Life Sciences Division

Monica Munoz-Torres, PhD
Project Manager for Scientific Software
Phoenix Bioinformatics
moni@phoenixbioinformatics.org
@monimunozto

Genome Annotation Workshop for EMBL-ABR

Apollo Workshop - Exercises
02 November 2017

The purpose of these exercises is to get annotators thinking about the best way to utilize the resources available in Apollo. They are not meant to be 'step-by-step' descriptions on how to conduct each editing operation, but rather prompts and hints to encourage annotators to come up with their own strategies.

A few things to remember:

- Bullet points (like this one) are meant to bring to your attention to an ACTION item.
- **"Stop and observe"** signs are meant to encourage you to pause and think before the next action takes place.
- Some browsers and some computers are pickier than others about using data from a PDF file. So, when asked to use the sequence provided, avoid using the sequence provided in this file. Instead, copy the appropriate sequence for each exercise, from the sequences I have provided in the flat text file labeled as:
  Apollo_Exercises_Sequences_EMBLABR-2017-11-02.txt
- For the purpose of these exercises, when conducting searches using the US National Center for Biotechnology information (NCBI) **Blast** service please be sure to *"exclude"* the honey bee genome from the database. Do this by typing the name of the species (*Apis mellifera*) in the **Organism** box, in the **Choose Search Set** area of the Blast search page. Then check the **Exclude** button to the right of the box.
- Also during these exercises, please refrain from using the tracks labeled with the word **NCBI** in the Apollo selection track.
- Lastly, for these exercises, let us pretend that the Gene Ontology (GO) terms associated with the homologs we are using to query the Honey bee genome (e.g. from FlyBase, NCBI, etc.) are indeed 'demonstrated' GO terms associated with the functions, processes, and cellular locations of the gene products we are annotating. That is, GO annotations for which supporting evidence (experimental or predicted) is available. And let us do the same for the PubMed Identifiers (PubMed IDs) associated with these GO terms.

Happy annotating!

Apollo Workshop -
Exercise 1

It has come to our attention that the honey bee *Apis mellifera* ortholog of the *small ribonucleoprotein particle protein SmD2* is fragmented into 2 or more genes in the current gene set (Official Gene Set v3.2). In *Drosophila melanogaster SmD2* is involved in mRNA splicing, via spliceosome.

- Use this fragment from the ortholog in the dwarf honey bee *Apis florea* (also available at NCBI as XP_003697276.1; http://www.ncbi.nlm.nih.gov/) to pull the corresponding location in *Apis mellifera*. That is, select the **BLAT Search** option from the **Tools** drop down menu in Apollo, and query the honey bee genome using the dwarf honey bee amino acid sequence. **Note:** do not include the definition line that appears at the top in your search. This search may take a few minutes to complete.

```
>A_florea Sm D2-like
MLNNIRNITYFSDQSSLTKPKSEMTPEELAKREEEEFNTGPLSVLTQSVKNNTQVLINCRNNKKLLGRVK
```

AFDRHCNMVLENVKEMWTELPRTGKGKKKAKPVNKDRFISKMFLRGDSVILVLRNPLATASGK

Stop and observe: the direction of transcription: does it go right to left? Or does it go left to right?

- Drag the honey bee gene model you found into the into the **User-created Annotations (Uc-A)** area and retrieve its sequence to query the public databases with it to inspect their structure.

- Use this product to query the non-redundant sequence collection at the US National Center for Biotechnology information (NCBI). You are looking for possible orthologs of this protein, using BLAST at http://blast.ncbi.nlm.nih.gov/Blast.cgi.

Stop and observe: In the NCBI pages, review the conserved domains, compare the alignments with other sequences and inspect their length differences.

Back in Apollo, find evidence in support of this model by closely inspecting the transcriptomes of **Forager** bees and **Nurse** bees as well as available ESTs. For now, please refrain from using previous bee gene models (e.g. OGSv1.0). You can see that RNA-seq evidence supports the current exons in this protein, as well as a possible additional exon at the 5' end of the resulting product.

- Check the integrity and accuracy of your finalized product:
    - At the NCBI Blast page http://blast.ncbi.nlm.nih.gov/Blast.cgi, use blastp to query the collection of proteins from Insecta. To do this, go to the **Choose Search Set** section of the blastp page, and begin typing the word **Insecta** on the **Organism** box. Auto-complete options will be shown. Then, add a box to this section using the "plus sign" button to the right of the box, and begin typing the words *Apis mellifera*; when the desired organism shows up, select it, and then tick the **Exclude** box to the right. This means that you will conduct the search without including any of the honey bee sequences in the database.

    - In the **Information Editor** box in Apollo, add the GO terms you can find from the *Drosophila* gene model entries.

    - Find the references for these GO terms – and any other ones you wish to add, as well as articles discussing this gene in honey bees (if any). Find their PubMed IDs from the NCBI database (http://www.ncbi.nlm.nih.gov/pubmed) and add them to the **PubMed ID** box in the **Information Editor** in Apollo.

