Exercise 2: Browser-Based Annotation and RNA-Seq Data

Jeremy Buhler

September 2, 2020

This exercise continues your introduction to practical issues in comparative annotation. You'll be annotating genomic sequence from the dot chromosome of *Drosophila mojavensis* using your knowledge of *BLAST* and the visualization tools provided by the GEP's Mirror of the *UCSC Genome Browser*. You'll also consider how best to integrate information from high-throughput sequencing of expressed RNA.

1 Getting Started

To begin, go to GEP's Mirror of the *UCSC Genome Browser* at `https://gander.wustl.edu/`. Select "Genome Browser" from the left-side menu. Enter "D. mojavensis" into the "Enter species or common name" field, and then choose the "Improved Dot (GEP/HWK2)" assembly for viewing. Finally, hit the *Go* button to start looking at the sequence.

The entire dot assembly is about 1.7 Megabases (Mb) in length; zoom out to see everything. This assembly is built from a set of overlapping fosmid clones prepared in 2009. We've added a variety of information to the genome browser to help you annotate, such as:

- gene-structure predictions from several different tools;
- repeats annotated using the *RepeatMasker* program;
- *BLAST* hits to *D. melanogaster* proteins;
- RNA-Seq data, which we'll describe in more detail later.

To keep the view to a manageable level, click on the "hide all" button beneath the genome browser image. Go to the bottom of the page and configure the display settings for the evidence tracks listed in the table below. Hit "refresh" after making these changes to update your display. Zoom out so that you can see the entire scaffold, which is about 1.7 Mb long.

Evidence Track	**Display Mode**	**Track Description**
D. mel proteins	pack	*BLASTX* alignments of *D. melanogaster* proteins
TopHat junctions	pack	Predicted introns based on spliced RNA-Seq reads
RNA-Seq Coverage	full	Number of RNA-Seq reads aligned to each genomic position
Genscan Genes	dense	Gene models predicted by *Genscan*
Nscan Genes	dense	Gene models predicted by *Nscan*
SNAP Genes	dense	Gene models predicted by *SNAP*
Geneid Genes	dense	Gene models predicted by *Geneid*
RepeatMasker	dense	Regions with similarity to *Drosophila* transposons
Simple Repeats	dense	Tandem repeats identified by *Tandem Repeats Finder*

2 Assessing *BLAST* Evidence in the Browser

Each labeled feature in the "D. mel proteins" track represents a *BLAST* hit to one annotated protein sequence from *D. melanogaster*. Individual HSPs (High-scoring Segment Pairs, a fancy name for aligned segments) in a *BLAST* hit are displayed as rectangular boxes spanning the bases of the assembly that matched; distinct HSPs from one protein are joined by a line with arrows indicating the strand of the match. To see the detailed *BLAST* output for a hit, including E-values and alignments for each HSP, click on the hit in the browser window. You can sort the HSPs by any of the columns in the summary view and click on any HSP to see its detailed alignment.

The *BLAST* hits are divided into three sets: high quality (E-value $\leq 10^{-50}$), medium quality (E-value $\leq 10^{-30}$), and low quality (E-value $\leq 10^{-10}$). You can show only one or two of these sets by clicking on the "D. mel proteins" track name in the configuration section below the display and checking or unchecking the box for each set.

You may notice that there are many similar *BLAST* hits with almost the same name, such as "Crk-PA," "Crk-PB," and "Crk-PC". These are the predicted protein products for different spliced isoforms of the gene *Crk*; the corresponding mRNAs are similarly named "Crk-RA," "Crk-RB," and "Crk-RC."

Now that you're oriented, it's time to get to work. Look at the high-quality *BLAST* hit to the protein CaMKI-PA. Clicking on this *BLAST* hit will give you detailed information about all of its constituent HSPs.

> **Question 1** *What is the distance in the assembly between the first and last HSPs (in their order on the scaffold) in this hit? How many HSPs does it contain? Are all the HSPs likely to be from a single gene that is homologous to CaMKI? Why or why not?*

As you saw in Exercise 1, NCBI *BLAST* tries to report every sufficiently high-scoring HSP, without considering whether the hits form a consistent gene model. You'll need to look carefully at the *BLAST* output to judge which HSPs in a hit belong together.

