DNA Subway…Red Line + Apollo

Exercise 2. Mastering Apollo- Building Gene Models

Learning Objectives:

Students should be able to

1. Take a DNA sequence to the end of the Red line.
2. Visualize genes and gene predictions using a genome browser.
3. Evaluate the strength of evidence for gene models.
4. Examine DNA sequences in Apollo
5. Use the Apollo 'exon detail editor'
6. Delete, split, and merge exon in Apollo
6. Build Gene Models in Apollo
7. Name Gene Models and upload them to DNA subway.

Pre-lab notes:

1. Please register at least 24 h in advance as a user for the DNA Subway that we will use in this lab. http://dnasubway.iplantcollaborative.org/
2. The following exercise was adapted from exercises generously provided by the developers of the DNA Subway: iPlant Genomics in Education Examples - http://gfx.dnalc.org/files/evidence/Worksheets/; accessed Nov. 2011.

Goal: Use DNA Subway and Apollo to build well-supported gene models

Introduction

In exercise 1, you learned about the DNA Subway Redline. In this exercise you will analyze more plant genome sequences on the Red Line but take things a bit further. All of the available evidence is analyzed in Apollo. Like DNA Subway, Apollo allows you to view the evidence for a particular gene. But Apollo is more than a DNA viewer; it is also a genome editor. In this exercise you will evaluate the evidence and build precise gene models. Apollo was developed through collaboration between the Berkeley Drosophila Genome Project and The Sanger Institute. Your primary goals in this exercise are to appreciate the basic Apollo toolkit and to develop a series of evidence-based gene models that can be uploaded from Apollo into DNA subway.

Part 1: Ride the Red Line- compile DNA evidence

I. **Create a Project**

1. *Enter* DNA Subway at http://www.dnasubway.org.
2. *Click* the red square to annotate a genomic sequence.
3. *Select* sample sequence Arabidopsis thaliana (mouse-ear cress) Chr5, 100.00 kb.
4. *Provide* a title (required), a project description (optional) and *click* Continue.

II. Mask Repeats to Speed Up Subsequent Analyses

1. *Click* RepeatMasker.
2. Once the bullet has finished blinking, *click* RepeatMasker again to *view* a listing of repetitive DNA sequences RepeatMasker has identified and masked.
3. How many and which types of repetitive DNA did RepeatMasker identify? (Use a search engine to search for unfamiliar attributes such as Copia or Harbinger.) What do the different attributes indicate? What is the range of repeat lengths? Can you identify any association between types and length ranges?
4. *Close* the table to return to DNA Subway.
5. *Click* Local Browser to view the results in a graphical interface.
6. *Maximize* the browser window.
7. *Change* Show 10 kb to Show 100 kbp in the Scroll/Zoom utility.
8. How many and which types of repetitive DNA does the browser display?
9. Which of the two views, table or graphics, would you find easier to work with?
10. *Close* the Local Browser screen to *return* to DNA Subway.

III. Predict Genes

1. *Click* Augustus.
2. Once Augustus has finished *click* FGenesH. Then, *click* SNAP. Finally, *click* tRNA Scan. (The Augustus, FGenesH and SNAP algorithms predict protein-coding genes; tRNA Scan identifies tRNA genes.)
3. *Determine* whether any of the 3 programs run significantly longer than any other?
4. Again, *view* the results in the table view and the Local Browser.
5. How many genes did the gene predictors predict? Which would you choose to answer this question, the table or the browser?
6. Do the different programs predict the same genes or can you identify differences among the predictions? Which do you think got it right?
7. *Close* the table and *browser* screens to *return* to DNA Subway.

IV. Search Databases for Gene Evidence

1. *Click* the BLAST buttons to search databases of known genes and transcripts such as cDNAs or ESTs (BLASTN) and proteins (BLASTX) for sequences that match the genomic DNA sequence.
2. To upload datasets of your own, click *Upload Data*, then *browse for DNA data*. (*Download* sample data from http://gfx.dnalc.org/files/evidence/Annotation.) Upload the at_est_evidence. Then click the R to run the User BLASTN.
3. *View* BLAST matches in the table view and the Local Browser.
4. For how many predicted genes did BLAST generate biological evidence?
5. *Close* the table and browser screens to *return* to DNA Subway.
6. *Generate* authoritative gene models in Part 2.

