



SUBWAY

## Fast-track to Gene Annotation and Genome Analysis

### 1.1 Introduction

DNA Subway is a bioinformatics workspace that wraps high-level analysis tools in an intuitive and appealing interface. By “riding” different lines (workflows) you can predict and annotate genes in up to 100,000 base pairs of DNA sequence (Red Line), prospect entire plant genomes for related genes and sequences (Yellow Line), and determine sequence relationships (Blue Line) .

### 1.2 Requirements

To use DNA Subway, you should have:

- Recent version of an Internet browser (Internet Explorer 7 or above; Firefox 3.5 or above; Safari 5 or above)
- Recent version of Java (Version 6, Update 18 for PCs; Java Version 6 for Macs)

### 1.3 User Accounts and Registration

Register or Enter as Guest

- Click “Register” to create a profile to save and share your work. Click “Log In” and enter your password.  
OR
- Click “Enter as Guest,” to try out the system. Guest accounts are temporary.

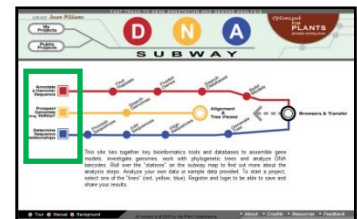
New users

- Click on a square to start an annotating or prospecting project.  
OR
- Click “Projects” to view public projects created by other users; select project link to open it.

### Contents

Section	Page
1.1 Introduction	1
1.2 Requirements	1
1.3 User Accounts / Registration	1
2.1 Use Red Line to Annotate Sequence	2
2.2 Create Project	3
2.3 Generate Evidence	3-4
2.4 Build Gene Models	5-11
2.5 Browse results	11-12
2.6 Prospect Genomes	13
3.1 Prospect Genomes Using TARGeT	14
3.2 Create Project	15
3.3 Search Genomes	15
3.4 View Alignment	16
3.5 View Tree	16
4.1 Determine Sequence Relationships	17
4.2 Create Project	18
4.3 Process Sequences	18
4.4 Add Sequences	19
4.5 Align Sequences	20
4.6 Generate Tree	21
4.7 Export to BOLD	21
4.8 Troubleshooting	22-23

Log in as a guest or registered user



Click any square to begin a workflow along a subway line

## 1.3 User Accounts and Registration (cont'd)

### Returning User

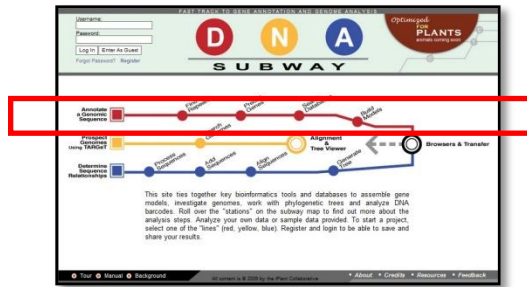
- Click on a square to start a new annotating or prospecting project.  
Or
- Click "Projects" to view your projects or public projects created by other users; select a project link to open a project.



View your saved projects or view projects that are public

Tip: Click "map" from the *Public Projects* or *My Projects* page to return to an overview of *DNA Subway*.

## Red Line: Annotate a Genomic Sequence



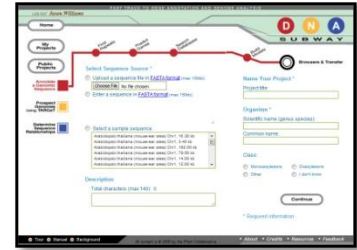
## 2.1 Annotate a Genomic Sequence Using Red Line

Annotation adds features and information to a DNA sequence – such as genes and their locations, structures, and functions. *DNA Subway's* annotation workflow is divided into five sections:

- **Create Project:**  
Upload a DNA sequence or use a sample. **(2.2)**
- **Generate Evidence:**  
Find repeats, predict genes, and search databases to build up information about your sequence. **(2.3)**
- **Build Gene Models:**  
Annotate your sequence by using the evidence generated. **(2.4)**
- **Browse:**  
View your work by using the local browser or the Phytozome genome browser. **(2.5)**
- **Prospect Genomes:**  
Transfer annotated data to the Yellow Line to look for related genes and proteins. **(2.6)**

## 2.2 Create Project

1. Click the red square to create a new annotation project.
2. Upload or enter a DNA sequence. Alternatively, select a sample sequence.
3. Enter required information. (*Scientific and common names are needed to set up the database.*)
4. Click "Continue" to load the project into *DNA Subway*.



Enter project information

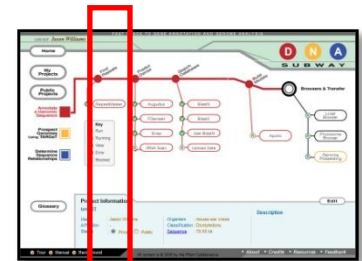
**Tip:** Once you have begun a project, the Project Information Panel at the bottom of the pages allows you to select the "Public" status radio button to share your work. Click "Edit" to change project title or description; then click "Save." (Note: all projects are private by default.)

## 2.3 Generate Evidence

At several stops on the Red Line you will generate mathematical evidence (using the repeat and gene prediction algorithms) and biological evidence (derived from BLAST homology searches).

### 2.3.1 Find Repeats

1. The "R" bullet to the left of RepeatMasker indicates that the program is ready to run. Programs labeled with an "X" bullet are disabled, and become available after prerequisite steps are completed.
2. Click "RepeatMasker" to identify and mask repetitive sequences.
3. The blinking "R" bullet indicates the program is running.
4. When the bullet changes to "V," click "RepeatMasker" again to view the results in a new window. (*The window is empty if no repeats were identified.*)
5. Close the result window to return to the Red Line.
6. After running RepeatMasker, all other analysis routines on the Red Line are ready.



RepeatMasker

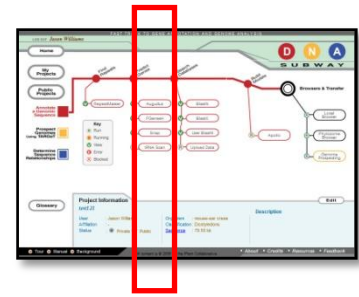
**Tip:** If the pop-up window does not close, use your browser "Back" button to return to *DNA Subway*. If this fails, close and reopen your browser. Then log into your account again, open your project using the "My Projects" button, and this should return you to where you left off. Guest users will need to start over if they close their browsers.



