Case Study: Etude Annotation

Welkin Pope

Oct 2010,
revised Oct 2011 for DNA Master Annotation Guide

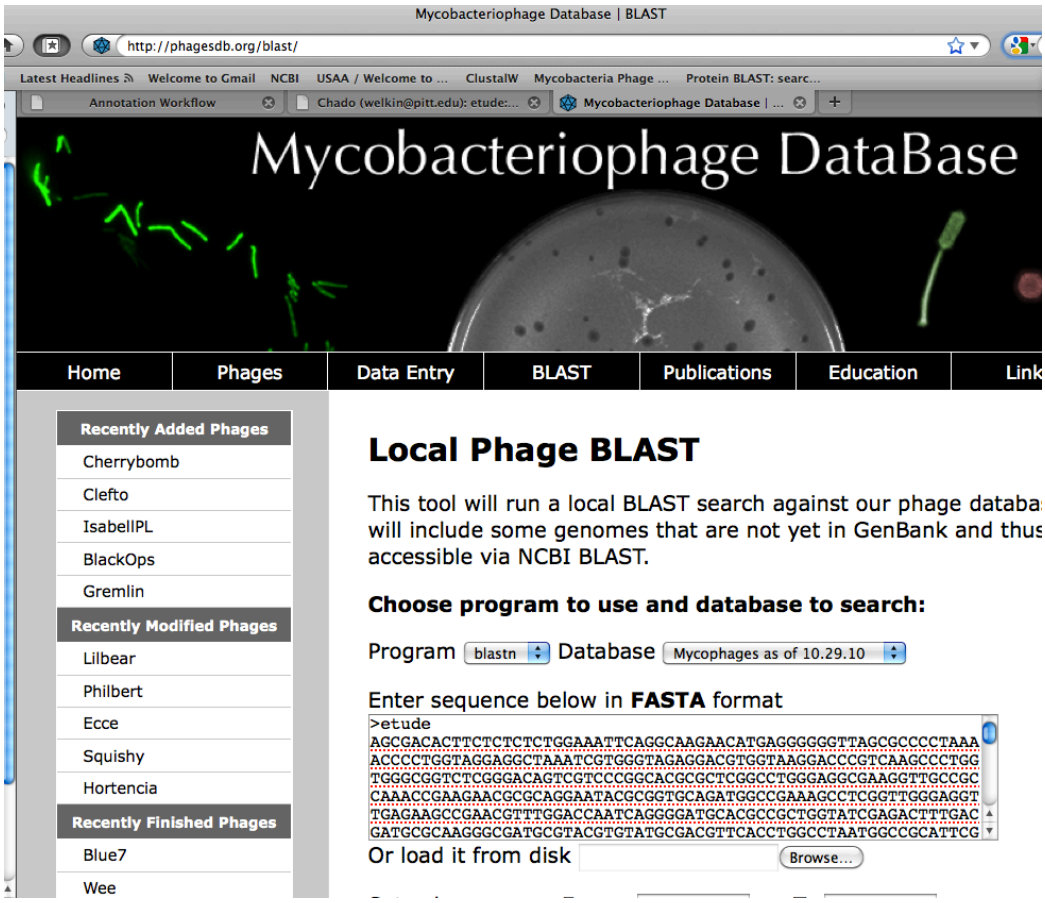
This is a step-by-step guide as to how I would annotate the practice genome Etude. This is not meant to replace any of the more detailed annotation guides, but is instead supplemental material to see how an experienced annotator assembles the necessary data and approaches the analysis. If any of the mechanics of any of the steps below seem unclear, please refer to the appropriate pages in your more detailed annotation guide.

It is also important to note that many of the bioinformatic decisions I make below are "best guesses" based on the information that I have at the time. Another annotator might make a different decision than I do (or I might make a different decision on a different day), and that is OK. Since I originally wrote this document last year, we've submitted another 140ish genomes to GenBank, and therefore there is a lot more data available at the time of this revision than there was last year. I've taken the new data into account and made a few comments specific to its analysis in italics in a few places throughout the case-study. Otherwise, only the instructions and the screenshots that pertain to DNA Master should have changed from last year's version.

Whole Genome Assessment

First, BLAST Etude against phagesdb.org.

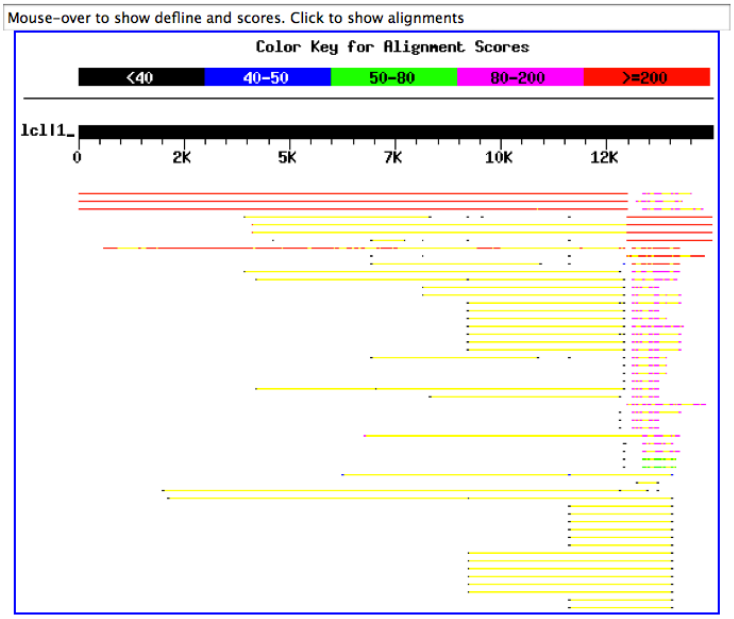
Go to the phagesdb.org BLAST page. Paste in the Etude sequence or browse to the FASTA file on your computer. Turn off the low complexity filter. Press BLAST.



When Etude is BLASTed against phagesdb.org, it appears that there is similarity to the Cluster L1 phages, and to the Cluster A3 phages.

Query= etude
(14,998 letters)

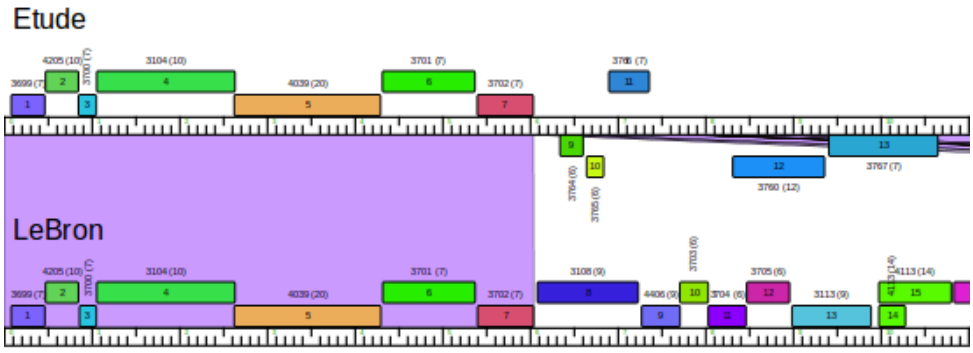
Distribution of 392 Blast Hits on the Query Sequence



Sequences producing significant alignments:	Score (bits)	E Value
UPIE Complete Sequence, 73784 bp including 10 bp 3' overhang (TC...	1.314e+04	0.0
LeBron	1.178e+04	0.0
JoeDirt Final Sequence, 74914 bp including 10 bp 3' overhang (TC...	1.169e+04	0.0
Microwolf Final Sequence, 50864 bp including 10 bp 3' overhang, ...	4022	0.0
Vix Complete Sequence, 50963 bp including 10 bp 3' overhang (CGG...	3998	0.0
JHC117 Final Sequence, 50877 bp including 10 bp 3' overhang, Clu...	3998	0.0
Bxz2	3998	0.0
Faith1 Complete Sequence, 75960 bp including 10 bp 3' overhang (...	1388	0.0
Rockstar Complete Sequence, 47780 bp including 10 bp 3' overhang...	232	2e-59
Peaches	212	2e-53
Eagle	204	4e-51
LHTSCC Complete Sequence (51813bp, including 10bp 3' overhang: C...	196	1e-48
George Final Sequence, 51578 bp including 10 bp 3' overhang, Clu...	137	8e-31

Now we pull up Etude in phamerator, next to its closest matches. I will use LeBron and Bxz3 for now, because these two phages have annotations in GenBank already—which means that when I look for individual genes using BLAST on the NCBI website, I should see these genes, and they will be genes that have already been curated and well-examined by the annotators. I will also check UPIE, JoeDirt, and Microwolf's draft annotations in phamerator.

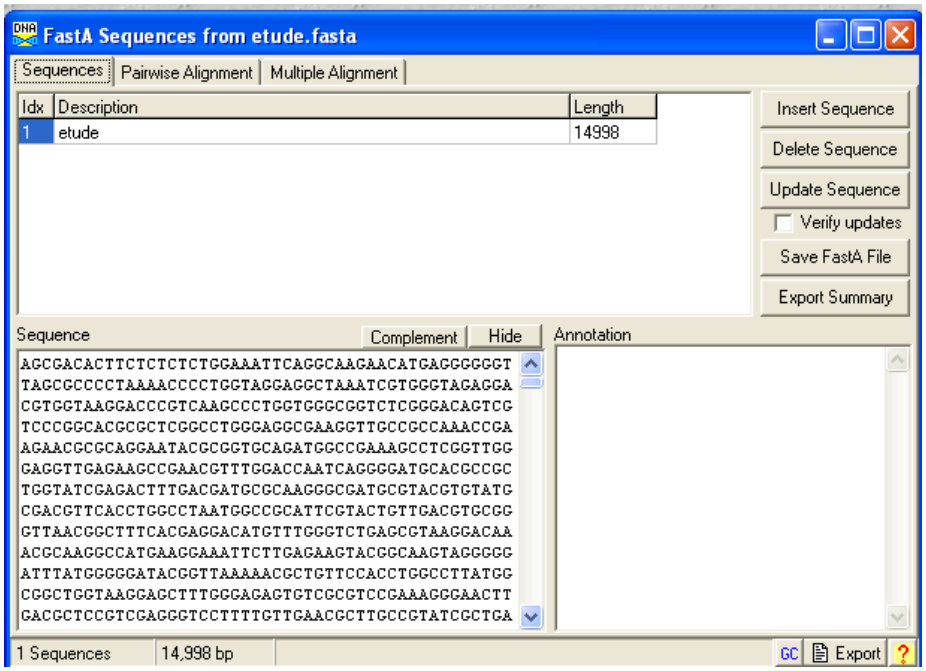
(Note: these phages all have final GenBank versions now 10/10/11—WP)



Notice how the purple between the two genomes indicates that the nucleotide sequence similarity is very high between Etude and LeBron for the first seven genes in both genomes. I will start with calling these seven genes.

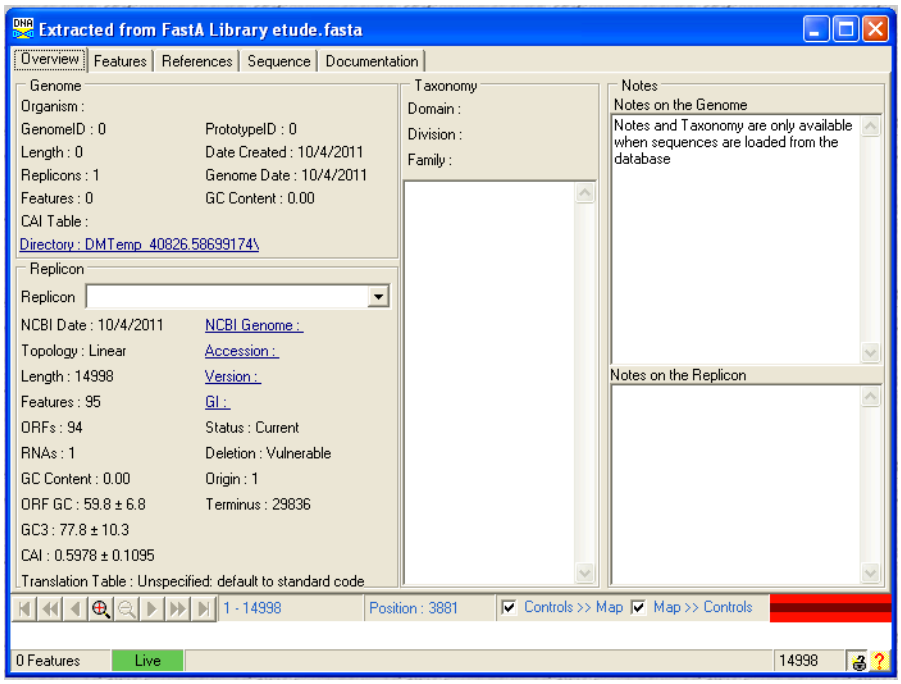
Next, I open DNA master, and load the Etude sequence from its fasta file.