Part 2: Synthesize Gene Predictions and Evidence into Gene Models

Prediction and evidence are good indicators for genes, yet the results of different algorithms don't always agree with each other – what do gene models look like that are supported by biological evidence? Can this information be associated with genomic DNA?

Technique 1: Edit Exons

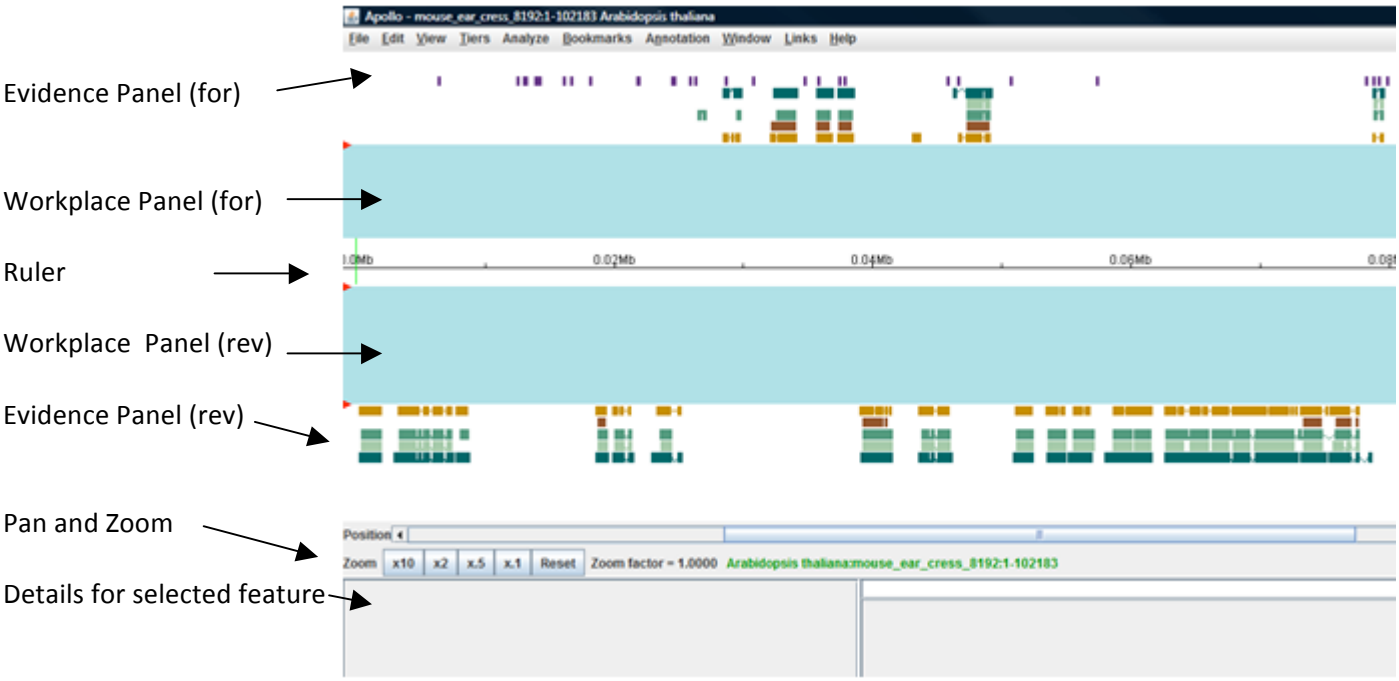
I. Open the Project Generated in Experiment 1 (if necessary)

1. *Click* My Projects.
2. *Click* the project that you generated in Part 1 above.

II. Build a Gene Model

1. *Click* Apollo.

You are loading the Apollo annotation editor. When Apollo loads you will see a horizontal ruler which represents your 100,000 bp. The panel above the ruler relates to the DNA strand in the forward direction and the panel below the ruler represents the reverse strand.



As you can see above there are workplace and evidence panels for both strands and a special area below to examine details.

2. *Click* Tiers and select Expand Tiers to *view* the entire evidence available. (Apollo initially collapses the different evidence types onto a single line each, regardless of how many pieces of evidence are available for each position.)