Stop and observe: After checking the current protein product against the sequences available at NCBI, the alignments show a discrepancy at the 5'end of the protein. Try answering the following questions:

- Can you spot the difference?
- Are you able to find the missing residues in the current *A. mellifera* genome assembly?
- Is there enough evidence to support this hypothesis?
- How many isoforms would you annotate in this case?

Apollo Workshop -
Exercise 2

It has come to our attention that in the European honey bee *Apis mellifera*, the ortholog of the *nuclear envelope spectrin repeat protein* (*Nesprin-1*), also known as *Muscle specific protein 300* (*MSP-300*) in Drosphila, is fragmented into several different genes in the official gene set (*Official Gene Set v3.2*).

In the fruit fly *Drosophila melanogaster*, *MSP-300* anchors nuclei to actin and has been reported to be essential for positioning of nurse cell nuclei during oogenesis, and thus production of mature ooctyes. More information available at http://en.wikipedia.org/wiki/Nesprin

- Use this fragment from the ortholog in the alfalfa leafcutter bee *Megachile rotundata* (also available at NCBI as XP_003705060.1; http://www.ncbi.nlm.nih.gov/) to pull the corresponding location in *Apis mellifera*. That is, select the **BLAT Search** option from the **Tools** drop down menu in Apollo, and query the honey bee genome using the leafcutter bee amino acid sequence. **Note:** do not include the definition line that appears at the top in your search. The protein product has over 12K amino acid residues in *M. rotundata*, so please be sure to begin with this smaller fragment. This search may take a few minutes to complete.

```
>M_rotundata nesprin-1-like (fragment)
MRIVEGRYSGSGGYWNVTILEGGDTGTWGYQKVGILERGGTRRWGYWNVGVPEGGDTGTWGYQKVGILER
GGTRRWGYWNVGVPEGGDTGTWGYQKVGILERGGTRRWGYWNVGVPEGGDTGTWGDSGAWGYWNVRILEG
GDTGTWRYQKVWSGGRTASSEELFQELDGRRNVGHFRSPTNPRPSSRASDSSFEESFERLVEEGELNGAK
VVKFEKITVRKSVREVAGTGVSHQRVLAETSRTPSEEHALEDSAYQSHSHGAPSHGSKSSSVTSFTRFPS
EESLSQRRGSSPQQHLGPDDRTPSEWYAEYHTQSFQNVAARIEYVRSKSEYDAHIAEIKDEQERVQKKTF
VNWINSYLSKRIPPLRVDDLIDDLKDGTRLLALLEVLSGEKLPVERGRNLKRPHFLSNANTALQFLQS
```

The results show that gene models GB40006-RA and GB40007-RA have significant sequence similarity with this gene model. You will see a yellow highlight surrounding the area of significant sequence similarity. Use the “highlighter” button to the right of the coordinates box to clear the highlight, or select the option from the **View** menu to do so.

If you had used the entire sequence for the *M. rotundata* protein, you would see high-scoring segment pairs (hsp) across honey bee gene models GB40006-RA, GB40007-RA, GB40008-RA, GB40009-RA, and GB40010-RA.

Stop and observe: the direction of transcription: does it go right to left? Or does it go left to right?

- Drag and merge GB40006-RA, GB40007-RA, and GB40008-RA into the **User-created Annotations (Uc-A)** area**.**

- Rename the resulting transcript to: nesprin-1-*username*-RA

- Tick (check box) to show the **Nurse RNA-seq reads** track in the Annotator Panel; you will see it listed under the **Tracks** tab. Observing these **Nurse RNA-seq reads**, we can see that there are no reads (i.e. transcription evidence) in support of the exon located at 483,800 - 483,705, while there are many reads in support of an intron passing across the region. Delete the exon (it was the first exon of GB40007-RA).

Closer inspection of RNA-seq reads also reveals the absence of transcription evidence in the region, which supports the idea that **there should not be** an exon in coordinates 494,235 - 491,660.

- Delete exon (it was the second exon of GB40006-RA).

It is now time to inspect the adjacent exons at the merging of GB40007-RA and GB40008-RA. Guided by RNA-seq reads (their length, not their exact coordinates), the phase and the structure of all exons downstream (in the three-prime direction - 3') of 454,041 (… 453,654), you may adjust the coordinates for the adjacent exons in this region, so they may have canonical splice sites.

- First, the last exon of GB40007-RA can be adjusted to the nearest canonical splice, located five-prime (5') in position 454,041.

Note that there is a gap between 453,759 .. 453,710.

- What used to be the first exon of GB40008-RA should also be adjusted to the nearest canonical site that extends the protein and is supported by RNA-seq reads. You could give it a try just dragging it by hand first. After trying a few sites, position 453,653 is a canonical splice site that conserves the phase of the rest of the product and matches the boundaries of RNAseq reads (reading frame is -2).