-> File ->Open -> FastA Multiple Sequence File



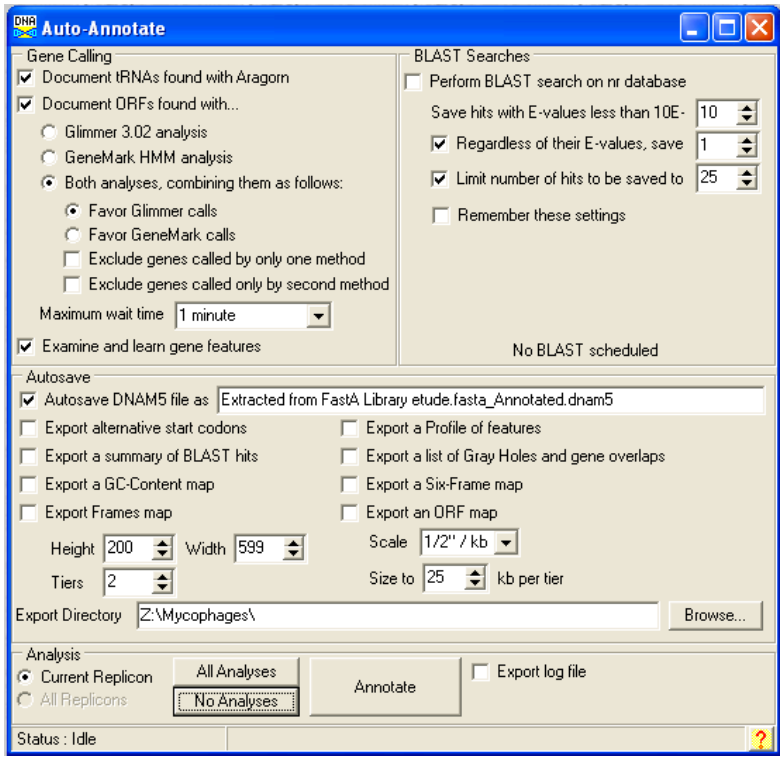
I then click “export” in the lower right corner, and “Create sequence from this entry only” from the menu that appears.

A DNA Master sequence file will be created:



This file is empty other than the imported sequence (viewable if you click the "Sequence" tab above.

Now I auto-annotate this file to generate and import the information from Glimmer, GeneMark, and Aragorn into the file. Click Genome->Annotation-> Auto-Annotate



Uncheck all the analyses buttons if necessary (click No Analyses)and then click "annotate". As Etude is a relatively short piece of DNA, I will check the "BLAST" box at the upper right.

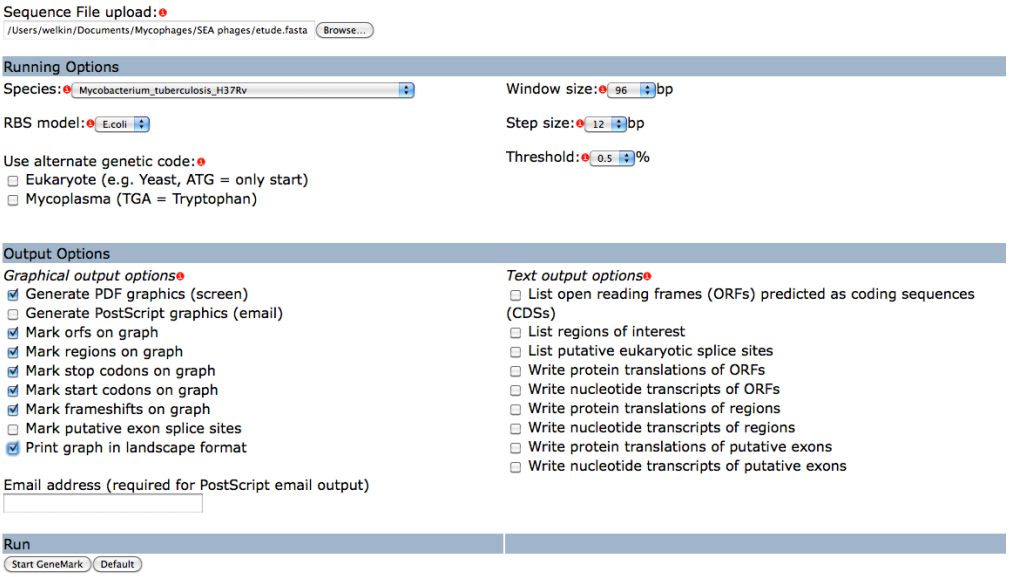
Once my genome is annotated and BLASTed, I will save it as Etude_annotated.dnam5.

Now I will generate the data from the programs outside of DNA master that I will need to review the auto-annotation.

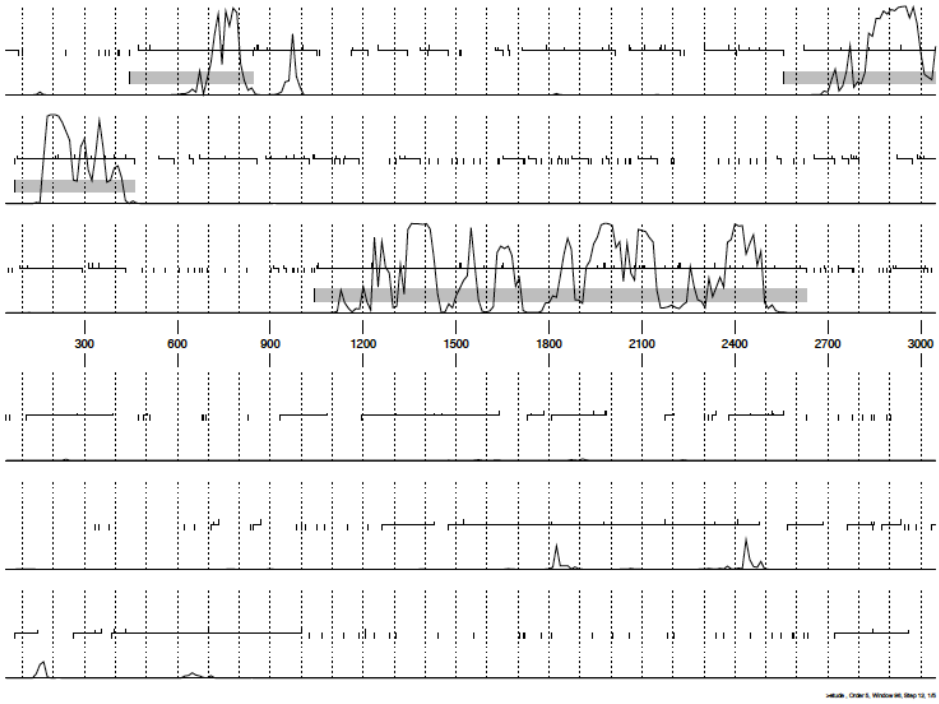
GeneMark TB:

GeneMark is located at http://exon.gatech.edu/genemark/genemark_prok_gms_plus.cgi

There is also a link from the phagesdb website. Upload the etude.fasta file and use the *Mycobacterium tuberculosis* coding model (either strain is fine).



Opening the .pdf of the GeneMark output should show you a 5 page document that begins like this:



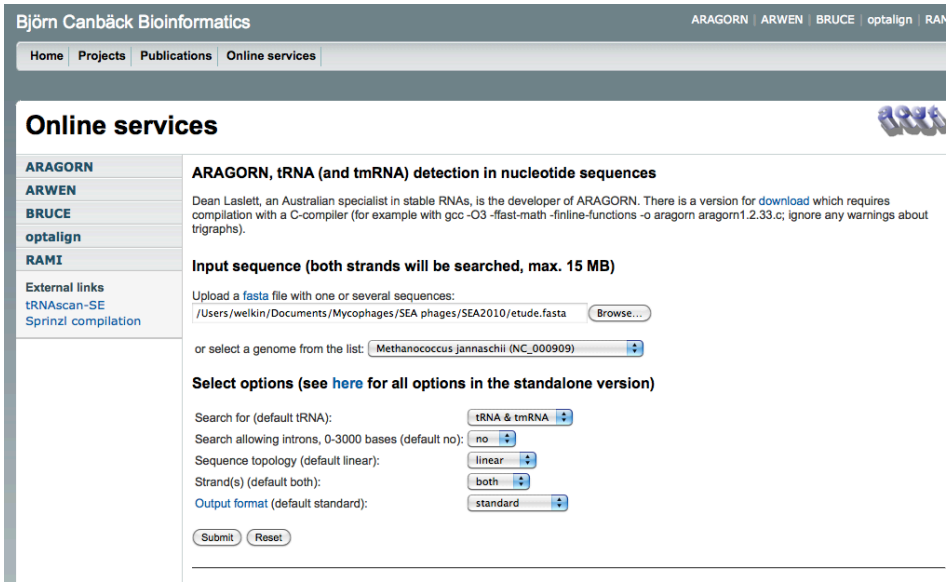
Here are the first four and start of the fifth genes in Etude as called by GeneMark. Each tier represents a different reading frame, with upticks from the center horizontal in each tier

representing start codons in that frame and downticks representing stop codons in that frame. ORFs of significant length are shown as horizontal lines in each tier. Coding potential is shown by the wiggly trace lines. Areas that GeneMark has designated "regions of interest" are shown with gray bars through the coding potential. Sometimes these regions are actually genes, sometimes not. I find it easier to just ignore the gray bars completely.

Aragorn:

Aragorn is found at: http://130.235.46.10/ARAGORN/or linked to from phagesdb.org.

Upload your FastA file, and change the default to "tRNA and tmRNA".



Then click "Submit". The output will look like:

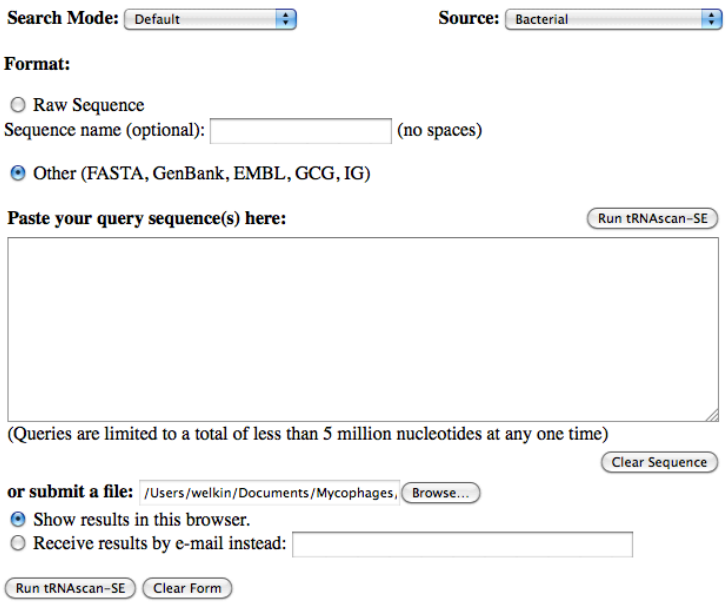
```
etude
14998 nucleotides in sequence
Mean G+C content = 60.2%

1.

              c
             a
           g+t
           g-c
           t-a
           c-g
           c-g
           t+g
           g-c     tg
          t   caccc  a
   taaa   a    !!!!!  g
  t    cgg     gtggg  c
  g    !!!     t    tt
  g    gcc     c
   caaa        t
          g+tg
          c-g
          g-c
          t-a
          c-g
          a-t
         c   a
         t   a
          caa


   tRNA-Leu(caa)
   75 bases, %GC = 56.0
   Sequence [6240,6314]


Primary sequence for tRNA-Leu(caa)
1    .    10    .    20    .    30    .    40    .    50
ggtcctgtaggcaaattggcaaagccgcgtcactcaaaatgacgtgtctg
tgggttcgagtcccaccgggactac
```

The web-based Aragorn output shows a single tRNA that has a correctly-trimmed 3' end. I will come back to this when I get to the tRNA in my draft annotation.

tRNAscan-SE:

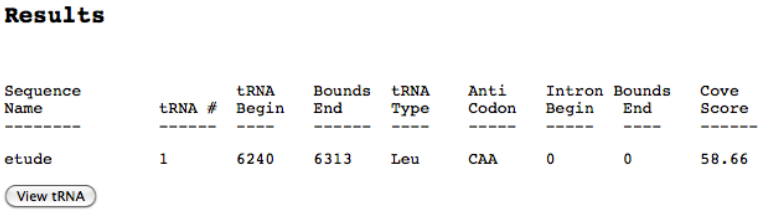
tRNAscan-SE is available at: http://lowelab.ucsc.edu/tRNAscan-SE/ or linked to from phagesdb.org

I change the source to "bacterial", and browse to my file:



Now I click "Run tRNAscan-SE.