3. *Zoom*, *pan* and *scroll* to nucleotide position 29,500-33,500 until you can comfortably view details for a gene on the forward strand in this location.
4. You should now be able to distinguish gene features such as exons and introns.
5. *Compare* the predictions with each other and with the BLAST evidence – what similarities and differences can you identify?
6. Specifically, which of the predictions appears supported by the biological evidence?
7. Discrepancies between the gene predictions and biological evidence consist in:
    i. misplaced splice sites (caused by the inability of BLAST to determine splice sites);
    ii. inaccurate transcriptional start and termination sites and therefore inaccurate 5'- and 3'-untranslated regions (caused by difficulties predicting first and last exons due to transcriptional start and termination sites not following easily discernable patterns).
8. The Augustus gene prediction has the same structure as the other predictions and the BLASTN evidence, however, it is longer than the other predictions and therefore exhibits stronger agreement with the BLASTN evidence.
9. *Double-click* the Augustus prediction and *move* it onto the workspace – this is the foundation for a model for the gene in this location.
10. For this model you should now be able to *distinguish* exons, introns, coding sequences and UTRs. [Box=exon/horizontal line=introns/filled box=protein-coding sequence CDS/open box=untranslated region, UTR/vertical green line=start codon/vertical red line=stop codon]]
11. *Double-click* and *move* the longest piece of BLASTN evidence onto the workspace
12. Yellow arrows indicate non-canonical splice sites (see side bar).
13. *Compare* the Augustus prediction and the BLASTN evidence. You will find that they share the same exon-intron structure, but differ in the overall lengths: the gene model starts and ends further down-stream than the BLASTN evidence.
14. Use Exon Detail Editor to adjust the lengths of the flanking exons of the model:
    i. *double-click* the gene model.
    ii. *right-click* (*command*- or *Apple-click* for many Mac users) the gene model;
    iii. *select Exon detail editor* in the pop-up window to open the Exon Editor;
    iv. the Exon Editor displays the sequences of the gene model and the BLASTN evidence side-by-side; a red frame highlights the gene model;
    v. *grab* and *hold* the edge at the beginning of the model's first exon and *move* it 34 nucleotides to the left to position it flush with the start of the BLASTN-match;
    vi. *click* the end of the gene model depicted in the schematic view at the bottom of the Editor window to edit that part of the sequence;
    vii. *grab* and *hold* the edge at the end of the last exon and *move* it 41 nucleotides to the right to *position* it flush with the end of the BLASTN-match;
    viii. *close* the Exon Editor.
15. Examine your gene model:
    i. Does it agree with the biological evidence?
    ii. Does it have a start and stop codon?
    iii. Are the splice sites ok?
16. Name your model and record your edits in the Annotation Info Editor:
    i. *right-click* (*command*- or *Apple-click* for many Mac users) the model;
    ii. *select* Annotation info editor in the pop-up window to open the Annotation Information window;
    iii. *replace* the Symbol ID for the gene and the transcript, with a gene name.
    iv. *click* Edit … comments to associate the gene and/or the transcript with notes that explain and justify your edits;

v. *click* Close.

17. To conclude your annotation for this gene's structure:
    i. *right-click* (*command*- or *Apple-click* for many Mac users) the BLASTN evidence on your workspace;
    ii. *select* Delete selection;
    iii. *delete* any other evidence or prediction from the workspace until only your gene final model remains;
    iv. *click* menu tab File and *select* Upload to DNA Subway.

III. Browse Your Gene Model

1. *Minimize* or *close* Apollo.
2. *Bring up* the DNA Subway window.
3. *Click* Local Browser to *browse* your gene model.