Use the key to determine the status of a *Subway* tool

### 2.3.2 Predict Genes

1. Click "Augustus," FGenesH," "Snap," or "tRNA Scan," to predict genes in your sequence. (You can run several programs simultaneously, but doing this may slow down the generation of your results.)
2. The blinking bullet changes to "V" when the results are ready to view. Click on the button of the program you used again to view the results in a new window. (The window is empty if no genes were identified.)
3. Augustus, FGenesH, and Snap files list sequence coordinates for each predicted gene. Features of each gene are identified as belonging to the same "parent:" gene, mRNA, exon, and CDS (coding sequence). Exons and CDS coordinates are mainly contiguous.



Predict genes using one or more algorithms

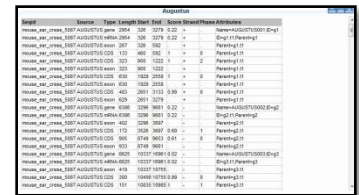
**Tip:** Only the Augustus gene prediction tool can predict 5' and 3' untranslated (noncoding) regions called UTRs, where first and last exons may extend beyond the CDS.

**Tip:** FGenesH can be used freely for samples included in DNA Subway. FGenesH can only be used on 20 uploaded samples per day by Registered users; Guest users cannot use FGenesH on uploaded samples. All other programs in DNA Subway are freely available to all users and all types of samples.

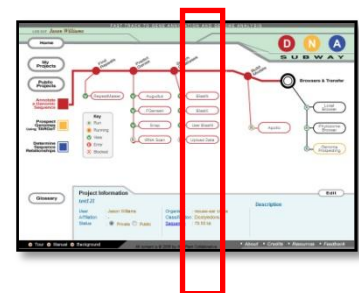
### 2.3.3 Search Databases

DNA Subway searches customized versions of UniGene and UniProt that contain only validated plant proteins, and are free of predicted or hypothetical proteins.

1. Click "BLASTN" to search the UniGene database for genes that are similar to your DNA sequence.
2. Click "BLASTX" to search the UniProt database for proteins that are similar to your DNA sequence. The BLASTX algorithm translates your sequence in all six reading frames into amino acid sequences and uses these to search UniProt.
3. Click "Upload Data" to search your own DNA or protein sequence data for matches.



Pop-up window with Augustus gene predictions



Search customized UniGene and UniProt databases

## 2.4 Build Gene Models

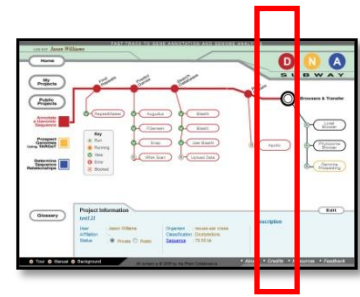
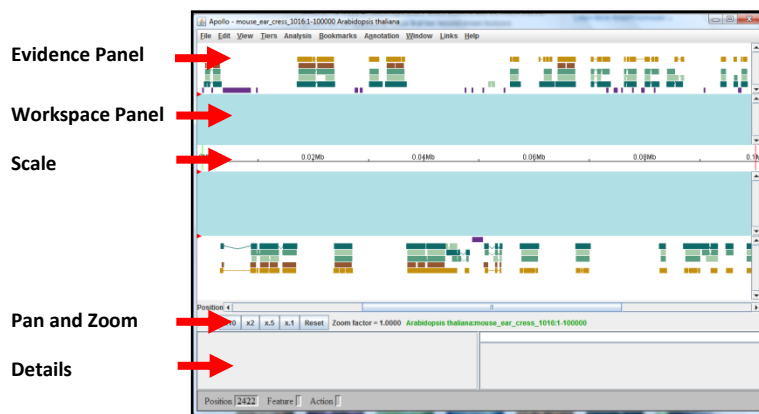
At this stop on the Red Line you will use the Apollo annotation editor to build gene models based on the evidence collected in section 2.3. (If you do not want to build gene models or want to quickly view your evidence, skip to section 2.5.)

### 2.4.1 Work with Apollo

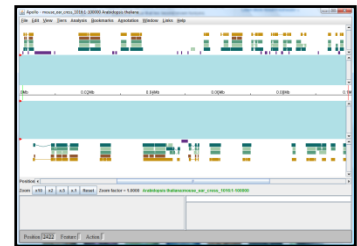
1. Click "Apollo" to launch Apollo using the Java webstart utility. You may need to respond to a dialog box asking for permission to download and/or run the application

The Apollo window is divided into a number of functional areas:

- A horizontal scale at the center shows the coordinates of your sequence.
- Panels above the scale relate to the DNA strand you are analyzing as the forward DNA strand, and panels below relate to the reverse strand.
- Adjacent blue panels are workspaces for building gene models on forward (top) and reverse (bottom) strands.
- Adjacent white panels display evidence on forward (top) and reverse (bottom) strands.
- Windows at the bottom of the page display detailed information about a selected feature.



**Tip:** Launching Apollo may take several minutes, as the application first downloads to your computer then loads your evidence into a new Apollo window. If Apollo does not open, close your browser and make sure that you are running the latest version of Java. Then reopen your browser and return to *DNA Subway*. Log into your account again, open your project using the "My Projects" button, and click "Apollo" again. Guest users will need to start over after closing their browsers.



**Apollo annotation editor  
(opens as a pop-up window)**

**Tip:** In the Apollo editor, the top half of the screen is the 5' DNA strand, the bottom half is the 3' strand.

### 2.4.1 Work with Apollo (cont'd)

#### 2. Adjust your workspace:

- Maximize the Apollo window, or stretch it by grabbing and pulling a corner.
- Focus on forward or reverse strand. In the menu bar at the top, *click* “View” tab and uncheck the strand (forward or reverse) that DOES NOT contain the gene you are working on. Note that the reverse strand reads from right to left.
- Grab and pull down the divider beneath the “Zoom” buttons to increase the evidence (white) and workspace (blue shaded) panels.
- Grab the red arrow, and pull up/down to increase the white evidence panel.

