DNA Subway…Red Line

Exercise 1. Using DNA Subway Red Line to generate evidence for a gene model.

Learning Objectives:

Students should be able to

1. Use various methods to visualize genes and gene predictions using DNA Subway.
2. Recognize the strengths of various gene prediction programs.
3. Run and gain an understanding of Repeat masker.
4. Retrieve DNA sequences from NCBI.
5. Distinguish between BLASTn and BLASTx
6. Evaluate the strength of evidence for gene predictions

Goal: Mastery of DNA Subway Red Line

Introduction

In your lectures, you learned about whole genome shotgun sequencing and mapping strategies for obtaining genomic sequences, but what happens once those sequences are obtained? They are basically long lists of A's, C's, T's and G's which need to be searched to find genes, repetitive DNA regions and other DNA landmarks. The process of starting from raw DNA sequences and then finding genes and repeat regions is called **genome annotation**. The amount of DNA sequence is huge, so these tasks are done computationally. The basic strategy is to run computer scripts that search for specific DNA sequences within a large region of submitted DNA sequence. A number of these computational tools have been linked together on a gene annotation website called the **DNA Subway**. Today, you will take 102,183 bp of *Arabidopsis thaliana* DNA from Chromosome I on the red line of the DNA Subway, and then you and your lab partner will be given a new piece of DNA to explore using the tools available on the DNA Subway.

Before embarking on the DNA Subway, download a recent version of JAVA onto your laptop. It must be at least Version 6, Update 18 for PCs or Java Version 6 for Macs. In addition, your browser must be Internet Explorer 7 or above, Firefox 3.5 or above or Safari 10 or above.

1. Go to http://dnasubway.iplantcollaborative.org/
2. Click on the blue Register link and fill in the registration information. Be sure to include Student/Educator information. Demographic information is also quite helpful for the developers of DNA Subway, and this information remains confidential. Be sure to use a user name and password that you will recall. Registration MUST be done at least 24 hours in advance!!! Make sure students register prior to the lab section!!!
3. Login to DNA Subway.
4. Click on the red square next to Annotate a Genomic Sequence.
5. Under Project Title, provide a title for this first excursion on the red line.
6. Under Organism, type *Arabidopsis thaliana* for Scientific name and *mouse-ear cress* for Common name. Click circle next to dicotyledons. Don't click Continue yet!!!!

You will next need to copy and paste the 102,183 bp of DNA sequence into the box labeled "Enter a sequence in FASTA format". This sequence was obtained from a Bacterial Artificial Chromosome (BAC) clone named F13B4. This DNA sequence is stored at the National Center for Biotechnology Information (NCBI) and is easily retrievable.

1. Go to http://www.ncbi.nlm.nih.gov/
2. From the pull down menu next to Search, choose Nucleotide
3. Type F13B4 into the open box below Nucleotide and click the Search icon.
4. A small number of results will be displayed, click on the link for "Arabidopsis thaliana chromosome I BAC **F13B4** genomic sequence, complete sequence."

This is the **GenBank accession** for the F13B4 BAC DNA sequence. The **accession number**, AC027134.4, is the unique identifier for this DNA sequence. All entries in GenBank have a unique accession number. Scroll down to find out which chromosome this DNA comes from, organism, the researchers who submitted this DNA sequence to GenBank, and some predicted genes. Then, keep scrolling and scrolling and scrolling just to get an idea of how long 102,183 bp of DNA is. Luckily, only one strand of DNA is shown in GenBank since knowledge of one strand will provide the sequence of the other reverse complement strand.

1. Return to the top of the accession and click on the FASTA link on the left side.
2. Now the sequence appears in FASTA format.
3. Use your cursor to select the **entire** sequence, including the heading line:

```
>gi|8576187|gb|AC027134.4| Arabidopsis thaliana chromosome I BAC F13B4
genomic sequence, complete sequence
```

4. Copy and paste the DNA sequence into the "Enter a sequence in FASTA format" box in the DNA Subway.
5. Click on Continue.

Once the project is created, the stops on the red line of the DNA Subway can be seen. We will take the local line and stop at almost every stop to be sure each step of Gene Annotation in understood. Hold on tight…we are on our way!

