

Sequence annotation

Steps:

- Identify Repeats
 - Dotter
 - RepBase/Censor
 - TREP
 - Blast2Sequences
- Identify Putative Genes
 - BLASTX – find possible proteins
 - BLASTn – find real ESTs that may match your proteins
 - FGENESH+ – polish your annotation
 - BLASTp – validate your predicted protein with ortholog from BLASTX
 - DNA Subway

Objective: Annotate genes and repetitive elements using a combination of gene-finding programs, and BLAST searches. Sequence text files for annotation are located on SmartSite in the Lab5 folder. These include a *T. monococcum* BAC and two *P. taeda* BAC sequences.

Open 322N09consensus.txt

1. Dotter

Dotter helps to identify repetitive elements present in a sequence.

1.1. Use **Dotter** to align the *T. monococcum* sequence with itself.

- Can you identify any repeat?
- Are they direct or inverted repeats?
- To facilitate annotation, let us **divide the sequence** into two separate regions: the one where repeats were identified, and the one with no repeats and where potentially genes will be found.

2. Annotating REPETITIVE ELEMENTS

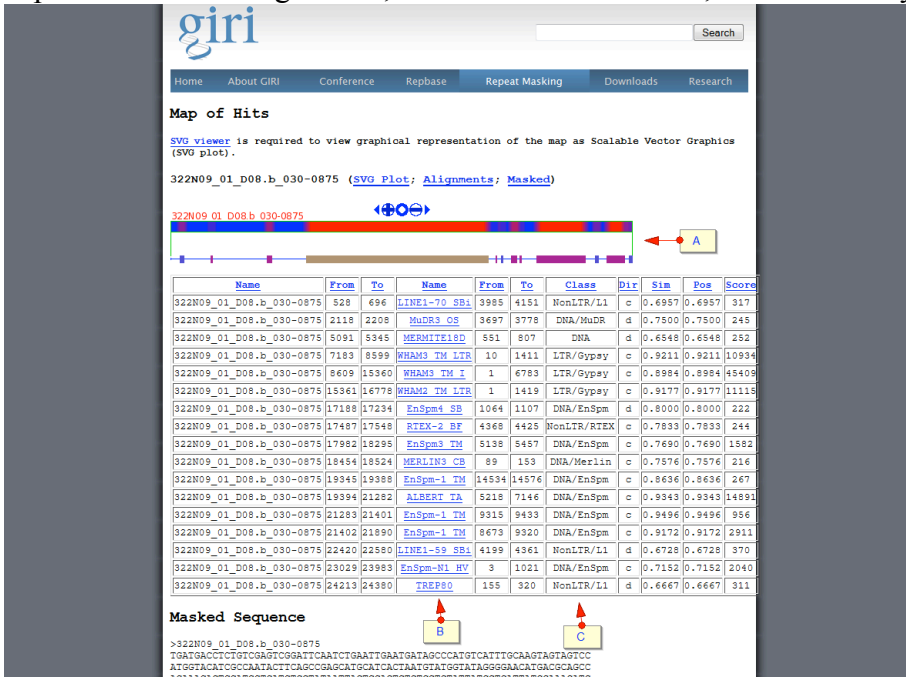
2.1. Censor is a tool that screens query sequences against a reference collection of repeats and masks homologous portions. It also is able to generate reports classifying all found repeats.

2.1.1. Go to <http://www.girinst.org/censor/index.php> and access the CENSOR tool. Use the provided wheat and loblolly BAC sequence to query for repetitive elements. Choose **A. All**

for sequence source, Paste the sequence into the **B. textbox** and hit the **C. Submit Sequence** button.



2.1.1. When the job is finished running, take note of the **A. alignments**, the **B. names**, the **C. classes** of the repetitive elements. Further down the page you will find a more detailed explanation of the alignments, the masked FASTA file, and a summary table.



2.2. Perform a **blastn** alignment with the **region of the sequence where the repeats were identified**, using the **nucleotide collection** database.