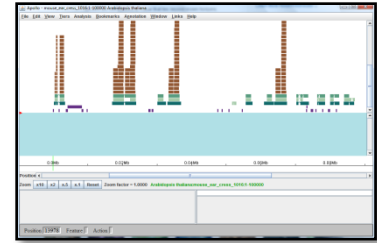
3. *Click* the “x2” button at the bottom to zoom in on your sequence. The structure of genes will become apparent – with alternating exons (boxes) and introns (lines). You can continue zooming until the center scale displays the nucleotide sequence. The green and red bars at the top and bottom of the screen are potential start and stop codons in each of the three reading frames on the forward (top) and reverse (bottom) strands.

4. *Click* “Reset” to zoom back to the original view.

5. In the menu bar at the top, *click* “Tiers,” and select “Show types panel” for a color-coded key to the evidence displayed in Apollo. Each “tier” contains a different type of evidence gathered in Part 2.3 – including:

- repeats (purple)
- gene predictions (green)
- BLASTX (brown)
- BLASTN (gold)

6. Return to the menu bar; *click* “Tiers,” and select “Expand all tiers.” This displays each piece of evidence on a separate line, allowing you to see all of the results of your analyses. The results of each prediction program display separately (Augustus, FGeneseH, Snap, and tRNA) in different shades of green. Notice that BLASTX and BLASTN can return many similar alignments for one gene or another.



**View of Apollo screen: forward stand, expanded tiers**

**Tip:** The default view collapses all tiers so that each type of evidence is displayed on a single line. This overview is useful but can be misleading.



**The Apollo Types menu**

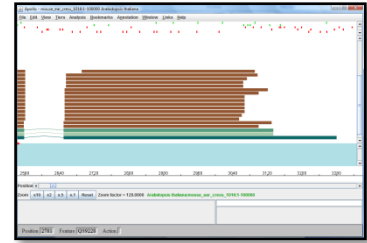
### 2.4.1 Work with Apollo (cont'd)

7. Click on any piece of evidence to display additional information in the detail windows at the bottom:

- The left window displays information about highlighted matches as a whole, such as: evidence type, name (source), and range (nucleotide coordinates). Importantly, for BLAST results the name field shows the organism in which the match was found. Click an evidence item to highlight it in the white evidence field above.
- The right window displays information about match details such as query frame (reading frame), genomic range (nucleotide coordinates), and genomic length for each exon of a multi-exon gene or prediction. Click an evidence item to highlight it in the white evidence field above.

8. Zoom in to focus on one set of gene models and BLAST evidence. Consider several reasons why various pieces of evidence may not align perfectly:

- The first and last exons are hardest to predict, because they have only one intron boundary.
- Gene predictors may miss exons of an alternatively spliced gene.
- BLAST evidence may come from different species, whose exons differ in length.
- The BLAST algorithm does not determine splice sites.
- BLAST results may include false matches.

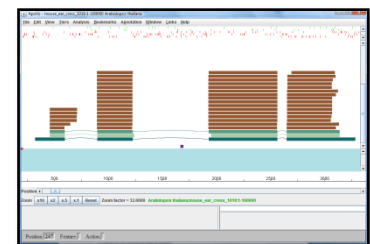


**Blast evidence may be incongruous**

### 2.4.2 Annotate a Gene

Using Apollo, you will create a basic gene model which generally entails using biological evidence from BLAST searches to confirm or adjust an algorithmic gene prediction.

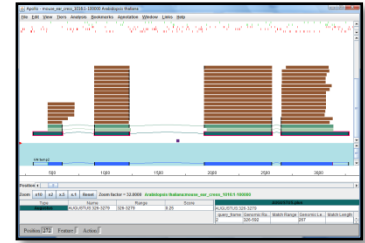
1. Scan your DNA sequence for a gene that has one or more predictions (FGenesH, Augustus, Snap), as well as BLASTN and/or BLASTX evidence. It may be on the forward or reverse strand.
2. Use the “Zoom” button to enlarge the gene and to clearly show its exon/intron structure.



**Zoom function allows whole “gene” overview**

### 2.4.2 Annotate a Gene (cont'd)

3. Compare predictions from FGenesH, Snap, and/or Augustus with evidence from BLASTN and BLASTX searches.
  - The gene prediction algorithms are quite good at identifying locations of genes and correct splice sites (exon/intron boundaries). However, they have difficulty identifying first and last exons, because these have only one splice site.
  - BLAST results are the most authoritative evidence for exons, because they are retrieved from databases for biological evidence from mRNA or protein sequence from a given organism. However, BLAST results often do not accurately reflect splice sites. Also, the exon structure may be different for a related gene from a different organism. Finally BLAST may return incomplete results as not all genes/proteins are represented in the database.
4. *Click* on any exon in the evidence panel to highlight “edge matches.” Black bars indicate exons in that share the same exon-intron boundaries. (However, these are not necessarily the correct splice sites.)
5. Select one gene prediction whose exon structure best overlaps with the BLAST evidence stacked above (or below) it. You will use this gene prediction as the basis for your model, because it is more likely to have correct splice sites.
6. To begin a model, *double-click* on the selected gene prediction to highlight all of its exons. Then, *click* and drag it onto the blue workspace. Note that your model may have a start codon (green bar) and stop codon (red bar). If either of these is missing, you may need to annotate them.
7. *Click* on any BLAST exons that do not stack on top of exons in the prediction, and move them into the blue workspace.
8. Next merge a BLAST exon with the gene prediction in the blue workspace:
  - *Double-click* to mark all the exons in the gene prediction. Then, *shift-click* on the BLAST exon.
  - *Right-click* (PC) or *command-click* (Mac) and *select* “Merge transcripts.” This creates a new gene model.

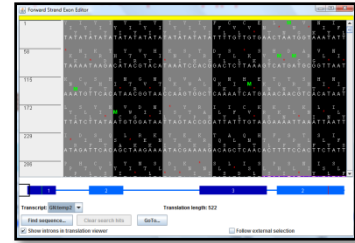


Gene model (blue) in workspace panel

**Tip:** If the gene prediction matches the biological evidence (BLAST search results) you could stop here and decide to adopt the gene prediction as a valid model for this gene. However, you should record and justify this decision in the “Gene Info Editor” (see 2.4.2.18). You may also wish to submit this new model to a BLAST search from within Apollo (see 2.4.2.17) and see whether you can corroborate your decision.