The first stop is **Find Repeats**. During this stop the submitted DNA sequence is scanned for repetitive sequences using the **RepeatMasker** computer script. It is essential that the search for genes is done in regions that do not contain repetitive DNA. For a large genome with lots of repetitive DNA, this would slow down the search process, and additionally, these regions are usually silenced. Thus the purpose of this first step is to find the repetitive DNA elements and mask them. RepeatMasker is designed to look for repetitive sequences that are universal, such as dinucleotide repeats $(AG)_n$, and repeat units that are species-specific, and thus it is important that the organism name is included when the project is created.

1. Click on Repeat Masker and let the program run (flashing R) on the F13B4 BAC sequence that was submitted.
2. When complete, the green view icon (V) will appear.
3. Now skip all the way to Local Browser to view the results; click on Local Browser, the black circle stop.
4. Wait a moment and a GBrowse window will appear over the DNA Subway.
5. Modify the Scroll/Zoom window so that it reads: Show 100 kb

GBrowse is one commonly used Genome Browser that allows the user to visualize the DNA sequence on a linear scale with DNA landmarks shown on the tracks below the kb marker. First note the yellow highlighted region which is the first 100 kb of the 102.183 kb region that is being analyzed. Scroll down until the Repeats track is in view. Purple marks can be seen along the 100 kb region. A cursor placed over each purple mark will indicate the type of repeat unit and the exact location. For instance, placing the cursor on the first purple mark shows that this repeat element is located between 6374 and 6458 bp of the submitted sequence. Click on the purple mark and then click on Show Details (wait a minute for the window to open) and the sequence is shown. This region of Low Complexity DNA contains TC repeats as well as strings of T's and C's. Click on the Back arrow at the top of your Browser to return to your GBrowse view. Check out a Simple Sequence Repeat to see what it is. Note that some regions of DNA have more repeats (10-20 kb) while other regions have no identified repeats (60-70 kb). Where would one expect to find more genes? RepeatMasker converts the identified repeat regions into strings of NNNN's that will be ignored during the next stop on the DNA Subway. Close the GBrowse window and return to the DNA Subway.

The next stop is Predict Genes. This stop consists of three different gene prediction computer scripts that scan the submitted DNA looking for open reading frames (contiguous triplet codons that do not contain stop codons), intron exon boundaries (recall that the DNA sequence for an intron begins with GT and ends in AG), transcription start sites (TSS), poly A addition sites (AATAAA) and other DNA landmarks that are common to genes. Two of the programs, FGenesH and SNAP are ***ab initio*** scripts, meaning they scan the DNA without consideration of experimental evidence (cDNAs and ESTs; see below). The third program, Augustus, is **evidence-based**, and tries to incorporate cDNA and EST data. Recall that cDNAs are DNA copies of mRNAs that were isolated from the organism, and thus represent spliced together expressed genes. ESTs (Expressed Sequence Tags) are cDNA sequences that are subject to one quick round of sequencing and usually just span a portion of the actual mRNA. The last program, tRNA Scan, does as its name suggests, it scans for tRNA sequences, which are highly conserved.

1. Return to DNA Subway and run all three gene prediction programs along with tRNA Scan simultaneously.
2. When all four are complete (green View icon), click on Local Browser to see the updated GBrowse window.

A large number of dark to light green boxes are now seen in the window…these are the predicted genes. Augustus and FGenesH are better at finding intron/exon boundaries while SNAP tends to report genes as single exons. Are there any tRNAs in this region?

The 100 kpb view is a bit overwhelming; to reduce the view, place your cursor over the arrow and select between 15 and 30 kb (yellow box); wait a minute and this region will fill the GBrowse window. Look at predicted gene Augustus004, FGenesH004 and SNAPGENE.3. All three programs predict the same gene with three exons and two introns. The gene is on the bottom strand, so the start codon is on the right and the stop codon is on the left. Click on the Augustus004 and a box appears that allows you to Show Details; click here and a view of the gene sequence can be seen. Be sure to scroll down until you can see the color-coded sequence. This provides an exact view of 5'-UTR (brown), exons (green), introns (orange) and 3'-UTR (pink). Remember that the 5' UTR and 3' UTR are part of the first and last exons, respectively. Now close the GBrowse window and return to the third stop on the DNA Subway.