When looking for homologous genes, you typically want to find a set of HSPs that are *collinear*, i.e. that can be sorted so that their starting coordinates are strictly increasing (or strictly decreasing) in *both* the query sequence and the assembly. We would expect that a *D. mojavensis* gene homologous to *CaMKI* would have exons in the same order as *CaMKI*'s exons in *D. melanogaster*, so matches between corresponding exons should form a collinear set of HSPs. It's worth mentioning that other *BLAST*-like tools, such as *WU-BLAST*, report only a collinear set of HSPs, rather than all sufficiently high-scoring HSPs; such tools make an annotator's life easier.

> **Question 2** *Based on your BLAST result, where in the assembly does there appear to be a single, consistent alignment to the CaMKI-PA protein, and on what strand does it lie?*

There are at least two other high-quality *BLAST* hits to the same region from different *D. melanogaster* genes: *CaMKII* and *Strn-Mlck*.

> **Question 3** *Of CaMKI, CaMKII, and Strn-Mlck, which gene is most likely to be the closest homolog to the feature in the region you identified in the previous question? Why? Why do you think the other two genes also match in this region? Do they have better matches elsewhere in the assembly?*

3 Digging Deeper: Exon Boundaries, and the Utility of RNA-Seq Data

Given the limitations of *BLAST* we've just seen, you can't be sure that it has found all and only the HSPs constituting the full alignment of a given protein. Hence, it isn't a bad idea to check your protein alignment with a second program. First, you'll need the sequence of the CaMKI-PA protein. You can get the sequence directly from the *BLAST* detail view by clicking on the FlyBase ID in the description at the top, then cutting and pasting the sequence from the "Sequence" tab of the resulting information screen. If you have an NCBI *BLAST* database of your proteins, you could also use the `blastdbcmd` command-line tool to extract the protein sequence you want.

BLAT, the Blast-Like Alignment Tool, is a sequence comparison tool written by the maintainers of the *UCSC Genome Browser*. While it is not quite as sensitive as *BLAST*, its speed and tight integration with the browser make it very useful for quickly aligning a known gene or protein sequence to an assembly. To align CaMKI-PA to our assembly with *BLAT*, select the "BLAT" menu item (under the "Tools" tab) at the top of the browser screen, paste the CaMKI-PA sequence into the box, and hit "submit". Find the result that matches the region that you believe to contain the *CaMKI* homolog in the assembly, and click to show this region in the browser. You should see a new browser track, labeled "Your Sequence from Blat Search", that shows the component HSPs of your *BLAT* hit in a form that can be visually compared to the HSPs of your *BLAST* hit.

> **Question 4** *Does the BLAT hit for this gene agree with the BLAST hit? Where and how does it differ? You may need to zoom in on the display to see the precise HSP boundaries.*

One major difference between the *BLAST* and *BLAT* results is that the latter has more distinct HSPs. It could be that *BLAT*'s HSPs are distinct exons of the gene, or that they are just alignment mistakes caused by, e.g., not extending a match through a homologous region with too many gaps. To correctly annotate this gene, you'll have to decide where the likely exon boundaries lie.

> **Question 5** *Does it make sense to divide some of the longer BLAST HSPs into multiple exons? What is your evidence for doing so (Hint: look for gaps in the BLAST alignment for each HSP.) Can you justify such a split in all cases for CaMKI in which BLAT disagrees with BLAST? Why might a single BLAST HSP span an intron?*

You have two sources of evidence for declaring exon boundaries besides the *BLAST* data. One is the output of model-based gene predictors, such as *Nscan* and *Genscan*. Gene predictors are software tools that use extensive statistics on how eukaryotic genes are structured, including typical base composition of coding exons and sequences found at splice junctions (`GT`/`AG`), to help guess which parts of a sequence represent the exons of a gene. Some predictors, such as *Nscan*, can also use *BLAST* output to help guess where exons occur. Each set of connected bars in a gene prediction track represents one predicted gene; each bar in the set is a predicted exon.

> **Question 6** *Do the results of the model-based gene finders agree more closely with a division into exons matching BLAST's or BLAT's results?*

The other source of information is RNA-Seq data. RNA-Seq results are generated by producing many short sequence reads (< 100 bases) from an organism's expressed mRNAs. Each read is aligned back to the genomic reference sequence, and the number of reads overlapping each base of the assembly is plotted in the "RNA-Seq Coverage" track. This data shows qualitatively where transcription is occurring in the genome – ideally in the exons of genes.