## 2.4.2 Annotate a Gene (cont'd)

9. The new model may contain several errors that need to be fixed:
  - A green or red arrow indicates the gene model is missing a start or stop codon.
  - A yellow arrow indicates an inaccurate splice site.
10. *Right-click* (PC) or *command-click* (Mac) and select “Exon detail editor” to see the nucleotide sequence of your model.
  - A thick blue bar over the sequence indicates an exon, and a thin blue line indicates an intron. (Multiple gene models will display as multiple lines of sequence.)
  - Each of the three reading frames is shown in a different shade of blue.
  - The diagram at the bottom of the window shows you where you are in the selected gene model as you scroll through the sequence with the slider bar to the right. A sliding bracket shows where you are in the main Apollo window.
  - *Note that both forward and reverse strand genes “read” left-to-right (5’ to 3’) in the Exon detail editor. This means that a reverse strand gene is displayed “backwards” from Apollo view.*
11. To correct a yellow arrow *at the beginning* of an intron, drag the exon boundary next to the closest splice signal (GT) *in the exon or intron*. The arrow will disappear in the gene model. Then, *right-click* (PC) or *command-click* (Mac) the gene model and select “Calculate longest ORF.”
12. To correct a yellow arrow *at the end* of an intron, drag the exon boundary next to the closest splice signal (AG) *in the intron or exon*. The arrow will disappear in the gene model. Then, *right-click* (PC) or *command-click* (Mac) the gene model and select “Calculate longest ORF.”



Exon detail editor

## 2.4.2 Annotate a Gene (cont'd)

13. To correct a green arrow, you have two options.

- To use a start codon (ATG) that lays *within* a current exon, select it from the rows of start codons above (or below) the evidence space (see 2.4.1.3.) and drag it onto your gene model; the arrow in the gene model will be replaced with a green line. Then, *right-click* (PC) or *command-click* (Mac) the gene model and select “Calculate longest ORF.”
- To use a start codon that precedes the first exon use the “Exon Detail Editor” to drag the sequence exon bar to that codon, and the arrow in the gene model will be replaced with a green line. Then, *right-click* (PC) or *command-click* (Mac) the gene model and select “Calculate longest ORF.”

14. To correct a red arrow, you have two options.

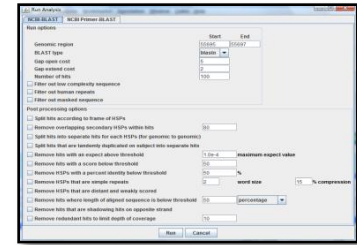
- To use a stop codon that lays within a current exon, select it from the rows of start codons above (or below) the evidence space (see 2.4.1.3.) and drag it onto your gene model; the arrow in the gene model will be replaced with a red line. Then, *right-click* (PC) or *command-click* (Mac) the gene model and select “Calculate longest ORF.”
- To use a stop codon that lies outside of an exon use the “Exon Detail Editor” to drag the sequence exon bar to that codon, and the arrow in the gene model will be replaced with a red line. Then, *right-click* (PC) or *command-click* (Mac) the gene model and select “Calculate longest ORF.”

15. Make multiple gene models if you have evidence of alternative splicing – different configurations of the same exons denoted by the BLAST results.

16. You can undo work in several ways:

- Click on “Edit” in the menu bar at top, and select undo to sequentially undo each action.
- *Right-click* (PC) or *command-click* (Mac) and select “Delete selection” to undo any highlighted exon or gene model in the blue workspace.

17. To quickly get more information about a model or exon, *right-click* (PC) or *command-click* (Mac) and click “Analyze region” to submit your highlighted selection to a BLAST search.



Apollo analysis window

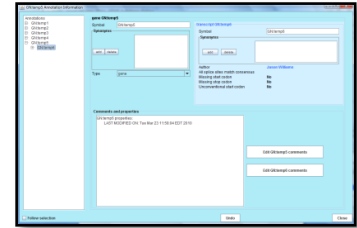
## 2.4.2 Annotate a Gene (cont'd)

18. Name your model, and add comments and annotations.

- Highlight a gene model, *right-click* (PC) or *command-click* (Mac), and select “Annotation Info editor.”
- Enter the same name in the gene “Symbol” field (left) and transcript “Symbol” field (right). This will replace a name that begins with “GN:temp” or “USER:0000.”
- Click on either of the “Edit comments” buttons at bottom left. Then click on the “Add” button, and write your comments in the pop-up window. These can include information about protein function obtained from a BLAST search or unresolved annotation problems. Click close to return to the information window.
- You can access all of your gene models in the list on the left side, or you can edit existing annotations.

19. In the menu bar, click “File,” then select “Upload to DNA Subway.” A dialog window will notify you when the upload is complete.

20. Close or set aside Apollo to return to *DNA Subway*.



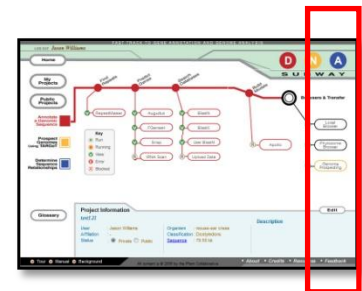
Apollo annotation window

## 2.5 Browse Results

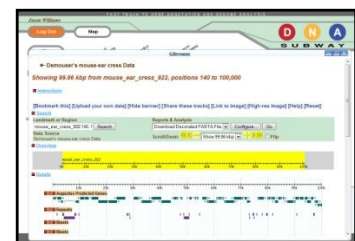
At this stop on the Red Line you can use a graphical interface to scan to the repeats, gene evidence, and gene models you developed at other stops.

### 2.5.1 Local Browser

1. Click “Local Browser” to load your work and open the browser window. This may take up to a minute or so, depending on the amount of evidence you have generated.
2. Click and drag the yellow box on the “Overview” scale to navigate to different areas of your sequence. To view your entire sequence, move the cursor to the zero-point, then click and highlight the entire scale.
3. Click and drag the blue box on the “Details” scale, and select to zoom in on any gene currently displayed.
4. Each type of evidence appears as a separate track, and models you have made appear in the track “User created models.”
5. Click the “-/+” toggle in the track name box to hide or show a track.
6. Click the track name box and drag it up or down to change the track order.



