recently as 2010. Today, we are poised, along with other groups around the world, to deliver multiple reference sequences for a range of conifers. The sea changes in sequencing technology, computational capacity, and software development have all come together to permit this rapid progress. Continued improvements in all sectors will permit refinements to these genome assemblies, and improved annotation and curation will provide useful genomic resources for basic and applied research.

We conclude this presentation with results of a number of recently published genome sequences, and the 0.9 draft of the loblolly pine reference sequence assembly. Interested readers may check project websites for continually updated results of the PineRefSeq project (see “website resources” on following pages).

Download: http://loblolly.ucdavis.edu/bipod/ftp/Genome_Data/genome/pinerefseq/Pita/v0.9/

BLAST: http://dendrome.ucdavis.edu/resources/blast/

Data Release Policy: http://www.pinegenome.org/pinerefseq/Data_Use_Policy.pdf