Stop and observe: We are being conservative about the length of the protein. There is actually another canonical splice site at 453,659; the only way of knowing whether we should extend it to 453,659 or not is by inspecting many other proteins to see whether those 2 amino acid residues are part of the protein.

- There are many exons in this model. So far, this newly created gene product is 7,600 amino acids in length. Use this product to query the non-redundant sequence collection at NCBI. You will be looking for possible orthologs of this protein, using Blast on the NCBI website at http://blast.ncbi.nlm.nih.gov/Blast.cgi

The search at NCBI retrieves homologs from other species of Apidae (*A. dorsata, A. florea, N. vitripennis, B. impatiens*) of similar length to the leafcutter bee's protein --> ~39Kbp, 12K aa.

Stop and observe: In the NCBI pages, review the conserved domains, compare the alignments with other sequences and inspect their length differences.

- Modify the exon coordinates at the flagged non-canonical sites at the junctions of gene models using RNA-seq to guide the changes. Try enabling the **Color by CDS** feature from the **View** menu to help you ensure that you are preserving the coding frame.

RNA-seq data show that the last exon in GB40008-RA may not be a real one (or perhaps, it is possible that there is more than one isoform (see position 431,750)). This is a possibility, since we know that here are 11 isoforms in the fruit fly *D. melanogaster*.

- With RNA-seq evidence as support, use Apollo's **edge-matching** functionality to modify the 5' end of the first exon of GB40009-RA so it matches the boundary of the transcripts shown below. The coordinates will move from 431,668 to that of the matching RNA-seq evidence at coordinate 431,683. The action you need to perform here is:
  - First click to select, and holding down the Shift key select the RNA-seq read; right-click over the exon, and choose the **Set as 5' end** option from the **right-click** menu. *Remember:* only those reads oriented in the same sense as the transcription will allow you to conduct the **Set as x' end** operation.

- In the Information Editor, name this isoform using the extension **-RA** and create a copy using the option **Duplicate** from the **right-click** menu. Name that second isoform as -RB.

- In the second isoform (-RB), delete the exon present at coordinate ~431,750 (it was the last exon of GB40008-RA). This reveals that the phase in the rest of the peptide is restored to that of original gene model. Use Apollo to copy the amino acid residues for the resulting protein (9,009 aa) and query the public databases again using BLAST at NCBI.

- Merge the resulting isoform -RB with GB40010-RA. Because Apollo recalculates reading frames and renames resulting models in the **User-created Annotations (Uc-A)** area, be sure to rename the resulting gene model to reflect that this is isoform B.

- At the junction of these two gene models, follow the RNA-seq reads to adjust the last exon of GB40009-RA to the nearest canonical splice site (it will be at coordinate 427,372), and the first exon of GB40010-RA to the nearest canonical splice site at 427,069, located immediately 3' of the assembly gap. This is the longest extension of the exon, supported by RNA-seq data.

You may also use additional evidence tracks to support these decisions: e.g. *Forager Illumina Bee Contigs*, *Mixed Antennae 454 Contigs*, and *Fgenesh++ with RNAseq training data*.

Stop and observe: In this peptide, at position 394,197 a GC splice donor (common in insects) is visible and heavily supported by the evidence. (GC/AG instead of GT/AG). Note this on the *Comments* box in the **Information Editor** and finish the annotation.

In the end, you should obtain a product of (approx.) 16,509aa in length. Don't worry if you are missing a residue or two, as long as you have recovered most of the protein. I will explain why this is possible during the exercises.

- Check the integrity and accuracy of your finalized product:
  - At the NCBI Blast page http://blast.ncbi.nlm.nih.gov/Blast.cgi, use blastp to query the collection of proteins from Insecta. To do this, go to the **Choose Search Set** section of the blastp page, and begin typing the word **Insecta** on the **Organism** box. Auto-complete options will be shown. Then, add a box to this section using the "plus sign" button to the right of the box, and begin typing the words *Apis mellifera*; when the desired organism shows up, select it, and then tick the **Exclude** box to the right. This means that you will conduct the search without including any of the honey bee sequences in the database.

  - In the **Information Editor** box in Apollo, enter (*Add*) the GO terms you can find from the *Drosophila* gene model entries.

  - Follow the bibliographic references for these GO terms – and any other ones you wish to add, as well as articles discussing this gene in honey bees (if any). Find their PubMed IDs from the NCBI database (http://www.ncbi.nlm.nih.gov/pubmed) and add them to the **PubMed ID** box in the **Information Editor** in Apollo.

Apollo Workshop -
Exercise 3

It has come to our attention that the honey bee *Apis mellifera* ortholog of the fruit fly (FlyBase) gene model CG31619 is fragmented into 2 or more genes in the current gene set (Official Gene Set v3.2). In *Drosophila melanogaster* CG31619 is involved in proteolysis, has metalloendopeptidase activity (molecular function), and belongs to the ADAMTS family of peptidases. ADAMTS stands for "A Disintegrin And Metalloproteinase with Thrombospondin Motifs".

