Original Paper

Analysis and functional annotation of expressed sequence tags from the diesel tree (*Copaifera officinalis*)

Sam R. Zwenger, Robert E. Reinsvold, Chhandak Basu*

School of Biological Sciences, 2480 Ross Hall, University of Northern Colorado, Greeley, CO, USA
* E-mail: chhandak.basu@unco.edu Tel: +1-970-351-2716 Fax: +1-970-351-2335

Received March 24, 2010; accepted June 28, 2010 (Edited by D. Shibata)

Abstract Copaiba (*Copaifera officinalis*) is a tropical plant that is also known as the 'diesel tree', previously noted for production of diesel-like oleoresin. The advancements in molecular tools such as expressed sequence tag (EST) library construction provide a novel opportunity for insight into the physiology of this tree. We generated a small set of ESTs for a young copaiba, sequencing and annotating a total of 613 unigenes. Of these, 84% showed similarity to the National Center for Biotechnology Information (NCBI) database. Annotation showed 70% of unigenes had at least one associated Gene Ontology (GO) term. We found a majority of ESTs to be associated with heat response genes. Based on these data, this EST library of *C. officinalis* represents a small but important step in helping to understand the general physiology and heat-response expression patterns. Additionally, this small collection of EST offers a modest starting point in helping to understand this enigmatic tropical plant.

Key words: Biofuel, *Copaifera,* diesel tree, EST library, heat stress.

The importance of tropical plants and their diversity in biochemical products is difficult to overstate (Russell et al. 2008). Many tropical plants have been used for agricultural purposes such as banana (*Musa acuminata*), papaya (*Carica papaya*), chocolate (*Theobroma cacao*) and sugar cane (*Saccharum* spp.). Additionally, important medicinal compounds have been discovered in tropical plant species such as periwinkle (*Catharanthus roseus*), digitalis (*Digitaria purpurea*) and cocaine (*Erythroxylum coca*).

Currently, there are many examples of commercial and industrial products that were either originally discovered in or are presently derived from tropical plants. This is in part due to the great diversity of plant species located in the diverse ecological niches in tropical environments (Mittermeier et al. 1998).

Nucleotide sequences for many tropical plants have been obtained, however most information is limited to food crops. For example, the papaya genome and sugarcane genomes have been sequenced. More insight into the physiology of tropical plants can be provided with expression profiling, even if a small number of expressed sequence tags (ESTs) are generated. Acquiring sequence information is also important to detect similarity of an understudied species among species for which sequence information currently exists.

An interesting tropical plant that deserves molecular investigation and lacks sequencing information is the genus *Copaifera* (Fabaceae), whose members include copaiba (diesel tree). Little molecular data (i.e., ESTs) have been published on this particular genus. With a better understanding of the genetic underpinnings of metabolic pathways in this species, we may be better able to understand the development and physiological conditions that lead to increased production and accumulation of the diesel-like resins (Calvin, 1980).

In addition to the fuel-like properties, the oleoresin has been used for numerous medicinal purposes (Coussio et al. 2001; Paiva et al. 2007). While its medicinal benefits might be garnered for a small proportion of people, it is unrealistic to harvest the oleoresin from these trees on a large scale. This hinders the use of the oleoresin for large populations for fuel purposes. To mitigate conflicts of sustainability, we sought genes from this tree to later express in algae or plants, which might be used to produce biofuels. To prepare for these downstream applications we had the *Copaifera* sequences ligated into the plant expression vector pCHF3 (supplementary file 1).

The sequences were searched against the National Center for Biotechnology Information (NCBI) database. Very few sequences were found pertaining to oleoresin production. To our surprise many sequences matched heat stress-related genes. A total of 613 unigenes, representing 716 ESTs were annotated according to Gene Ontology terminology (supplementary file 2).

This article can be found at http://www.jspcmb.jp/

Statistical analysis for each GO term (supplementary file 3) and InterProScan results (supplementary file 4) are also provided.

