Getting Started: Gene Annotation in *Zea mays_B73v5*
Note: Documenting your annotation process for each gene you annotate is very important.

1. **Find a gene you are interested in annotating**

    a. This Google Doc lists genes of interest selected for this workshop.
    b. If you wish to select your own, you can use Plant Reactome Database.
    c. Start looking up background info on your selected gene and gene product.
    d. You can used Gramene.org or MaizeGDB.org to find information about the gene
        i. Identify the functions of the gene
            1. What protein does the gene produce?
            2. What does the protein or gene product do?
            3. When is this gene expressed during maize development?

2. **Look at different predicted model tracks in Apollo[1]**

    a. Log in Apollo
    b. Bring up predicted gene model track: evidence models
        i. Look at all the predicted models to see if there is a consensus on model length, exon placement, etc.
        ii. Choose a model that will serve as the basis of your annotated gene model
            1. Typically, the longest model that has features of the other models within it is a good one to start with
        iii. Click and drag the model into the yellow editor box.
        iv. If the model is less than four transcripts you can annotate them one by one.

    c. Look at Hierarchy Evidence list to identify the best evidence sources to curate your gene model
        i. Click on evidence tracks available to identify support to your predicted gene model: est2_genome_isoseq, GMAP_B73v4_mapped_to_v5, protein2genome_Os, B73_RNAseq[2]_merged.bam, etc.

    d. Look at orthologous[3] protein sequences on NCBI[4]
        i. Right click on one of the introns from the gene model and get the sequence. Copy the peptide sequence and BLAST[5] in NCBI's protein database.
        ii. Note the average sequence length for the protein sequences in the search results.

iii. Note the important conserved domains within the protein sequences
    1. Click on graphic summary and Show Conserved Domains"

e. Look at good orthologous sequence in a species that is closely related to *Zea mays*
    i. Closely related species are Poales[6] which include: *Sorghum, Oryza, Brassica*, among others. Other less related species such as *Arabidopsis, Nicotiana, etc.* can be useful, too.
    ii. A good sequence is not labeled as "predicted" or "partial" and the domains are not broken up or incomplete.

3. Completing the manual annotation of the gene model by naming it:

Comment to validate your annotation, even if you made no changes to an existing model
Add a comment to inform the community of unresolved issues you think the model might have

a. In Apollo, right click annotated model, edit information
    i. For both Gene and mRNA Name fields, enter in a *Zea mays*_B73v5[7] identifier followed by an underscore and a two-letter (your initials) suffix, e.g. Zm00001e04987_mk
    ii. **Note**: Isoform number can be left out of the gene name and should be included in the mRNA name (e.g. Zm00001e04987_T002_mk).
    iii. Enter in Gene Ontology IDs that describe your gene[8]
    iv. Add important information about your model in the comment box (e.g. If the model has proper start/stop/splice sites, any changes you made, evidence you use to curate the model, if it is partial curation, potential concerns, brief BLAST results summary)

Defined Terms and Helpful Tips:

1. **Apollo** –Apollo or Web Apollo is a plugin for JBrowse that allows us to instantly view the gene models and genome edits being made by other users.
2. **RNAseq data** – RNA Sequencing is a technique that lets you to look at RNA expression. First, RNA is collected from an organism and converted back it into DNA (this DNA is referred to as cDNA). The cDNA is then sequenced, and then the sequenced cDNA reads are mapped to the genome. Now, we can see where on the genome this RNA is coming from, and we can use the RNAseq data to help us annotate the gene correctly.
3. **Orthologous sequence** - Orthologs are similar genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function.
4. **National Center for Biotechnology Information** - NCBI is a website that provides access to huge gene sequence and protein sequence databases that we can use to find orthologous sequences and gene information.
5. **BLAST** – Basic Local Alignment Search Tool is an algorithm for comparing amino-acid sequences of proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify other sequences that resemble the query sequence.
6. **Poales** – The order of grasses that *Zea mays* belongs to. The Poales are a large order of flowering plants in the monocotyledons, and includes families of plants such as the grasses, bromeliads, and sedges
7. **Zea mays_B73v5** identifier follows a specific format: Zm00001eXXXXXX.Y where XXXXXX is the gene number and Y is the isoform or transcript number. The gene number can be determined by the models in evidence models track.
8. **GO terms** – A GO term is an identifier that describes a characteristic of a gene. For example, the GO term "GO:0000016" represents "lactase activity". So, if a gene is labeled with GO:0000016, then you know it has lactase activity. Gene ontology is an attempt to unify gene nomenclature across all species because the same genes in different organisms can have wildly different names even though they have identical functions.