Technique 2: Fix Start Codons

1. *Navigate* to nucleotide position 14,000-18,500.
2. *Identify* the differences among the predictions and the BLAST evidence.
3. Specifically, what start and end points for the gene do the different prediction and evidence items indicate?
4. Discrepancies between the gene predictions and biological evidence consist in:
    i. misplaced splice sites;
    ii. inaccurate transcriptional start and termination sites and therefore inaccurate 5'- and 3'-untranslated regions
    iii. missing or misplaced translational start and/or stop codons (caused by BLAST matches that may come from different species whose exons differ in length, or because Apollo automatically displays the longest open reading frame (ORF) as the coding sequence).
5. *Move* the Augustus gene prediction and the BLASTN evidence for this gene onto the workspace; a*djust* the 5'- and 3' ends of the model as described in Technique 1.II.14. [Box=exon/horizontal line=introns/filled box=protein-coding sequence CDS/open box=untranslated region, UTR/vertical green line=start codon/vertical red line=stop codon]
6. *Examine* the model's beginning: Does it have a start codon? *Zoom* in to the first third of the first exon (position 14060 through 14200) to *answer* this question.
7. To define a start codon for your model:
    i. *zoom* into the first exon;
    ii. *evaluate* whether the biological evidence (BLASTX) provides evidence for a start codon;
    iii. if the biological evidence does not provide a position for a start codon *choose* the first ATG/methione instead;
    iv. *move* your cursor to the upper edge of your screen;
    v. *grab* and *hold* the first green rectangle located within the first exon;
    vi. *move* the green rectangle all the way down onto your model to insert it as a new start codon.
8. To finalize your annotation:
    i. *zoom* out and verify your model (Technique 1.II.15.);
    ii. *record* your edits and *name* your model (Technique 1.II.16.); [Annotation Info Editor is set to accept the same name for a gene and its transcript. However, to name alternative

transcripts for the same gene append the gene name in the transcript field with "-transcript 1," "-transcript 2", etc.]

iii. d*elete* from the workspace any evidence or predictions other than your final model for this gene (Technique 1.II.17.);

iv. *upload* your result to DNA Subway (Technique 1.II.17.).

Technique 3: Delete Exons

1. *Navigate* to nucleotide position 46,500-51,500.
2. *Identify* the differences among the predictions and the BLAST evidence.
3. Specifically, what is the number of exons for the different predictions and evidence items?
4. Discrepancies between the gene predictions and biological evidence consist in:
   i. misplaced splice sites;
   ii. inaccurate transcriptional start and termination sites and therefore inaccurate 5'- and 3'-untranslated regions;
   iii. inaccurate gene structures (caused by missed or superfluous exons or introns in predictions and/or BLAST matches).
5. *Move* the Augustus gene prediction and the BLASTN evidence onto the workspace.
6. *Compare* the Augustus-derived gene model and the BLASTN evidence. You will find that the model's leading exon is not supported by BLAST evidence. To remove it:
   i. *click* the first exon in the gene model.
   ii. *right-click* (*command*- or *Apple-click* for many Mac users) the model;
   iii. *click* Delete selection.
7. *Adjust* the 5'- and 3' ends of the model by using Exon Detail Editor to match it to the BLASTN evidence as described in Technique 1.II.14. above.
8. To finalize your annotation:
   i. *zoom* out and verify your model (Technique 1.II.15.);
   ii. *record* your edits and *name* your model (Technique 1.II.16.);
   iii. d*elete* from the workspace any evidence or predictions other than your final model for this gene (Technique 1.II.17.);
   iv. *upload* your result to DNA Subway (Technique 1.II.17.).

Technique 4: Split Exons

1. *Navigate* to nucleotide position 18,500-21,000.
2. *Identify* the differences among the predictions and the BLAST evidence.
3. Specifically, what is the number of exons for the different predictions and evidence items?
4. Discrepancies between the gene predictions and biological evidence consist in:
   i. misplaced splice sites;
   ii. inaccurate transcriptional start and termination sites and therefore inaccurate 5'- and 3'-untranslated regions;
   iii. inaccurate gene structure.
5. *Move* the Augustus gene prediction and the BLASTN evidence for this gene onto the workspace; a*djust* the 5'- and 3' ends of the model as described in Technique 1.II.14
6. *Compare* the gene model and the BLASTN evidence. You will find that the gene model shows one long leading exon where the BLASTN evidence has two. To split this exon:
   i. *zoom* into the first exon in the gene model;
   ii. *click* the first exon in the gene model.

