Annotation Walkthrough Workshop **NAME:**
BIO 173/273 – Genomics and Bioinformatics
Spring 2013
Developed by Justin R. DiAngelo at Hofstra University

A Simple Annotation Exercise

Adapted from: Alexis Nagengast, Widener University, Jacob Jipp and Marian Kaehler, Luther College, Genomic Annotation Exercise and Gary Kuleck, Loyola Marymount University, Sample Annotation Report, and A Simple Annotation by John Braverman, St. Joseph's University and Chris Shaffer of the GEP.

Introduction

This tutorial takes you through a series of basic steps that have been found to work well for annotation of species closely related to *Drosophila melanogaster*. It provides a technique that can also be the foundation of annotation in other, more divergent species, but in those cases other special techniques will probably be needed. The example given uses a gene that is conserved between *D. melanogaster* and *D. biarmipes*. By utilizing tools such as BLAST and leveraging the large amount of research data on *D. melanogaster*, you will be able to identify the gene in *D. melanogaster* that has the highest degree of sequence similarity to the *D. biarmipes* feature you are annotating. Then, knowing the predicted protein sequence of that gene and proceeding exon-by-exon, you will be able to use BLAST and the UCSC Genome Browser to help you identify the exact coordinates of each exon. At the conclusion of this exercise you will have successfully created a gene model for *D. biarmipes*.

"Assignment"

You are assigned the following fosmid and gene to annotate:

Species: *Drosophila biarmipes*
Project: contig61
Region / Assembly: Aug 2012, GEP/Dot

Annotation Worksheet

Work in pairs and turn in one worksheet **electronically** per group with the file name as: AnnotEx<*yourinitials*>BioinfoSp13.

Part 1: Computer Set-Up

1. Open a web browser (Internet Explorer, Mozilla Firefox, Chrome, or Safari)
   Three different windows/tabs will be utilized for this lab. Each website will have a specific function throughout the course of this workshop. It is important to use the website that is directed for each step.

2. Open the following three websites (in separate tabs):
Throughout this worksheet, these sites will be referred to by the names in quotes.

- "Gander" = GEP UCSC Genome Browser Mirror at http://gander.wustl.edu/cgi-bin/hgGateway

- Using the software created by the UCSC Genome Bioinformatics Group, the GEP has setup genome browsers with different Drosophila annotation projects. The official UCSC Genome Browser (http://genome.ucsc.edu) is an up-to-date source for genome sequence data integrated with a large collection of experimental and computational results for many different species. It has downloadable files and is designed to be easy-to-use.

- "FlyBase" = http://www.flybase.org/blast/

- A comprehensive site containing genetic information and biological information about *Drosophila melanogaster* and other Drosophila species. FlyBase is developed and maintained by a consortium of researchers at Harvard University, Indiana University, and Cambridge University in the UK.

- "[NCBI] BLAST" = http://blast.ncbi.nlm.nih.gov/Blast.cgi

- Web interface to the NCBI BLAST program. BLAST is a tool designed for identifying local regions of similarity between two different sequences. A variety of inputs (including nucleotide, translated nucleotide, and protein sequences) can be used as query to search an extensive database of genomes in order to find similar sequences. Sequence similarities are given as an output based upon the statistical significance of the alignment.

3. Spend a few minutes to search the various sites. Try to become familiar with the different attributes of the sites and the differences between them. Also explore the NCBI and FlyBase BLAST sites for information on the different BLAST programs. **Complete TABLE 1**.

TABLE 1. Types of sequences used by each BLAST program:

BLAST program	Query sequence	Database sequence
blastn		
blastp		
blastx		
tblastn		
tblastx		

Part 2: Initial Fosmid Observation

1. You will be analyzing contig61 from the *D. biarmipes* genome Aug 2012 (GEP/Dot) assembly. Using the Gander Genome Browser site, enter this information and click on submit. **Take a screen shot of the graphical output window and paste below.**

2. If you scroll down you will notice many different settings, which allow you to configure the display options for many different "evidence tracks" (These are simply different features of the genome and predictions generated by different bioinformatics programs). Click "hide all." Next, change the settings so they match those below:

GENSCAN genes = pack
RNA-Seq Alignment Summary = show
RepeatMasker = dense

Click on the "refresh" button and you should be able to see the gene predictions. The color bars correspond to the different gene predictions: The thick, solid bars are the exons and the thin lines connecting the solid bars are the introns. **Take a screen shot of the graphical output window and paste below.**

3. Now that the gene predictions are shown, you should be able to identify how many predicted genes are present in the fosmid.