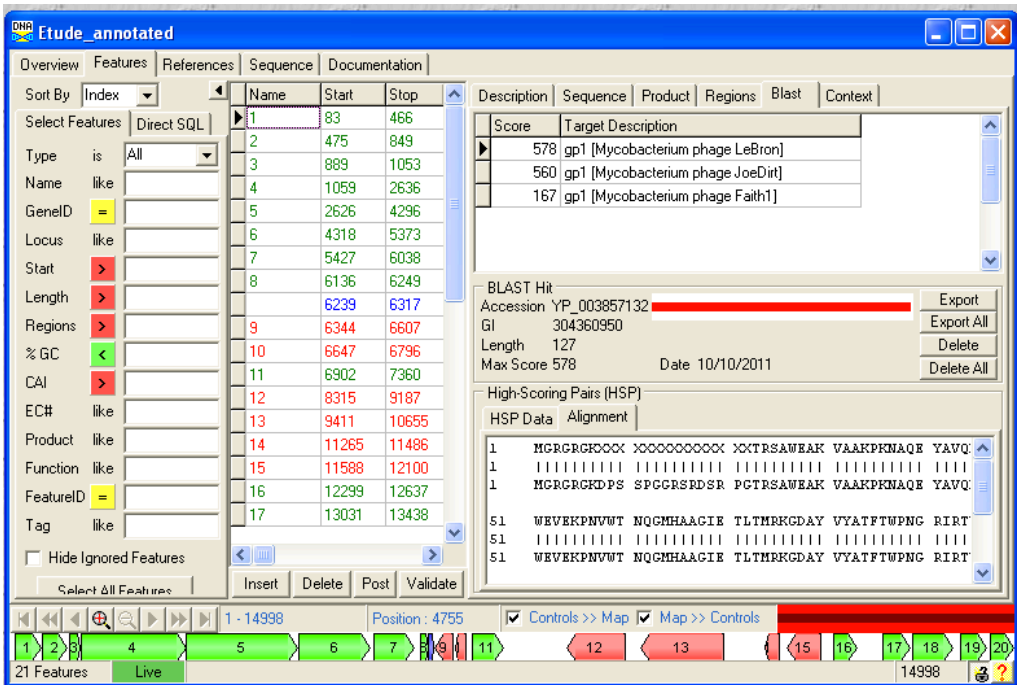
The results are similar to Aragorn, however tRNAscan SE has called the tRNA with different start/stop coordinates:



I will evaluate these when I reach the tRNA in the genome sequence.

At this point, I might also make a genome map, however, this genome is so small I can visualize the entire thing at once in the interactive map in DNA Master, so I am going to skip the additional map.

Whole Genome overview:



My auto-annotation has 21 called genes and 1 tRNA. If I click on the BLAST tab for gene 1, I can see the scores of the alignments from GenBank and the actual alignment.

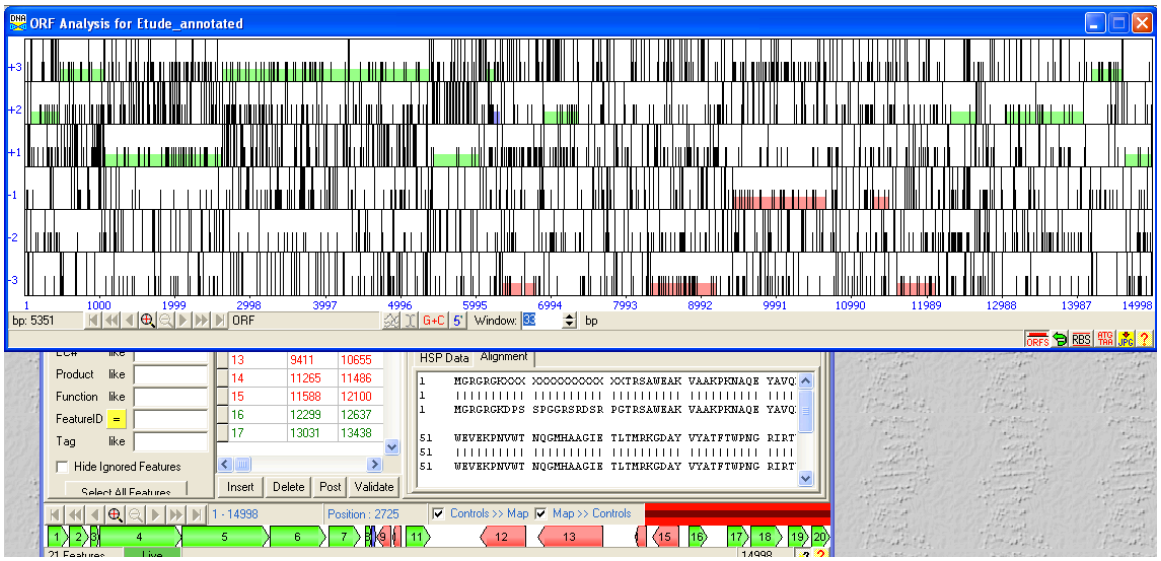
I can tell from the interactive map at the bottom of my Etude sequence file that there are some large gaps in my genome between genes 11 and 12 and 13 and 14. I will take a closer look at these areas when I reach them in my annotation. There is also a gene overlap with gene 8 and the tRNA. This will also need to be resolved.

Refining my annotation

I now open the Frames window:

Click →DNA→Frames

→Toggle on the features listed in my features table by clicking the "ORFs" button in the lower right-hand corner.



As you mouse over the Frames window, the lower left box will display in real-time the base pair coordinate of your cursor.

My screen isn't quite big enough to comfortably display both the frame window and my sequence window, so I will be flipping back and forth between the windows.

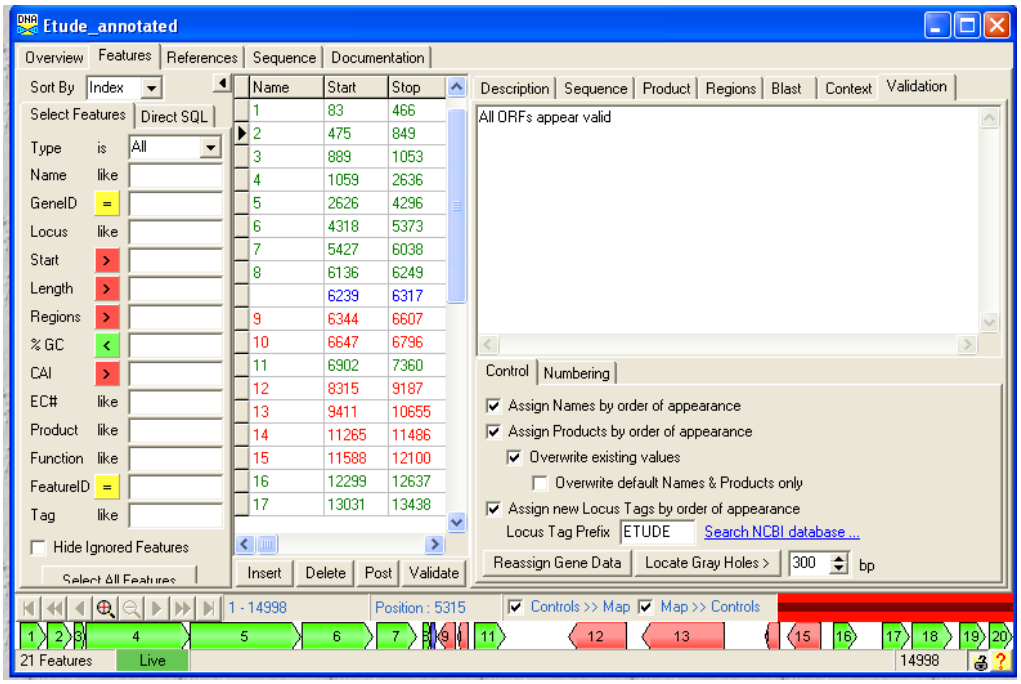
In the Frames window, I see my forwards transcribed ORFs in green, the reverse in red, and my tRNA in blue. The 6 horizontal tiers represent the six-translational frames, and the full vertical lines within the tiers are stop codons while the half-vertical lines are start codons.

Note about GeneMarkTB vs the DNA Master frames window: while both of these programs display data in six tiers representing the six-translational frames, they are not in the same order. The top three forwards frames match between program, the bottom three reverse frames do not. If we label the tiers in each program 1-6 from top to bottom, then 1:1 DM to GTB; 2:2 DM to GTB; 3:3 DM to GTB; 4:5 DM to GTB; 5:6 DM to GTB; and 6:4 DM to GTB.

Before I start going through the genes one by one, I am going to validate the gene calls and renumber the genes. The auto-annotation function of DNA Master does not assign gene numbers to tRNAs, however, we do count them in our genbank files. To keep everything consistent, it is easiest to renumber genes initially and then again at the end of the gene identification process. As you grow more confident in your annotation ability, you may also want to renumber periodically as you add and delete genes from the annotation.

To renumber (and assign locus tags):

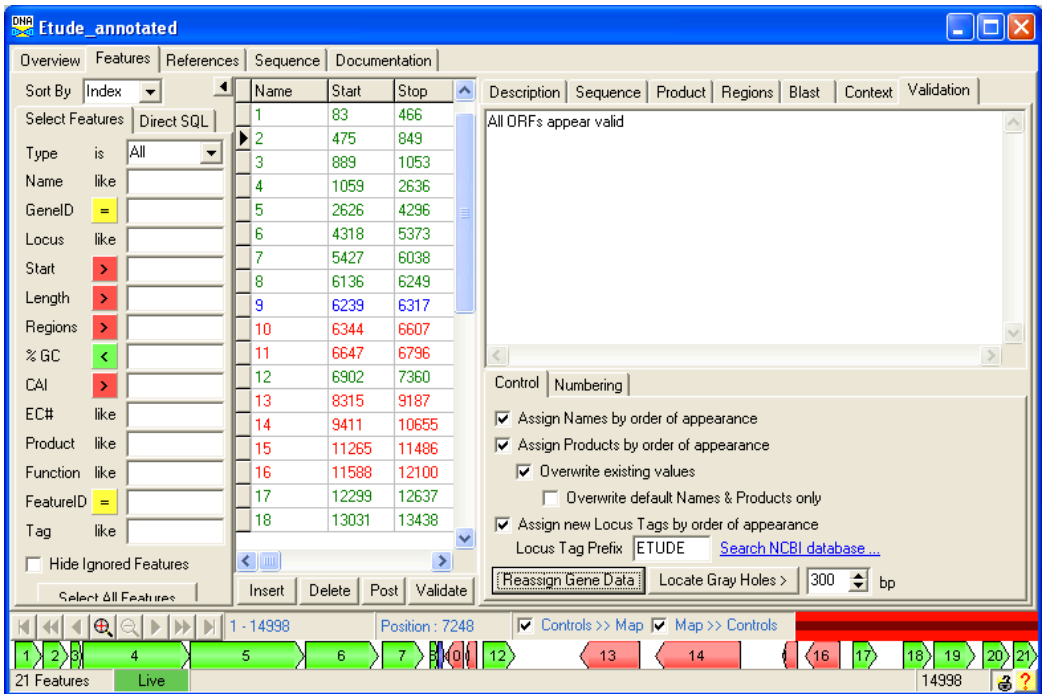
Click → Validate (at the bottom of the central column)



Check the box marked “Overwrite existing values”

Write the phage’s name in the Locus Tag Prefix field. GenBank uses locus tags to assign a unique id to every gene in the database. We prefer to create our GenBank submission files with locus tags comprised of the phage’s name and gene number already assigned, to prevent GenBank from assigning every gene a random number.

Click→Reassign Gene Data



The genes have been renumbered (notice the tRNA is now gene 9, whereas before it didn't have a number), and the locus tags have been adjusted on the description tab.

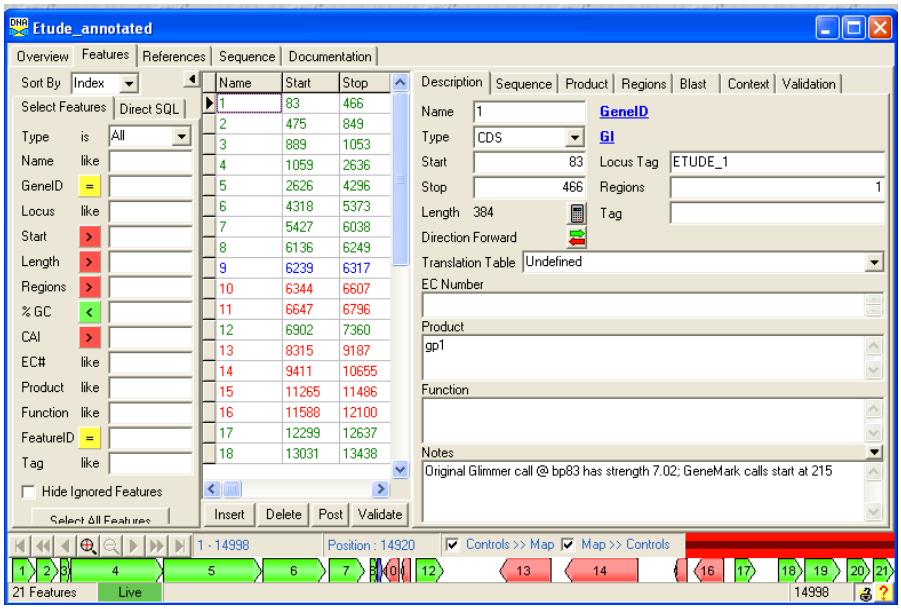
I will now start with Gene 1.