We know that the homolog of CG31619 in the giant honey bee *Apis dorsata* is similar to the ADAMTS proteins.

- Use this fragment from the ortholog in the giant honey bee *Apis dorsata* (also available at NCBI as XP_003697278.1; http://www.ncbi.nlm.nih.gov/) to pull the corresponding location in *Apis mellifera*. That is, select the **BLAT Search** option from the **Tools** drop down menu in Apollo, and query the honey bee genome using the giant honey bee amino acid sequence. **Note:** do not include the definition line that appears at the top in your search. This search may take a few minutes to complete.

```
>Apis_dorsata ADAMTS-like protein partial
RENPRYWLQPCENPSDFRSEQCAAFDDVPYSGQLLKWYPHYDPSRPCALICRGEQSVENTGNRLRQE
TSVEKTLPHDATDALQLDSEETIVVQLADKVEDGTKCYTDSMDVCINGECMKVGCDLRVGSNKNTDPCGV
CGGNGSSCQSRYSWSLESISACSKSCGGGFKIAMAVCKAIGPDESVVDDSYCDPDNRPEKTLMPCNTHPC
```

Stop and observe: Look carefully through the resulting gene models. What is the direction of transcription? Does it go right to left? Or does it go left to right?

- Drag each of the honey bee gene models you found into the **User-created Annotations (Uc-A)** area and retrieve their sequences to query the public databases with them to inspect their structure.
- Use this product to query the non-redundant sequence collection at the US National Center for Biotechnology information (NCBI). You are looking for possible orthologs of this protein, using BLAST at http://blast.ncbi.nlm.nih.gov/Blast.cgi.

Stop and observe: In the NCBI pages, review the conserved domains, compare the alignments with other sequences and inspect their length differences.

Once you inspect the expected conserved domains, it will be apparent that the models need to be brought together. Back in Apollo, find evidence in support of this operation by looking through the transcriptomes of **Forager** bees and **Nurse** bees as well as available ESTs. For now, please refrain from using previous bee gene models (e.g. OGSv1.0). The available evidence will show support of splice site corrections and possible excess of exons.

- When satisfied with the final model, check the integrity and accuracy of your finalized product:
  - At the NCBI Blast page http://blast.ncbi.nlm.nih.gov/Blast.cgi, use blastp to query the collection of proteins from Insecta. To do this, go to the **Choose Search Set** section of the blastp page, and begin typing the word **Insecta** on the **Organism** box. Auto-complete options will be shown. Then, add a box to this section using the "plus sign" button to the right of the box, and begin typing the words *Apis mellifera*; when the desired organism shows up, select it, and then tick the **Exclude** box to the right. This means that you will conduct the search without including any of the honey bee sequences in the database.

- In the **Information Editor** box in Apollo, add the GO terms you can find from the *Drosophila* gene model entries.

- Find the references for these GO terms – and any other ones you wish to add, as well as articles discussing this gene in honey bees (if any). Find their PubMed IDs from the NCBI database (http://www.ncbi.nlm.nih.gov/pubmed) and add them to the **PubMed ID** box in the **Information Editor** in Apollo.

Apollo Workshop -
Exercise 4

It has come to our attention that the honey bee *Apis mellifera* ortholog of *RNA Polymerase II 140 KD* is fragmented into 2 or more genes in the current gene set (Official Gene Set v3.2). In *Drosophila melanogaster RNA Pol II* is involved in transcription from the RNA polymerase II promoter and has a DNA binding and DNA-directed RNA polymerase activity (molecular functions).

- Use the fragment provided below from the giant honey bee *Apis dorsata* gene model (also available at NCBI with identifier XP_006610182.1; http://www.ncbi.nlm.nih.gov/) to pull the corresponding location in *Apis mellifera*. That is, select the **BLAT Search** option from the **Tools** drop down menu in Apollo, and query the honey bee genome using the giant honey bee amino acid sequence. **Note:** do not include the definition line that appears at the top in your search. This search may take a few minutes to complete.

```
>Apis_dorsata RNA pol II subunit RPB2-like partial
MYSLEEDQYDDEDAEEISSKLWQEACWIVINAYFDEKGLVRQQLDSFDEFIEMSVQRIVEDSPQIDLQAE
AQHTSGEIENPVRHLLKFEQIYLSKPTHWEKDGAPSPMMPNEARLRNLTYSAPLYVDITKTIVKDGEDPI
ETQHQKTFIGKIPIMLRSKYCLLAGLSDRDLTELNECPLDPG
```

Stop and observe: the direction of transcription: does it go right to left? Or does it go left to right?

- Drag the honey bee gene models you found into the into the **User-created Annotations (Uc-A)** area and retrieve their sequence to query the public databases with it to inspect their structure.

