Annotations to Reference Genome of Quercus Lobata

Tanveer Salim

MCDB 187AL

Major: MCDB, B.S.

Abstract

Oak trees are vital to the health of ecosystems, and serve as a valuable commodity in the hardwood industry. However, climate changes due to industrialization may disrupt the oak ecosystem. To provide more information on oak genomics, the Apollo genome-editing software was used to construct gene models from the reference genome of *Quercus lobata* to search for genes that allow the tree to adapt to climate change. Four gene models were identified in the region of the genome we studied. The peptide sequences were derived from the gene models, and using Blast, and at least six homologous peptide sequences were found. The peptide sequences of the gene model, together with the homologous sequences were used to construct a Multiple Sequence Alignment and Phylogenetic Tree. These were used to measure the degree of sequence conservation between the homologs and our gene model. Since all four peptide sequences derived from our gene models had excellent conservation with their homologs, we were confident that the gene models were constructed correctly. Secondly, the peptide sequences of our models were pasted to InterProScan to make inferences about their functions. Peptide sequences derived from three of the gene models were found to be affiliated with protein binding using InterProScan. A protein with 402 residues was found to be affiliated with DNA and RNA binding. Lastly a 230 amino acid sequence was found to be affiliated with calcium ion binding.

Background and Introduction

Valley Oak (Quercus lobata) is an economically, ecologically, and culturally valuable resource (Sork et. al 2016). Oak trees such as Valley Oak are used as inexpensive, renewable raw material for hardwood lumber (Luppold and Bumgardner 2013) found throughout hunting and range territories (Standiford and Howitt 1993; Kroeger et al. 2010). Ecologically, Valley Oak is also a source of nutrition for wildlife in those regions (Dahlgren et al. 1997; Herman et al. 2003). Oak ecosystems stabilize the habitats of aquatic and terrestrial animals (Dosskey et al. 1997; Kroeger et al. 2010). Culturally, oak trees are used to beautify urban areas and as a source of food by native cultures (Pavlik et al. 2006; Anderson et al. 2013).

All of these are reasons to invest in the maintenance of oak ecosystems. To preserve their ecosystems, oaks are planted by selective harvesting and by their own natural reproduction. However, a problem with using selective harvesting is rapid climate change (Spittlehouse and Stewart 2004; Millar et al. 2007; Aitken and Whitlock 2013; Aitken et al. 2008). Climates around the world are quickly changing due to industrialization, and this may disrupt the health of oak ecosystems. To account for this, researchers wish to better understand the genes that control the oak’s adaptation to climate change. This would allow genetic engineers to design seeds that are resistant to extremes in climate change (Sork et al 2016). To make this easier, researchers have published drafts of the nuclear and chloroplast genomes of Quercus lobata (e.g., Derory et al. 2006; Gugger et al. 2016a; Spiess et al. 2012; Sork et al. 2016).

Reference genomes allow scientists to determine gene models and understand the evolution of the species (Sork et al 2016). For the sake of the preservation of oak ecosystems, a reference genome of Quercus lobata would facilitate the identification of genes that specialize in the plant’s adaptation to climate change (Sork et al 2016). Although past work such as Sork et al. 2016 have published drafts of the annotated reference genome for Quercus lobata, revisions and extensions to the annotations must be made. This paper determines annotations to specific loci from the new genome. To annotate them, the loci were analyzed using the Apollo genome-annotation software. Peptide sequences encoded in these genes were analyzed by first finding homologous peptide sequences using NCBI’s Protein BLAST. The homologous sequences were then used to construct phylogenetic trees and Multiple Sequence Alignment using Clustal Omega. Lastly, the protein structures of homologous peptide sequences were compared using SWISS-MODEL. When combined, all these tools allowed for the inference of the function of the peptide sequence.

Methods

Constructing a Gene Model with Apollo

To determine gene models, the mRNA reference sequence was examined using Apollo. Within the chr8 700,001..795,000 genomic region, the “Unique and MultiMapping FPM” was selected from the “Tracks” menu to display The Valley Oak genomic RNA reference sequence. Peaks on this reference map indicated the frequency at which the exon sequence was expressed in Valley Oak. To construct gene models, “IsoSeqTopTier” was selected from “Tracks” to display pieces of transcripts. Transcripts in the “IsoSeqTopTier” space that aligned with the RNA reference sequence in the “Unique and MultiMapping FPM” were dragged to the “User-Created Annotations” space and then merged together. If there were gaps in the constructed gene model, then the genomic sequence was selected, copied, and then pasted into the Augustus gene prediction tool. Augustus used an algorithm to predict its own gene model based on the genomic sequence. Augustus’s gene model was then compared to the constructed gene model for precision. If there were sequences in the gene model made by Augustus that patched the gaps in the Apollo constructed gene model, such edits were made to the gene model on Apollo accordingly. The Peptide sequence icon was selected, giving the peptide sequence of the gene model (Lewis et al 2002).