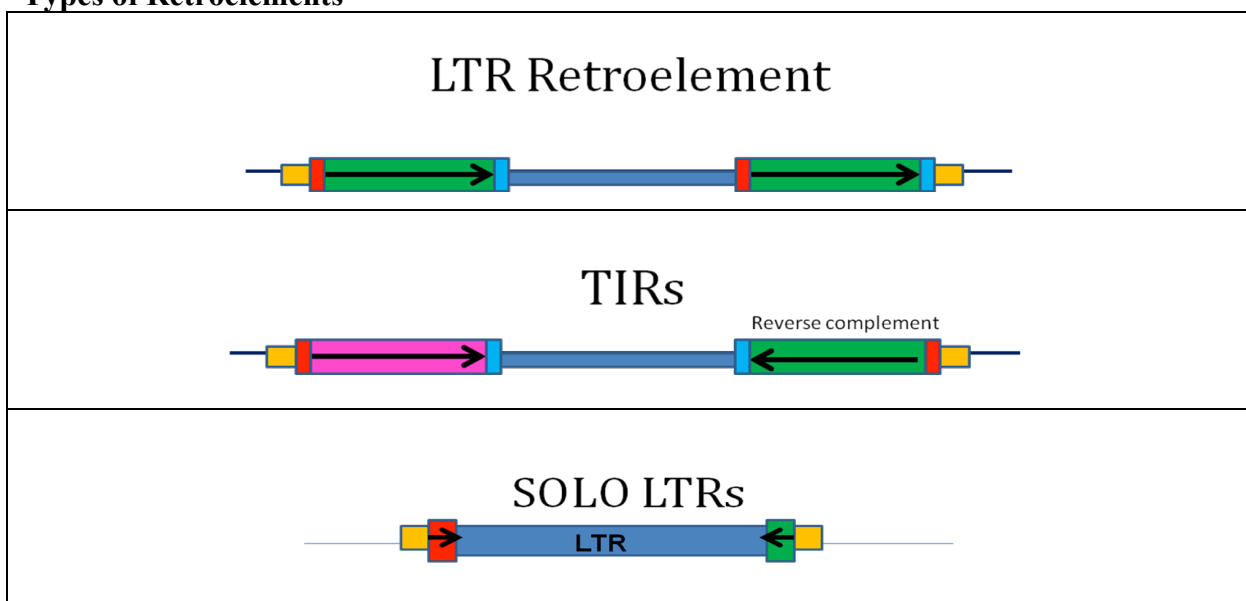
- Click on the best **blastn** alignment. What is the accession/version number?
- Open the flat file in a new tab. Go to the 'features' section to see the annotation of the element present in this region of the subject sequence. What is it? Using the coordinates of the alignment between your query sequence and the subject sequence from the database, find and highlight this element in your sequence in Word.
- Divide now the **region of the sequence where the repeats were identified** into two approximate halves, and align the two halves using **blast2sequences** to help yourself identify the LTRs of the repetitive element found. Mark the LTRs with **bold letters**. Can you identify the host duplication and the inverted repeats flanking the LTRs? Highlight them in your sequence in Word. Highlight the inverted repeats flanking the LTRs with the same color used to highlight the complete repetitive element, but use a different color to highlight the host duplication, since it is not part of the repetitive element.
- Annotate the repetitive elements in the *T. monococcum* sequence in Word.

2.2.1. TREP: BLAST repeats

Go to <http://wheat.pw.usda.gov/ggpages/Repeats/blastrepeats3.html>. This specialized repeat database allows you to discover and annotate the repetitive elements present in a sequence. Copy the **region of the sequence where the repeats were identified** and paste it in the **TREP** window. Select **blastn** program and **Cereal repeat sequences, complete set** database. Click on **Search**.

- Do you get any significant hit? What is it? Does it agree with your previous finding?

Types of Retroelements



3. Annotating GENES

3.1. Perform a **blastn** alignment with the **region of the sequence with no repeats**, using the **est_others** database.

- How many exons would you predict the gene present in this region has? Highlight them in your sequence in Word.

3.3.1. Gene prediction programs

3.3.1.1. FGENESH

Go to

<http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind> .

Copy the **region of the sequence with no repeats** and paste it in the **FGENESH** window.

Select **Monocot plants** as organism, and click on **Search**. Take a look at the predicted genes by clicking on **Show picture of predicted genes in PDF file**.

- How many genes are predicted? How many exons?
- Using a combination of your **blastn** search and the gene finding programs, identify start codon, splicing sites, exons, stop codon, and PolyA, and highlight them in your sequence in Word.

3.3.1.2. You can go to **Gene Sequer** at <http://www.plantgdb.org/PlantGDB-cgi/GeneSequer/PlantGDBgs.cgi>, and **GENSCAN** at <http://genes.mit.edu/GENSCAN.html>, and compare the results between the different gene prediction programs.

- How many genes/exons are identified by each of the programs?