Gene 1:

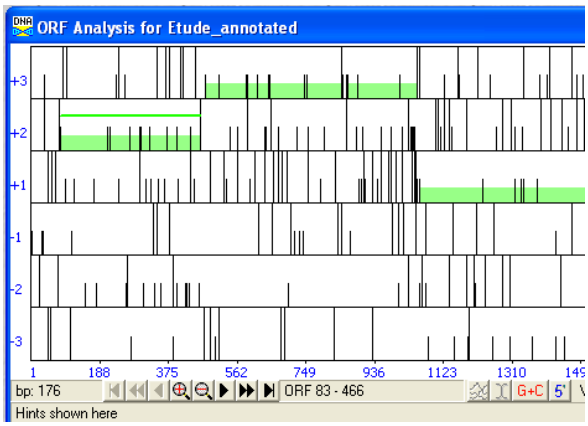
We need to decide: if this gene is a gene, if it is really gene 1, and where its start is. To do this, I will examine five pieces of data: coding potential in GeneMark TB, Glimmer/GeneMark auto-annotation calls, ribosome binding site (RBS) scores, gene gap/overlap with preceding gene, and BLAST alignment with previously annotated genes.

Coding Potential: From looking at the GeneMark TB output, it appears that the coding potential starts in this genome around 200 or so bp. There aren't any ORFs upstream of the called Gene 1 with coding potential, so I am confident that this gene is, in fact, Gene1.

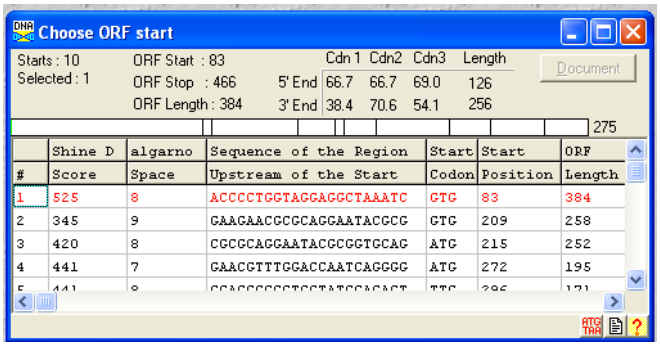
Glimmer/GeneMark: From the Notes field in my auto-annotation, I can tell that both GeneMark and Glimmer have called this gene; Glimmer starting at bp 83, and GeneMark starting at bp 215. This means that the GeneMark call does not encompass all of the coding potential as shown by the GeneMark TB output. However, both programs called the gene, and there is good coding potential in the GeneMark TB output. So I am confident that this ORF is a gene, and now just need to resolve what the start coordinate should be.



RBS scores: Click on the first highlighted green bar in the "Frames" window (you may want to zoom in by clicking the magnifying glass with the + icon at the lower left in the window). A thin green line will appear:



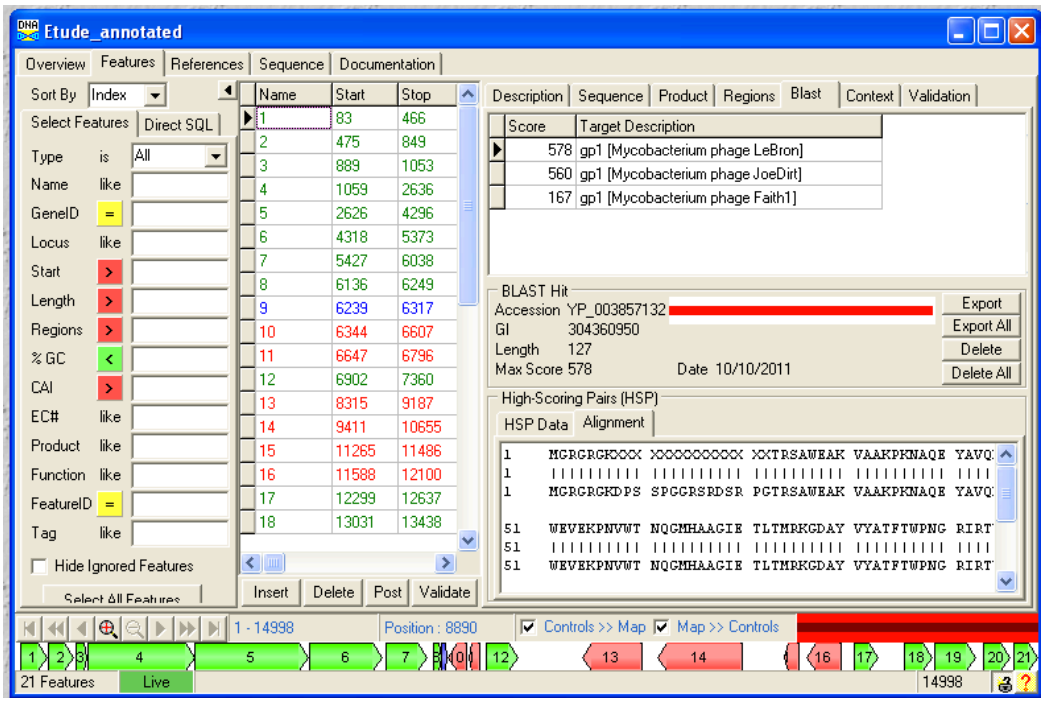
Then click the "RBS" button at the lower right of the Frames window. A new window will pop-up.



The start at position 83, the one called by Glimmer, has a higher Shine-Dalgarno (RBS) score (525) than the GeneMark start at position 215 (420). The start at position 83 is yields the longest possible gene as well.

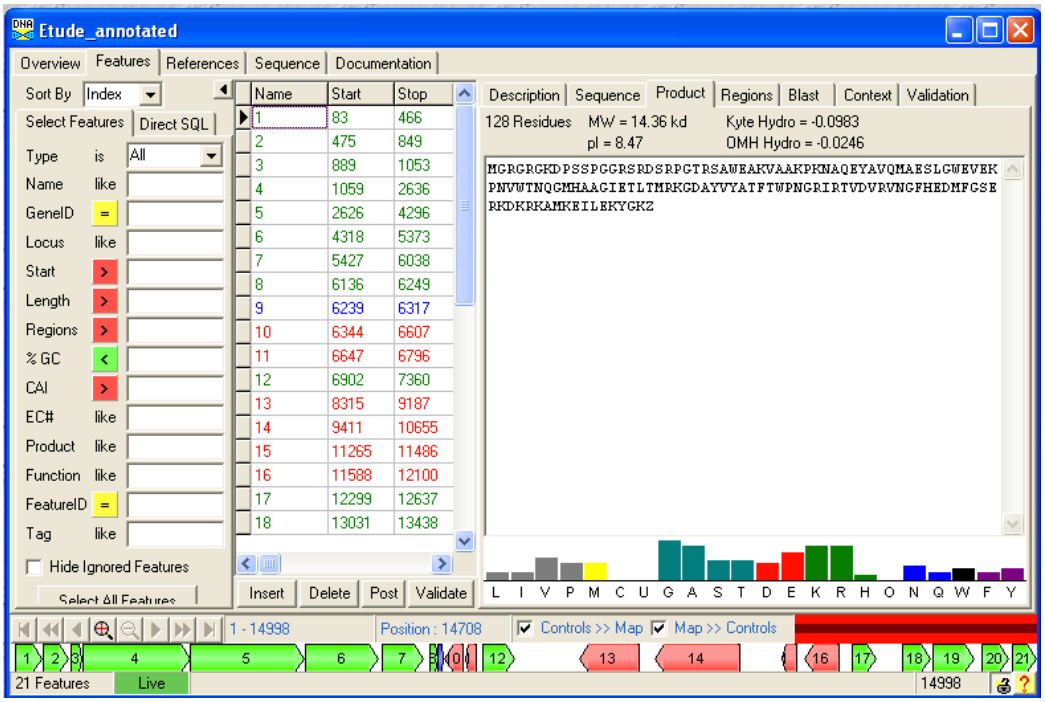
Gap/Overlap: Since it is gene 1, we can omit determining the gap or overlap with the upstream gene (as there isn't one!)

BLAST data: If I click on the BLAST tab (see below), I can see that the genes in GenBank that align well with my gene. Our top hit is (as expected from our Phamerator view) to LeBron gene 1. More importantly, when we look at the alignment, we see that "Query 1" aligns with "Sbjct 1". This means that we selected the same start codon that was chosen in the LeBron annotation. While it is not necessary to pick the same start codon as another closely related previously curated phage, it is necessary to examine all possible start codons, weigh the data, and make the best choice. Knowing that we have selected the same start codon as a similar phage makes it more likely that we have not accidentally missed a more appropriate start codon and that our reasoning is supported by previous examination of other annotators.

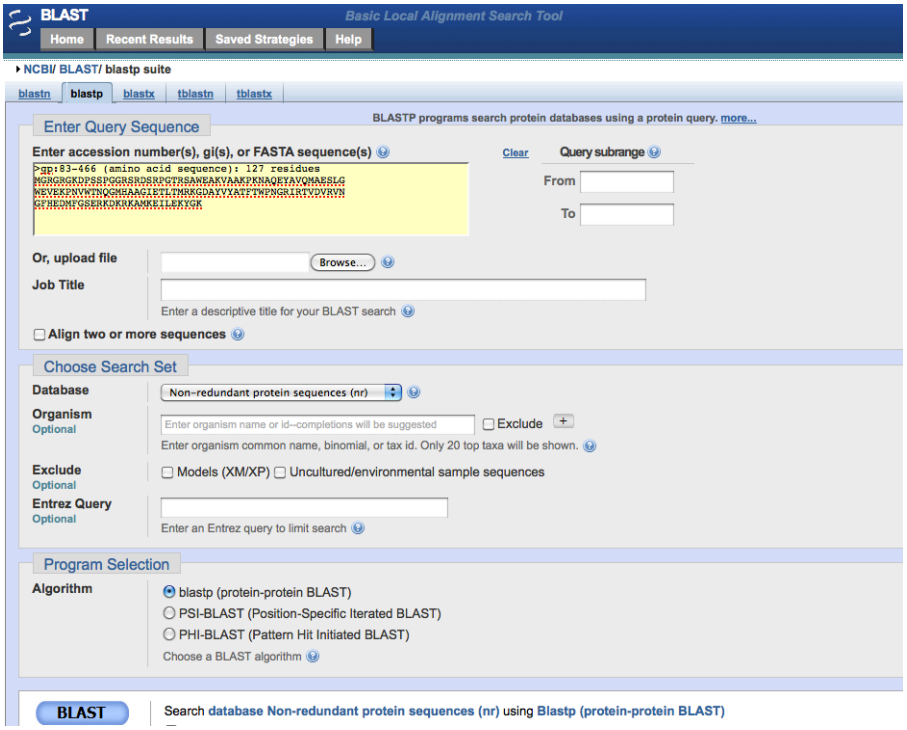


At this point, I am ready to make my decision. The BLAST data, RBS score, GeneMark TB coding potential, and Glimmer output all suggest that the best start for this gene is bp83.