Use browsers to examine results



Local browser pop-up window

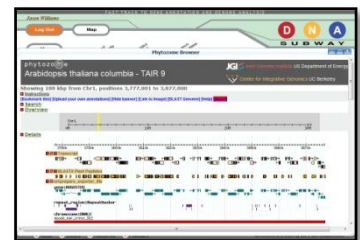
### 2.5.1 Local Browser (cont'd)

7. Use the “Tracks” and “Display Settings” at the bottom of the window, and click “Update Image,” to further customize your browser window.
8. Close or reduce the browser window to return to *DNA Subway*.

### 2.5.2 Phytozome Browser

If your project used a sample sequence, at this stop you can view your evidence and gene models in the context of the assembled plant genomes at the Department of Energy’s *Phytozome* site.

1. Click “Phytozome Browser” or export and upload your evidence and gene models to Phytozome. *(If you don’t see a ‘Phytozome Browser’ button, you are working with a sequence for which DNA Subway does not provide this export function.)*
2. As another instance of the GBrowse (Generic Genome Browser), *Phytozome* has the same “look and feel” as the *DNA Subway* Local Browser.
3. The “Transcript” track displays Phytozome annotated genes, while the “url” track displays your evidence, gene models, and shows your results; including any gene models you may have generated.
4. Use the “Tracks” and “Display Settings” at the bottom of the window, and click “Update Image,” to further customize your browser window. *(The more tracks you activate the longer it takes to re-load the data.)*
5. Close or reduce the browser window to return to *DNA Subway*.

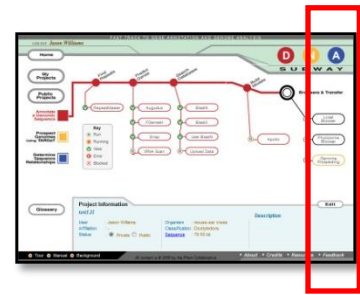


Phytozome browser window

## 2.6 Prospect Genomes

The “Genome Prospecting” button sends evidence or a gene model to the Yellow Line, where it is used to search for similar sequences within plant genomes.

1. Click “Genome Prospecting,” then “Continue” in the pop-up window to open the local browser with your evidence and gene models.
2. Click on any exon, BLAST result, or repeat feature, then click “Send to TARGeT” in the pop-up window. (Alternatively, highlight a region of up to 10,000 base pairs, click it, and then click “Send to TARGeT” in the pop-up window.)



Transfer to Yellow Line



Initiate transfer from local browser

Continue on to Yellow Line to Prospect Genomes

FAST TRACK TO GENE ANNOTATION AND GENOME ANALYSIS

LOG OUT Jason Williams

Home My Projects Public Projects

Search Genomes

Annotation a Genomic Sequence Prospect Genomes Using TARGeT? Determine Sequence Relationships Glossary

Mosses: Physcomitrella patens, Lycophytes: Selaginella moellendorffii, Algae: Chlamydomonas reinhardtii

Dicotyledons: Arabidopsis lyrata, Arabidopsis thaliana, Carica papaya, Cucumis sativus, Glycine max, Manihot esculenta, Medicago truncatula, Populus trichocarpa, Ricinus communis, Vitis vinifera

Monocotyledons: Brachypodium distachyon, Oryza sativa japonica, Sorghum bicolor, Zea mays

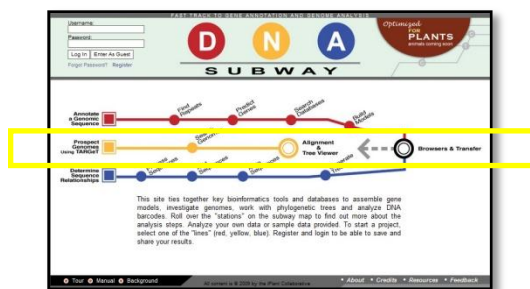
Key: Run, Running, View, Error, Blocked

Alignment & Tree Viewer: Alignment Viewer, Tree Viewer

Project Information: 1112w1, User: Jason Williams, Affiliation: -, Status: Private, Organism: Oryza sativa, Protein name: Ping Transposase, Class: DNA Transposase, Function: Transposition

Tour Manual Background About Credits Resources Feedback

## Yellow Line: Prospect Genomes



ID	Project Title	Description	Organism Type	Status
10100011	10001	10001	10001	10001
10100011	10002	10002	10002	10002
10100011	10003	10003	10003	10003
10100011	10004	10004	10004	10004
10100011	10005	10005	10005	10005
10100011	10006	10006	10006	10006
10100011	10007	10007	10007	10007
10100011	10008	10008	10008	10008
10100011	10009	10009	10009	10009
10100011	10010	10010	10010	10010

View your saved projects or view projects that are public

**Tip:** Click “map” from the *Public Projects* or *My Projects* page to return to an overview of *DNA Subway*.

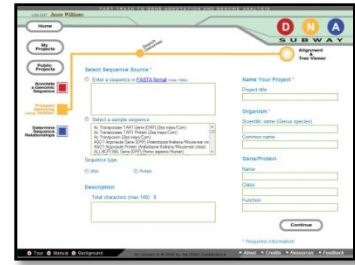
### 3.1 Prospect Genomes Using TARGeT through Yellow Line

Genome prospecting uses a query sequence (DNA or protein of up to 10,000 base pairs/amino acids) to find related sequences in specific genomes or in a database. A major purpose of genome prospecting is to identify members of gene or transposon families. *DNA Subway* uses the TARGeT workflow, which integrates BLAST searches, multiple sequence alignments, and tree-drawing utilities. *DNA Subway's* genome prospecting workflow is divided into four sections:

- **Create Project:**  
Select or enter a DNA or protein query sequence. **(3.2)**
- **Prospect Genomes:**  
Search one or more plant genomes for matching sequences. **(3.3)**
- **View Alignment:**  
Align matching sequences and compare to a consensus sequence. **(3.4)**
- **View Tree:**  
Display sequence relationships among matching sequences that may indicate phylogeny. **(3.5)**

### 3.2 Create Project

1. Enter *DNA Subway*.
2. Click the yellow square to create a new annotation project.
3. Upload or enter a DNA or protein sequence. Alternatively, select a sample sequence.
4. Enter required information. (*Scientific and common names are needed to setup the database.*)
5. Click "Continue" to load the project into *DNA Subway*.