- Use this product to query the non-redundant sequence collection at the US National Center for Biotechnology information (NCBI). You are looking for possible orthologs of this protein, using BLAST at http://blast.ncbi.nlm.nih.gov/Blast.cgi.

Stop and observe: In the NCBI pages, review the conserved domains, compare the alignments with other sequences and inspect their length differences. Closely inspecting the domains:
  - Are the expected conserved domains intact in each model?
  - Will bringing these models together improve the integrity of the RNA Pol II subunit RPB2 homolog in honey bee?

- Merge the models together, and re-arrange the boundaries at the joints of the merge region so that they are canonical splice sites on each side.

Stop and observe: Try answering these questions:
  - Did you spot the sequence error visible in the DNA-Track?
  - What effect do you think this had on the automated annotation process?

- When satisfied with the final model, check the integrity and accuracy of your finalized product:
  - At the NCBI Blast page http://blast.ncbi.nlm.nih.gov/Blast.cgi, use blastp to query the collection of proteins from Insecta. To do this, go to the **Choose Search Set** section of the blastp page, and begin typing the word **Insecta** on the **Organism** box. Auto-complete options will be shown. Then, add a box to this section using the "plus sign" button to the right of the box, and begin typing the words *Apis mellifera*; when the desired organism shows up, select it, and then tick the **Exclude** box to the right. This means that you will conduct the search without including any of the honey bee sequences in the database.

- In the **Information Editor** box in Apollo, add the GO terms you can find from the *Drosophila* gene model entries.

- Find the references for these GO terms – and any other ones you wish to add, as well as articles discussing this gene in honey bees (if any). Find their PubMed IDs from the NCBI database (http://www.ncbi.nlm.nih.gov/pubmed) and add them to the **PubMed ID** box in the **Information Editor** in Apollo.

Apollo Workshop -
Exercise 5

It has come to our attention that the honey bee *Apis mellifera* ortholog of *Ceramidase* is fragmented into 2 or more genes in the current gene set (Official Gene Set v3.2). *Ceramidase* is an enzyme, which cleaves fatty acids from ceramide, producing sphingosine (SPH), which in turn is phosphorylated by a sphingosine kinase to form sphingosine-1-phosphate (S1P). Ceramide, SPH, and S1P are bioactive lipids that mediate cell proliferation, differentiation, apoptosis, adhesion, and migration.

- Use the fragment provided below from the European bumble bee *Bombus terrestris* gene model (also available at NCBI, identifier XP_003397164.1; http://www.ncbi.nlm.nih.gov/) to pull the corresponding location in *Apis mellifera*. That is, select the **BLAT Search** option from the **Tools** drop down menu in Apollo, and query the honey bee genome using the European bumble bee amino acid sequence. **Note:** do not include the definition line that appears at the top in your search. The complete protein product is far larger, so please be sure to begin with this smaller fragment. This search may take a few minutes to complete.

```
>B_terrestris Ceramidase-like
GTTTAAGAGTGTTCGCGCCAATTGTTCGCGGCGAGACTGGCCGTGCAGACCAGCTGTTATAGCCGCGTCT
CCGCTCTGTCTCTGCTGATCCATCGATCACCTACGCATCGATCCCTCGTTCGATCAACGTGGTCATGAGC
TGGAGCGTTTGAGCGCCGCTATCAGACTGGCGGCAGAGAAAAACTGAATGGAGGCACCGGCAGTTGGACG
CTTTAGAATCCTTGCGTTGTTGACGATATGGCTGGTCCAGCTTGCGGTGCCCGGCGCCATCGCGTCTTAC
AGCATCGGGGTGGGCAGAGCAGATGCTACAGGACCCGCCGCTGAAATTGTTTTTATGGGCTACGCGAAGA
TCGATCAAAAAGGATCAGGAATCCATCTTCGAACATTCTCCCGCGCATTCATCATCGACGATGGCGAGGA
GAGGTTCGTCTTCGTCAGCGTGGATAGCGCCATGATAGGAAACGGCGTTCGTCAAACGGTGTTGCAGAAT
CTTGAAAAGGAGTTTGGCAGCCTGTACACAGAGAAAAATGTGATGATCAGTGCAACTCACTCGCACTCCA
CACCCGGTGGATTCATGTTGCACATGTTGTTCGATATTACGACATTCGGTTTCGTTCAAGAGACCTTCGA
TGCTATGGTCAAGGGAATCACGAAGAGTATTCAACGTGCTCACTATGCCATAGTTCCAGGCAGAATATTC
ATCACCCATGGAGAAGTTCATGGTGTGAACATTAATAGAAGCCCATCCG
```

Stop and observe: the direction of transcription: does it go right to left? Or does it go left to right?

- Drag the honey bee gene model(s) you found into the into the **User-created Annotations (Uc-A)** area and retrieve its/their sequence to query the public databases with it/them to inspect their structure.