Materials and methods

Library creation and cDNA sequencing

Seeds from *C. officinalis* were obtained from the University of Puerto Rico, San Juan. These were then germinated in approximately 0.5 cubic meter of Miracle Grow, 5–7–5 potting soil (Scotts, USA) in controlled conditions (30°C, 80% relative humidity, ~630 lux) at the University of Northern Colorado, Greeley, Colorado, USA. Leaf material (leaf and stem) was harvested from 18-month-old trees, snap frozen and ground in liquid nitrogen. Total RNA was extracted using Trizol Reagent (Invitrogen, USA) according to manufacturer's protocol and shipped on dry ice to Advanced BiotekServices (San Diego, USA) for 'construction of uncut and directionally cloned expression library'. cDNA inserts were ligated into the plant expression vector pCHF3 (Borevitz et al. 2000), which has the CaMV35S promoter and spectinomycin and neomycin selectable markers (Supplementary file 1; drawn using PlasMapper (Dong et al. 2004)). Library complexity was estimated to be 2.1×10^6 (data provided by Advanced Biotekservices). Single pass sequencing of the cDNA clones were performed with 96 well plates on the ABI 3730 xl sequencer by Lucigen Corporation (Middleton, Wisconsin, USA). Sequences were obtained by using the primer

5′-TTACAAGCACAACAAATGGTCAAGAA-3′.

cDNA sequence annotation

A total of 1008 sequences were analyzed with Sequencher software (Gene Codes Corporation, USA) using default settings. The pCHF3 vector ends were removed from the sequence of interest by comparing partial flanking vector sequences with the trace files. For this we used a minimum overlap of three bases and 80% minimum match. To improve quality of sequences, end trimming was performed with removal of uncalled bases at 3′ and 5′ ends. Further manual curation was not performed. A total of 716 ESTs were used for assembly and library analysis and only ESTs that were >200 bp were used for further analysis. The resulting ESTs were assembled into 545 fragments and 68 contigs, which yielded 613 unigenes. Contigs consisted of 2 ESTs (50 contigs), 3 ESTs (12), 4 ESTs (2) and 5 ESTs (2). The two largest contigs were composed of 9 and 8 ESTs, respectively.

Sequence similarity search by the blastx program was performed against the database of non-redundant protein sequences provided by NCBI (nr). In order to facilitate batch handling of sequence data the Blast2Go (B2G) software suite (http://www.blast2go.org/) was implemented (Götz et al. 2008). The B2G software implements batch blast, mapping (retrieving GO terms associated with each blast hit), Gene Ontology annotation (giving functional terms to each query) based on their function, statistical testing, and InterProScan tools. For our analysis a cutoff E-value of 1.0E-5 was used to select for significant matches, although nearly all sequences were below this cutoff value (see supplementary file 3).

Results

A total of 716 ESTs with an average insert size of 1.5 kb were used to form 68 contigs and 545 singletons. 84% of unigenes showed positive hits with the NCBI database. In general, shorter sequences tended to not match database sequences. The top hits for blastx (as of Jan. 5, 2010) results are shown in Figure 1. There were 143 matches to *Vitis vinifera* and 137 top hits for *Glycine max*. The next three top hits for species distribution was *Ricinus communis, Populus trichocarpa* and *Medicago trunculata*. It is difficult to determine if the species distribution for copaiba sequences was truly reflective of sequence similarity or if it largely depended on the number of sequences for that species within the NCBI database.

In most cases, fewer than 5 copaiba sequences were found to have top hit matches to any single species. In most cases the unigenes could be mapped and annotated (Figure 2). However, 89 sequences did not have significant matches from blastx and therefore could not be mapped or annotated. Of the sequences with blastx hits, 31 could not be mapped and 61 could not be annotated. Sequences with annotations were assigned gene ontologies according to biological process, molecular function and cellular component (Figure 3). For gene products a higher level indicates a more specific description of gene ontology terms. A higher level of annotation therefore corresponds to a more specific description of gene products. Copaiba translated sequences were distributed across many levels of GO annotations. A total of 2,046 annotations were made with a mean level of 4.92. More than 275 annotations were given for GO level 5 for the cellular component and less than 90 were given for molecular function.

The GO level 2 was used in annotating our data and constructing pie charts in Figure 4. Each category of biological process, cellular component, and molecular function has associated parent-child terms, which places a particular translated EST. A total of 839 annotations were observed as a biological process. More than half (58%) of these were annotations for either a metabolic or cellular process. 12% (103) annotations were related to a response to stimulus while 6% (51) and 3% (24) of annotations were for developmental processes and regulating biological processes, respectively. Less than 1% (7) annotations were for growth.