Enter project information

**Tip:** The Gene/Protein section is not required, but allows you to enter detailed information about the gene or protein you are working on.

### 3.3 Search Genomes

1. Select a genome to query. You can search several genomes simultaneously, but this may slow down your results. Consider limiting yourself to querying one genome at a time.
2. Click "Run."
3. The blinking "R" bullet indicates the program is *running*.
4. The bullet changes to "R" again when your results are ready to view. This also indicates that you can change your genome selection and re-run the search.
5. Select "Alignment viewer" or "Tree viewer" to see the results of your search.



Select species to search



**Tip:** In the Project Information Panel at the bottom of the page you can select the "Public" status radio button to share your work. Click "Edit" to change project title or description; then click "Save." (Note: all projects are private by default.)

### 3.4 View Alignment

1. Click “Alignment Viewer” to view the results of the search results aligned against each other. This will open the multiple alignment editor and analysis tool *Jalview* in a new window.
2. The menu bar at the top of the pop-up window provides access to editing, analysis, and tree-drawing tools – click the “Help” tab to access a detailed “Documentation.”
3. Close or set aside the alignment window to return to the Yellow Line.



Use the alignment viewer to see sequence comparison

**Tip: Interpreting the alignment results:**

Amino acid sequences are displayed in One-letter code. Short dashes signify gaps inserted by the program to optimally the sequences.

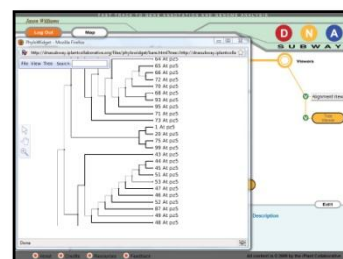
“**Conservation**” bar at the bottom of the window provides for each amino acid position in the alignment the number of amino acids with conserved chemical properties.

“**Quality**” bar displays the likelihood of observing mutations (if any) in each column of the alignment. A high quality score suggests that there are no mutations in the column, or that most mutations observed are neutral.

“**Consensus**” bar lists for each position in the alignment the amino acid that is most prevalent in this column. If more than one residue occurs equally often in a column, a “+” symbol is used in the display because it is not possible to display multiple characters in the single character space available.

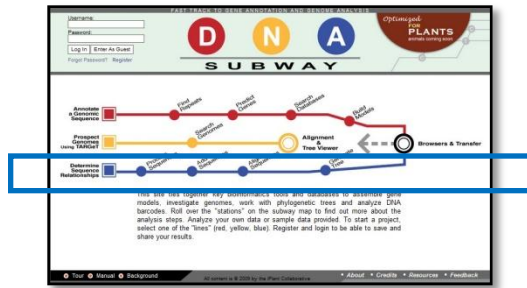
### 3.5 View Tree

1. Click “Tree Viewer” to open *PhyloWidget* in a new window. This application displays your alignments in a tree format, which shows one solution of how the matching sequences are may be related to one another. Like a pedigree tree, the most closely related sequences are clustered together on nearby branches, while less related sequences are located on more distant branches.
2. In the menu bar to the left:
  - Click the “zoom” icon, then move your cursor over the area you are interested in. *Left-click* and move your cursor up to zoom in, or down to zoom out.
  - Click the “hand icon, then grab the tree to move it.
  - Click “arrow” icon, then *click* on a match name, or tree node, or branching point (dot) to edit. Click “Node Edit” and select “Name” to enter a new name for that node. Explore other menu items to customize your tree presentation.
3. In the menu bar at the top:
  - Click “View,” then *select* a “Layout” to change how the tree is displayed. Explore other menu items to customize your tree presentation.
  - Click “File,” and *select* “Export Image” to save your tree as a pdf or other file format.
4. Close or set aside the tree window to return to the Yellow Line.



Use the *PhyloWidget* to view a tree of sequence similarity

## Blue Line: Determine Sequence Relationships



View your saved projects or view projects that are public

**Tip:** Click "map" from the *Public Projects* or *My Projects* page to return to an overview of DNA Subway.

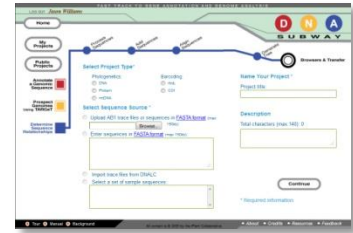
### 4.1 Determine Sequence Relationships Using Blue Line

You can analyze relationships between DNA sequences by comparing them to a set of sequences you have compiled yourself, or by comparing your sequences to other that have been published in databases such as *GenBank* (National Center for Biotechnology Information). Generating a phylogenetic tree from DNA sequences derived from related species can also allow you to draw inferences about how these species may be related. By sequencing variable sections of DNA (barcode regions) you can also use the Blue Line to help you identify an unknown species, or publish a DNA barcode for a species you have identified, but which is not represented in published databases like BOLD (Barcode of Life Online Database – [www.boldsystems.org](http://www.boldsystems.org)). DNA Subway's sequence relationship workflow is divided into 6 sections:

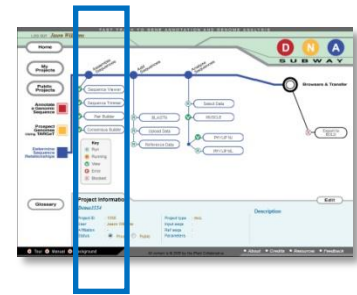
- **Create Project:**  
Upload sequence trace file(s), DNA sequence(s), or use a sample. **(4.2)**
- **Process Sequences:**  
View your trace files, trim N's from sequence, build pairs of sequences, build a consensus sequence from pairs. **(4.3)**
- **Add sequences:**  
Use BLAST to import additional sequences, or upload / import additional sequences. **(4.4)**
- **Align Sequences:**  
Select the sequences you wish to align, and generate a multiple alignment using MUSCLE. **(4.5)**
- **Generate Tree:**  
Generate a phylogenetic tree using PHYLIP. **(4.6)**
- **Transfer:**  
Publish sequences to BOLD. **(4.6)**

## 4.2 Create Project

1. Click the blue square to create a new project.
2. Select the type of sequence data you will be working with (DNA, Protein, or mtDNA). If you are doing a DNA barcoding project, there are two default settings (*rbcl* and *COI*) which will load default sample sequences.
3. Upload DNA sequence trace file(s) in AB1 format. If you have sequenced DNA with Genewiz, through the DNA Learning Center, click 'Import trace files from DNALC.' Alternatively, past in FASTA formatted sequence or select a sample sequence.
4. Name your project (required) and enter a description if desired.
5. Click "Continue" to load the project into *DNA Subway*.