Searching for Homologous Sequences Using BLAST

This peptide sequence was homologous sequences from other species. Photos of the Color Key, Putative Conserved Domains, and Top six Query sequence matches each from six unique species were taken. The Color Key is a tool to determine if the Apollo gene model is constructed correctly. The more complete the alignment of a sequence is relative to subject homologous sequences, the better the indication that the gene model is complete. Secondly, the more red the bands of the Apollo gene model and the subject sequences are, the higher the quality of alignment. This is a second indication that the constructed Apollo gene model is complete. The Putative Conserved Domains page displays the conserved domain that is most likely the functional domain the Apollo constructed gene model belongs to. Query sequences were another means to check accuracy of constructed Apollo gene model. If the top six Query sequence results from six unique species do not match the input sequence, the Apollo gene model was most likely inaccurate (Castresana et al 2007).

Using Clustal Omega to Check Accuracy of Gene Model

The FASTA peptide sequences from these six unique species were pasted to Clustal Omega along with the peptide sequence of the constructed gene model. Clustal Omega generated a Multiple Sequence Alignment Page that was used to compare conservation of peptide sequence amongst the seven total species. Asterisks were drawn beneath the columns of sequences of the seven species when there was perfect conservation. The more asterisks available in the Multiple Sequence Alignment Page, the better the match. And this was a sign the constructed gene model was done correctly. A second tool Clustal Omega offered for checking the accuracy of the constructed gene model was the Phylogenetic Tree. If species of the same genus amongst the seven species were not drawn close to each other in the Phylogenetic Tree, this was a sign the gene model was designed incorrectly (Sievers et al 2011).

Using Biozentium’s SWISS model To Determine Structure of Peptide Sequence

After the Apollo constructed gene model passed all these tests for accuracy, precision, and completeness, the peptide sequence of the Apollo constructed gene model was pasted into Biozentium’s SWISS MODEL search bar. SWISS then searched for a structure posted on the SWISS website that exactly matched the Apollo constructed gene model. If this was not available, SWISS instead displayed structure results for the topmost homologous peptide sequences. The results of the topmost homologous peptide sequence matches could then be used to infer the actual structure of the constructed Apollo Gene Model. Since the structure of a protein determines its function, establishing the structure of the constructed Apollo gene model is a major source of evidence that was used to infer the function of the gene model (Guex et al 2005).

Using String-DB To Verify Function of Gene

The peptide sequence of the gene construct was pasted to the String-DB database and analyzed relative to Arabidopsis lyrata as it is one of the most genetically similar species to Quercus lobata found within the database. Results in “Predicted Functional Partners” and “Functional Enrichments in Your Network” would provide evidence as the functional role of the gene construct.

Using UniProt to Understand The Function of the Gene Construct and Localization

Although the above tools were designed to determine the names of the functions of the gene constructs, UniProt was used to understand their meanings. The name of the function was pasted in the UniProt, and UniProt would display research papers describing the functions and cellular localization of the gene.

Research

The 1693 residue gene construct spanned from region [chr8:742485-762148] (Fig. 1A). To demonstrate the accuracy of the construct, Multiple Sequence Alignments(Fig. 1B) and Phylogenetic Trees(Fig. 1C) were constructed. A color key map was also constructed through Protein Blast to determine the model’s accuracy. InterproScan results for the construct were made to help make an inference (Fig. 1D). Since InterproScan generates multiple inferences, Amigo(Fig. 1E) was also used to compare and confirm possible functional predictions. Both InterproScan and Amigo (Fig. 1F) predicted that the construct was involved in proteosome activator activity. Lastly, UniProt was used to determine cellular localization. The proteasome activator subunit complex 4 is located in the nucleus and cytosol (Ustrell et al 2002).

The 333 residue gene construct spanned from region [chr8:730430-734747] (Fig. 2A). The Multiple Sequence Alignment (Fig. 2B) also had excellent matches with the six species homologs. The Phylogenetic Tree (Fig. 2C) demonstrates that Cork Oak and the English Walnut are closely re;ated to the model. The color key demonstrated that the model aligned very well with homologous species from Protein Blast (Fig. 2D). The InterproScan (Fig. 2E) result was vague: protein binding. When supplemented with AmiGo (Fig. 2F), it is clear the best inference is that the construct plays some role directly with DNA or mRNA. According to UniProt, the protein it encodes is localized in the nucleus.

The 302 residues gene construct spanned from region [chr8: 716680-724153] (Fig. 3A). The Multiple Sequence Alignment demonstrates the construct aligns very well with other homologs (Fig. 3B). And the Phylogenetic Tree (Fig. 3C) demonstrates the construct is closest to Cork Oak, which indicates the construct was made correctly. The Color Key (Fig. 3D) for the construct, like all the previous results, showed strong alignment as well. According to InterProScan (Fig. 3E) and AmiGO (Fig. 3F), the gene construct plays a role in DNA and RNA binding. According to UniProt, the protein is localized only in the nucleus.