A total of 939 annotations were observed for cellular component. More than 90% (852) of these were associated with one of three components; organelle (238), cell (308) and cell part (306). About 2% (20) were located in an extracellular region and 3% (29) were associated with a macromolecular complex.

Sequences annotated as having molecular function consisted of eight categories with a total of 535 annotations. Nearly half (46%) of these were observed as having some type of binding (e.g. nucleotide binding) activity. A large portion (36%) was annotated as having catalytic activity. The remaining six annotation categories were for structural molecule activity (30), translation regulator activity (15), transporter activity (25), molecular transducer activity (5), transcription regulator activity (17) and enzyme regulator activity (4).

Discussion

In GO terminology, strict vocabulary is used to place protein names with more specificity (termed 'child') below a less specific protein category (termed 'parent'). This not only allows for rigorous control of language but also aids in relating each sequence to a GO match of a specific category. Further, the three main GO categories

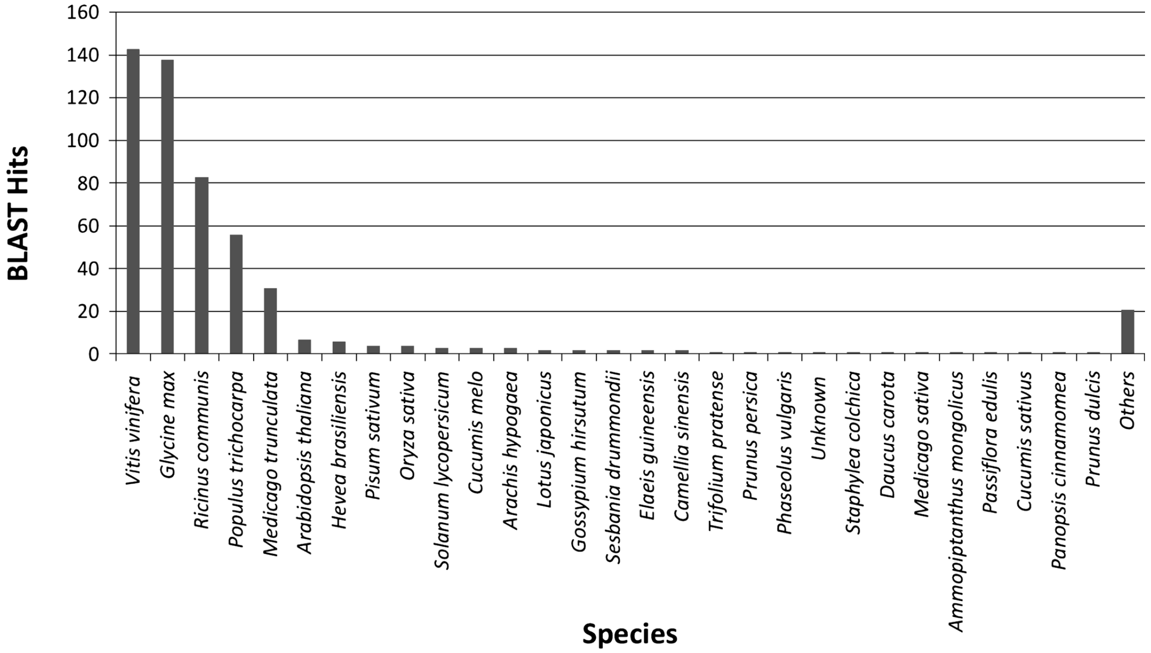

Figure 1. Top blastx hits for copaiba sequences arranged according to their frequency of blastx hits. Most sequences were closely related to *Vitis vinifera*, *Glycine max*, and *Ricinus communis*.