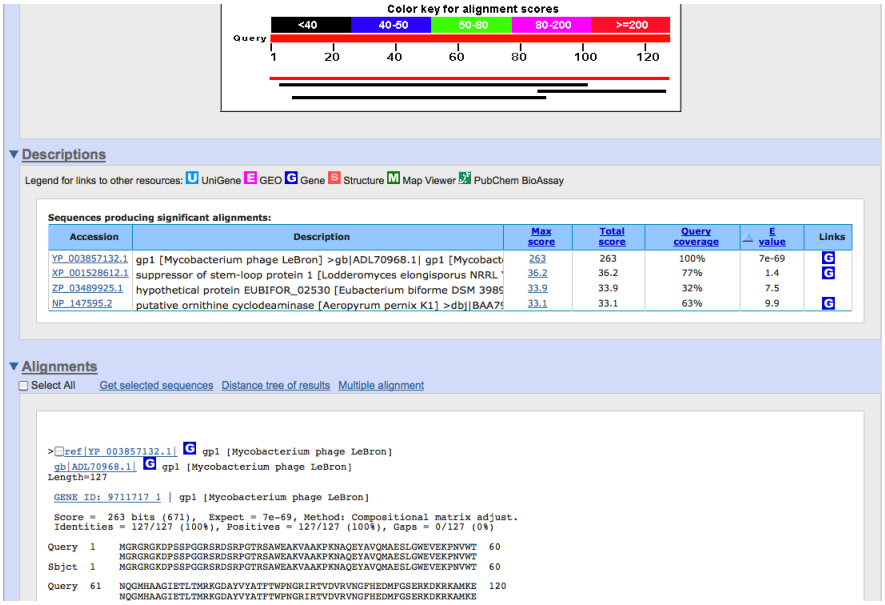
Now we need to decide if this gene has a function. Click the “product” tab



Copy this sequence and paste into NCBI's BLASTP page: http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome



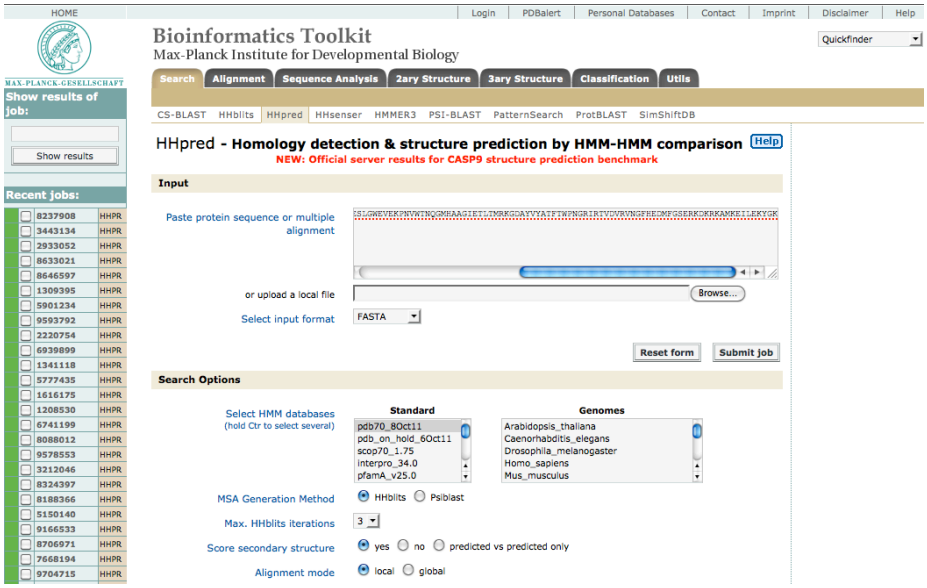
By pressing the "BLAST" button at the bottom of the page, we get the following result:



Our top hit is (as expected from our Phamerator view) to LeBron gene 1. LeBron is the only hit listed above that has an acceptable E value, and therefore we discount the other hits and their possible functional assignments.

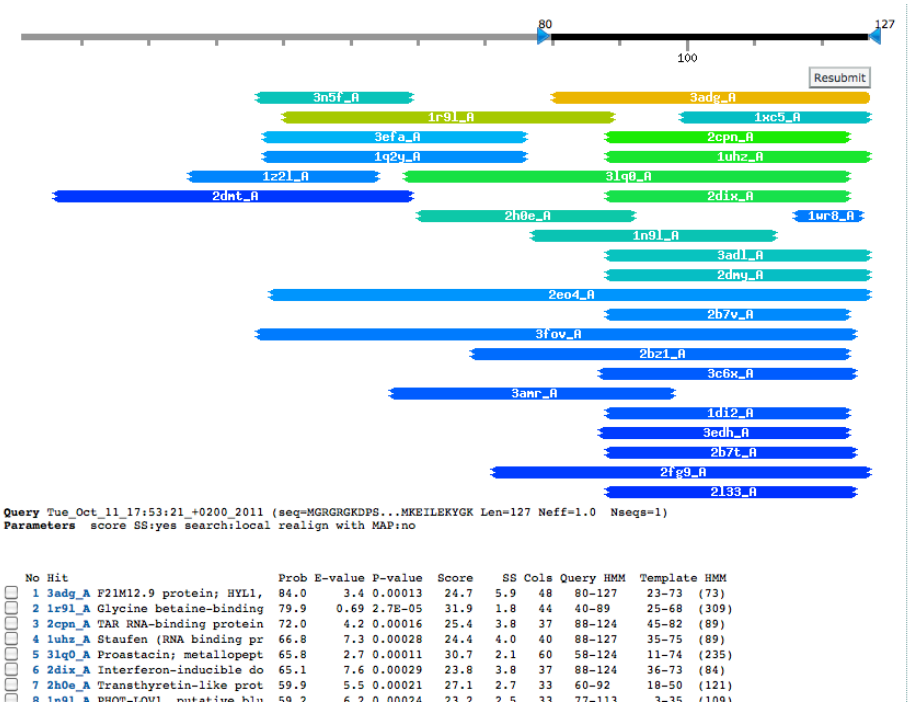
HHPred: (http://toolkit.tuebingen.mpg.de/hhpred)

Copy the amino acid sequence into HHPred's sequence field. Make sure to remove the "Z" at the end (DNA Master represents stop codons with a Z in the product field).



Click "Submit job" (at the far right, just above the beige bar labeled "Search Options")

Eventually, you will see a results that look like this:



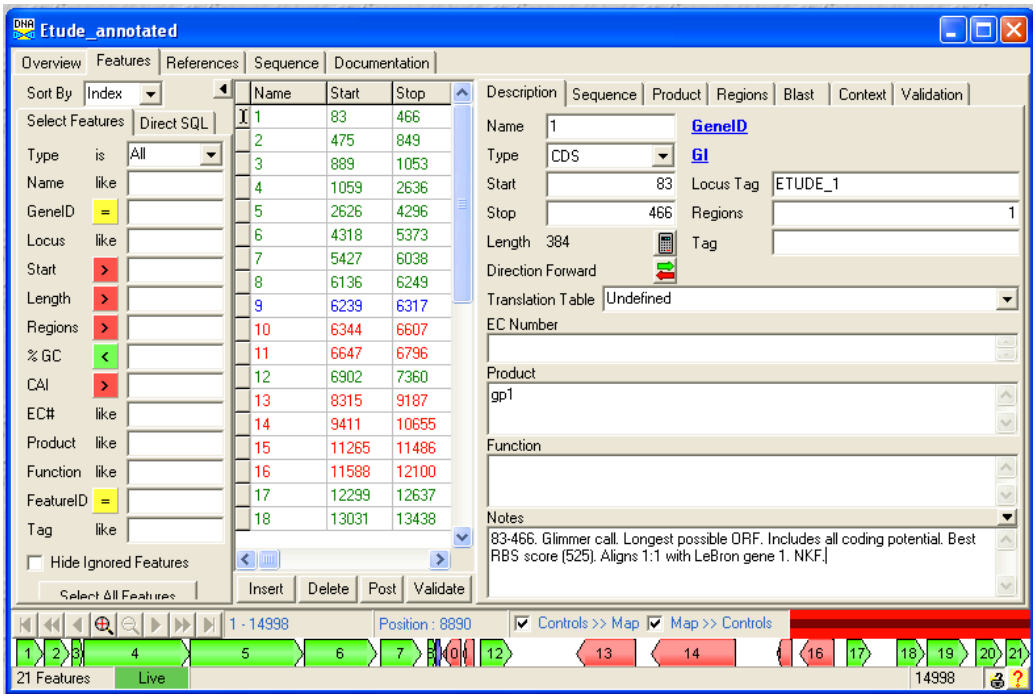
Only one of these alignments has a Probability score of above 80, and it from a small portion of our query to a protein in *Arabidopsis thaliana*. We will consider this not a match.

Finally, the Hatfull-labeled maps also suggest that there is no known function for gene 1 in LeBron.

Gene 1 has no known function (NKF).

Our last task is to add our annotation rationale to the Notes field for this gene.

In the Click on the Description tab in the right-hand section of the Features tab.
In the Notes field, add your notes.
Things to include: The gene coordinates for your gene call. Is this the longest possible gene for this gene call? Is this the Glimmer/GeneMark call? What is the gap or overlap between this gene's start and the previous gene's stop? Does this start have the best RBS score? Does this gene match anything in GenBank when you BLAST it? If so, what? What is the alignment between the start that you chose and the closest GenBank match? Is there a known function?



It is important that you physically type in the gene coordinates into your comments, as in some cases I have received files in which people believed they had changed their start coordinates and were not actually able to. In case there are any discrepancies between the gene coordinates that you think you are choosing and the gene coordinates that are actually saved into the file, it is important that you write what your gene coordinate choices are into the notes here.

It is also important that you report the BLAST alignments here, for two reasons: It is possible to generate spreadsheets of the gene data fields, including the notes, that can be very useful for genome checking. These spreadsheets will not include data from the BLAST tab. And if you ever accidentally lose your BLAST data (say, from parsing your

documentation, or from corrupting your file) you'll have a record of what the alignment was without having to BLAST your entire genome again.

Post your changes to the notes field, either by clicking "Post", or by moving to gene 2 by clicking on the corresponding row in the central column.

Gene 2:

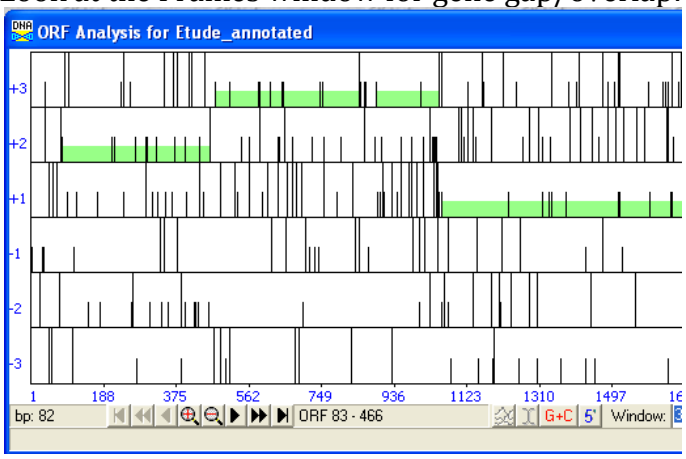
On the feature tab, click "2" in the central column.

In the Notes for gene 2, we can once again see that Glimmer and GeneMark have disagreed on the start for the gene (475 for Glimmer and 514 for GeneMark). However, as with gene 1, we can see that both programs have called the gene, and that there is good coding potential for the gene in the GeneMark TB output. So we will agree that this is a gene, and now just need to resolve its start.

Now we check our three criteria for start selection: coding potential, gene gap/overlap and RBS scores.

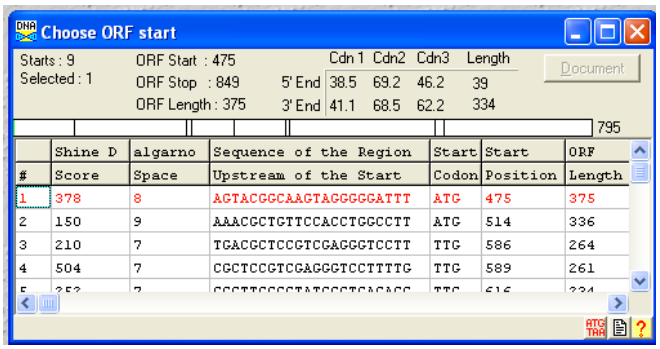
Coding Potential: the trace for the GeneMark TB coding potential doesn't start to rise until about bp 600 or so, so both the Glimmer and GeneMark start codons encompass all the coding potential.

Look at the Frames window for gene gap/overlap:



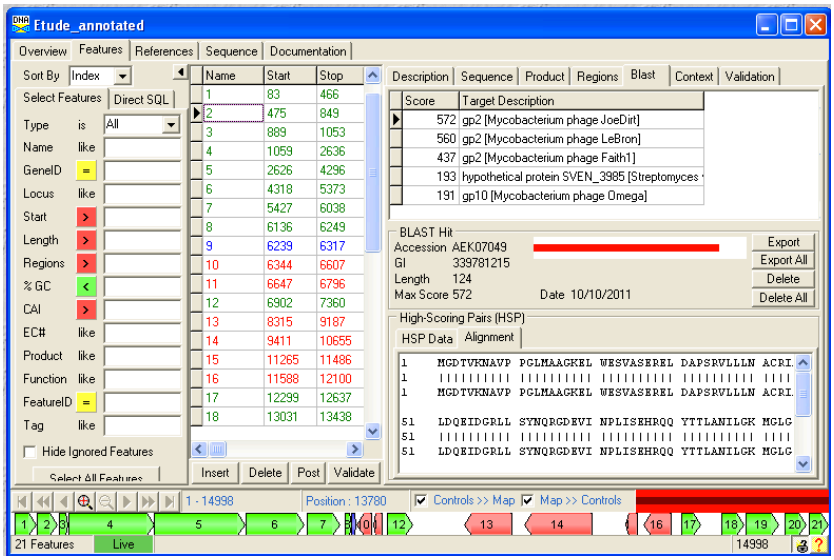
We can see here that the called start (the Glimmer start) represents the longest possible start for this gene—any extension would run into the upstream stop codon. There is no gene overlap, and there is a 9bp gap.

The RBS scores: Click in the box with the second green highlighted bar, and then click the RBS button on the lower right side of the frames window.



Again, the Glimmer call at bp475 has a higher score than the GeneMark call at bp514.

