Progress on Enhanced EST Resources for Loblolly Pine (*Pinus taeda* L.) and Other Conifers

W. Walter Lorenz[1], Virgil E. Johnson[1], Saravanaraj Ayyampalaym[2], Chun Liang[3], and Jeffery F. D. Dean[1]

A project is underway in collaboration with the US Department of Energy Joint Genome Institute (JGI) to deepen coverage of the loblolly pine transcriptome and also provide expressed gene sequence information for conifers in other phylogenetic clades. In an initial study, normalized and non-normalized cDNA libraries prepared using shoot tissues pooled from three loblolly pine genotypes were analyzed in two runs of a GS-FLX sequencer, which yielded approximately 876,000 reads and 200 Mb of new sequence. *De novo* assemblies of the transcriptome data from the two individual libraries, as well as the combined datasets, were performed using the SeqMan NGen assembler, and the results were compared to assemblies generated by the JGI bioinformatics pipeline, as well as the miraEST assembler. Additional assemblies were generated by combining the GS-FLX datasets with the Sanger-based ESTs available in GenBank, and also by assembling the new sequence to templates from the unigene set identified by NCBI for loblolly pine. Analyses suggested that normalization of the cDNAs had a tendency to skew representation of certain gene families more than others, but overall increased the rate of new gene discovery. The new assemblies will soon be made available via the new release of ConiferEST (Liang et al. 2007; http://www.conifergdb.org/coniferEST.php).

Future plans for the CSP project include production of an additional 1.5 Gb of loblolly pine cDNA sequence information using the GS-XLR (Titanium) platform, as well as an additional 5.5 Gb of transcribed sequence data from 11 other conifer species. When completed, this project should yield a reasonably comprehensive picture of the true depth and complexity of the loblolly pine transcriptome, and will open many new opportunities to answer important fundamental questions about conifer genomes and phylogenies.

MATERIALS AND METHODS

Conifer tissues were collected and total RNA prepared as previously described (Lorenz et al. 2006). Normalized and non-normalized shoot tip cDNA libraries were prepared by JGI personnel using the SMART PCR cDNA synthesis (Clontech, Mountain View, CA) and Trimmer normalization (Evrogen JSC, Moscow, Russia) kits. DNA sequencing on the GS-FLX platform (Roche Applied Science, Indianapolis, IN) was performed at JGI following standard protocols.

Sequence datasets for the non-normalized (CFCP) and normalized (CFCN) cDNA libraries were assembled separately, together, and in combination with ca. 228,000 *P. taeda* Sanger ESTs downloaded from GenBank. Initial assemblies were performed using the mer-based Seqman NGen, Ver 2.0 (DNASTAR Inc., Madison, WI), and parameter settings were determined after iteratively processing multiple assemblies of the combined dataset and varying minimum match identity percentage, match size, and match spacing. Other settings were left at the recommended

[1] Warnell School of Forestry and Natural Resources, University of Georgia, Athens, GA
[2] Department of Plant Biology, University of Georgia, Athens, GA
[3] Department of Botany, Miami University, Oxford, OH

default values. Data from the GS-FLX was initially trimmed for adaptor and polyA sequences by the JGI bioinformatics pipeline; however, additional quality trimming was performed in NGen after analysis of the initial assemblies. Comparative assemblies of the GS-FLX data and the GS-FLX plus Sanger data were performed using miraEST (Chevreux et al. 2004) at the default settings. Templated assemblies were created using NGen to run the combined dataset against the NCBI *P. taeda* Unigene set. Contigs containing >1 read were compared to the NCBI NR database using BLASTX. Returns from the BLASTX analysis were mapped and annotated for gene ontology (GO) terms using Blast2GO (Conesa et al. 2005).

RESULTS AND DISCUSSION

First-flush candle tissues (elongating apical shoot tips) from three loblolly pine genotypes (CCLONES 40430, 40368, and 41586) were pooled for this pilot study. These same genotypes were previously used in an NSF-funded project that created an EST collection from transcripts expressed in loblolly pine roots undergoing various biotic and abiotic challenges. As shown in Table 1, roughly equivalent numbers of reads were returned from a full-plate GS-FLX analysis of each library. Average read lengths from this platform were about 250 nucleotides, so approximately 105 Mb of total sequence was obtained for each library.