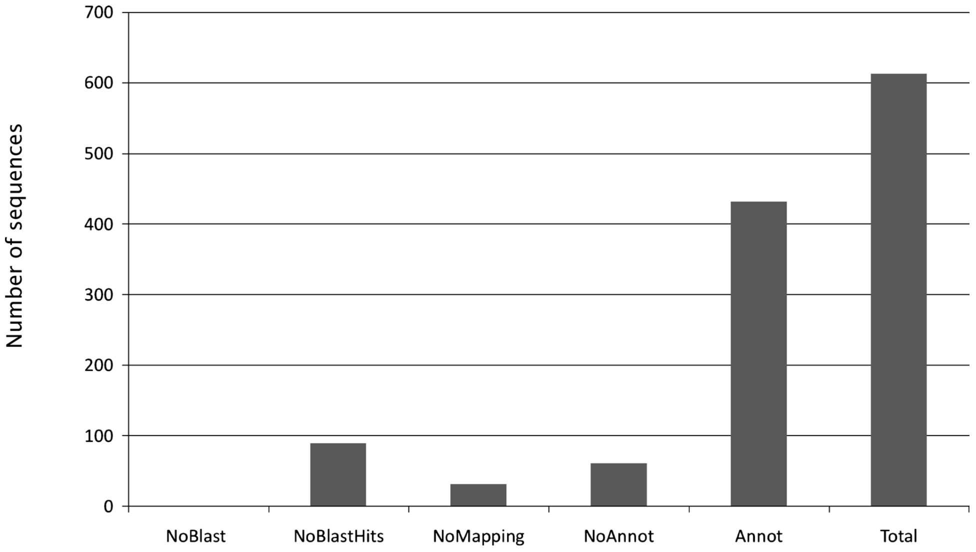

Figure 2. The distribution of gene ontology annotations varied, with more than half matching with previously annotated sequences. Mapping consists of retrieving GO terms associated with each blast hit and annotation is the process of giving functional terms to each query (NoBlast: no blast was performed; NoBlastHits: sequences returning no blast hits; NoMapping: mapping step could not be performed; NoAnnot: no annotation could be performed; Annot: sequences that could be annotated; Total: total number of all sequences for which blast, mapping, and annotation were performed).

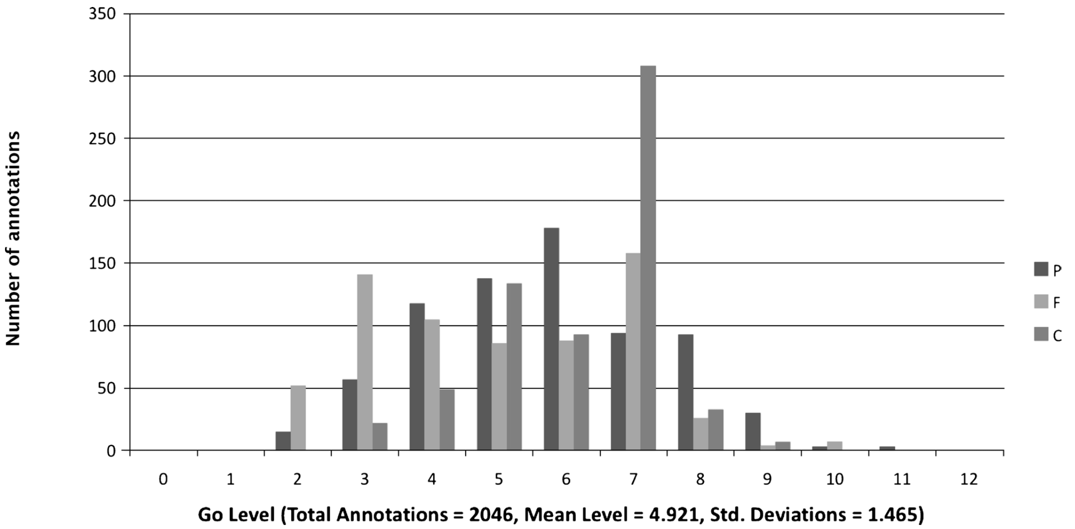

Figure 3. GO level distribution for annotations of copaiba unigenes. A total of 2,046 annotations were given across all GO domains (P=biological process, F=molecular function, and C=cellular component). A single GO term can be given to more than one sequence, and hence 2,046 terms were given to 716 sequences. The GO terminology provides a description for gene products with higher levels providing more specific terms. (Note: Figures 1, 2 and 3 were redrawn based on the Blast2Go software data and figures output).

(cellular component, biological process, and molecular function) can show repeated annotations of the same sequence.

