PepsiCo releases annotated gene set and associated files for OT3098 v2 genome in partnership with GrainGenes

Mandy Waters, PepsiCo, Minneapolis, MN, USA.

Description of data release
PepsiCo is releasing a set of predicted transcripts and translated protein sequences mapped to the OT3098 v2 reference. Researchers are able to access the annotation file as a track on the genome browser. Additional files, including a GFF, gene nucleotide fasta, ORF translated peptide fasta, and a text file with top BLAST hits, are also available via the download site on GrainGenes.

List of files:
PepsiCo_OT3098_v2_gene_annotations.gff3: gff file with predicted genes (details below on prediction pipeline and nomenclature)
PepsiCo_OT3098_v2_predicted_genes_nucleotide_seq.fa: fasta file with nucleotide sequences for predicted genes
PepsiCo_OT3098_v2_predicted_genes_protein_seq.fa: fasta file with amino acid sequences for genes where an open reading frame could be identified
PepsiCo_OT3098_v2_predicted_genes_blast_results.txt: text file with up to the top 5 blast results for each gene

Use:
Researchers are free to use and publish with all OT3098 genomic resources shared on GrainGenes. Given that no direct publication will be submitted for these sets of data, we choose to opt out of the Toronto Agreement, so researchers can freely use these resources as they become available:
• Genome Browser: https://wheat.pw.usda.gov/jb?data=/ggds/oat-ot3098v2-pepsico
• BLAST: https://wheat.pw.usda.gov/blast/ (select "PepsiCo OT3098 Hexaploid Oat v2 pseudomolecules (2021)" under the "Oat Selections")
• Data Download: https://wheat.pw.usda.gov/GG3/graingenes-downloads/pepsico-oat-ot3098-v2...

Citation:
If you use these resources, please cite:
"Avena sativa – OT3098 v2, PepsiCo, https://wheat.pw.usda.gov/jb?data=/ggds/oat-ot3098v2-pepsico"

Pipeline information:

Written by Dr. Marissa Macchietto, Bioinformatics Analyst, Minnesota Supercomputing Institute

Ab initio gene prediction with Augustus

Oat genes were predicted per-chromosome using Augustus 3.2.3 using wheat gene training models and an intron hints file produced by IsoSeq transcripts alignments from funnannotate. Augustus protein sequences were extracted from each chromosome GFF file using custom perl script and run through blastp (ncbi_blast+/2.8.1) against the blast nr database (settings: -evalue=1e-10, -max_target_seqs=10, -qcov-hsp_perc=70) to 1) find genes with evidence and to 2) help determine their identities.

Augustus predictions supported by IsoSeq evidence

To find Augustus predictions that have IsoSeq evidence, per-chromosome gene predictions were overlapped with IsoSeq loci using bedtools intersect (v 2.29.2) requiring IsoSeq loci to be covered by Augustus predictions by at least 20% (-F 0.2). All IsoSeq predictions that were not covered by Augustus by at least 20% were considered “missing”.

Combining and renaming per-chromosome Augustus predictions supported by IsoSeq evidence

An in-house Python script was used to filter out all Augustus predictions that had no IsoSeq evidence and to combine all the remaining predictions across chromosomes together implementing the oat naming conventions for gene/mRNA identifiers as outlined by The International Oat Nomenclature Committee (e.g., gene 1 on chromosome 1A would be: AVESA.00001b.r1.1Ag0000001). This script also wrote out all the protein sequences for the Augustus predictions that had IsoSeq evidence with the new naming conventions. Since Augustus does not provide unique exon and CDS attributes in their GFF file, unique exon and CDS attributes were added to the GFF file and the CDS and exon order of appearance in the GFF file was changed depending on the strand orientation of the gene feature using another in-house python script. For example, positive sense genes would have CDS/exon numbers increasing as you read down the GFF file, whereas negative sense genes would have CDS/exon numbers decreasing as you read down the GFF file.