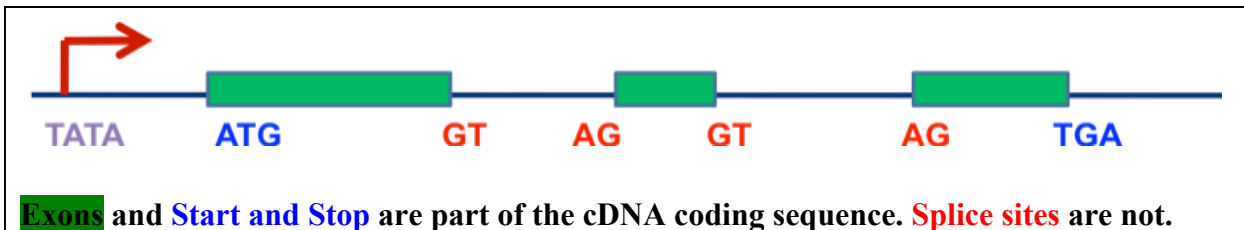
3.3.2. Gene annotation

3.3.2.1. BLASTP (protein blast) and BLASTX

Go to <http://www.ncbi.nlm.nih.gov/BLAST/> and then to **BLASTP (protein blast)** to perform a search with the translated protein (after translation using **GeneTool**), or to **BLASTX** to perform a search using the protein database with the cDNA.

- Do you get any significant hit? What is it?
- Do you find any conserved domain? What is it?
- Annotate the gene in the *T. monococcum* sequence in Word

Basic Gene Structure



3.3.2.2 DNA Subway

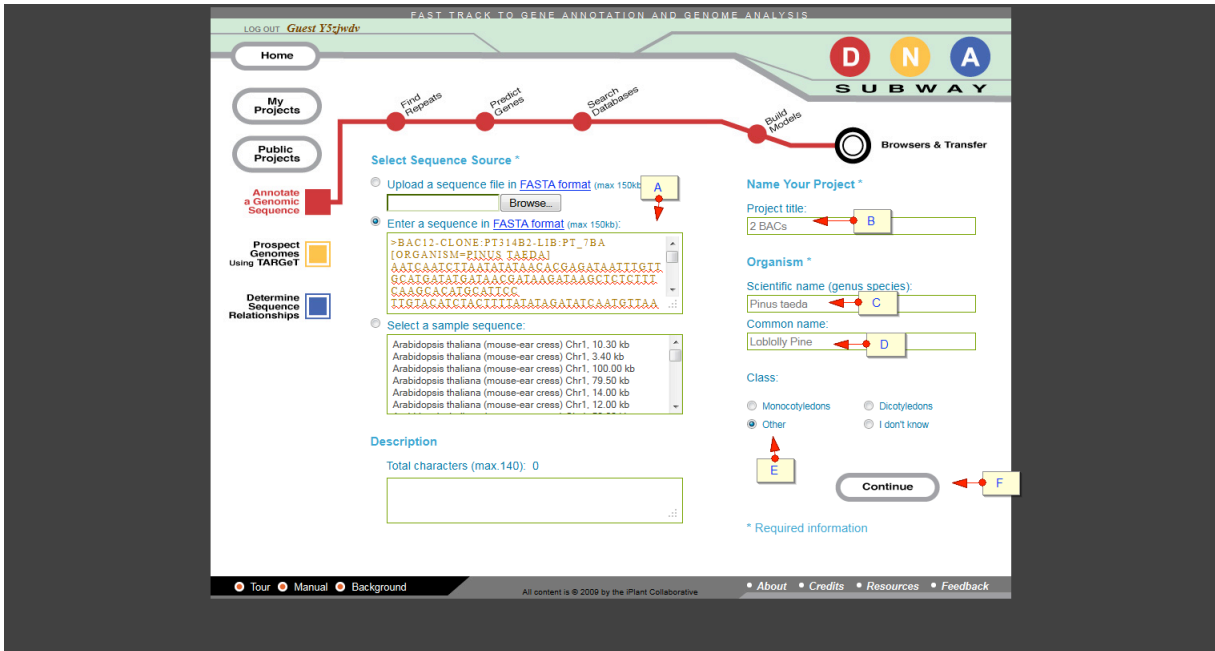
Go to <http://dnasubway.iplantcollaborative.org/> and then click on **A. Enter as Guest** in the top left corner.