The EST library presented in this paper provides limited yet interesting insight into the transcriptional composition of *C. officinalis*. Contig0007 consisted of 8 ESTs and closely matched (E-value 3E-65) an early light inducible protein. This is important since the abundance of an EST can help understand relative proportions of transcripts. Hits associated with other fundamental metabolic activities were also observed. The singleton P7G08 was closely related (E-value 2E-16) to a enoyl-acyl carrier protein reductase. The enoyl-acyl carrier protein reductase is a crucial accessory enzyme for *de novo* fatty acid synthesis. The acyl carrier protein (ACP) is a key cofactor in a number of acetyl condensation reactions, but ACP must be continually reduced in order for the reactions to proceed (Slabas and Fawcett 1992). Enoyl-ACP reductase acts as the terminal enzyme of ACP reduction, producing a saturated acyl-chain that can continue in the fatty acid condensation reaction. The insertion and expression of the *Brassica napus* enoyl-ACP reductase gene into *E. coli* showed the feasibility of using this gene as a target for fatty acid biosynthesis manipulation (Kater et al. 1991).

A major finding of this work was that when grown at 30°C and ~630 lux, copaiba expresses a multitude of heat stress-related genes. This was based on blastx hits and the fact that 103 of the biological process GO annotations were associated with response to stimuli (e.g. abiotic factors). The false discovery rate (FDR) for 'response to abiotic stimulus' (5.32E-08) and 'response to stress' (1.53E-07) indicated an overrepresentation of these terms from the *Copaifera* test set when compared to the *Arabidopsis* reference set (see supplementary file 4). This is surprising since most tropical plants are accustomed to living at these (or higher) temperatures. A contributing factor might have been the humidity of the growth chamber, which tends to decrease water transpiration from stomata.

Many contigs generated closely matched heat shock proteins. Contig0010, which had a high similarity to a heat shock protein (3E-70) consisted of nine ESTs. Contig0022 and contig0031 consisted of three ESTs and contig0049 consisted of two ESTs and all generated similarity hits to heat shock proteins. Contig0067 and contigs0071 both consisted of two ESTs and were observed to have blastx hits matching heat shock proteins. A large amount of singletons also had blastx hits relevant to heat stress or matching heat shock proteins.

Response to abiotic stimulus was not limited to heat stress related genes. Response to oxidative stress was observed with multiple blast hits that matched cytochrome c oxidase and peroxidase. Contig0043 showed high similarity (1E-109) to the ascorbate peroxidase in *Citrus maxima*. Ascorbate peroxidase (AsPX), an enzyme unique to plant and algae, functions as a vital protectorate from highly reactive hydrogen peroxide and hydroxyl radicals, particularly within the chloroplast (Asada 2006). Higher levels of AsPX in fruit have been associated with longer post harvest shelf life and increased antioxidant activity (Lester and Hodges 2008).

Gramosa and Silveira (2005) performed a GC/MS analysis on *C. officinalis* specimens from Brazil. They found more than 40 different compounds, with an abundance of γ-muurolene, and β-caryophyllene in the leaves. They captured a different terpene profile from separate parts of the plant (e.g., leaves, seed, bark, etc.). Since our library was taken from developing leaf and stem, we expected to find correlating terpene synthase expression. However, this was not necessarily the case and may have been partly due to small library size and age of the specimen tissue was collected from.

In some cases it is important to closely examine blast hits for less similar hits. For example, although no terpene synthase matches were observed, we did observe the presence of a sequence similar to the rubber elongation factor (REF), previously described in the rubber tree (*Hevea brasiliensis*) (Priya et al. 2006). This protein facilitates the interaction between a growing *cis*-polyisoprene unit (i.e., rubber), from isopentenyl pyrophosphate (IPP), and a prenyltransferase (Dennis and Light 1989). REF is believed to affect the stereochemistry of IPP, allowing for a *cis* addition to the growing *cis*-polyisoprene units. Its absence prevents further addition of IPP onto *cis*-polyisoprene by prenyltransferase (Dennis and Light 1989). The presence of a potential homolog to REF in *C. officinalis* provides key insight into possible molecular pathways involved in the production of large terpenes seen in extracted oleoresin.