Enter project information

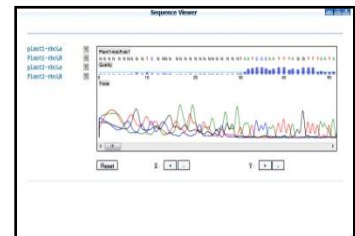


Sequence Processing Tools

## 4.3 Process Sequences

### 4.3.1 View Sequences

1. Click "Sequence Viewer" to display the sequences you have input in the project creation section. If you uploaded trace files, click on the file names to view the trace files. If you did not upload trace files, you can scroll to see the whole sequence uploaded.
2. Use the "X" and "Y" buttons to adjust the level of zoom. You can undo zooming by pressing the "Reset" button.



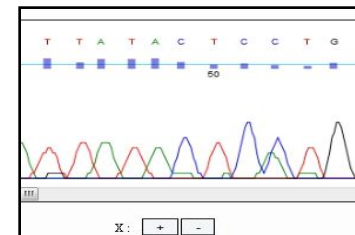
Sequence Viewer Window

### 4.3.2 Trim Sequences

1. Click "Sequence Trimmer" to automatically remove "N's" on the 5' and 3' end of your sequence.
2. If you wish to view trimmed sequences again, click on the file name.

### 4.3.3 Build Pairs

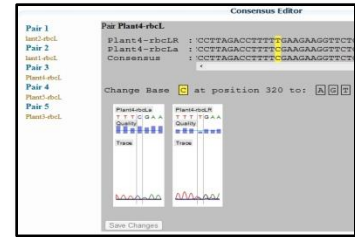
1. If you have multiple reads of a sequence, pair the sequences by checking the box to the right of a sequence. After checking a single sequence, a dialogue box appears when you check the next sequence, confirming that you wish to designate the sequences as a pair.
2. Be default, DNA Subway assumes that all reads are in the forward orientation, and displays an 'F' to the right of the sequence. If any sequence is not in that orientation, click the "F" to reverse complement the sequence. The sequence will display an "R" to indicate the change.



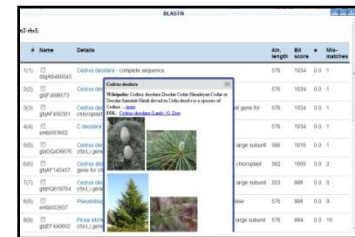
Quality scores (blue bars) in relation to the significance threshold (blue line).

#### 4.3.4 Consensus Editor

1. Once you have created sequence pairs, *click* “Consensus Editor” to make a consensus sequence from both sequences in the selected pairs.
2. To examine the consensus sequence *click* “Consensus Editor” again, and then *click* on the pair name you wish to examine.
3. If there are any differences in nucleotides in the first and second sequence, these will be highlighted yellow in the consensus editor window. *Click* on any highlighted nucleotide in the consensus sequence if you wish to change that nucleotide. If trace files were used as sequence sources, then the trace information for those loci in both sequences will be displayed allowing edits to be made based on quality score information. Remember to *click* “Save Changes” if edits were made.



**Consensus Editor highlights discrepancies between paired sequences**



**BLASTN displays top sequence matches from NCBI as well as Wikipedia information.**

**Tip:** The BLAST search can only return results of species that are actually in the database. Even though you may find sequence match that has few if any mismatches, this match is not guaranteed to be a definitive identification of your species. Sequence quality, completeness of the database, and other factors will determine the level of confidence that can be placed in identifying a species based on DNA sequence.

#### 4.4 Add Sequences

##### 4.4.1 BLASTN (BLASTX if using protein data)

1. *Click* “BLASTN” to display your sequences (or consensus sequences) and search the NCBI Database for matches.
2. To perform a BLASTN on individual sequences, *click* the “BLAST” link displayed next the individual sequence name.
3. The BLASTN search should return the top search results for your query sequence. The species names for your BLAST hits should be hyperlinked to available images of that species from Wikipedia, you can also *click* the hyperlinks to get to the Wikipedia article for that species.
4. If you wish to add any of the sequences obtained from the BLAST search, select the sequences by checking the grey box to the left of the species name. Once you have made your selections, save them by scrolling down and *clicking* “Add BLAST hits to project.”

##### 4.4.2 Upload Data

1. *Click* “Upload Data” to add additional sequence data to your analysis, without starting a new project. Use “Upload Sequence(s)” to upload AB1 trace files or FASTA formatted sequences locally stored on your computer; Use “Enter Sequences(s)” to paste or type sequences in FASTA format.



**Add additional sequence data from sources including BOLD and GenBank.**

#### 4.4.2 Upload Data (cont'd)

2. If you would like to import sequences from non-local sources you can use "Import Sequence" to search a sequenced database using a sequence identifier. For GenBank sequences you can search by identification numbers (GI or Version). Search BOLD by species name, or search the DNALC sequence database by tracking number for sequences you processed with GeneWiz through the DNALC system.

#### 4.4.3 Reference Data

1. Click "Reference Data" (optional) to include additional sequences. Depending on the project type you have created, you will have access to additional sequence data that may be of interest. For example, if you are doing a DNA barcoding project using the *rbcl* gene, samples of *rbcl* sequence from major plant groups (Angiosperms, Gymnosperms, etc.) will be provided. Choose any data set to add it to your analysis; you will be able to include or exclude individual sequences within the set in the next step.

### 4.5 Align Sequences

#### 4.5.1 Select Data

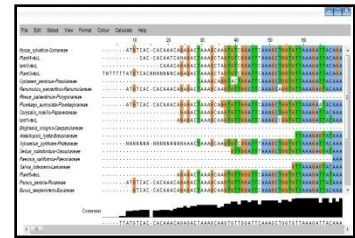
1. Click "Select Data" to display all the possible sequences you have brought into your analysis so far, including "user data" (the data you originally started your project with) and any additional or reference data. Check off individual sequences you wish to include in your multiple alignment. You may also use the "Select all" feature to include all sequences. You may come back to this step as often as you wish to select the sequences you want to include in a multiple alignment or tree.
2. Click "Save" in the blue dialog box that appears when you make any selections.