Examine the BLAST tab:

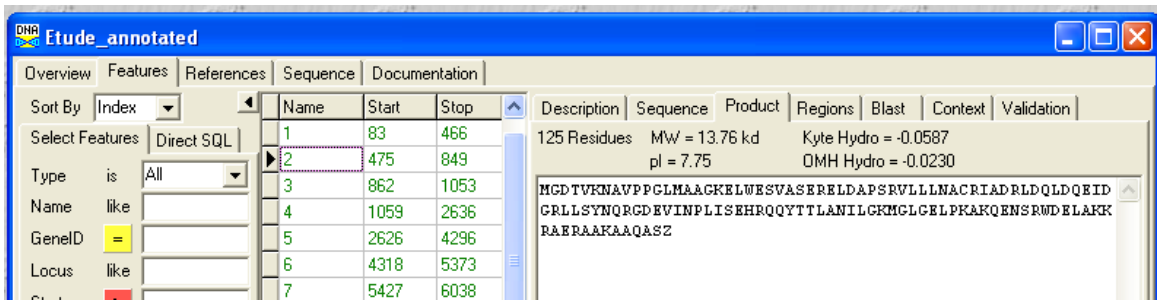


Once again, the best match aligns 1:1 with JoeDirt and LeBron.

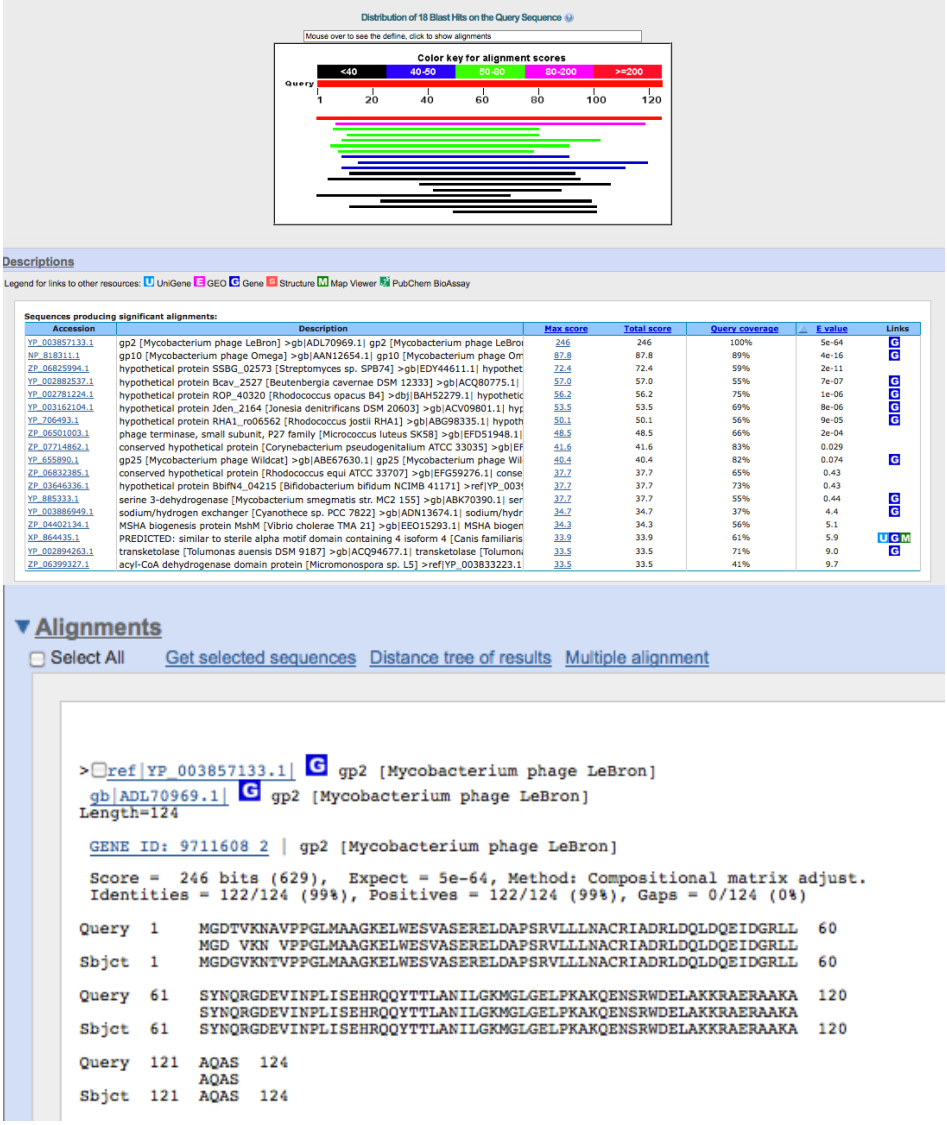
So we will pick the Glimmer call at 475 as our gene start, and enter the appropriate description into the notes. Now that we have an upstream gene, we will also write the gap/overlap in bp of this gene with the previous one.

Functional assignment: If we BLASTP this gene outside of DNA Master on the NCBI website, or examine Phamerator, or the Hatfull-approved genome maps with functions, we will see this gene is the small subunit of the terminase.

Click on the product tab in DNA Master (or on the gene in Phamerator

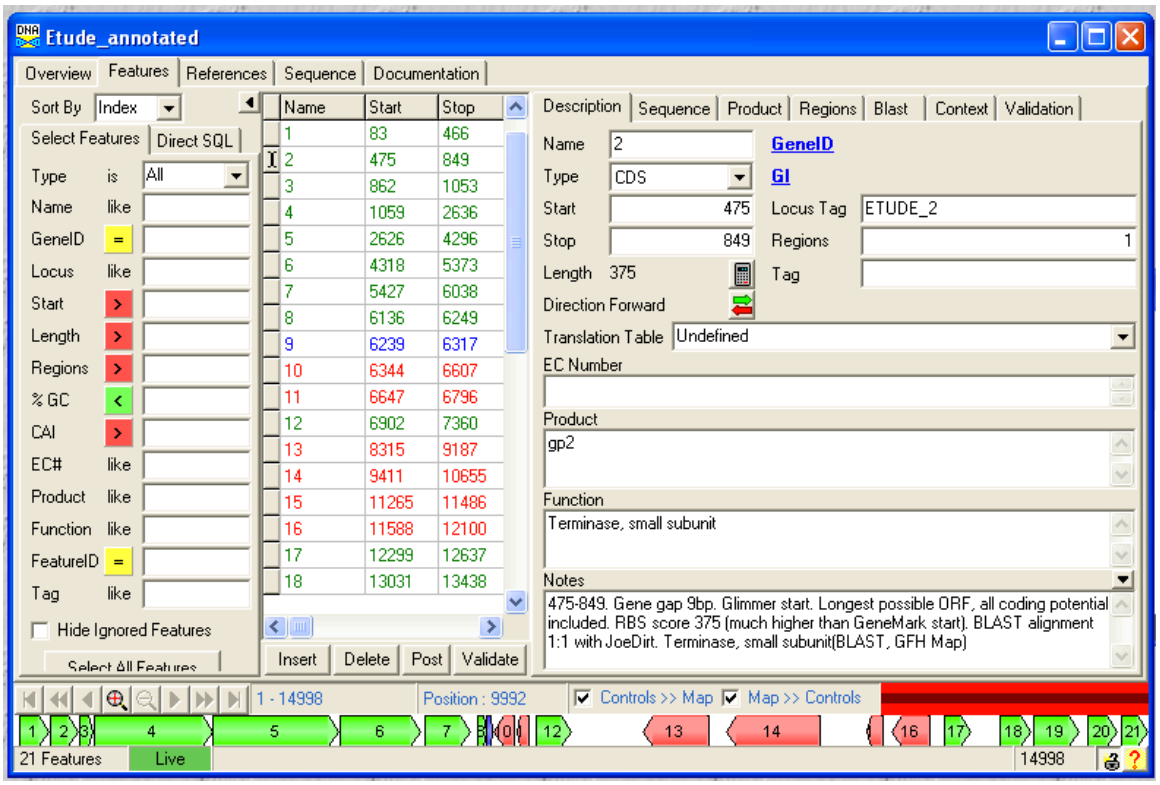


As in gene 1, copy and paste the amino acid sequence into the NCBI BLASTP page. The BLAST result show more hits this time, not only LeBron, but multiple other phages. This gene is the small subunit of the terminase, and so we must add the function into our annotation. The LeBron alignment is still “query 1 to sbjct 1”, indicating that gene 2 of LeBron and our gene 2 of Etude use the same start codon.

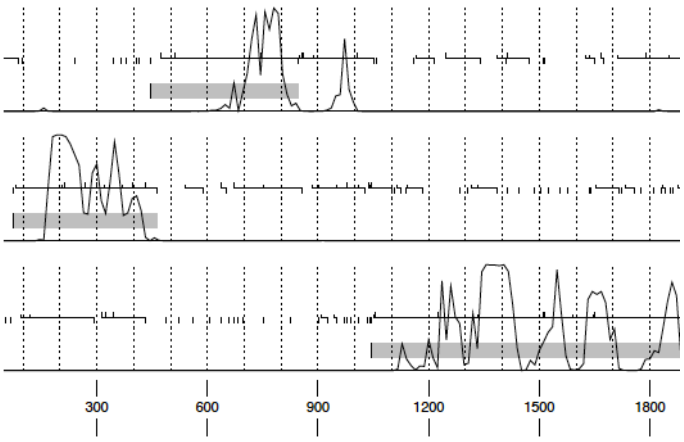


We are confident enough with the BLASTP and map assignments that it is not necessary to run HHPred.

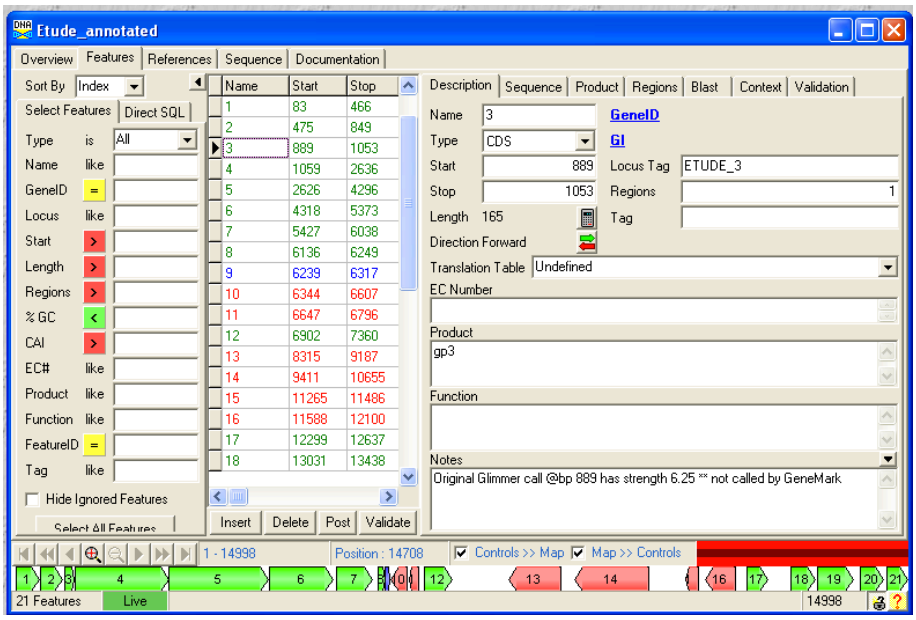
Add detailed annotation notes as in gene 1. Make sure that you include the gene gap/overlap, the functional assignment, and the source of your functional assignment. Since this is a Hatfull Map approved function, we will add it to the Function field as well:



Gene 3: Look at the coding potential trace in the GeneMark TB output. The coding potential after Gene 2 shows a smaller peak in the top tier following the gene 2 peak.

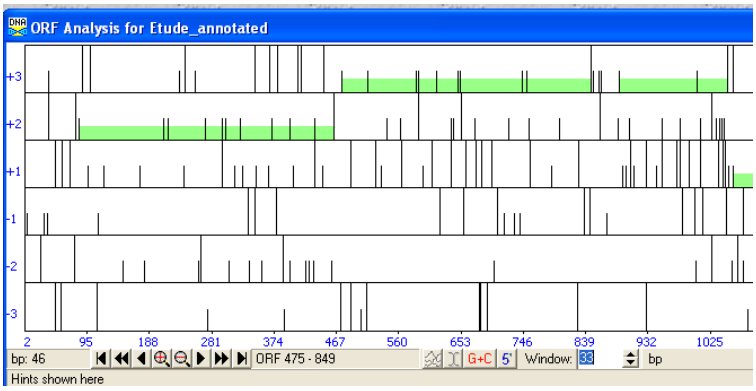


Glimmer has decided to call this ORF a gene, GeneMark has decided to omit it.



Since we see some coding potential in this frame and not in any others, and this ORF nicely fills a gap in the genome between gene 2 and gene 4, we are going to call this gene.

Now we need to pick a start codon. Examine the frames window:



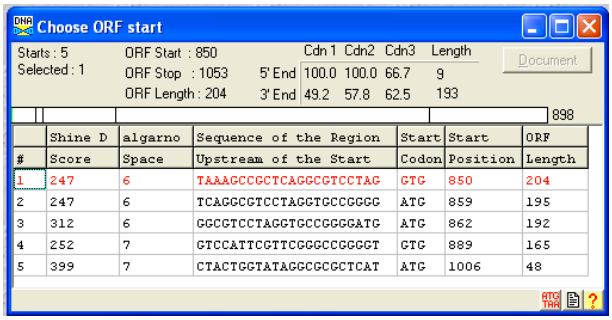
There are at least four possible starts for gene 3 that will not overlap with gene 2 (we can't overlap gene 2 at all in this case, as genes 2 and 3 are in the same frame. Gene 2's stop codon would prevent translation of gene 3 from any earlier start.)