- Use this product to query the non-redundant sequence collection at the US National Center for Biotechnology information (NCBI). You are looking for possible orthologs of this protein, using BLAST at http://blast.ncbi.nlm.nih.gov/Blast.cgi.

Stop and observe: In the NCBI pages, review the conserved domains, compare the alignments with other sequences and inspect their length differences. Closely inspecting the domains:

  - Are there any conserved domains?
  - Are they intact in each model?
  - After inspecting the biological evidence in the area, do you think these models should be brought together? And if so, how?

- Merge the models together, and re-arrange the boundaries at the joints of the merge region so that they are canonical splice sites on each side. Then try answering the following question:
  - Are you able to establish whether you should annotate UTRs to this gene model hypothesis?

- Once you are satisfied with the final model, check the integrity and accuracy of your finalized product:
    - At the NCBI Blast page http://blast.ncbi.nlm.nih.gov/Blast.cgi, use blastp to query the collection of proteins from Insecta. To do this, go to the **Choose Search Set** section of the blastp page, and begin typing the word **Insecta** on the **Organism** box. Auto-complete options will be shown. Then, add a box to this section using the "plus sign" button to the right of the box, and begin typing the words *Apis mellifera*; when the desired organism shows up, select it, and then tick the **Exclude** box to the right. This means that you will conduct the search without including any of the honey bee sequences in the database.

    - In the **Information Editor** box in Apollo, add the GO terms you can find from the *Drosophila* gene model entries.

    - Find the references for these GO terms – and any other ones you wish to add, as well as articles discussing this gene in honey bees (if any). Find their PubMed IDs from the NCBI database (http://www.ncbi.nlm.nih.gov/pubmed) and add them to the **PubMed ID** box in the **Information Editor** in Apollo.

Apollo Workshop -
Exercise 6

It has come to our attention that the honey bee *Apis mellifera* ortholog of the *Dopamine receptor 2 (D2R)* is fragmented into at least 2 genes in the current gene set (Official Gene Set v3.2).

This gene encodes the D2 subtype of the dopamine receptor, which is coupled to Gi subtype of G protein-coupled receptor. This G protein-coupled receptor inhibits adenylyl cyclase activity. In mice, regulation of D2R surface expression by the calcium sensor NCS-1 in the dentate gyrus (part of the hippocampus) is involved in exploration, synaptic plasticity, and memory formation. In flies, activation of the D2 autoreceptor protected dopamine neurons from cell death induced by a toxin mimicking Parkinson's disease pathology.

- Use the fragment provided below for the Dopamine Receptor 2 gene model from the leafcutter ant *Atta cephalotes* (available at NCBI, with identifier XP_012061111.1; http://www.ncbi.nlm.nih.gov/) to pull the corresponding location in *Apis mellifera*. That is, select the **BLAT Search** option from the **Tools** drop down menu in Apollo, and query the honey bee genome using the leafcutter ant amino acid sequence. **Note:** do not include the definition line that appears at the top in your search. This search may take a few minutes to complete.

```
>gi|801401472|ref|XP_012061111.1|  PREDICTED:  dopamine  receptor  2
isoform X2 [Atta cephalotes]
MNDSEIYLLNWEEEVHALNNDLDRSFYNANYPSFNHTYEDLWELATDRIGLAIVLLLFSVATVFGNSLVI
LAVFRERHLHTATNYFVTSLACADCLVGLVVMPISAVYEVLENRWLFTTDWCDVWRSLDVLFSTASILNL
CVISLDRYWAITDPFTYPTRMSRKRAAILIAIVWICSSAISFPAIAWWRAVRTEQVPEDKCPFTENLGYL
IFSSTISFYLPLFVMVFTYYRIYRAAVIQTRSLKLGTKQVMMASGELELTLRIHRGGTTTNTDARHLFRT
ASSTPEDLQDLEEPLTALHNNGLTRVPSARINNKQHLGKNFSLSRKLAKFAKEKKAAKTLGIVMGVFIIC
WLPFFVVNLWSGFCTRCIWQEEIVSAAVTWLGWINSGMNPVIYACWSRDFRRAFVRILCYCCPRKMKRRY
QPAFRCKPSQKFASGRYYSAYSLHHVRGSRENSGEQTYI
```

Stop and observe: the direction of transcription: does it go right to left? Or does it go left to right?

As suspected, there are two gene models in the honey bee set with significant sequence similarity hits when querying the genome using the D2R from the leaf cutter ant. The highest scores for sequence similarity are located in Group15.19.

- Drag the honey bee gene models GB50154-RA and GB50155-RA into the **User-created Annotations (Uc-A)** area and retrieve their sequence to query the public databases with it/them to inspect their structure.