Some researchers have taken on broader studies of metabolite production. Medeiros and Vieira (2008) have provided work on a related species, *C. multijuga*, which they found abiotic and biotic factors contributed to oleoresin production (e.g. termites, age and size of tree). A separate study found little relationship between soil types (variations in nitrogen and moisture) and leaf sesquiterpene content in leaves from *C. multijua* (Nascimento and Lagenheim, 1986). Similar results were found by Feibert and Langenheim (1988). However, they attribute higher sesquiterpenes in leaves mostly to herbivory (response to stimuli).

Although the *Copaifera* cDNA library sheds little insight into the biochemical pathways and genes involved in oleoresin synthesis, this library might be important in understanding *Copaifera* developmental and heat stress-related gene expression. This might allow for future research to incorporate novel heat stress-related genes transgenically expressed in non-*Copaifera* species. Based on the diversity of sequences obtained from this tropical tree, further research may be able to include studies on heat, drought, and other predicted climatic responses on tropical species.

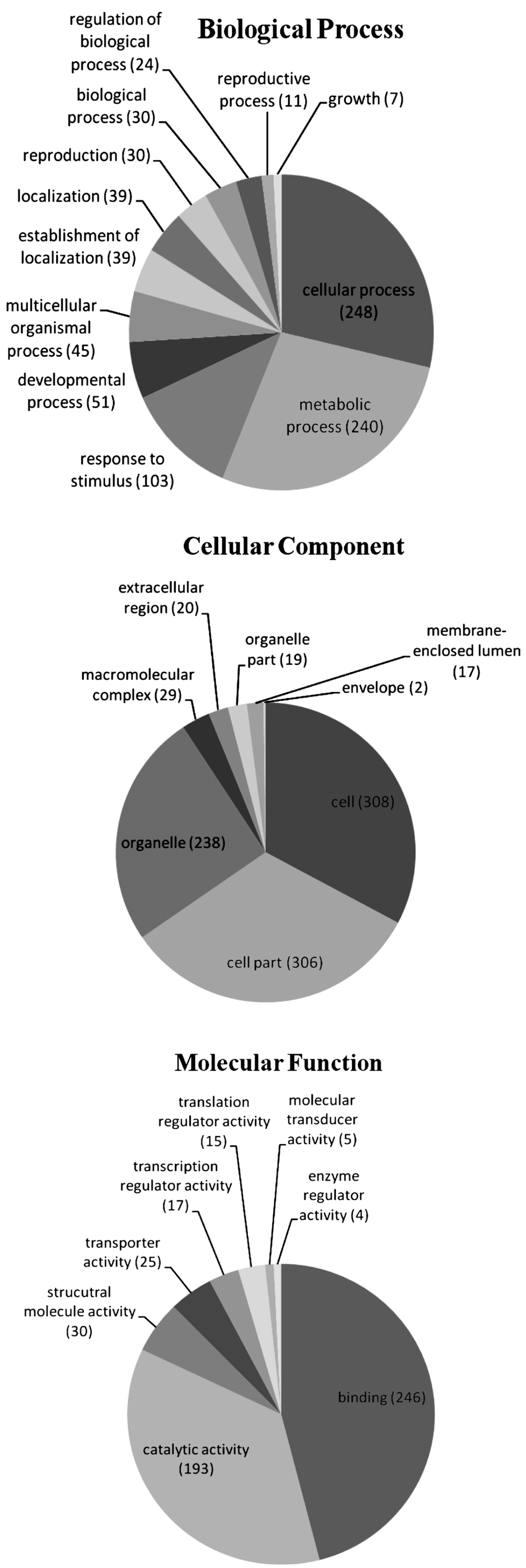

Figure 4. Distribution of sequences when compared to GO database. All blastx and GO results are provided in supplementary file 2. More than half of the biological processes category consisted of annotations for cellular or metabolic processes. A large portion of unigenes categorized as cellular component were either cell, cell part or organelle. The molecular function category consisted of relatively few annotations for enzyme regulator activity and transporter activity.

