SAVE THE REDWOODS LEAGUE RESEARCH GRANT FINAL REPORT

In fulfillment of your Save the Redwoods League Research Grant agreement, please submit a document containing the following information:

1. Grant details
   Grant Number: 075
   Project Title: Using modern technology to resolve an ancient mystery: unraveling the hexaploid origin of the coast redwood

   Principal Investigator's name: David Baum, Alison Scott
   Grantee Institution: University of Wisconsin, Madison

2. Summary

Background
Coast redwoods (*Sequoia sempervirens*) are well-known for their great height (over 100m) and advanced age (over 2,000 years), but perhaps less so for being the only hexaploid conifer (having six copies of each chromosome; 2n=6x=66). Though these colossal ancient trees are limited to the foggy coastal forests of central and northern California and southwestern Oregon, the redwood fossil record suggests a broader historical range across the Northern hemisphere. How, when, and where polyploidization took place remains a mystery, though diverse genome donors and polyploidization mechanisms have been proposed.

Sequoia is a monotypic genus, meaning it contains only one species. Its closest living relatives are the giant sequoia, *Sequoiadendron giganteum*, of the Californian Sierra Nevada and the dawn redwood, *Metasequoia glyptostroboides*, found in China. Cumulatively, *S. sempervirens, S. giganteum*, and *M. glyptostroboides* are known as the redwood clade.

Despite their restricted current distributions, the fossil record shows a variety of redwood-like lineages scattered across the Northern Hemisphere, including overlap of different species. This historical overlap suggests that hybridization among redwood lineages was possible, and may have contributed to polyploidy in coast redwood. Our project uses a combination of molecular sequence data and the rich fossil record to shed light on the evolutionary history of this enigmatic lineage.

Methods

The accessibility of next-generation sequencing technology has drastically increased the ease of generating sequence data while decreasing the cost. Though sequencing the entire genome of an organism is possible, such an endeavor is impractical in coast redwood due in part to its polyploid status and enormous genome size (nearly 10 times larger than the human genome). However, it is now feasible to use transcriptome sequencing to obtain reliable sequences of genes that are shared among the redwoods. A transcriptome contains only the portions of the genome that are "turned on" in the organism, so it is a much smaller amount of genetic material to sequence.

To generate a transcriptome of coast redwood and its close relatives, we extracted RNA from living foliage. As RNA is an unstable molecule, RNA was then translated to create a cDNA library. The resulting cDNA will be fragmented and prepared for sequencing, then run on an Illumina next-generation sequencing platform. The resulting sequence fragments were assembled into contiguous sequences (contigs). To further clarify polyploidization mechanisms, a subset of 908 contigs were chosen for targeted sequencing of genomic DNA to isolate all genome copies from polyploid coast redwood.

Results

Transcriptome sequencing of the three redwoods yielded molecular data for up to 70,000 expressed sequences per species. A preliminary analysis based only on genes with a single detected copy per species showed that, for a great majority of genes, copies from coast redwood and giant sequoia are sister to one another. This pattern implies either autopolyploidy or allopolyploidy involving only genome donors within the California redwood clade. Subsequent phylogenetic comparison among gene copies in coast redwood and its close relatives will reveal whether hybridization with an extant lineage contributed to polyploidy in the *Sequoia* lineage.

What new questions does this research raise? The first question, soon to be answered by new data, is: where did the three parental genomes of coast redwood come from? There must have been at least two episodes of polyploidization to go from diploid to hexaploid (with either a triploid or tetraploid step in between). When and where did these polyploidization events take place? Have any additional polyploidizations happened more recently that we don't yet know about – perhaps in a highly clonal population of coast redwood? How is the coast redwood "using" all its duplicated genes – if it's using them at all?

How can this research help us save more redwoods? This project resulted in the development of genomic tools for redwoods, which have many potential conservation applications. For example, we are now investigating the targeted sequencing data to find genes that are variable within species. With these markers, we can designate groves of particular conservation concern and improve management strategies, perhaps by identifying adaptive variation in redwood populations, which may later serve as source populations for seed banks and plantations.

3. Full report

Background/Rationale

Coast redwoods (*Sequoia sempervirens*) are well-known for their great height (over 100m) and advanced age (over 2,000 years), but perhaps less so for being the only hexaploid conifer (2n=6x=66). Though these colossal ancient trees are limited to the foggy coastal forests of central and northern California and southwestern Oregon, the redwood fossil record suggests a broader historical range across the Northern hemisphere. How, when, and where polyploidization took place remains a mystery, though diverse genome donors and polyploidization mechanisms have been proposed.

Sequoia is a monotypic genus. Its closest living relatives are the giant sequoia, *Sequoiadendron giganteum*, of the Californian Sierra Nevada and the dawn redwood, *Metasequoia glyptostroboides*, found in China. Cumulatively, *S. sempervirens*, *S. giganteum*, and *M. glyptostroboides* are known as the redwood clade. Despite their restricted current distributions, the fossil record shows a variety of redwood-like lineages scattered across the Northern Hemisphere, including overlap of different species. This historical overlap suggests that hybridization among redwood lineages was possible, and may have contributed to polyploidy in coast redwood.