- Use each of these sequences separately to query NCBI's nr (non-redundant sequence collection) for orthologs, using BLAST at http://blast.ncbi.nlm.nih.gov/Blast.cgi Remember that for the purpose of this exercise, we will *"exclude"* the honey bee genome from the database when conducting searches on the NCBI Blast pages. Do this by typing the name of the species (*Apis mellifera*) in the **Organism** box, in the **Choose Search Set** area of the Blast search page. Then check the **Exclude** button to the right of the box.

Stop and observe: As a G protein–coupled receptor (and member of the opsin family), D2R has a seven-transmembrane domains (7TM). When looking at the BLAST results for GB50154-RA, you can see this domain identified in the panel with information from the Conserved Domain Database (at the top of the BLAST results page). Further inspection of the 7TM domain shows that it seems to be incomplete.

Although no conserved domains are identified, GB50155-RA covers a portion of the 3' end of the D2R sequences available in NCBI. It is evident we are on to the right path, and now we need to bring these two models together. But not just yet!

Stop and observe: In the absence of more gene models in the region, we must turn to using the available experimental evidence. In this case, there is deep coverage of the region with RNA-seq reads.

- Tick (click on the box) to show the **Nurse RNA-seq reads** track in the Annotator Panel; you will see it listed under the **Tracks** tab. And use these reads to assist you in changing the boundaries as you go.

By looking through the alignments of the BLAST hits (on the NCBI pages), it seems that the last string of conserved amino acids (aa) across taxa is located about 250 aa into the sequence we obtained for GB50155-RA. After that, the significant sequence similarity drops. Look through the alignments in NCBI to identify this string.

- Back in Apollo, use the string of conserved aa sequence from the honey bee model (`ASGELELTLRIHRGGG`) to query the genome again, so you can locate it in the browser. From the **File** menu on the upper left corner in the Apollo window (shaded in blue, next to the logo), choose the **Add sequence search track** option and paste the string. Choose the appropriate option for the search and click on the **Search** button. Zoom to base level to inspect the match for the string you just used (it is visualized as a yellow rectangle and black arrow in a new track).

Detailed inspection of this region, the alignments in NCBI, as well as the FASTA sequences of the proteins with significant sequence similarity for our query will give us a hint about how to locate and annotate the "gap" that seems to exist between the two gene models we have identified.

- Again in the NCBI results page you generated, look at the alignment results comparing the positions that your honey bee gene models cover on the same sequence; for example, compare the alignments of GB50155 and GB50154 with the D2R sequence of *Apis cerana* (Eastern honey bee). Note that there is a string of approximately 130 amino acids where no sequence similarity is found with either honey bee model.

Back in our genome browser, position 795,505 is the last nucleotide in the conserved string from GB50154-RA. This exon is being translated in the +2 frame – be sure to use the **Color by CDS** option from the **View** menu to differentiate the frames. Then, approximately around position 795,511, we can observe that the amino acid sequence from the +3 frame is actually quite similar to the sequences available in NCBI (`NTDARHLFRTPSSTPEDLQD…`). This denotes a shift in the coding frame!

We are able to see deep coverage of RNA-seq reads over this region, and all across to the region where GB50155 is located. It is evident that the gene models need to be brought together, but more edits must be done in order to recover the amino acid sequence that most closely resembles those of the known gene models from closely related species, if possible.

- Merge the gene models to bring them together.

- To fix the shift in frame, you will need to conduct a sequence alteration. Position the cursor in coordinate 795,506 (the first nucleotide after the conserved region), and remove the next 2 nucleotides so the CDS can be shifted to the next frame. Select the entire gene model to see that the sequence has been extended.

- Using the RNAseq reads, refine the boundaries of the offending exon by extending it all the way to the "splice" area – apparent as lighter green portions of the RNAseq reads. Use the 'Shift' key to select both the annotation and the first of the RNA-seq reads that crosses the "splice" area, then right click (over the gene model in the **U-cA**) to bring up the menu with the **Set as 3' end** option. Select the entire gene model and check its sequence to corroborate that the sequence has been extended.

RNA-seq reads around the beginning of the second portion of this gene (first exon of the original GB50155-RA model) show that that first -now offending- exon can be removed. RNAseq reads also show that there is an additional exon in the region 824,881 and 825,230. Other evidence actually shows the possibility of this entire region being an intron. Let's annotate the isoform that includes the exon.

- You may drag a "forward strand" read and adjust the boundaries appropriately for this isoform. Notice that after adjusting the boundaries, the warning of a non-canonical splice site goes away. Then, select the entire gene model to observe the final changes in the amino acid sequence. The final product should be 458 amino acids long.

- Use the RNAseq reads located at the 5' end (N-terminus) and 3' end (C-terminus) of the protein to see if you can annotate UTRs.

- You may then compare it with the public databases to corroborate accuracy and integrity of the protein, and to add all pertinent details and editing comments in the **Information Editor**.
    - At the NCBI Blast page http://blast.ncbi.nlm.nih.gov/Blast.cgi, use blastp to query the collection of proteins from Insecta. To do this, go to the **Choose Search Set** section of the blastp page, and begin typing the word **Insecta** on the **Organism** box. Auto-complete options will be shown. Then, add a box to this section using the "plus sign" button to the right of the box, and begin typing the words *Apis mellifera*; when the desired organism shows up, select it, and then tick the **Exclude** box to the right. This means that you will conduct the search without including any of the honey bee sequences in the database.