Now we use our five pieces of data to determine which start of the four possible starts we like the best.

Coding Potential: the earliest blip in the GeneMark TB coding potential trace is about 900 bp, so all four starts encompass all the coding potential.

Gene gap/overlap: The best start with regards to gene packing is the longest start, which leaves no gap between genes. However, the tandem starts (start 2 and start 3) leave a fairly small gap as well, either 10 or 13 bp.

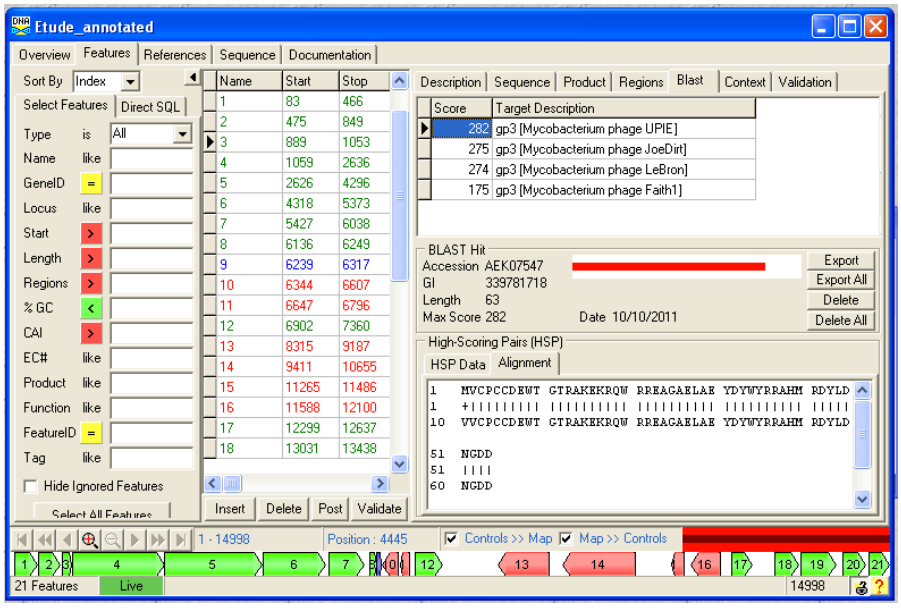
RBS scores:



According to DNA master, starts 1 and 2 have exactly the same score, which is only very slightly lower that start 4 (the Glimmer start). Start 3 has the best score of 312.

At this point, we need to weigh whether we minimize the gap (start 1 at bp 850) or pick the best RBS score (start 3 at bp 862). An interesting biological discussion!

I will check the BLAST results, too:

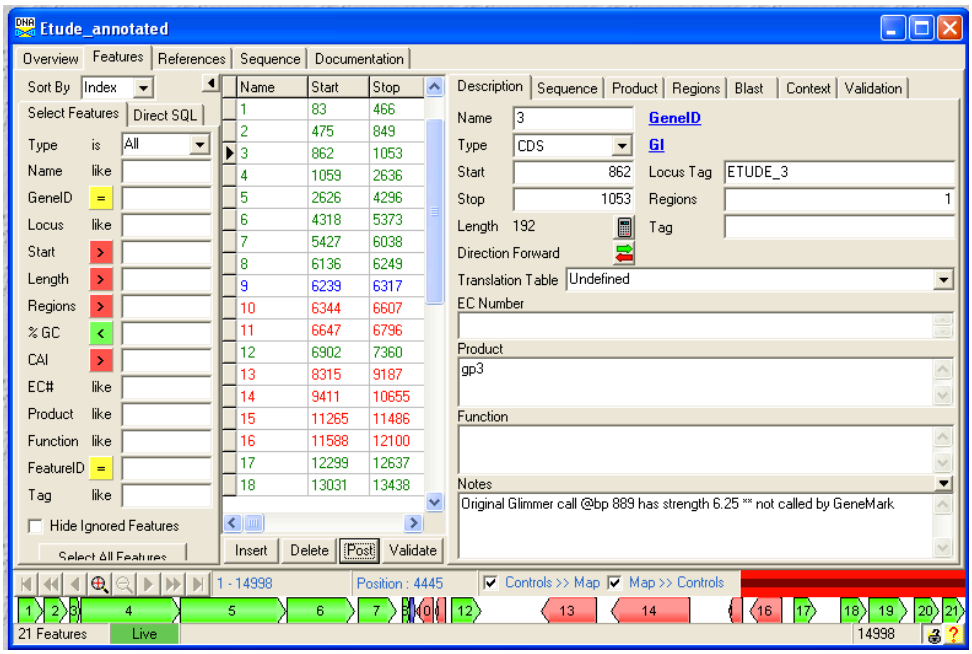


This indicates that we are 9 amino acids short of the called start in the best BLAST hit, that of UPIE. 9 amino acids is 18 bp, or equivalent to the start at 862.

Without wet lab evidence, we can't really say. So at this point, I will choose the best RBS score at (bp 862) and the BLAST alignment. I could just as easily choose to minimize the gap, as the RBS score is not too terrible for start 1 and we frequently see phage genes that start and stop practically on top of each other in this manner. Both are valid choices.

Change the gene start in the description tab:

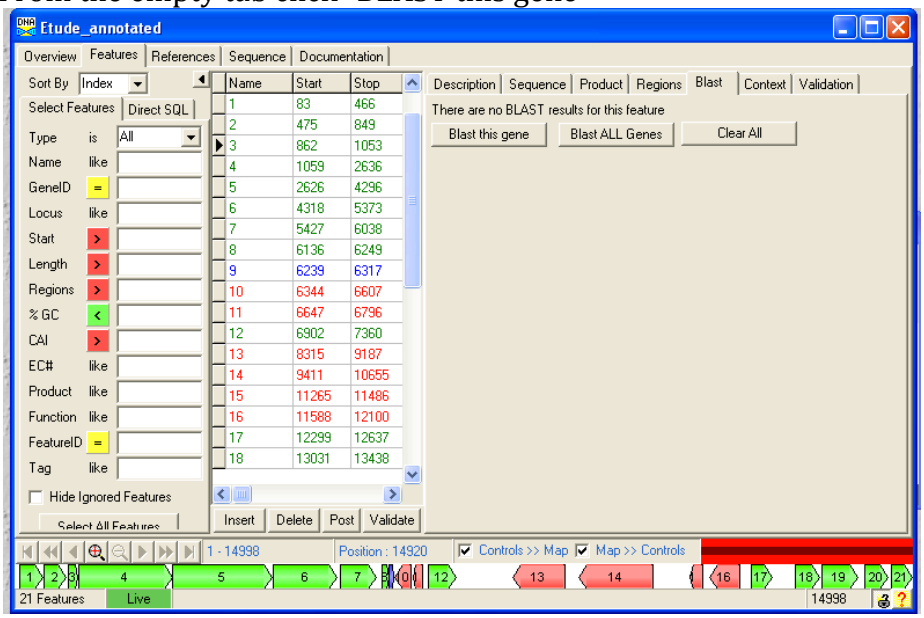
→Write the new coordinate in the start field, and then click the calculator icon (the button directly under the field labeled "Stop") to change the gene length and post the change to the database.



Now reBLAST the gene: click the BLAST tab.

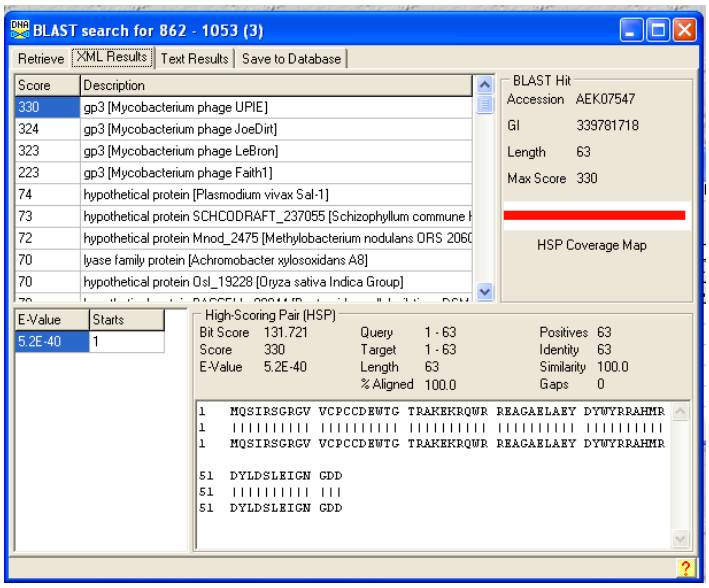
Click “Delete all”. Click yes in the box that pops up that asks you if you really want to do this. This should empty the tab of BLAST data.

From the empty tab click “BLAST this gene”

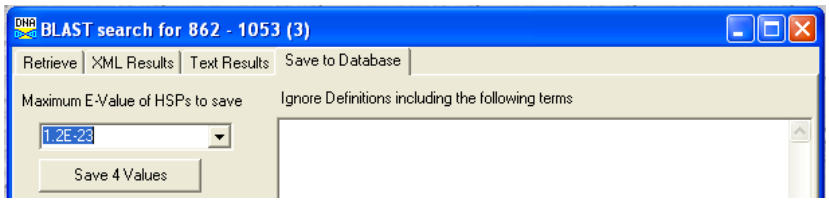


Note: If you click “Clear All” in the empty BLAST tab, you will delete the BLAST data for ALL of your genes.

A new window will appear, showing your BLAST request status. When it finishes, it will load your data and look like this:

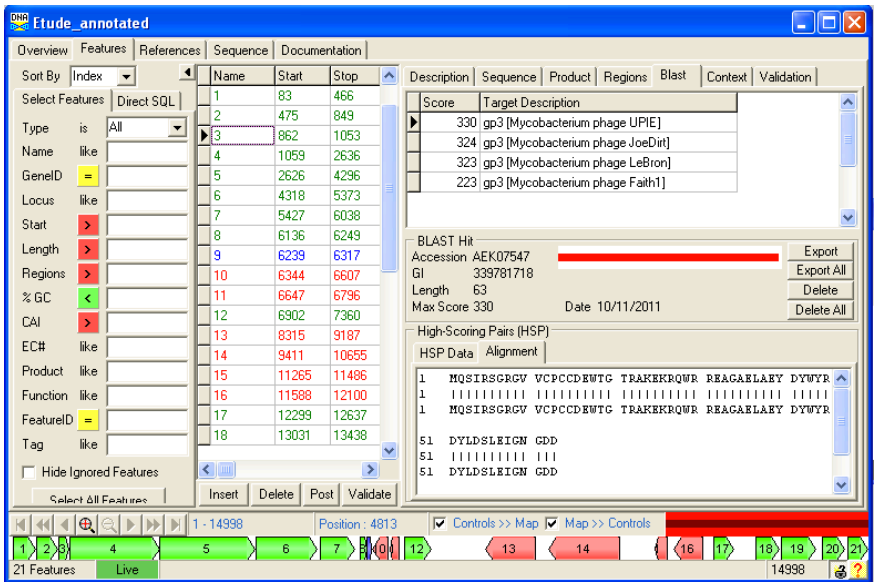


Click the tab that says “Save to Database”



Only the top four hits have reasonable E values, so we will only save those four to our genome database. I will click “save 4 values”. Then I close the window.

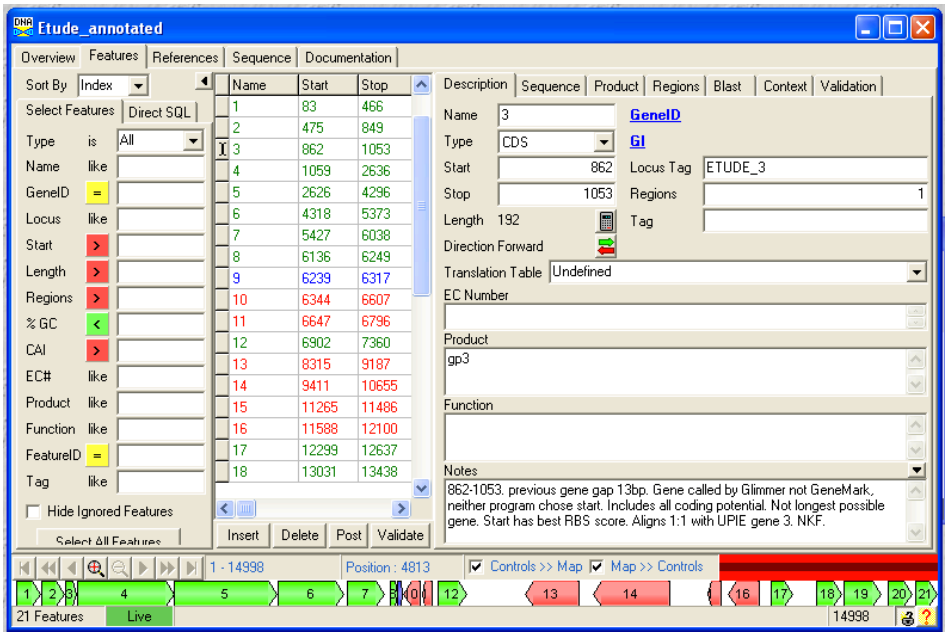
Back in my main genome window, the BLAST data has been altered for gene 3 (I had to click gene 2 and then back to gene 3 for it to load into the window):



Now when we BLAST the amino acid sequence of gene 3, it matches the UPIE annotation gene 3, with the “query 1” aligning with the “sbjct 1”. This makes me feel better about the start that I chose, for even though both starts were good choices, it is nice to be consistent with a similar genome. That way, in the future, any wet bench data that we get about this gene from one phage will be easily applied to similar genomes.