iii. *right-click* (*command*- or *Apple-click* for many Mac users) in the first exon approximately at the position where you wish to split it;
iv. *select* Split exon to split the first exon into two fragments;
v. *double-click* the gene model.
vi. *right-click* (*command*- or *Apple-click* for many Mac users) the gene model;
vii. *select* Exon detail editor in the pop-up window to open the Exon Editor;
viii. the Exon Editor displays the sequences of the gene model and the BLASTN evidence side-by-side; a red frame highlights the gene model;
ix. *maximize* the Exon Editor window;
x. *find* the gap in the highlighted sequence at the spot at which the background color in the former first exon changes – this is the position where the exon has been split;
xi. *grab* the 3'-edge of the first exon fragment and move it to the left and up to position it flush with the end of the first BLASTN exon;
xii. *grab* the 5'-edge of the downstream fragment and *move* it to the right and down to position it flush with the beginning of the second BLASTN exon;
xiii. *close* the Exon Editor.

7. You will find that by splitting the first exon into two you generated a non-canonical splice site. To adjust the splice site:
   i. *double-click* the gene model.
   ii. *right-click* (*command*- or *Apple-click* for many Mac users) the gene model;
   iii. *select* Exon detail editor in the pop-up window to open the Exon Editor;
   iv. *adjust* the beginning of the gene model's second (new) exon to start following (in 3'-direction) the nearest AG;
   v. *close* the Exon Editor.
8. To finalize your annotation:
   i. *zoom* out and verify your model (Technique 1.II.15.);
   ii. *record* your edits and *name* your model (Technique 1.II.16.);
   iii. d*elete* from the workspace any evidence or predictions other than your final model for this gene (Technique 1.II.17.);
   iv. *upload* your result to DNA Subway (Technique 1.II.17.).

Techniques 5 & 6: Merge Exons and Build Alternative Transcripts

1. *Click* Apollo, *expand* all tiers and *navigate* to nucleotide position 89500-92,500.
2. *Identify* the differences among the predictions and the BLAST evidence.
3. Specifically, do evidence items indicate contradicting structures for this gene?
4. Discrepancies between the gene predictions and biological evidence consist in:
   i. misplaced splice sites;
   ii. inaccurate transcriptional start and termination sites and therefore inaccurate 5'- and 3'-untranslated regions;
   iii. contradicting gene structures (caused by missed alternative splice forms in gene predictions);
5. *Move* the Augustus gene prediction and the longest BLASTN transcript evidence that resembles the model (5 exons, Exon #4 about 60 nt) onto the workspace; a*djust* the 5'- and 3' ends of the model as described in Technique 1.II.14.
6. *Record* your edits and *name* the model as described in Technique 1.II.16. above.
7. *Delete* the BLASTN evidence from the workspace.

8. *Compare* the gene model with the various biological evidence items. You will find that some BLASTN evidence shows Exon #4 to be about 110 nt long as opposed to 58 nt in the first model.
9. To build an alternative transcript for this gene:
    i. *double-click* the first model;
    ii. *right-click* (command- or Apple-click for many Mac users) the first model;
    iii. *select* Duplicate transcript to generate the foundation for an alternative transcript.
10. *Move* the BLASTN evidence that contains five exons with an Exon #4 of about 110 nt in length onto the workspace.
11. *Extend* the 3'-end of Exon #4 in the alternative model to the 3'-edge of the BLASTN evidence using Exon Detail Editor.
12. To update the open reading frame/coding sequence:
    i. *double-click* the new model; then
    ii. *right-click* (*command*- or *Apple-click* for many Mac users) the model;
    iii. *select* Calculate longest ORF.
13. Delete the BLASTN evidence from the workspace.
14. Record your changes and name the alternative gene model.
15. *Compare* the biological evidence with the two gene models. You will find that some BLASTN evidence shows a large fourth exon that encompasses Exon #4 and Exon #5 in the current two models.
16. To build a third alternative transcript:
    i. *duplicate* the first model again;
    ii. *shift-click* the fourth and fifth exons in the third model;
    iii. *right-click* (*command*- or *Apple-click* for many Mac users) one of the exons;
    iv. *select* Merge exons.
17. *Update* the third model's open reading frame/coding sequence.
18. *Record* your changes and *name* the new alternative gene model.
19. *Compare* the biological evidence with the two gene models. You will find that some BLASTX PROTEIN evidence shows a large second exon that encompasses Exon #2 and Exon #3 in the previous two models. However, the problem with using this information to build a fourth alternative transcript is that no biological evidence is available that would allow you to determine what other exons would be part of this fourth transcript – therefore you should not build a fourth alternative model without further evidence.
20. To finalize your annotation:
    i. *zoom* out and verify your models (Technique 1.II.15.);
    ii. d*elete* from the workspace any evidence or predictions other than your final models for this gene (Technique 1.II.17.);
    iii. *upload* your results to DNA Subway (Technique 1.II.17.).

