Practical Course in Genome Bioinformatics

26.2.2016

Exercises - Day 6

http://ekhidna.biocenter.helsinki.fi/downloads/teaching/spring2016/

Answer the 5 questions (**Q1-Q5**) according to the work report instructions and include **Figures 1 and 2** in your work report according to the instructions below.

1. Preliminaries

We use a piece of web service software called Web Apollo for manual annotation. Web Apollo homepage can be found at: http://apollo.berkeleybop.org/

Open your Firefox internet browser (you can use Internet Explorer if you want but some functions may appear differently). Navigate to Web Apollo (OLD) demo page at http://icebox.lbl.gov/WebApolloDemo/sequences

NOTE: Only use this OLD demo of Web Apollo to do these exercises. If you navigate to Web Apollo from their new webpage, you will be guided to the new (and more "crowded") demo of Web Apollo 2.

Log in with the demo user credentials:

User name: **demo**
Password: **demo**

You will see a list of genomic DNA sequences of *Apis mellifera* (honey bee) opening on the page. Here you can e.g. search sequences and areas of interest and filter the sequences by sequencing length.

Next, we search a specific sequence named ”Group12.6” by typing it to the text field above the sequence name header and clicking search. The sequence is found and we can see it is of length 307788 bases.

2. Web Apollo annotation view

Click the sequence name of ”Group12.6” and the annotation view for this sequence opens in a separate tab.

NOTE: If you close the just-opened window by accident at any point or e.g. the internet connection is lost, you can always navigate back to this view by replicating the instructions above.

In this view, one can zoom into the actual reference sequence and click on and off various lines of evidence to guide the annotation of different areas of the reference sequence. We next annotate one predicted gene in this sequence:

1. Click on the evidence track called ”Augustus Set 12” on the left column. Gene predictions from Augustus become visible in the annotation view. Zoom into a prediction called ”au12.g2718.t1”. One can view details of the entire annotation by right clicking at the tip of the prediction arrow.

2. Next, drag-and-drop the gene prediction into the ”User-created Annotations” area with light-yellow background above. Right-click the annotation from the arrow tip and select ”Edit information (alt-click)” in the menu.

3. Set name field in the mRNA column of the menu to your student ID for this course, the same you use to login to taito-shell.csc.fi e.g. ”lhbio15” or your own CSC account. You can also enter your first name if you like to.
   NOTE: This step only is done to separate these demo annotations from each other. In your own annotation projects this field should be left untouched at this point. Close the ”Edit annotation” menu from the small cross in the top right corner.

4. If the screen is already or becomes full of student annotations at this point, you can toggle between collapsed and full view of the annotation area from the “User-created annotations” menu located on the left hand side of the light-yellow annotation area. Try it now.

5. Click on five more lines of evidence from the ”Available Tracks” menu on the left: ”NCBI RefSeq Protein Coding Genes”, ”Forager RNA-Seq reads”, “Forager RNA-Seq heatmap”, “Nurse RNA-Seq reads” and ”Nurse RNA-Seq HeatMap”. Save a screenshot of the view showing the collapsed view of the User-created annotation area and the opened tracks of evidence to your local computer. This is to be included as your **Figure 1** for the work report.

Q1: What can you tell from the NCBI RefSeq Protein Coding Genes track for ”au12.g2718.t1”?

Q2: There’s almost no overlap in either the “Forager” vs. “Nurse” RNA-seq reads or their HeatMaps. Why might this be?

3. A closer look into the new annotated gene

We next seek for verification for the annotation that we added based on Augustus gene prediction.

1. Get the cDNA sequence of the annotated region by right clicking the annotation arrowhead. Select ”Get sequence”.

2. Click and select the ”cDNA” radio button and copy the entire sequence with its header into a text editor of your choice (WordPad is recommended).

3. Navigate to NCBI web BLAST at http://blast.ncbi.nlm.nih.gov
4. Select ”nucleotide blast” on the opening page (center left side).

5. Paste or upload your annotation to the ”Enter query sequence” window and click ”BLAST” button on the bottom of the page.

Q3: Look at the best BLAST hit. Are there similarities between the best BLAST hit and the NCBI RefSeq Protein Coding Genes evidence track for the annotated region in the Web Apollo annotation view? If so, what does this mean?

We next seek even more evidence for a good annotation by aligning the annotated sequence to the BLAST hits.

1. Download the two best BLAST hits to your local computer and open them in a text editor (e.g. WordPad).
   - Select the two top hits by clicking the check boxes before the hit names in the hit list
   - Click “Download” and select “FASTA (complete sequence)”
   - Save the file as **blast_hits.fasta**

2. Paste your annotated sequence on top of these sequences in the text editor and save the file as **sequences.fasta** . If you need to fetch the sequence again from the Web Apollo annotation view, please do so.

3. Navigate to the MAFFT web tool (**M**ultiple **A**lignment using **F**ast **F**ourier **T**ransform) at http://www.ebi.ac.uk/Tools/msa/mafft/

4. Paste or upload the **sequences.fasta** file to the ”STEP 1 – Enter your input sequence” window. Then select ”ClustalW” as the OUTPUT FORMAT in ”STEP 2”. When done, click submit in ”STEP 3”.

5. On the result page, take screenshot of the alignment from the beginning so that the start of the matching region (”*”-characters) can be seen in the screenshot. Include this screenshot as **Figure 2** in your report.

Q4: What is the reason that the beginning (~300 bp) of the annotated sequence evidently does not align to the best BLAST hit (or the second blast hit for that matter)?

4. Finishing and evaluating the annotation

Navigate back to the Web Apollo annotation view. Right click on your annotation and select ”Edit information (alt-click)”. Set the shortened name of the best BLAST hit, “LOC100577596” that is, in the “Description” field of the annotation. Add a comment of your choice to the “Comment” field.

Try to think the comment content from the viewpoint of the person who is going to be checking the manual annotations. Then close the edit window. Re-open the edit window to be cer

Finally, click on the evidence tracks ”Official Gene Set v1.0” and ”Official Gene Set v3.2” from the Available Tracks menu on the left.

Q5: What could be the reason that *Official Gene Set v1.0* track shows no gene in the predicted gene location ”au12.g2718.t1” but *Official Gene Set v3.2* does?

NOTE: The user created annotations stay in the demo database for some time after we close Web Apollo. It is likely that the annotations are removed on a regular basis, e.g. nightly or weekly, by the demo system admins to keep the database clean. (At least the annotations from the last year’s course were not visible anymore!)

Please return the work report of these exercises to juhana.kammonen@helsinki.fi as PDF at latest on **Fri 25 March 2016**. General instructions for the report can be found on the course homepage:

http://ekhidna.biocenter.helsinki.fi/downloads/teaching/spring2016/