Stop #3 allows the user to Search Databases. As mentioned previously, many cDNAs have been sequenced and are available as GenBank accessions stored at NCBI. In addition, protein sequences are also available as GenBank accessions. Two types of searches are commonly performed for Genome Annotation using the Basic Local Alignment Search Tool or BLAST. BLASTN uses the DNA sequence submitted to DNA Subway as a query against all the non-redundant sequences in GenBank. For the DNA Subway, this search is limited to the organism that is the source of DNA, in this case *Arabidopsis thaliana*. BLASTX takes the DNA submitted to DNA Subway, and translates all six possible reading frames. (Recall that there are three possible reading frames on the top strand and three possible reverse complement reading frames on the bottom strand.) The resulting amino acid sequences are used as a Query against the GenBank protein accessions.

1. Run BLASTN and BLASTX; when View Results icon comes up, click on Local Browser to see results in GBrowse. You do not have data to upload, so don't click on this section of the DNA Subway.

Two new tracks can be seen, the BLASTN and BLASTX results. BLASTN is a full-length cDNA that is nearly identical to AUGUSTUS004. This is not surprising considering that AUGUSTUS used cDNA data as evidence for gene prediction. Since AUGUSTUS uses the cDNA data, it can identify the untranslated regions at the beginning and end of a gene; this is not done by FGenesH or SNAP, the *ab initio* programs that are solely dependent on computational searches. The cDNA that matches AUGUSTUS004 encodes histone H3. Do you recall the function of histone H3? Would you expect this protein to be found in other eukaryotic organisms?

The second track is the BLASTX results. These are GenBank proteins that match the open reading frames that can be translated from the sequence in DNA Subway. Similar proteins are found in *Oryza sativa* (rice), *Nicotiana tobacum* (tobacco), *Lolium multiflorum* (annual ryegrass), *Lilium longiflorum*

(Easter lily), *Gossypium hirsutum* (cotton), and *Chlamydomas reinhardtii* (a green algae). Although histone H3 would be expected to be in all eukaryotes, DNA Subway restricts the search to members of the plant kingdom.

Gene prediction can get complicated. To get an idea of some of the difficulties, look at a different portion of the same 102 kb region.

1. On the ruler at the top, select 65 – 75 kb.
2. Focus on AUGUSTUS021, FGENESH021 and SNAPGENE.17, .18, and .19.
3. Scroll down to look at cDNA evidence from the BLASTN search.

Which program gave results that most closely match the cDNA evidence? (SNAP!). Thus, you can see that different programs succeed in different regions and for this reason, multiple prediction programs are used.

Gene prediction is also quite difficult if there are few cDNAs submitted into GenBank. To demonstrate this, you will now take some muskmelon DNA sequence on the DNA Subway.

Follow the steps above for BAC HM854820. Use the BAC name as your search query in NCBI nucleotide and copy and paste this ~40 kb DNA sequence into a new gene annotation red line project. The Scientific Name is *Cucumis melo*, and the Common name is muskmelon. Since you know what each stop on the DNA Subway does, go to all three stops before looking at GBrowse. Select 8 to 22 kb for a closer view, and look at AUGUSTUS003 and the related genes predicted by FGenesH and SNAP…what a mess! This illustrates why extensive cDNA collections need to be sequenced in order for reliable gene annotation.

A list of Arabidopsis BAC sequences is shown below:

F14L17

F25P22

F1M20

MDF20

MIK19

MJB24

You and your lab partner should now take one of these sequences on the DNA Subway and fill out the following table using MSWord; just cover the first 50 kb of your BAC. The first row of the table is filled out for F13B4 1- 10 kb. There are also a number of questions that need to be answered.

	Low complexity	Simple repeat	other	Augustus	FGenesH	SNAP	tRNA	cDNA BLASTN	Protein BLASTX
1-10 kb F13B4	1	0	0	3	3	2	0	3 All unknown	0 listed
1-10 kb Your BAC									
10-20 kb									
20-30 kb									
30-40 kb									
40-50 kb									

Review Questions:

- Why does RepeatMasker need to be the first step on the DNA Subway?
- What sequence is repeated for RepeatMasker19 on the F13B4 BAC sequence?
- What is the difference between evidence-based and *ab initio* gene prediction programs?
- What is the query and what is searched using BLASTN? What is the query and what is searched in BLASTX?
- A region of the genome is transcribed into a long non-coding RNA. When used as a query, would matches be expected in BLASTN, BLASTX or both? Why?
- Genome sequencing is being performed for many economically important species. Besides deciphering the genome sequence, large collections of cDNAs are sequenced as well. Why is this important?