Construction of a cDNA or EST library provides information on the transcriptional state of an organism. Millions of plant ESTs have already been sequenced and uploaded to http://www.ncbi.nlm.nih.gov/dbEST/, providing researchers with an inordinate amount of genomic data, without the burden and expense of full genome sequencing. They have been important in understanding plants in relation to invasive species (Wang et al. 2006), weediness (Broz et al. 2007), (Anderson et al. 2007), biotic stress responses (Zhang et al. 2007), and flowering times (Carlson et al. 2006). Copaiba ESTs are lacking in this database and the research presented here might contribute to filling this gap. Copaiba is a tropical tree and primarily found in warm and humid regions. To understand the oil production in the tree, it is essential for us to understand the responses of the tree in heat stress. We hope, the heat stress related ESTs identified in this project will be beneficial for scientists to pursue further research on copaiba stress physiology and oil production in a changing environment.

Conclusions

The copaiba library presented here was ligated into the plant expression vector pCHF3 and sequencing plates were obtained for future work. All clones within this library are available to the scientific community. Ideal platforms for creating transgenic plants have been suggested before (Yuan et al. 2008). We suggest using a weedy species, which will not replace a food supply like the way corn ethanol does. Transgenic work on overexpressing these genes in *Arabidopsis* and other plants are currently underway in our lab. The complex biochemical pathways are one of the major hurdles in manufacturing terpenes in bacterial/plant cell cultures (Roberts 2007). Isopentenyl diphosphate and dimethylallyl diphosphate are limiting steps in further, more complex terpene synthesis in transgenic plants (Mahmoud and Croteau 2001).

Further research might also include full sequencing of copaiba genome and custom printing of a microarray slide based off the sequences from this library. A larger scale sequencing project may reveal information on genes in the terpene biosynthesis family and additional genes in the copaiba developmental pathways. Our hope is that the research presented here will contribute to understanding of the physiology of diesel-like resin production in copaiba. Since many plant metabolic pathways are being mapped, we thought the same should be done for pathways that generate potential biofuels. Future sequencing and analysis should focus on characterizing adult trees and their transcriptome, if the determination of novel terpene synthases is to be pursued.

Acknowledgements

This project would not have been possible without the help of Dr. James Ackerman at the University of Puerto Rico (UPR), San Juan, Puerto Rico. We also thank Dr. C. Neal Stewart Jr., University of Tennessee, who supplied plant seeds. We are thankful to Dr. Ana Conesa at the Bioinformatics Department of the Centro de Investigacion Principe Felipe, Spain for her help with the B2G program. This work was funded by Colorado Office of Economic Development through Colorado Bioscience Grant and the Graduate Assistance in Areas of National Need (GAANN) Fellowship from the U.S. Dept. of Education.

Data deposition

All copaiba ESTs are deposited in dbEST (The Expressed Sequence Tags database) and can be found at http://www.ncbi.nlm.nih.gov/dbEST/ (search term: copaifera; GW315375-GW316090).

References

Anderson JV, Horvath DP, Chao WS, Foley ME, Hernandez AG, Thimmapuram J, Liu L, Gong GL, Band M, Kim R, et al. (2007) Characterization of an EST database for the perennial weed leafy spurge: An important resource for weed biology research. *Weed Sci* 55: 193–203

Asada K (2006) Ascorbate peroxidase—a hydrogen peroxide-scavenging enzyme in plants. *Physiol Plant* 85: 235–241

Blüthgen N, Brand K, Cajavec B, Swat M, Herzel H, Beule D (2005) Biological profiling of gene groups utilizing Gene Ontology. *Genome Inform* 16: 106–115

Borevitz JO, Xia Y, Blount J, Dixon RA, Lamb C (2000) Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. *Plant Cell* 12: 2383–2394

Broz A, Broeckling CD, He J, Dai X, Zhao PX, Vivanco JM (2007) A first step in understanding an invasive weed through its genes: an EST analysis of invasive *Centaurea maculosa*. *BMC Plant Biol* 7: 25

Calvin M (1980) Hydrocarbons from plants: Analytical methods and observations. *Naturwissenschaften* 67: 525–533

Carlson JE, Leebens-Mack JH, Wall PK, Zahn LM, Mueller LA, Landherr LL, Hu Y, Ilut DC, Arrington JM, Choirean S, et al. (2006) EST database for early flower development in California poppy (*Eschscholzia californica* Cham., Papaveraceae) tags over 6,000 genes from a basal eudicot. *Plant Mol Biol* 62: 351–369

