Genome annotation instructional guide worksheet for students

Getting Started: Gene Annotation in *Diaphorina citri*

Note: Documenting your annotation process for each gene you annotate is very important.

Before beginning the work for a step, make sure you understand the listed concepts well enough to answer the background questions. Explanations of some important terms are located in the footnotes.

Step 1 - Identifying Your Gene

Important Concepts: Literature Searches, Databases, Homology (Orthology/Paralogy)

Background Questions:

_____is a website where you can search for scientific publications about a topic of interest.

Orthologs are evolutionarily related genes separated by a _____ event.

Paralogs are evolutionarily related genes separated by a _____ event and perhaps a _____ event.

True or False: Paralogs are always within the same genome.

Paralogs and orthologs are both types of _____.

Method

1. Find a gene you are interested in annotating
    1. Select a gene of interest from the list on Google Docs.
    2. Start looking up background info on your selected gene and gene product.
        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 insect development?

ii. Look at orthologous[1] protein sequences on NCBI[2]

1. Search for protein name in NCBI's protein database

2. Note the average sequence length for the protein sequences in the search results

3. Note the important conserved domains within the protein sequences

a. Click on one of the sequences to go to its protein page

b. On the right-hand side under "Analyze this sequence" click "Identify Conserved Domains"

Step 2 - Finding Orthologs

Important Concepts: BLAST

Background Questions:

The acronym BLAST stands for _____ _____ _____ _____ _____.

The sequence used to search a BLAST database is the _____.

What does the e-value of a BLAST result mean?

Where in a BLAST report can you find information about conserved domains?

Method:

2. Look for a good orthologous sequence in a species that is closely related to *Diaphorina citri.*

    1. Closely related species are hemipterans[3] which include: *Acyrthosiphon pisum, Bemisia tabaci, Cimex lectularius,* and *Halyomorpha halys*, among others. Other non-hemipteran species such as *Tribolium castaneum, Drosophila melanogaster, Aedes aegypti,* and *Anopheles gambiae* can be useful, too.
    2. A good sequence is not labeled as "predicted" or "partial" and the domains are not broken up or incomplete.
    3. Search NCBI protein database with the protein name + a species name

        i. Go to protein page if there is one. (If there isn't one, use a different species to search with.)

        ii. BLAST[4] sequence to analyze its quality

        1. Click "Run BLAST" on the right-hand side of protein page

        2. Type "Insecta" into the Organism box and select it from the dropdown menu and hit BLAST.

        3. BLAST results give domain info and compare the sequence to other sequences within the database.

Step 3 - Locating Your D. citri Gene

Important Concepts: FASTA format, reciprocal BLAST, BLAT

Background Questions:

Sequences in FASTA format have a header line beginning with _____.

True or False: A FASTA file can contain either DNA or protein sequences.

How do you perform a reciprocal BLAST search?

What is the purpose of a reciprocal BLAST search?

The sequence search tool in Apollo uses BLAT, which requires a more / less exact match than BLAST (circle one).

Method:

3. Take your sequence and BLAST into MCOT[5] database available on citrusgreening.org
    1. Acquire your protein sequence in FASTA[6] format by going back to its protein page and selecting the FASTA tab at the top. Copy the sequence.
    2. Go to citrusgreening.org > Vector > BLAST > *Diaphorina citri* databases. Then select "*Diaphorina citri* MCOT proteins" from the Database dropdown menu.
    3. Select "blastp (protein to protein db)" from the Program dropdown menu.
    4. Paste your orthologous protein sequence into the text box and click BLAST.
        i. MCOT database will bring up protein predictions within the *D. citri* genome that are similar to the sequence you searched.
        ii. The best MCOT match is the top one in the results
    5. Look for the best MCOT ID within the *D. citri* genome browser
        i. Copy and paste the MCOT ID into the index on JBrowse[9]
        ii. OR search the genome for the predicted MCOT sequence.
            1. Get sequence of MCOT ID by clicking it in the BLAST results and then clicking "View matched sequence." Copy the sequence.
            2. Go to the *D. citri* genome browser, click "Tools" in the top left, then "search sequence" and paste it.
            3. Select "Blat protein" from the dropdown menu
            4. Check "Search all genomic sequences" and hit Search

5. Note: This Blat search feature will not work if sequence is in FASTA format. Remove the >identifier so only the sequence remains.

6. Results will show you different regions of the genome are similar to the MCOT sequence you entered. Select the top one and go to that region of the genome.

