Methods

• Provide the reader enough details so they can understand and replicate your research

The Valley Oak genomic region from chr8 700,001..795,000 were annotated using Apollo.

• Explain how you studied the problem, identify the procedures you followed, and order these chronologically where possible

• Explain new methodology in detail; otherwise name the method and cite the previously published work

• Include the frequency of observations, what types of data were recorded, etc.

• Be precise in describing measurements and include errors of measurement or research design limits

COLOR KEY- Red is strong indication of high quality alignment. And the more complete it is relative to the completeness of the subjects, the better the indication that the gene model is complete. SWISS MODEL-Homologous peptide sequences have previously-written annotations on what their structure is, so you can use that to infer what the protein-of-interest’s structure is. Since the structure of a protein determines its function, understanding the structure is one of the major pieces of evidence used to infer the function of the gene model.

HOW IT WORKS: PASTE THE PEPTIDE SEQUENCE OF YOUR CONSTRUCTED GENE MODEL INTO THE SEARCH FIELD. SWISS WILL CHECK IF A FULL STRUCTURAL MODEL FOR THE SEQUENCE WAS ALREADY MADE. IF NOT, will look for topmost homologous peptide sequences and show structural model results for those to help you infer the structural model of your input peptide sequence.

BLAST Query results for genes—If query sequences do not match input peptide sequence that was pasted in BLASTp, the constructed gene model is most likely inaccurate. Use query results to check if your constructed gene model was correctly constructed. Also the Query Sequence should begin with Methionine and end at a similar end peptide number that the subjects ended at. Otherwise, Apollo gene model is incomplete.

Augustus-Another tool that can be used to help you construct a gene model. Paste the genomic sequence of your constructed gene model from Apollo. Augustus will present its own constructed version of the gene model based on the genomic sequence you pasted. Compare this gene model to the one on Apollo and make edits to the gene model as deemed necessary.

Augustus—uses a math algorithm to predict a gene model sequence from the genomic sequence. But it more sense to trust the actual experimental-based data from IsoSeqTopTier. It is only if we have a gap in experimental data for the gene model to we fill in the missing patch based on Augustus. Hopefully there is a transcript that can be dragged to User-Annotation space that patches according to what Augustus predicted.

HOW IT WORKS: PASTE GENOMIC SEQUENCE OF CONSTRUCTED GENE MODEL

Phylogenetic Tree-Can be used to infer quality of gene model. If species with the same genus are not closely drawn to each other in the gene, this is a strong sign the constructed gene model made on Apollo is wrong.

Multiple Sequence Alignment: If there were poor matches amongst the peptide sequences of supposed homologous sequences, this was a sign the constructed gene model was done incorrectly.

Putative Conserved Domains-Gives best conserved domain and the most likely functional domain the gene model belongs to.

To determine gene models, the mRNA reference sequence was consulted 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. Next, the merged gene model was right-clicked and the option for “Gene Sequence” was selected. If there were gaps in the constructed gene model, then first, 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. This peptide sequence was pasted into NCBI’s Protein BLAST to search for conserved 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 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. After the Apollo constructed gene model passed all these tests for accuracy, precision, and completeness, the peptide sequence of the 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.