The screenshot shows the DNA Subway homepage. At the top, there is a navigation bar with the letters 'D', 'N', and 'A' in colored circles, and the word 'SUBWAY' below them. To the right, it says 'Optimized FOR PLANTS animals coming soon'. On the left, there is a login form with fields for 'Username:' and 'Password:', and buttons for 'Log In', 'Enter As Guest', 'Forgot Password?', and 'Register'. A red arrow labeled 'A' points to the 'Enter As Guest' button. Below the login form is a subway map diagram with three lines: red, yellow, and blue. The red line includes stations: 'Annotate a Genomic Sequence', 'Find Repeats', 'Predict Genes', 'Search Databases', and 'Build Models'. The yellow line includes: 'Prospect Genomes Using TARGet', 'Search Genomes', 'Alignment & Tree Viewer', and 'Browsers & Transfer'. The blue line includes: 'Determine Sequence Relationships', 'Assemble Sequences', 'Add Sequences', and 'Analyze Sequences'. A text block below the map explains the site's purpose: 'This site ties together key bioinformatics tools and databases to assemble gene models, investigate genomes, work with phylogenetic trees and analyze DNA barcodes. Roll over the "stations" on the subway map to find out more about the analysis steps. Analyze your own data or sample data provided. To start a project, select one of the "lines" (red, yellow, blue). Register and login to be able to save and share your results.' At the bottom, there are links for 'DNA Subway Training', 'Tour', 'Manual', 'Background', 'About', 'Credits', 'Resources', and 'Feedback'.

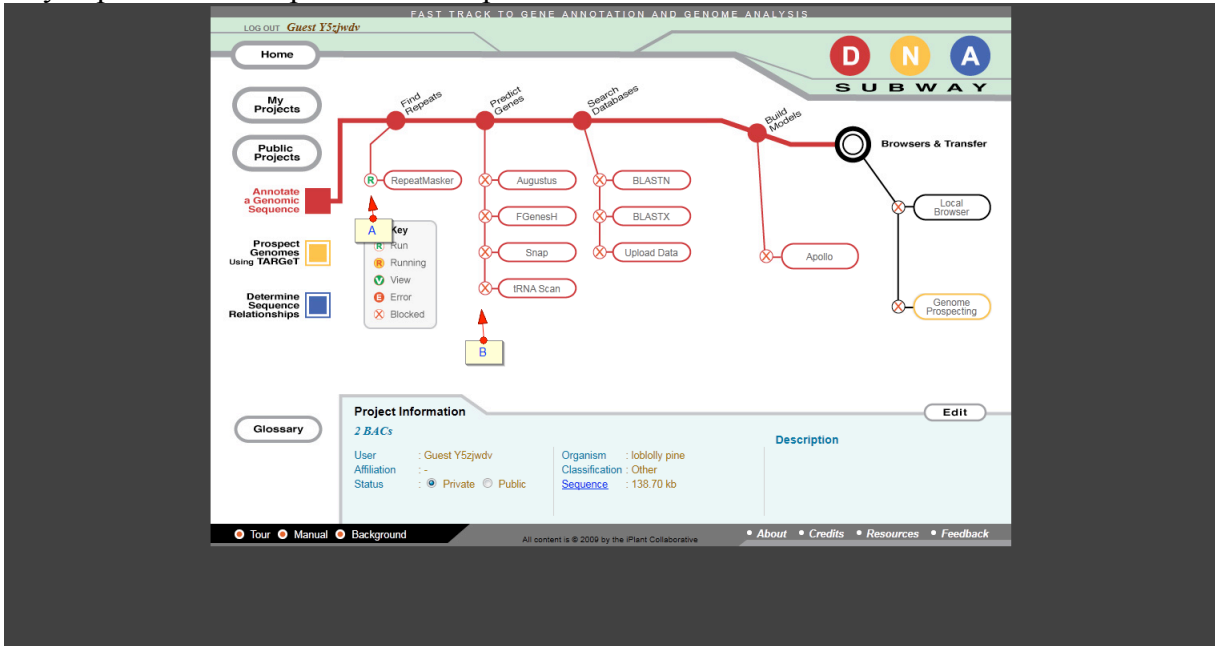
On the right hand side, click on **B. Annotate a Genomic Sequence**.