Assembly	Total Seqs	Assembled Seqs	Contigs >1 EST	Ave. Length	ESTs/ Contig	Contigs >2kb	Largest Contig	% Assembled
CFCP (N)	464,532	391,212	37,690	456	10	122	4,002	84.20
CFCN (N)	411,672	347,453	37,488	460	9	10	2,420	84.40
CFCP (J)	464,708	419,852	32,091	478	13	257	4,167	90.00
CFCN (J)	411,558	363,781	33,562	470	11	11	2,405	88.40
Comb. (N)	876,198	756,253	57,303	486	13	209	3,998	86.30
Comb. (M)	876,198	793,559	56,440	501	14	425	4,814	90.70
Mapped (N)	1,104,078	945,058	65,647	707	14	1244	4,581	85.60
Mapped (M)	1,104,450	991,914	69,214	760	14	1625	4,824	90.00

Table 1. Statistics for *de novo* assemblies compiled using different assembler programs. Input datasets included non-normalized (CFCP), normalized (CFCN), non-normalized plus normalized (Comb.), and non-normalized plus normalized plus GenBank sequences (Mapped). Assembler programs included NGen (N), miraEST (M), and the JGI pipeline (J).

As expected, mapping *de novo* assemblies to the existent Sanger-based EST sequence in GenBank increased the average number of reads per contig and significantly increased the contig average length. However, under these conditions, the number of contigs containing more than one EST continued to increase and nearly 10% of the reads remained as singletons. This data confirms that pines manifest a large and diverse transcriptome, and a significant amount of the transcribed gene space remains poorly assessed. However, three additional loblolly pine cDNA preparations will be sequenced using the GS-XLR platform at JGI as part of this project (Table 3), and the data resulting from those analyses will provide us with a much clearer understanding of the depth and complexity of the pine transcriptome.

Normalization procedures (e.g. Bogdanova et al. 2008) are often used in the production of cDNAs for gene discovery efforts as a means to increase the rate of discovery by minimizing the sequence redundancy. However, it has been widely reported that the gene families in pine are larger and more highly conserved than what is typically seen for gene families in angiosperm plants (Ahuja and Neale 2005), which raised concerns that normalization could impact the representation in certain gene families more than others. For this pilot study, JGI personnel prepared two libraries from the same RNA preparation, one using normalization procedures and the other not, so that we could make an attempt to judge whether normalization would have negative consequences on pine gene family representation.

Gene Description	CFCN BLAST Hits	CFCP BLAST Hits	Fold Difference
Actin/Actin-related	8	17	1.9
Aquaporin	3	9	2.7
Cellulose synthase/like	26	46	1.6
Chalcone synthase	7	20	2.5
Chlorophyll a/b-binding protein	4	17	3.8
Cyclophylin	1	10	8.9
Dehydrin	13	23	1.6
Glycine-rich RNA-binding protein	1	10	8.9
Histones 3&4	35	30	0.8
Homeobox	14	15	1.0
Myb/Myb-related	9	12	1.2
NBS/LRR	13	42	2.9
Sucrose synthase	3	14	4.1
Ubiquitin/Ubiquitin-related	17	50	2.6
WRKY	2	6	2.7
Zinc-finger	16	23	1.3

Table 2. The impact of cDNA library normalization on new gene discovery and the potential for skewing by gene family.

Tentative annotations for assemblies from the two individual libraries were obtained using a BLASTX comparison with the NCBI NR database, and assemblies returning the same annotation classes ($<1x10^{-20}$ e-value) were roughly binned as possible members of the same gene family. A selection these general gene family bins comprising classes either known to have extensive family structure or that were identified as transcription factors were selected for further examination. Table 2 shows the data recovered for a selection of annotation classes having functions of wide interest to conifer researchers and for which the numbers of returned hits were large enough that differences between the two libraries might be considered significant. For most annotation classes there was little evidence for the type of skewing that might be of concern, and as evidenced by the lower numbers of hits for the CFCN library, normalization did change the cDNA representation in a manner that should result in increased rates of gene discovery. However, it can also be seen that certain annotation classes (e.g. cyclophilins and glycine-rich RNA-binding proteins versus histones 3&4) behaved very differently, and it remains to be seen whether this will be reflected in the gene family structures that are eventually recovered.