How many predicted genes are present in the fosmid?

Are the genes oriented in the same direction? How can you tell?

4. If you look at the bottom of the genome browser graphic, you can see the regions that have been annotated by RepeatMasker.

Approximately (a simple estimate is fine) how much of the fosmid sequence appears to be repeats?

Where in the fosmid are these repeat sequences primarily located?

5. Click on the feature contig61.2.
What is the value of the "Position" field? What does the "Position" field represents?

What is the Genomic Size? What does this mean?

What is the Strand? What does this mean?

6. Identify the likely *Drosophila melanogaster* ortholog. To do this, copy the predicted protein sequence (this will be your query sequence) of your particular *D. biarmipes* predicted gene from Gander and perform a blastp search against the *D. melanogaster* annotated proteins (AA) on the FlyBase website (select the "GenBank protein sequences" from the database drop-down menu). A list of hits in a color-coded chart and as a BLAST Hit Summary Table should come up as the result. A small E-value indicates high degree of similarity between the two sequences.

Based on the color-coded diagram, how many proteins appear to have highly similar sequences when compared to your query protein sequence?

How many different protein records are similar to your query protein? What are their E-values?

7. Scroll down (if you click on the gene in the chart or click on the 'score' on the table, it should automatically transfer you to the corresponding alignment result on the page) and examine the alignments. A Refseq prefix of NP_ indicates that the protein sequence record is supported by experimental evidence while an XP_ prefix indicates that the record is derived from computationally predictions. Note that there are also generic GenBank records without the NP_ or XP_ prefixes. In general, because the staff at NCBI has manually reviewed the RefSeq records, we would generally prefer to use RefSeq records instead of the generic GenBank records in our annotations.

Are the matching records curated or automated? Why would this matter?

8. Determine the best match for your query sequence.

Take a screen shot of the information after the ">" and paste it below.

What is the name of your gene?

What is the % of identical matches?

What is the % of positive matches?

How many gaps are in the alignment?

Briefly explain what these three alignment terms (% identities, % positives, gaps) mean:

Part 3: Gene Structure Research

1. Open another tab and go to http://gep.wustl.edu. Under "Projects," and "Annotation Resources," click on "Gene Record Finder." Type in the name of your gene (**note: gene names in Drosophila is case sensitive!**) and hit "Find Record."

What FlyBase release is this based on and when was it last updated?

2. Under "Gene Details," click on the FlyBase ID.
Which FlyBase release is this based on and when was it last updated?

Which species is this information based on?

What is the function of your gene?

What chromosome is it on?

3. Go back to your Gene Record Finder results window. Click on the "Transcript Details" tab.
Take a screen shot of the Transcript Details and paste it below.

How many isoforms does your gene have?

How many exons does your gene have?

4. Under "mRNA Details," click on the FlyBase ID for your gene.
What is the Evidence Rank for this isoform?

What sort of data would support this?

What is the length of the processed transcript?

5. Scroll down to "Other Products of this Gene."

What is the length of the polypeptide derived from your transcript?

6. Go back to your Gene Record Finder results window and click on the "Polypeptide Details" tab.
Take a screen shot of the Polypeptide Details and paste it below.

How many coding exons are in the A isoform of *Actbeta*? Is this number the same as the number of exons in the transcript? If not, explain why these numbers are different.

Part 4: Annotation

****For each exon you will need to complete the following steps:****

You now have the gene ID and protein sequence from *D. melanogaster*. You must locate and evaluate how similar this sequence is to the *D. biarmipes* contig. Essentially, you will need to compare each coding exon (also known as Coding DNA Sequences or CDS) from *D. melanogaster* to the *D. biarmipes* genomic DNA and identify the region of the *D. biarmipes* genome with the closest match.