#### 4.5.1 Multiple Alignment with MUSCLE

1. Click "MUSCLE" to run the MUSCLE multiple alignment software. This software will align all sequences that were included in the "Select Data" step.
2. Click "MUSCLE" again to open the created multiple alignment. An alignment that is suitable for creating a phylogenetic tree will have an overall high consensus score (represented by the height of the black bars on the lower portion of the alignment window).



Select the sequences you wish to include in your multiple alignment.



Untrimmed multiple alignment created by MUSCLE

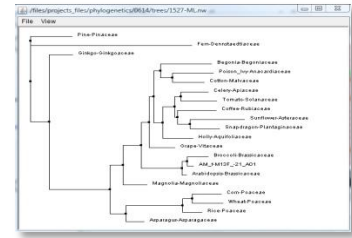
**Tip:** If the alignment is poor (lots of gaps, some sequences do not share identity with other sequences) you may wish to consider going back to the "Select Data" step and removing those sequences.



Trimmed multiple alignment created by MUSCLE (with trimming dialog box visible)

#### 4.5.1 Multiple Alignment with MUSCLE (cont'd)

3. You will need to “trim” the alignment so that all sequences are of the same length before building a phylogenetic tree. To trim, *Click* the position number (or the whitespace above the topmost nucleotide at the position you wish to trim). A dialogue will appear allowing you to trim all sequences to the right or left of that position. Do this for the 5’ and 3’ area of your sequences, then *Click* “Submit trimmed alignment.”



**Maximum Likelihood Tree**

**Tip:** A phylogenetic tree is a graphical representation of relationships between taxonomic groups. In this experiment, a *gene tree* is determined by analyzing the similarities and differences in DNA sequence.

## 4.6 Generate Tree

---

### 4.6.1 Neighbor Joining Tree

1. Click on “PHYLIP NJ” to generate a phylogenetic tree using the Neighbor Joining method. A tree will open in a new window; and the MUSCLE alignment used to produce it will open in another window.

### 4.6.2 Maximum Likelihood Tree

1. Click on “PHYLIP ML” to generate a phylogenetic tree using the maximum likelihood method. A tree will open in a new window; and the MUSCLE alignment used to produce it will open in another window.

## 4.7 Exporting Sequences to BOLD

---

This feature is not operational in the current version of subway. However, you can upload sequences to the Barcode of Life Online Database at:

[www.boldsystems.org/](http://www.boldsystems.org/)

## 4.1 Troubleshooting

---

### 1. Fields or borders are misaligned

- Make sure you are using a recent version of an internet browser: Internet Explorer 7 or above, Firefox 3.5 or above, Safari 10 or above.

### 2. “Invalid file format” upon sequence submission

- Make sure you submit only one sequence and that it is in FASTA format. (To learn more about FASTA, *click* the link above the “Upload” or “Enter” form.)

### 3. Pop-up windows don’t open

- Check to see if the pop-up window is open behind your current window.
- Check your browser settings and insure that pop-up windows are allowed.

### 4. Pop-up windows don’t close

- Use your browser “Back” button to return to *DNA Subway*.
- Restart your browser and return to *DNA Subway*. Log into your account again, open your project using the “My Projects” button to return to your project. (Guest users will need to start over after closing their browsers.)

### 5. Gene prediction program FGenesH is not available after running RepeatMasker

- Registered user: Per license agreement with FGenesH owner Softberry ,Inc., registered users are permitted to use FGenesH on uploaded sequences up to 20 times daily. If you exceed the 20 time limit, you will be unable to run this algorithm until the next day.
- Guest user: Per license agreement with FGenesH owner Softberry, Inc., guest users are not permitted to use FGenesH on uploaded sequences.
- FGenesH usage is unlimited on the sample sequences already included in *DNA Subway*.

### 6. A gene prediction program or a BLAST search doesn’t generate any evidence

- No genes or related sequences were found in the sequence that was searched.
- Run a different program or search.
- Use a longer sequence of DNA.

### 7. Gene prediction program or BLAST searches returns an error “E”

- Re-run the program or search.
- Restart your browser and return to *DNA Subway*. Log into your account again, open your project using the “My Projects” button, and *click* “Apollo”. Guest users will need to start over after closing their browsers.)
- If the error persists use the “Feedback” button at the bottom of *DNA Subway* to inform the site administrator about this error. Be specific in your description of the error and the operation that produced the error. Include your project ID, your location, the date and time the error occurred.

## 8. Apollo doesn't start up

- Check to see if Apollo is open behind your current window.
- Re-start Apollo and wait several minutes. (Other operations that may run on your computer, such as virus scans, may slow Apollo considerably.)
- Restart your browser and return to *DNA Subway*. Log into your account again, open your project using the "My Projects" button, and *click* "Apollo" (Guest users will need to start over after closing their browsers.)
- Close *DNA Subway*, go to java.com and verify that your computer is running the latest version of Java. Then restart your project as above.
  
- If the error persists use the "Feedback" button at the bottom of *DNA Subway* to inform the site administrator about this error. Be specific in your description of the error and the operation that produced the error. Include your project ID, your location, the date and time the error occurred.

## 9. Can't upload annotations from Apollo to *DNA Subway*

- Insure you are connected to the Internet; then try to upload again.
- Save annotations to your desktop. Then, close and re-open Apollo. Your annotations may automatically load into *DNA Subway*. If you don't see your annotations, load them from your saved desktop file.
- If the error persists use the "Feedback" button at the bottom of *DNA Subway* to inform the site administrator about this error. Be specific in your description of the error and the operation that produced the error. Include your project ID, your location, the date and time the error occurred.

## 10. "Search Genomes" on the Yellow Line doesn't generate any results.

- Try searching a different genome. Some genomes may not have any similar genes.

## 11. "Search Genomes" on the Yellow Line returns an error "E"

- Re-run the search.
- Restart your browser and return to *DNA Subway*. Log into your account again, open your project using the "My Projects" button to return to your project. (Guest users will need to start over after closing their browsers.)
- If the error persists use the "Feedback" button at the bottom of *DNA Subway* to inform the site administrator about this error. Be specific in your description of the error and the operation that produced the error. Include your project ID, your location, the date and time the error occurred.