To resolve the origin of hexaploidy in coast redwood, we propose to use transcriptome sequencing to identify low-copy nuclear genes in *S. sempervirens*, followed by targeted sequence capture of these genes in *S.*

sempervirens and its close relatives. Unraveling the evolutionary history of the coast redwood may help us to understand its current ecological adaptations and evaluate how this keystone species may respond to future changes in climate. Additionally, the genomic tools that we will develop will provide opportunities for many additional genetic studies in redwoods.

Methods

As proposed in the grant application, we used transcriptome sequencing to obtain reliable sequences of genes that are shared among the redwoods. To generate transcriptome data from coast redwood and relatives, we extracted RNA from living foliage following standard protocols. Extraction protocols were modified to mitigate the problems caused by secondary compounds in redwood foliage (good for the trees but bad for molecular work!) Once we were confident in the quality and quantity of our RNA extractions, cDNA libraries were synthesized and normalized. The cDNA synthesis is due in part to the unstable nature of RNA; once a cDNA library is constructed it can be stored with little risk of degradation. Normalization allowed us to identify sequences from lowly-expressed transcripts, while reducing the frequency of highly-expressed transcripts. Our cDNA libraries were then fragmented and prepared as Illumina libraries following standard manufacturer protocols. This process involves ligating on special adapters to the cDNA fragments, and adding barcodes to different accessions to allow pooling of samples in a single sequencing lane. Samples were sequenced on an Illumina HiSeq 2000 with 100bp paired-end reads.

The resulting raw reads were demultiplexed (sorted by barcode) and filtered to remove low-quality sequences using CLC Genomics Workbench. The resulting sequence fragments were assembled into contiguous sequences (contigs) using both CLC Genomics Workbench and Trinity. We compared the resulting assemblies against each other using BLAST to identify genes conserved among the three redwood taxa. With this reduced dataset, we then compared the redwood-conserved genes to the Norway Spruce genome to identify orthologs present in outgroup conifers (as opposed to redwood-specific genes). We used RepeatMasker to remove low complexity stretches of sequence and interspersed repetitive DNA, which can cause problems in the hybridization process.

At this stage we wanted to isolate all genome copies from polyploid coast redwood, and we used a different technique than initially proposed. The grant proposal described sequencing 96 genes using pooled PCR. Instead, we decided to use targeted sequence capture to sequence 900+ genes. This technique can target more genes at a time than PCR, and is better suited for high-throughput applications. Targeted sequence capture combines genomic DNA, prepared as above for Illumina sequencing, with RNA "baits" matching your genes of interest in a hybridization process. In this technique, called in-solution hybridization, genomic DNA libraries and RNA baits are added to a master mix (similar to PCR). The mixture is heated to allow denaturation of both libraries and baits, then cooled as RNA baits anneal to their complementary DNA sequences. Post-hybridization, the captured portions of the DNA library are selected using streptavidin-coated magnetic beads. As RNA baits are biotinylated, the biotin-binding properties of streptavidin allow unbound DNA to be washed away, while retaining only the DNA hybridized to an RNA bait. The end product of this protocol is an Illumina library enriched for genomic DNA matching the targeted sequences.

We chose 908 orthologs for targeted sequencing. RNA baits were designed based on the transcriptome sequence and synthesized by Mycroarray. We opted for baits 120 base pairs long, as longer baits tolerate more mismatches in hybridization. The baits covered our target sequence with 2x tiling, ensuring we obtain sequence covering the full length. A first-run test panel of our baits included eight genomic DNA accessions: three individuals of *Sequoiadendron giganteum*, two *Sequoia sempervirens* (including Korbel KT), one *Metasequoia glyptostroboides*, one *Thuja occidentalis*, and one *Fitzroya cupressoides*. By including multiple members of the same species and some more distantly related members of the Cupressaceae, we could identify which markers are variable within vs. among species, and see how much sequence divergence is tolerated by this method (i.e. how broad, phylogenetically speaking, are the baits we designed?) As the baits are based on transcriptome data, they only cover exonic (coding) regions. The resulting sequence data, however, includes both exons and their flanking introns. As genomic DNA was broken down into fragments, many of these fragments that are captured during hybridization and subsequently sequenced will contain introns adjacent to an exon bait.

Libraries of genomic DNA for the aforementioned eight accessions were hybridized with baits covering 908 genes. Subsequent hybridized "enriched" libraries were sequenced on an Illumina MiSeq with paired-end 300 base pair reads. Similar to the transcriptome analyses, raw sequencing reads were demultiplexed and filtered to remove sequences with poor quality scores. Sorted reads were then assembled into contigs, using transcriptome sequences as a reference. Based on preliminary analyses from the test panel, a second set of baits were designed, removing targeted sequences overrepresented in the sequence data. The second round of targeted sequence capture will result in sequences evenly distributed across targets, and will allow for more thorough reconstruction of copy variants (alleles or duplicate genes) for each target. We anticipate new targeted sequence results in Spring 2015.