Answer to selected questions in handout.

Part 1: Ride the Red Line- compile DNA evidence

II.3. 32 repetitive DNA segments; Low complexity DNA < 200 nt; Simple Repeats < 200 nt; Harbinger and Copia are transposons whose lengths range from several hundred up to 3000 nt.

II.8. The same as in the table.

II.9. The answer to this question will probably depend somewhat on the purpose of your work. The browser view is useful to get a quick overview; exact position data and lengths are easier to discern from the table.

III.3. The Augustus run takes significantly longer to complete than FGenesH or SNAP.

III.5. Although the graphical browser provides a quick way to see the predictions in context it may be faster to *copy* each table into a text processor, *replace* the word gene and *record* the number of replacements made – this is the number of genes found by the predictor. Augustus predicted 22 genes, FGenesH 22 genes, and SNAP 36 genes. In many locations all three programs predicted genes but, overall, there appear to be significant differences between the results returned by the three algorithms.

III.6. Predictions are deviating a lot. Like weather predictions, it is not possible to say which got it right without material evidence for the presence of genes such as transcripts (mRNA) or proteins.

IV.4. BLASTN identified evidence for 24 genes. Again, *copying* each table into a text processor and *conducting* a few of smart replacements, will reveal this number in seconds – and eliminate counting errors. BLASTX identified evidence for eight locations. Because of multiple BLASTX evidence items found in identical locations it is faster to determine BLASTX evidence using the graphical browser. BLAST is not programmed to discover splice sites and may align sequences that have already been aligned to a previous exon; this often leads to faulty splice sites. BLASTX results may not always synch well with gene structures indicated by the gene predictors and/or by BLASTN evidence. The major reasons for that are that a) protein matches from another organism may deviate in their amino acid sequence from what the sequence in the species whose DNA you analyze; b) protein matches may be from paralogs of genes elsewhere in the genome and therefore deviate from the genes under examination..

Part 2: Synthesize Gene Predictions and Evidence into Gene Models

Technique 1: Edit Exons

II.5. FGenesH and SNAP generated identical predictions for a spliced gene with 8 introns and 9 exons. The Augustus prediction and BLASTN evidence indicate the same gene structure but the two flanking exons extend further outward than the FGenesH and SNAP predictions. The lengths of the flanking exons differ between Augusts and the BLASTN evidence. BLASTX evidence confirms the presence of the different exons, however, it is out of synch in a few places with the BLASTN evidence and the predictions. However, as explained in the Results section for Experiment 1.IV.4., BLASTX evidence needs to be taken with a grain of salt.

II.6. The BLASTN evidence supports the Augustus prediction more than the other two predictions.

Technique 4: Split Exons

3. The predictions predict 2 exons, the BLASTN evidence indicates 3 exons.

Techniques 5 & 6: Merge Exons and Build Alternative Transcripts

3. The gene predictors predict similar structures of 3 exons (SNAP) and 5 exons (Augustus, FGenesH), respectively. BLASTN evidence indicates contradicting structures showing three different exons: one of about 60 nt length, one of about 100 nt length, and a third of about 770 nt length that merges exons IV and V predicted by Augustus and FGenesH as distinct exons into one large terminal exon instead.

Review Questions

- Which of the gene predictors is 'evidence based'? What does that mean?
- Which combination of zoom buttons (10X, 2X, 0.5X, 0.1X) would generate a small zoom in? a small zoom out?
- What tool in Apollo is used to make subtle changes to the length of an exon?
- Which is better at identifying the precise ends of the introns- gene predictors or BLAST?
- What is the fate of gene models generated in Apollo?