    - In the **Information Editor** box in Apollo, add the GO terms you can find from the *Drosophila* gene model entries.

    - Find the references for these GO terms – and any other ones you wish to add, as well as articles discussing this gene in honey bees (if any). Find their PubMed IDs from the NCBI database (http://www.ncbi.nlm.nih.gov/pubmed) and add them to the **PubMed ID** box in the **Information Editor** in Apollo.

Apollo Workshop -
Exercise 7

It has come to our attention that the honey bee ortholog of *adenomatous polyposis coli (APC) - like* is fragmented into possibly 4 or more genes in the current gene set (Official Gene Set v3.2). In *Drosophila melanogaster adenomatous polyposis coli (APC)-like* is involved in beta-catenin binding and microtubule binding. *APC* is a negative regulator that controls Beta-catenin concentrations and interacts with E-cadherin, which are involved in cell adhesion. *APC* is classified as a tumor suppressor gene. In humans, mutations in the *APC* gene may result in colorectal cancer.

- Use the fragment provided below from the Florida carpenter ant *Camponotus floridanus* gene model (available from NCBI with ID EFN68235.1; http://www.ncbi.nlm.nih.gov/) to pull the corresponding location in *Apis mellifera*. That is, select the **BLAT Search** option from the **Tools** drop down menu in Apollo, and query the honey bee genome using the carpenter ant amino acid sequence. **Note:** do not include the definition line that appears at the top in your search. This search may take a few minutes to complete.

```
>Camponotus floridanus APC-like partial
GRVLCGSAASVGVEGVGGAWGAQRRQSPPPLASRRLESAADAVVAAAAAAAVATTAASFNPVTQRCAKKG
AVDAVAPRRIYPGPPLVAETSKEDAMEPEAKEEDERRSSTPSPEFYLRRSKRYNEDGSSEEETKQDSISL
PNHRLPSSYRGTWPIRRDVWANQQIGFPAQQSTSLAAHNDVASVMSFSSSSNSAGLDSHSVELQGHRRLG
AKVDVVYNLLGMLEKNGRDDMSTTLLSMSTSLDSCLVMRQSGCLPLLVQLIHAPRQDPDTRDRAMQALHN
VVHAKSDERAGRREARVLRFLEQLRDYCQSLRMSLERGQSMDDLERGHPAATIAALMKLSFDEAHRHAMC
QLGGLHAVAELIEMDHLAHGSESDDQNCITLRRYAGMALTNLTFGDGNNKALLCSFKEFMKALVSQLKSP
SDDLRQVTASVLRNLSWRADSSSKQTLREVGAVTGLMKAAMEGRKESTLKSILSALWNLSAHCSTNKVDI
```

Stop and observe: the direction of transcription: does it go right to left? Or does it go left to right?

- Drag the honey bee gene model(s) you found into the into the **User-created Annotations (Uc-A)** area and retrieve its/their sequence to query the public databases with it/them to inspect their structure.

- Use this product to query the non-redundant sequence collection at the US National Center for Biotechnology information (NCBI). You are looking for possible orthologs of this protein, using BLAST at http://blast.ncbi.nlm.nih.gov/Blast.cgi.

Stop and observe: In the NCBI pages, review the conserved domains, compare the alignments with other sequences and inspect their length differences. Closely inspecting the domains:
  - Are there any conserved domains?
  - Are they intact in each model?
  - After inspecting the biological evidence in the area, do you think these models should be brought together? And if so, how?

- Merge the models together, and re-arrange the boundaries at the joints of the merge region so that they are canonical splice sites on each side.

- Once you are satisfied with the final model, check the integrity and accuracy of your finalized product:
  - At the NCBI Blast page http://blast.ncbi.nlm.nih.gov/Blast.cgi, use blastp to query the collection of proteins from Insecta. To do this, go to the **Choose Search Set** section of the blastp page, and begin typing the word **Insecta** on the **Organism** box. Auto-complete options will be shown. Then, add a box to this section using the "plus sign" button to the right of the box, and begin typing the words *Apis mellifera*; when the desired organism shows up, select it, and then tick the

Exclude box to the right. This means that you will conduct the search without including any of the honey bee sequences in the database.

- In the **Information Editor** box in Apollo, add the GO terms you can find from the *Drosophila* gene model entries.

- Find the references for these GO terms – and any other ones you wish to add, as well as articles discussing this gene in honey bees (if any). Find their PubMed IDs from the NCBI database (http://www.ncbi.nlm.nih.gov/pubmed) and add them to the **PubMed ID** box in the **Information Editor** in Apollo.