Results

Transcriptome sequencing of the three redwoods yielded molecular sequence data for up to 70,000 expressed sequences per species. A preliminary analysis based only on genes with a single detected copy per species showed that, for a great majority of genes, copies from *S. sempervirens* and *S. giganteum* are sister (Figure 1).

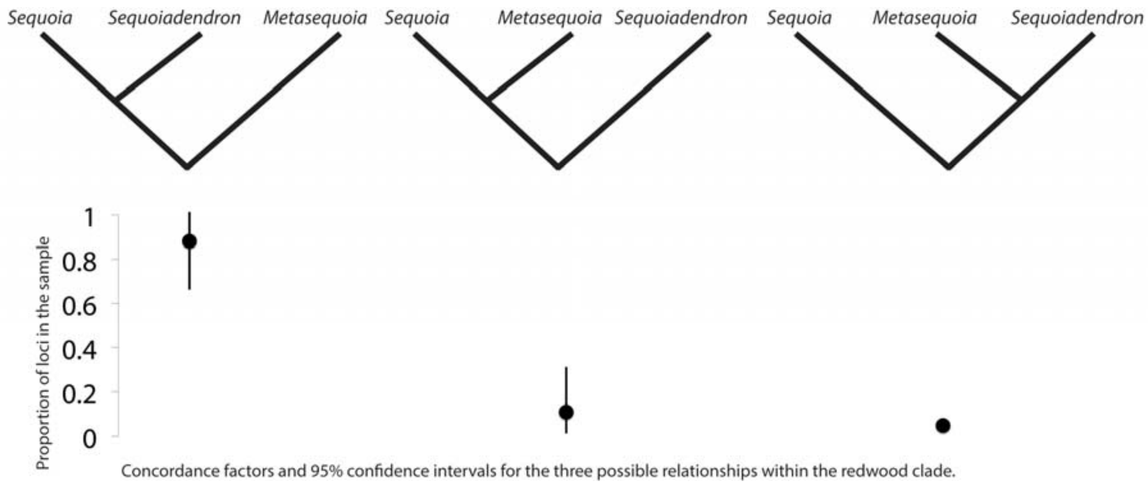

Figure 1: Preliminary condordance analysis based on twenty orthologs present in a single expressed copy. Phylogenetic trees were constructed with MrBayes for each gene. Bayesian concordance analysis with BUCKy combines data from multiple genes and estimates how much of the genome supports each relationship, based on the sampled individuals.

This pattern implies either strict autoployploidy (polyploidization within a single lineage), allopolyploidy involving hybridization with genome donors within the California redwood clade, or autoallopolyploidy combining the two mechanisms. Discordance among gene trees can be explained by incomplete lineage sorting.

Discussion/ Implications for conservation

Interestingly, the data so far do not support previous hypotheses of hybridization with distant relatives of *Sequoia*, such as suggested genome donors *Taxodium* and *Cryptomeria*. Instead, our data suggest that polyploidization happened, at its broadest, within the California redwoods. To be sure, we have to construct gene trees that include multiple copies (homeologs) from hexaploid *Sequoia*. So, to further clarify polyploidization mechanisms, we rely on a subset of genes chosen for targeted sequencing of genomic DNA to isolate all genome copies from *S. sempervirens*. Subsequent phylogenetic comparison among gene copies in *S. sempervirens* and its close relatives will reveal whether hybridization with an extant lineage contributed to polyploidy in the *Sequoia* lineage.

These same markers, useful for answering evolutionary history questions, also have applications for conservation. We consider genetic diversity a crucial element in conservation planning. The genomic data we generated from *S. sempervirens* and *S. giganteum* have great potential for the development of gene-based markers, such as the targeted sequences we apply here. Using gene-based markers permit us to look not just at neutral genetic diversity (in the introns) but also to potentially identify adaptive variation. One application of these markers is a current collaboration between A. Scott and R. Dodd at UC Berkeley, which uses targeted sequence capture to identify genetic variation among groves of giant sequoia. In particular, we are targeting sequences that may be involved in drought tolerance, as we anticipate these genes are variable due to different climatic conditions across the Sierra Nevada. The estimation of adaptive genetic potential from genetic markers could serve as an alternative to the common garden approach, helping to identify groves of *S. giganteum* that could serve as seed sources for seed banks and plantations. Future applications could look at genes involved with water management in coast redwood, important in light of the recent fog reductions.

Expenditures

The majority of expenses were in sequencing, library preparation, and RNA bait synthesis ($10,847). Additional expenses included reagents and consumables for RNA and DNA extraction ($2,267) and computational hardware and data storage/service fees ($1,859). The cost of the second "filtered" round of RNA baits exceeded the grant balance, and was subsidized by departmental funding to A. Scott (~$1,100).