Step 4 - Starting Your Annotation Model

Important Concepts: Official Gene Set, *de novo* Transcriptome, Evidence Tracks

Background Questions:

True or False: An Official Gene Set (OGS) contains only manually annotated genes.

A *de novo*-assembled transcriptome is a genome independent / dependent evidence source. (Circle one.)

True or False: Mapped transcripts from a *de novo*-assembled transcriptome retain their original sequences in Apollo.

Name three types of evidence that can be used when manually annotating genes.

1.

2.

3.

Method:

4. Look at different predicted model tracks
    1. Bring up predicted gene model tracks OGS, MCOT models, NCBI 101, Maker, Augustus

        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

Step 5 - Manually Correcting Your Model

Important Concepts: Gene structure; RNA-Seq reads

Background Questions:

Summarize the central dogma of molecular biology.

The consensus sequence for the first two bases of an intron is _____.

The consensus sequence for the last two bases of intron is _____.

What amino acid marks the start of translation in almost all proteins?

The portion of the transcript upstream of the translation start site is called the _____ UTR.

The portion of the transcript downstream of the translation stop site is called the _____ UTR.

Explain the concept of reading frames with respect to DNA/RNA.

RNA-Seq data is particularly useful for determining the location of _____.

Method:

5. Start annotation
    1. Check for proper start/stop sites. Start site is represented by an "M" and stop site represented by an asterisk (*).
    2. Check exon boundaries for appropriate splice sites (5' intron splice site is GT and GC, 3' intron splice site is AG)
        i. Look at RNAseq/mapped reads[7]: whole adult, whole egg, and whole nymph
        ii. Analyze RNAseq data to determine if exons are appropriately located
            1. RNA sequencing strands won't go further than boundaries of exon
            2. Beware of artificially duplicated exons[8]
    3. It's good to have UTR at the beginning and the end of a gene model
    4. Further analyze gene model using NCBI BLAST
        i. Right click on model and click "get sequence"
        ii. Copy protein sequence and paste it into NCBI Protein BLAST.
        iii. Enter "insecta" into the Organism box and hit BLAST
        iv. Analyze BLAST results
            1. How well does your annotated sequence compare to the other orthologs?
            2. Domains: Analyze domains within your annotated sequence and compare to the domains you saw in orthologous sequences.
            3. Length: Is your annotated sequence similar in length to the other proteins in the BLAST results

4. Overall sequence quality can be determined by
    a. High query coverage (>80% good; >50% acceptable)
    b. High Identity (ID)
    c. High score (In general, higher is better)
    d. Which organisms are matching to your sequence the best
        i. If your sequence is aligning well to orthologs from hemipterans, then that is good.

Step 6 - Documenting your Work

Important Concepts: Nomenclature, Gene Ontology (GO Terms)

Background Questions:

How do you determine what to name a gene?

Name the three classes of Gene Ontology (GO) terms.

1.
2.
3.

Method:

6. Finish the gene model by naming it
    1. In Web Apollo[9], right click annotated model, edit information
        i. For both Gene and mRNA Name fields, enter in a DcitrG identifier (e.g. DcitrG063270.1.1)[10]

Note: Isoform number can be left out of the gene name and should be included in the mRNA name.

ii. Enter in Gene Ontology IDs that describe your gene[11]

iii. Add in the comments important info about your model (e.g. If the model has proper start/stop/splice sites, any changes you made, potential concerns, brief BLAST results summary)

iv. Keep log of model:

1. What is the name of the model?

2. What scaffold number is it on?

Defined Terms and Helpful Tips:

1. Orthologous sequence - Orthologs are similar genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function.

2. 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.

3. Hemiptera – the order of insects that *Diaphorina citri* belongs to.

4. 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.

5. MCOT is a database of protein predictions that have been mapped to the *D. citri* genome browser. These predictions are based off of actual transcript sequencing and computer predictions, so they are good jumping off points for annotating a new gene.

6. FASTA – A format used for nucleotide or amino acid sequences that allows a sequence to be preceded by important information such as a name/identifier, sequence length, reference ID, etc. A sequence that is in FASTA format begins with a right caret (>) followed by details/information, then there is a line break, then the actual nucleotide or protein sequence begins.

7. RNAseq data – RNA Sequencing is a technique that lets you look at RNA expression. First, RNA is collected from an organism and converted back 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.