ConiferGDB (http://conifergdb.org) is an online resource for investigating gene structure and transcript variation, and the current version displays contig assemblies for the Sanger-based EST sequences from loblolly pine that have been deposited in GenBank (Fig. 1). The GS-FLX reads from JGI are currently being mapped to new and existing assemblies in ConiferGDB and will be released for public viewing and use in the near future.

Figure 1. Screenshot of the ConiferGDB Viewer displaying an assembly of Sanger ESTs from GenBank.

The original proposal submitted to the JGI Community Sequencing Program (CSP) for an expanded conifer EST resource requested sequencing of materials from 23 species of conifers and would have resulted in an estimated 1.2 Gb of sequence information. As JGI has transitioned away from Sanger-based DNA sequencing platforms the final project workplan was revised to include fewer species (12 in addition to *P. taeda*), but will yield substantially more sequence information (ca. 5.5 Gb). Table 3 lists the species targeted for the revised project and the expected minimum sequence yields based on current technology. Each major conifer family (Araucariaceae, Cephalotaxaceae, Cupressaceae, Pinaceae, Podocarpaceae, Sciadopityaceae, Taxaceae) is represented by at least one species, and *Gnetum gnemon* from the Gnetaceae has been included in an effort to generate additional data with which to test the molecular phylogenies suggesting evolutionary relationships that place the Gnetales as a sister group to

Pinaceae. All sequencing work in this project should be completed in the Fall of 2009, and hopes are that reliable gene assemblies can be made publicly available by the end of the year.

Sample Name	Species	Platform	Source Tissues	Expected Yield
PtaVasc	*Pinus taeda*	GS-XLR	Multiple vascular tissues	500 MB
PtaSeed	*Pinus taeda*	GS-XLR	Perturbed seedlings	500 MB
PtaSusp	*Pinus taeda*	GS-XLR	Perturbed suspension cells	500 MB
PmeShoot	*Pseudotsuga menziesii*	GS-XLR	Mixed shoot tissues	500 MB
PpaShoot	*Pinus palustris*	GS-XLR	Mixed shoot tissues	225 MB
PlaShoot	*Pinus lambertiana*	GS-XLR	Mixed shoot tissues	225 MB
PabShoot	*Picea abies*	GS-XLR	Mixed shoot tissues	225 MB
CatShoot	*Cedrus atlantica*	GS-XLR	Mixed shoot tissues	225 MB
SseShoot	*Sequoia sempervirens*	GS-XLR	Mixed shoot tissues	225 MB
ChaShoot	*Cephalotaxus harringtonia*	GS-XLR	Mixed shoot tissues	225 MB
SveShoot	*Sciadopitys verticillata*	GS-XLR	Mixed shoot tissues	225 MB
PmaShoot	*Podocarpus macrophylla*	GS-XLR	Mixed shoot tissues	225 MB
TbaShoot	*Taxus baccata*	GS-XLR	Mixed shoot tissues	225 MB
WnoShoot	*Wollemia nobilis*	GS-XLR	Mixed shoot tissues	225 MB
GgnShoot	*Gnetum gnemon*	GS-XLR	Mixed shoot tissues	225 MB

Table 3. List of conifer species and sequencing targets remaining for the JGI Conifer EST project.

REFERENCES

Ahuja MR, Neale DB (2005) Evolution of genome size in conifers. Silvae Genet 54: 126-137

Bogdanova EA, Shagin DA, Lukyanov SA (2008). Normalization of full-length enriched cDNA. Mol BioSystems 4: 205–212

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai S (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 14:1147-59

Liang, C., Wang, G., Liu, L., Ji, G., Fang, L., Liu, Y., Carter, K., Webb, J.S. and J. F. D. Dean (2007) ConiferEST: an integrated bioinformatics system for data reprocessing and mining of conifer expressed sequence tags (ESTs). BMC Genomics, 8:134.

Lorenz WW, Sun F, Liang C, Zhao X, Kolychev D, Wang H, Cordonnier-Pratt M-M, Pratt LH, Dean JFD (2006) Water stress-responsive genes in loblolly pine (*Pinus taeda* L.) roots identified by analyses of expressed sequence tag libraries. Tree Physiol. 26:1-16