1. In the "Polypeptide details" tab, click on the first CDS to display the amino acid sequence for that exon. Copy this sequence (with the FASTA identifier).

2. In the NCBI BLAST tab, under the Basic BLAST heading, click “blastx” and paste the amino acid sequence for the particular exon into the “subject sequence” text box. Check the box that indicates that you want to “align two or more sequences.”

3. Go back to the Gander Genome Browser tab and return to the fosmid ID page. **Make sure the arrow at the left of the fosmid is consistent with the direction of transcription for your gene** (i.e. positive strand points to the right, negative strand points to the left). To view the entire sequence, enter “contig61” into the “position/search” box and then click on the “jump” button. Click the “DNA” link at the top of the page in the blue bar. Click on the “get DNA” button to retrieve the entire DNA sequence for your fosmid.

4. Select the entire DNA sequence (Ctrl-A on MS Windows, ⌘ A on Mac OS X) and copy it (Ctrl-C on MS Windows, ⌘C on Mac OS X) onto the clipboard. Return to the NCBI BLAST tab and paste the sequence into the “query sequence” text box. Under the “algorithm parameters” section, make sure to change the compositional adjustments to “no adjustment” and turn off the low complexity filter. Then click on the “BLAST” button to align these two sequences against one another. This blastx search will allow you to determine the approximate location of the particular exon location on the entire contig. The reading frame will indicate which orientation the exon lies relative to the fosmid. A -1, -2, or -3 indicates it reads right to left while +1, +2, +3 indicates it reads from left to right.

Which frame is your exon in?

What are the approximate coordinates of your exon? (This is what I like to call “crude mapping”)

5. Return to the Gander Genome Browser, use the back button of your web browser to go back to the graphical view of contig61. Under the “Mapping and Sequencing Tracks” section, select “full” for the “Base Position” track. Under “Experimental Tracks,” select “dense” for the “Predicted Splice Sites” track and click on refresh. Enter the coordinates of your exon into the “position/search” box and click on jump. Click on “base” to see the nucleotides.

Directions for zooming: Click on any number on the base coordinate. This will zoom in 3X and center on that location. Continue to click on the coordinates until you can see the three open reading frames and the nucleotide sequence in the Base Position track. To the left of the Base Position track is an arrow. Verify that this arrow is pointed in the correct orientation with respect to the orientation of the gene. Clicking on an individual nucleotide will toggle on and off the three different reading frames for translation.

Using your blastx results, navigate to the translation start site of this gene. You should notice a green block at the indicated translation “start” location. This green box represents a start codon

for the gene (Methionine or M) and the end of the 5' untranslated region (5'UTR) of the mRNA. Record the location of the start codon in **Table 3** located at the end of this workshop **(this is what I like to call "fine mapping").**

Note 1: we are only identifying the translated exon sequences within the mRNA using this annotation strategy. We will not try to annotate the 5' and 3' untranslated regions in this workshop.

Note 2: Gander shows the positive strand of the contig sequence by default (as indicated by the arrow next to the base position track that points from left to right). To see the reverse complement of that sequence and still maintain the contig coordinates, click on the arrow to reverse the direction.

6. Navigate along the sequence in the appropriate direction and make sure you do not see any "red" amino acid boxes (which indicate stop codons) in the open reading frame you have identified previously. (Note: red boxes in a different row are OK; these are in a different reading frame and do not impact the translation.)

Based on what you see as you scan the exon, what are the potential results of a frame shift mutation (when a nucleotide is added or deleted)? What are the potential consequences of these results?

7. Using the blastx alignment results, navigate to the approximate ending coordinate of the exon. Verify that you have identified the beginning of the intron. The beginning of the intron (splice donor site) sequence usually starts with a "GT" (In some rare cases, an intron will begin with "GC"). Record the base pair location where the exon ends.