> **Question 7** *Does the RNA-Seq coverage information better support BLAST's or BLAT's putative division of the gene into coding exons? Why? Is there evidence of possible coding exons not present in the BLAST or BLAT hits?*

A useful adjunct to the raw RNA-Seq coverage data is the list of *junctions* proposed by the *TopHat* program, which processes the list of aligned reads. A read that is split across two adjacent coding sequences may indicate that there is an intron between them. *TopHat* annotates a junction wherever one or more reads show evidence for such a split, with the two sides of the junction showing the presumed 5' and 3' boundaries of the intron.

> **Question 8** *Which numbered junction(s) support the gene model suggested by the RNA-Seq coverage data?*

> **Question 9** *Given all available evidence, how would you annotate the region of the assembly around your main BLAST hit to CaMKI? Describe the evidence supporting your annotation.*

4 More Uses and and Limitations of RNA-Seq Data

Let's try to annotate another region of the assembly. Find a *BLAST* hit that suggests a gene annotation for the region between (roughly) base 1,170,000 and base 1,210,000 of the assembly. Note that you may need to look at lower-quality *BLAST* hits. Use the technique of re-aligning the protein back to the assembly with *BLAT* to check your result.

> **Question 10** *Which gene seems to be present at this location? How does the BLAT alignment compare to the original BLAST hit in terms of predicted exons? Which alignment is in better agreement with the model-based gene-structure predictions?*

Given the rather large discrepancy between the observed *BLAST* hits at this locus and the predicted gene structure and *BLAT* hits, we'd like to look to the RNA-Seq data to resolve the gene structure. Recall that RNA-Seq reads are generated by sequencing expressed mRNAs, so we expect these reads to match the transcribed portions of the genome. Ideally, there should be many reads that match transcribed genomic sequence, while few or no reads match any other sequences.

Because RNA-Seq reads are not limited to protein-coding sequence, they indicate not only coding exons but also the extents of a gene's 5' and 3' untranslated regions, or UTRs, which are the parts of the mRNA that lie on either side of the coding sequence. Information about UTRs is generally *not* available from most computational gene predictors, which predict these sequences poorly if at all. (Indeed, all the gene prediction tracks on the browser except *Nscan* show only putative coding exons; *Nscan* indicates UTRs by smaller rectangular boxes at the 5' end of its predictions.) Hence, RNA-Seq is a valuable complement to gene predictors and other tools that focus on the coding exons alone.

Question 11 *What extents of the 5' and 3' UTRs for this gene that are supported by the RNA-Seq data? Where in the assembly do these UTRs seem to occur? Does this information concord with the BLAT alignment's view of the putative coding exons?*

Looking carefully at the RNA-Seq coverage in this region shows that not all coverage is likely evidence of exons. Unfortunately, you'll see that some non-transcribed sequences can look similar enough to transcribed sequences to make annotation based on RNA-Seq challenging.

To answer the following questions, you may find it helpful to modify the way the RNA-Seq coverage track is displayed in your browser. By default, the vertical scale of the track varies automatically to accommodate the highest peak in the displayed region. If that peak is very high (say, thousands of counts), it may make it difficult to view lower but still quite significant counts (say, 25-100) at other parts of the track. To disable this *auto-scaling* behavior, click on the "RNA-Seq Coverage" link under the "Genes and Gene Prediction Tracks" section. Change the "Data view scaling" field to "use vertical viewing range setting", and then change the "min" field to "0" and the "max" field to "100". Click on the "Submit" button to return to the genome browser view. These display settings will cause all peaks of height 100 or more to be displayed with the same (maximum) height.

Question 12 *Where within the span of the BLAT hit for the proposed gene is the RNA-Seq coverage maximal? Does this location correspond to a likely exon, based on the BLAST and BLAT hits and the predicted gene structures? What genome feature seems to have generated the read matches here?*

Question 13 *Why might an RNA-Seq read be generated from a repetitive element in your sequence? Would you trust that a read matching such an element is accurately aligned? Why or why not? Why is the coverage in repeat regions sometimes so high?*

Question 14 *Which RNA-Seq junctions annotated in this region support the BLAT alignment's idea of exon boundaries? Are there any conflicting junctions? Can you explain any of them with reference to the repeats in this region?*

Question 15 *Finally, how would you annotate this region, based on all the available evidence in your genome browser?*