The 230 residue gene construct spanned from [chr8: 711214-715487] (Fig. 4A). The gene construct in the Multiple Sequence Alignment (Fig. 4B), just as for all the previous gene constructs, aligned strongly with homologous sequences. The Phylogenetic Tree (Fig. 4C) is aligned closest to Cork Oak. The Color Key (Fig. 4D), just as for all the previous gene constructs, showed that the gene construct had excellent alignment with other homologs. And both InterProScan (Fig. 4E) and AmiGo (Fig. 4F) indicate the gene plays a role in calcium ion binding.

Discussion

The gene construct in region [chr8:742485-762148] was found to have proteasome activator complex activity. To learn more about the function, the String-DB database was used. The sequence was compared with Arabidopsis Lyrata in the database. The top four predicted functional partners were proteasome subunit beta type. Functional enrichments in the network were also involved in the proteasome pathway, with a false discovery rate of less than 10^-19. Thus, String-DB gave very strong supporting evidence in support of the proposition that the construct coded for a proteasome subunit complex.

To discover the role proteasome activator subunit complex 4 plays in Quercus Lobata, the complex was researched on UniProt. Proteosome Activator Subunit Complex 4 (PSME4) recognizes acetylated histones and promotes ATP and ubiquitin-independent degradation of core histones during spermatogenesis (pollen generation) and DNA repair after double strand breaks. PSME4 also recognizes and binds acetylated histones through bromodomain-like (BRDL). PSME4 binds to the core proteasome through the C-terminus and opens the closed structure of the proteasome through an active gating mechanism. Finally, PSME4 is a component of the proteasome found in the male reproductive organs of plants. PSME4 binds to acetylated histones and facilitates the degradation of the histones. Thus, PSME4 participates in the exchange of histones during spermatogenesis ([Blickwedehl](http://www.uniprot.org/uniprot/?query=author:%22Blickwedehl+J.%22) et al 2008; Blickwedehl et al 2012; Qian et al 2013; Ustrell et al 2002).

The gene construct in region [chr8:730430-734747] plays some role directly with DNA or RNA. When the peptide sequence was analyzed through String-DB, the “Analysis” section indicated that the most likely function of the gene was RNA transport from nucleus. (false discovery rate 4.68e-23). According to UniProt, the gene construct plays a role in the maintenance of the THO Complex Subunit 3. This protein complex is thought to couple mRNA transcription, processing, and its nuclear export. Moreover, it also ensures small interfering RNAs function properly. The complex is required for the generation of messenger ribonucleoproteins (mRNPs). THO Complex Subunit 3 may transport long RNA molecules so it can serve as a template for secondary siRNA production. Finally, THO Complex Subunit 3 may even facilitate the transport of siRNA precursors to ARGONAUTE catalytic center (Jauvion et al 2010; Yelina et al 2010).

According to UniProt results, chromosome region [chr8: 716680-724153] encoded for a KH-domain-containing of the gene SPIN1. The domain-containing protein plays a role in controlling the time when the plant flowers. The protein inhibits photoperiodism during both long and short days. The protein also plays roles in DNA and RNA binding (Vega-Sanchez et al 2008). The KH domain-containing protein SPIN1 is involved in the Reactome pathway.

The gene construct in chromosome region [chr8: 711214-715487] was found by AmiGo and UniProt to most likely encode for a protein that serves as a calcium ion sensor in the pollen tubes of Quercus Lobata. Uniprot thus recommended giving the protein the name “Probable calcium-binding protein CML21”. In cells, calcium serves as second messengers and are involved in cellular signaling. The protein is most likely responsible for the regulation of calcium homeostasis and calcium-dependent tip growth and oscillation in pollen tubes. The protein may also play a role in the inhibition of pollen tube elongation (Yang and Zhou et al 2009).

Works Cited

Figure 1A Construct is bolded in red.

Aitken, Sally N., and Michael C. Whitlock. “Assisted Gene Flow to Facilitate Local Adaptation to Climate Change.” *Annual Review of Ecology, Evolution, and Systematics*, vol. 44, no. 1, 2013, pp. 367–388., doi:10.1146/annurev-ecolsys-110512-135747.

Aitken, Sally N., et al. “Adaptation, Migration or Extirpation: Climate Change Outcomes for Tree Populations.” *Evolutionary Applications*, vol. 1, no. 1, 2008, pp. 95–111., doi:10.1111/j.1752-4571.2007.00013.x.

Anderson, Kat. *Tending the Wild: Native American Knowledge and the Management of California's Natural Resources*. University of California Press, 2013.