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles, M (2005) Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674–3676

Coussio JD, Ciccia GN, Silva GL (2001) Profisetinidin type tannins responsible for antioxidant activity in *Copaifera reticulata*. *Pharmazie* 56: 573–577

Dennis MS, Light DR (1989) Rubber elongation factor from *Hevea brasiliensis*. Identification, characterization, and role in rubber biosynthesis. *J Biol Chem* 264: 18608–18617

Dong X, Stothard P, Forsythe IJ, Wishart DS (2004) PlasMapper: a web server for drawing and auto-annotating plasmid maps. *Nucleic Acids Res* 32: W660–664

Feibert EB, Langenheim JH (1988) Leaf resin variation in

Copaifera langsdorfii: relation to irradiance and herbivory. *Phytochemistry* 27: 2527–2532

Götz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* 36: 3420–3435

Gramosa NV, Silveira ER (2005) Volatile constituents of *Copaifera langsdorffii* from the Brazilian Northeast. *J Essent Oil Res* 17: 130–132

Kater MM, Koningstein GM, Nijkamp HJ, Stuitje AR (1991) cDNA cloning and expression of *Brassica napus* enoyl-acyl carrier protein reductase in *Escherichia coli*. *Plant Mol Biol* 17: 895–909

Lester GE, Hodges D (2008) Antioxidants associated with fruit senescence and human health: Novel orange-fleshed non-netted honey dew melon genotype comparisons following different seasonal productions and cold storage durations. *Postharvest Biol Tec* 48: 347–354

Mahmoud SS, Croteau RB (2001) Strategies for transgenic manipulation of monoterpene biosynthesis in plants. *Trends Plant Sci* 7: 366–373

Medeirosa RS, Vieira G (2008) Sustainability of extraction and production of copaiba (*Copaifera multijuga* Hayne) oleoresin in Manaus, AM. *Braz Forest Ecol Manag* 256: 282–288

Mittermeier RA, Myers N, Thomsen JB, da Fonseca GAB, Olivieri S (1998) Biodiversity hotspots and major tropical wilderness areas: Approaches to setting conservation priorities. *Conserv Biol* 12: 516–520

Nascimento JC, Langenheim JH (1986) Leaf sesquiterpenes and phenolics in *Copaifera multijuga* on contrasting soil types in a central Amazonian rain forest. *Biochem Syst Ecol* 14: 615–624

Paiva LA, Gurgel LA, Silva RM, Tome AR, Gramosa NV, Silveira ER, Santos FA, Rao VS (2002) Anti-inflammatory effect of kaurenoic acid, a diterpene from *Copaifera langsdorffi* on acetic acid-induced colitis in rats. *Vascul Pharmacol* 39: 303–307

Priya P, Venkatachalam P, Thulaseedharan A (2006) Molecular cloning and characterization of the rubber elongation factor gene and its promoter sequence from rubber tree (*Hevea brasiliensis*): A gene involved in rubber biosynthesis. *Plant Sci* 171: 470–480

Roberts SC (2007) Production and engineering of terpenoids in plant cell culture. *Nat Chem Biol* 3: 387–395

Russell AM, Myers N, Thomsen JB, da Fonseca GAB, Olivieri S (1998) Biodiversity hotspots and major tropical wilderness areas: Approaches to setting conservation priorities. *Conserv Biol* 12: 516–520

Slabas AR, Fawcett T (1992) The biochemistry and molecular biology of plant lipid biosynthesis. *Plant Mol Biol* 19: 169–191

Wang YC, Yang CP, Liu GF, Jiang J, Wu JH (2006) Generation and analysis of expressed sequence tags from a cDNA library of *Tamarix androssowii*. *Plant Sci* 170: 28–36

Yuan JS, Tiller KH, Al-Ahmad H, Stewart NR, Stewart CN, Jr. (2008) Plants to power: Bioenergy to fuel the future. *Trends Plant Sci* 13: 421–429

Zhang J, Liu T, Fu J, Zhu Y, Jia J, Zheng J, Zhao Y, Zhang Y, Wang G (2007) Construction and application of EST library from *Setaria italica* in response to dehydration stress. *Genomics* 90: 121–131