The screenshot shows the 'MY PROJECTS' page on the DNA Subway website. At the top, there is a navigation bar with the letters 'D', 'N', and 'A' in colored circles, and the word 'SUBWAY' below them. On the left, there is a 'LOG OUT Guest Y5jwdy' link. Below that, there are buttons for 'Home', 'My Projects', and 'Public Projects'. The 'My Projects' button is highlighted with a purple underline. To the right of the 'My Projects' button, it says 'MY PROJECTS' and 'You have no projects, yet. To start a project, click one of the squares on the left.' Below this, there are three colored squares: a red square labeled 'Annotate a Genomic Sequence', a yellow square labeled 'Prospect Genomes Using TARGet', and a blue square labeled 'Determine Sequence Relationships'. A red arrow labeled 'B' points to the red square. At the bottom, there are links for 'Tour', 'Manual', 'Background', 'About', 'Credits', 'Resources', and 'Feedback'.

On this next page, paste in the **A. textbox** the 2 BAC sequences, give the **B. project a name**, fill in the organisms' scientific name **C. Pinus Taeda**, Common name **D. Loblolly Pine**, and Class **E. Other**. Press the **F. continue** button when finished.

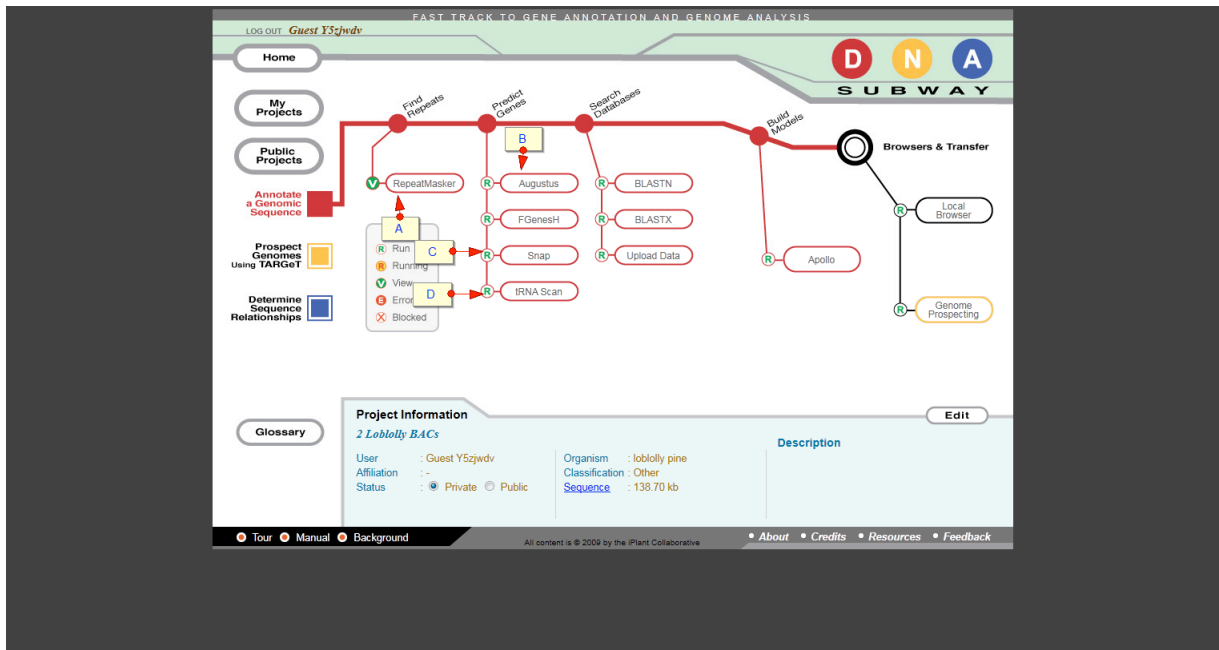


Now, we can run each step in the pipeline as they become available. For example, we can see a **A. green R** next to RepeatMasker in the circle under the first heading, *Find Repeats*. This means the RepeatMasker step is ready to run. You can see under the next heading, *Predict Genes*, the circles have a **B. red X** meaning that those steps can not be ran currently, since they depend on the RepeatMasker step.



Let's go ahead and click on the **A. RepeatMasker** button to start running that process. Once it has finished running, the R will change to a V so that you may click on the box to view the results of that individual step. Also, the next steps under *Predict Genes*, have changed from a

red X to a green R indicating they are ready to be ran. Once the RepeatMasker step is finished, click on the **B. Augustus**, **C. Snap**, and **D. tRNA Scan** boxes to start running those steps.



Continue stepwise through the DNA Subway pipeline by running **A. BLASTN** and **B. BLASTX**. Use Apollo to Build models and visualize your results. You can then view final results in the Local Browser.