Castresana, Jose. “On Homology Searches by Protein Blast and the Characterization of the Age of Genes.” *BMC Evolutionary Biology*, BioMed Central Ltd, 4 Apr. 2007, bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-7-53.

DAHLGREN, R.A., and X. HUANG. “Oak Tree and Grazing Impacts on Soil Properties and Nutrients in a California Oak Woodland.” *SpringerLink*, Kluwer Academic Publishers, link.springer.com/article/10.1023/A:1005812621312.

Derory, Jérémy, et al. “Transcriptome Analysis of Bud Burst in Sessile Oak (Quercus Petraea).” *New Phytologist*, vol. 170, no. 4, 2006, pp. 723–738., doi:10.1111/j.1469-8137.2006.01721.x.

Dosskey, Michael G., et al. “Riparian Buffers for Agricultural Land.” *Iowa State University Digital Repository*, lib.dr.iastate.edu/for\_pubs/9/?utm\_source=lib.dr.iastate.edu%2Ffor\_pubs%2F9.

Guex, Nicolas, and Manuel C. Peitsch. “SWISS‐MODEL and the Swiss‐Pdb Viewer: An Environment for Comparative Protein Modeling.” *ELECTROPHORESIS*, Wiley-Blackwell, 14 Apr. 2005, onlinelibrary.wiley.com/doi/full/10.1002/elps.1150181505.

Gugger, Paul F., et al. “Association of Transcriptome-Wide Sequence Variation with Climate Gradients in Valley Oak (Quercus Lobata).” *Tree Genetics & Genomes*, vol. 12, no. 2, 2016, doi:10.1007/s11295-016-0975-1.

Herman, et al. “NITROGEN DYNAMICS IN AN ANNUAL GRASSLAND: OAK CANOPY, CLIMATE, AND MICROBIAL POPULATION EFFECTS, Ecological Applications.” *DeepDyve*, Ecological Society of America, 1 June 2003, www.deepdyve.com/lp/ecological-society-of-america/nitrogen-dynamics-in-an-annual-grassland-oak-canopy-climate-and-IXZ8SWzCXG.

Kroeger. “An Economic Analysis of the Benefits of Habitat Conservation on California Rangelands.” *California Rangeland Conservation Coalition*, [www.carangeland.org/images/An\_Economic\_Analysis\_of\_the\_Benefits\_of\_Habitat\_Conservati\_3\_.pdf](http://www.carangeland.org/images/An_Economic_Analysis_of_the_Benefits_of_Habitat_Conservati_3_.pdf).

Lewis, et al. “Apollo: a Sequence Annotation Editor.” *Genome Biology*, BioMed Central, 23 Dec. 2002, genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-12-research0082.

Luppold, William G., and Matthew S. Bumgardner. “Factors Influencing Changes in U.S. Hardwood Log and Lumber Exports from 1990 to 2011.” *BioResources*, vol. 8, no. 2, June 2013, doi:10.15376/biores.8.2.1615-1624.

Millar, Constance I., et al. “Climate Change And Forests Of The Future: Managing In The Face Of Uncertainty.” *Ecological Applications*, vol. 17, no. 8, 2007, pp. 2145–2151., doi:10.1890/06-1715.1.

Millar, Constance I., et al. “Climate Change And Forests Of The Future: Managing In The Face Of Uncertainty.” *Ecological Applications*, vol. 17, no. 8, 2007, pp. 2145–2151., doi:10.1890/06-1715.1.

Pavlik, Bruce M. *Oaks of California*. Cachuma Press, 2006.

Sievers, Fabian, et al. “Fast, Scalable Generation of High‐Quality Protein Multiple Sequence Alignments Using Clustal Omega.” *Molecular Systems Biology*, EMBO Press, 1 Jan. 2011, msb.embopress.org/content/7/1/539.print.

Sork, V. L., et al. “First Draft Assembly and Annotation of the Genome of a California Endemic Oak Quercus Lobata Nee (Fagaceae).” *G3&Amp;#58; Genes|Genomes|Genetics*, Dec. 2016, doi:10.1534/g3.116.030411.

Spittlehouse, D & B Stewart, Robert. (2003). Adaptation to climate change in forest management. British Columbia Journal of Ecosystems and Management. 4.

Spiess, Nadine, et al. “Ecophysiological and Transcriptomic Responses of Oak (Quercus Robur) to Long-Term Drought Exposure and Rewatering.” *Environmental and Experimental Botany*, vol. 77, 2012, pp. 117–126., doi:10.1016/j.envexpbot.2011.11.010.

Standiford, Richard B., and Richard E. Howitt. “Multiple Use Management of California's Hardwood Rangelands.” *Journal of Range Management*, vol. 46, no. 2, 1993, p. 176., doi:10.2307/4002277.