8. Be wary of exon duplications that do not belong. Right click an exon you suspect of being a duplication to get its sequence, go to Tools > Search sequence, leave "Search all genomic sequences" unchecked, and search the exon sequence. If it is a duplication, you will have more than one hit. Right click and delete the duplicated exon. Run an NCBI BLAST search of the new model sequence to confirm improved annotation.

9. JBrowse and Web Apollo – JBrowse is a genome browser that works directly in your web browser, and we use it to visualize the psyllid genome and create gene models. Web Apollo is a plugin for JBrowse that allows us to instantly view the gene models and genome edits being made by other users.

10. DcitriG identifier follows a specific format: DcitrGXXXXX.Y.Z where XXXXX is the gene number, Y is the version number and Z is the isoform number. The gene number can be determined by the models in OGS2.0 track. If your gene model is the same as an OGS2.0 model, then use that OGS2.0 identifier for your model's gene/mRNA name. If your annotated model is different from and replacing an OGS model in the same location, then use the same gene number for your gene/mRNA name and change the version number. If there is no OGS2.0 model in your model's location, then the gene number should fall between the two OGS2.0 models that flank your model's location. (For example, the OGS model to the left of my annotated model is DcitrG063270.1.1, and the OGS2.0 model to the right is DcitrG063275.1.1. I will name my annotated model DcitrG063273.1.1 because 73 falls between 70 and 75.)

11. 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 is predicted to have 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.

12. Most importantly, when in doubt…BLAST it and find out!

Answer Key:

Step 1

__PubMed_____is a website where you can search for scientific publications about a topic of interest.

Orthologs are evolutionarily related genes separated by a ___ speciation_____ event.

Paralogs are evolutionarily related genes separated by a _____duplication_____ event and perhaps a __speciation_____ event.

True or False: Paralogs are always within the same genome.

Paralogs and orthologs are both types of ___homologs_____.

Step 2

The acronym BLAST stands for ___Basic_____ ____Local_____ ____Alignment_____ ____Search_____ ___Tool_____.

The sequence used to search a BLAST database is the ___query____.

What does the e-value of a BLAST result mean?

It number of hits expected by chance to match the query as well (receive a similar score) as a that particular hit. This probability is affected by the length of the query sequence and the size of the BLAST database.

Where in a BLASTp report can you find information about conserved domains ?

At the top of Graphic Summary page

Step 3

Sequences in FASTA format have a header line beginning with _ >__.

True or False: A FASTA file can contain either DNA or protein sequences.

How do you perform a reciprocal BLAST search?

Use a protein from species #1 as the query against species #2. Take the top hit from species #2 and use it as the query to BLAST against species #1.

What is the purpose of a reciprocal BLAST search?

If two proteins are reciprocal best hits they are predicted to be orthologs.

The sequence search tool in Apollo uses BLAT, which requires a more / less exact match than BLAST (circle one).

Step 4

True or False: An Official Gene Set (OGS) contains only manually annotated genes.

A *de novo*-assembled transcriptome is a genome independent / dependent evidence source. (Circle one.)

True or False: Mapped transcripts from a *de novo*-assembled transcriptome retain their original sequences in Apollo.

Name three types of evidence that can be used when manually annotating genes.

1. Mapped de novo assembled transcripts
2. Mapped RNA-Seq Reads
3. Computationally predicted transcripts

There may be other correct answers.

Step 5

Summarize the central dogma of molecular biology.

DNA to RNA to protein

transcription translation

The consensus sequence for the first two bases of an intron is _GT_.

The consensus sequence for the last two bases of intron is __AG_____.

What amino acid marks the start of translation in almost all proteins?

Methionine (M)

The portion of the transcript upstream of the translation start site is called the __5'____ UTR.

The portion of the transcript downstream of the translation stop site is called the __3'___ UTR.

Explain the concept of reading frames with respect to DNA/RNA.

It takes three nucleotides to encode a single amino acid. Depending on which nucleotide you start with when translating you can have three reading frames on each strand. DNA has six reading frames. An mRNA has three reading frames because it is single . Any insertion or deletion that is not a multiple of 3 nucleotides will affect the reading frame.

RNA-Seq data is particularly useful for determining the location of ____intron/exon boundaries_____.

Step 6

How do you determine what to name a gene?

A gene should be named based on its orthologs. For insects, it is usually best to use the name of the gene from *Drosophila melanogaster*, the most widely used insect model.

Name the three classes of Gene Ontology (GO) terms.

1. biological process
2. cellular component
3. molecular function