At the end of the exon, note whether the coding sequence ends within or at the end of a codon. If the exon ends in the middle of the codon, record on the table how many nucleotides in the codon are coded for by the exon. These extra nucleotides between the last complete codon and the donor site (donor phase) will combine with the extra nucleotides between the splice acceptor site and the first complete codon in the next exon (acceptor phase) to form a complete codon. The donor and acceptor phases must be compatible with each other (i.e. the sum of the donor and acceptor phases must be either 0 or 3) in order to maintain the open reading frame. I think of this as how much of an "overhang" is left by one codon in the 5' exon and then how much "overhang" must be present in the 3' exon to make a full codon.)

8. Continue this procedure for each exon in the gene. The reading frame for each exon might not be the same (your blastx results should have told you which frame each exon is in). This is OK as long as an intact codon is accounted for at the exon-exon junctions. Remember that after the initial exon, each subsequent exon does **not** need to begin with a start codon. Instead, an "AG" acceptor site will indicate the end of the preceding intron. Within the final exon, you should see a red box indicating a stop codon. The stop codon is necessary to stop translation, and marks the beginning of the 3' untranslated region (3'UTR) of the mRNA.

How do the exon/intron boundaries correlate with reading frames?

TABLE 3. ANNOTATION DATA FOR GENE ISOFORM _____

Exon Number	Start Coordinate	Stop Coordinate	Bases needed to complete the codon
One			
Two			
Three			
Four			
Stop Codon coordinates			

Part 5: Checking your work with Gene Model Checker (GMC)

Gene Model Checker, or GMC, can be found at http://gander.wustl.edu/~wilson/genechecker/index.html, or from the GEP home page under Projects -> Annotation Resources. This tool was made by the GEP to help identify and resolve common annotation errors. Passing one or more tests on GMC does *not* mean your gene annotation is correct, nor does it mean that you have indeed found an ortholog to a *D. melanogaster* gene.

GMC mostly checks to make sure your gene model follows the basic biological rules of a gene: it will make sure you have start and stop codons, and that they are directly attached to the first and last exons. It makes sure each internal exon has a canonical acceptor and donor sites. GMC also checks that the model has the same number of exons as the putative *D. melanogaster* ortholog.

To use GMC, you'll need your entire contig's DNA sequence saved in a .txt or .fasta file (you can find this on the course BB site), your exon coordinates, and the stop codon's coordinates.

Once you have all that information, enter it into GMC. Tell GMC what gene and isoforms you annotated, whether it was in a negative or positive reading frame, and if it was only part of the gene or all of it. It will look like this.

GMC is not very user-friendly. It can be picky about how you input certain information. For example, if entering "Contig 61" or "contig 61" instead of "contig61" gives an error. Sometimes GMC will explain the error and what you need to do to correct it, but not always. Input the information as it is shown here, and it should not give you problems.

Note that gene names in Drosophila are **case-sensitive**. Hence you will need to be careful when entering the gene name in GMC (and in the Gene Record Finder). For genes with multiple exons, separate the coordinates with a comma and a space. In general, the exons should be ordered from 5' to 3' (this means that for genes on the minus strand, the coordinates should be in descending order).

The GMC will report an error if it detects overlapping coordinates in the set of exon coordinates. If it gives you an error, sometimes rewriting the coordinates works.

Note the "Ortholog in *D. melanogaster*" field. It requires the isoform name (not just the gene name). Protein names in Drosophila have a -P suffix followed by one or more letters that identifies the isoform. For example, the name "Actbeta-PA" corresponds to the protein product derived from the A isoform of the *Actbeta* gene. Because different isoforms of the same genes could have different coding regions, this suffix is an essential part of your annotation.

After entering the information correctly, click "Verify Gene Model."

What results did you get? Take a screen shot of the initial output. Paste it here.

If you receive all passes, and perhaps a few legitimate "skips," that means your gene model is satisfies the basic biological rules for a gene. If, instead, you see "Fails" in red in the "Status" column, you should try to correct these errors. Ask for help if needed. **When you get to this point, get my attention and ask me to come over to your computer and I will show you a couple things about GMC that you'll need to know for when you do your own annotation!**