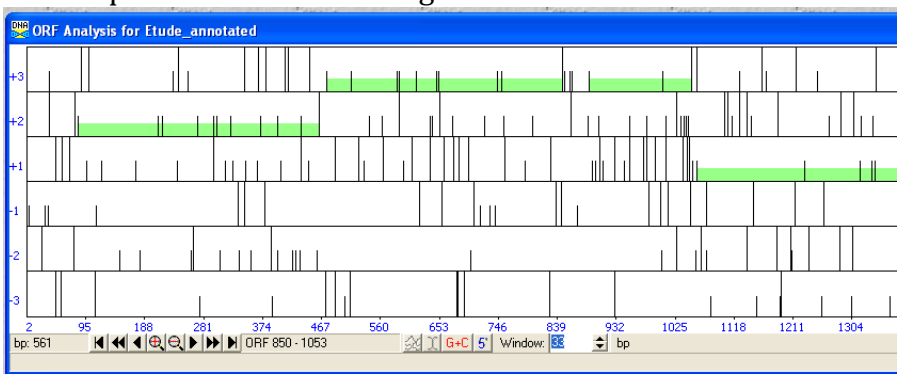
Check for functions via BLAST, Phamerator, HHPred, and the Hatfull Maps. None of these return a known or likely function (the best HHPred match is to a zinc-finger protein in Homo sapiens).

Add detailed annotation info to the notes on the Description tab:



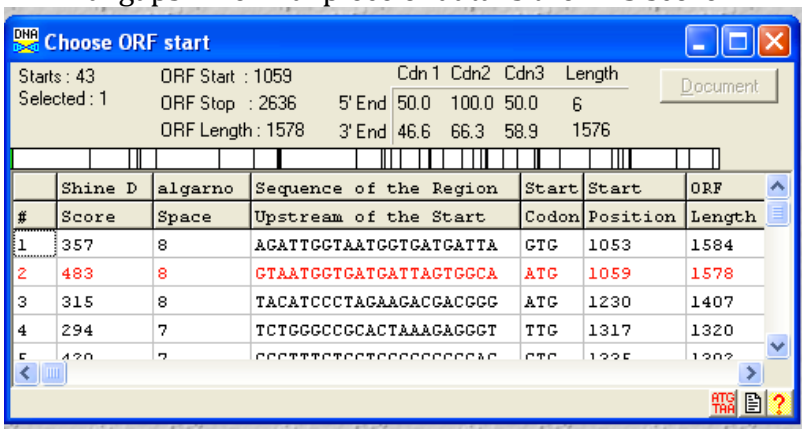
Gene 4:

Back to the coding potential trace in GeneMark TB. Gene four is in frame three. It looks like the coding potential starts around bp ~980 or 990. According to the frames window, there are two possible starts for this gene:



1053 and 1059 (the start called by both Glimmer and GeneMark).

We already know that both starts encompass all the coding potential, and that both have minimal gaps. The final piece of data is the RBS score:



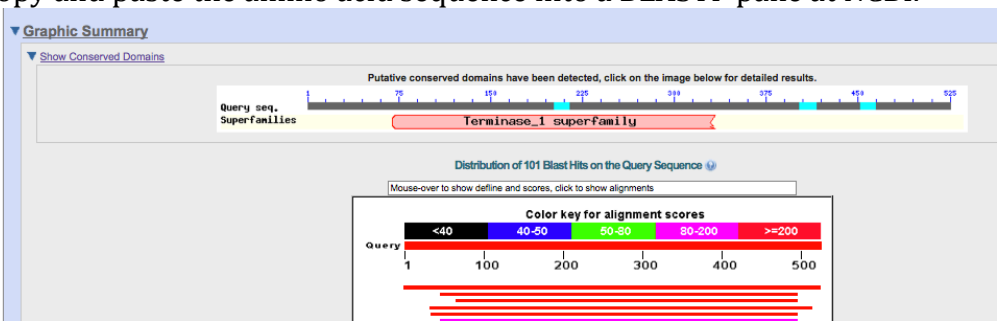
Here we come to one of those gray areas in annotation. The RBS score of the first start is lower, but not much lower. The gap between genes is smaller with the first start, but not much smaller. And both algorithms selected the second start.

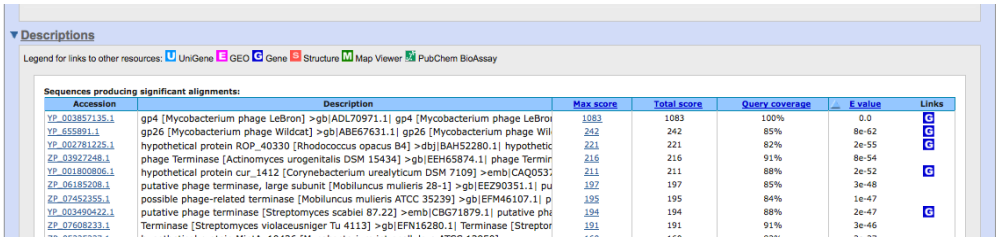
BLAST: The data from the BLAST tab indicates that the algorithms have selected the same start as the genes already in GenBank.

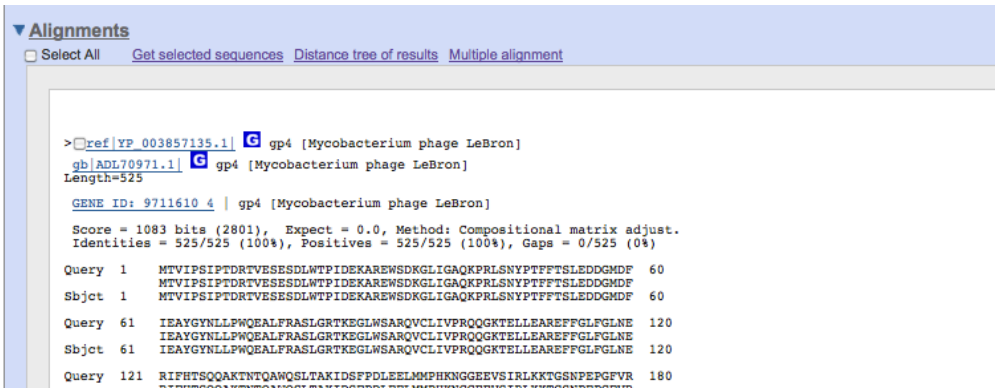
So I am going to pick the Glimmer/GeneMark start.

Functional assignment:

Copy and paste the amino acid sequence into a BLASTP pane at NCBI.





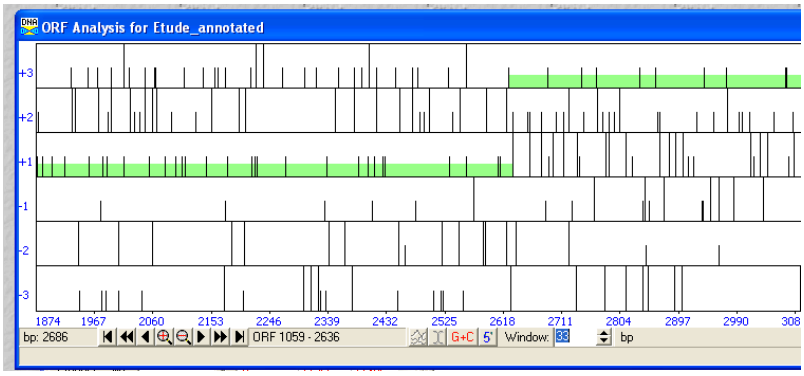


Gene 4 is the large subunit of the terminase, which is part of the DNA packaging machinery (helps to stuff the DNA into the new phage head). We got a conserved domain hit and numerous phage hits. Once again, our best match is to LeBron, and we once again align perfectly, with the Query 1 matching the Sbjct 1. The assignment is supported by the Hatfull Maps, and running HHPred is not necessary.

I will make the appropriate notes in the Notes field and Function field.

Gene 5:

Gene 5 is the easiest gene by far that we have looked at. Both Glimmer and GeneMark call this gene, the GeneMark TB coding potential starts around ~2650, and there is only one start codon (at 2626) that neither overlaps gene 4 too much and encompasses all the coding potential. In fact, there is only one start codon in the correct frame for this gene. Notice there is a small 10bp overlap between genes now. This is OK, overlaps need to be much larger before we discount them.



BLAST tab results indicate that this gene is a perfect match 1:1 with LeBron gene 4.

Done!

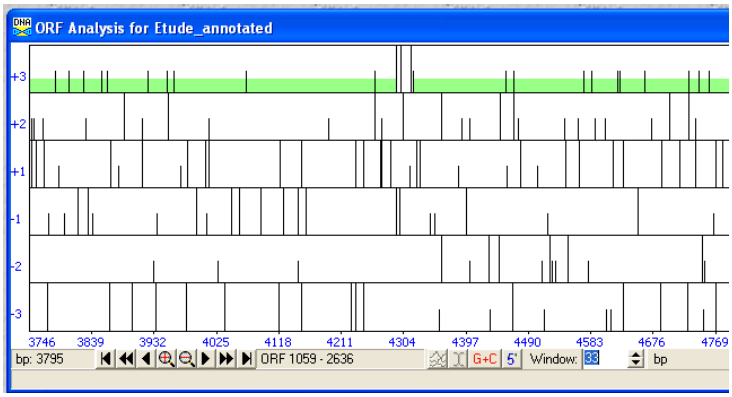
Paste the amino acid sequence into a BLAST p pane at NCBI.

This protein is a phage portal protein and forms a dodecameric ring at the vertex of the capsid that the DNA is threaded through and that the tail then joins to. We once again match the LeBron gene call perfectly. The Hatfull Maps support this assignment.

I will write the appropriate notes in the Notes field and Function field.

Gene 6:

Again, both the Glimmer and GeneMark calls agree on a single start that does not have any close starts near it in the same frame and is the longest possible start for this gene.



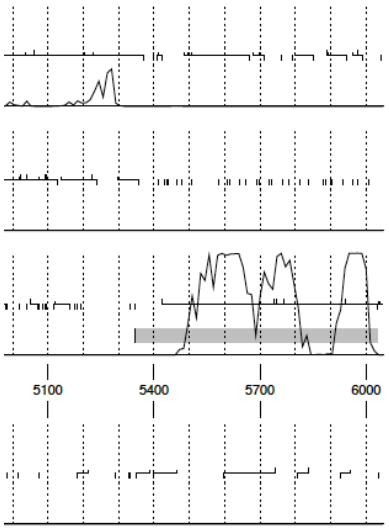
The BLAST tab data shows a 1:1 alignment with LeBron's gene 6. The Hatfull Map shows that this gene is the capsid maturation protease (frequently found after the portal gene in phage genomes). This protease cleaves the scaffolding protein in the immature capsid (also called the procapsid) and allows the phage capsid to expand to its mature size during assembly and DNA packaging.

I will write the appropriate Notes in the Notes field, and Function field.

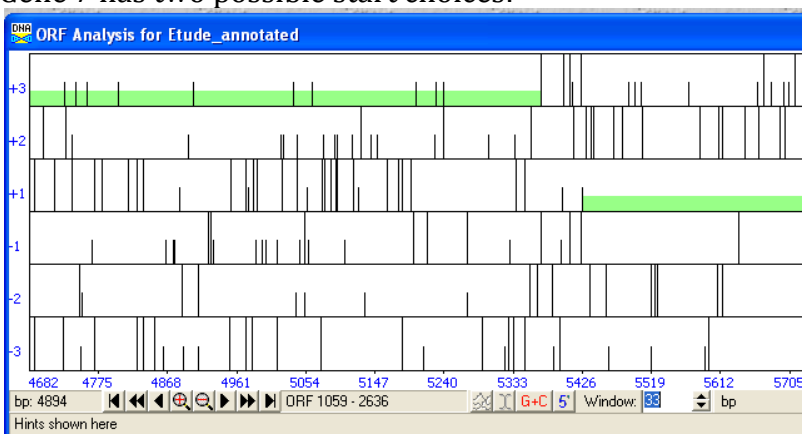
BLASTing the sequence at NCBI shows that we once again match LeBron gene 6's start exactly.

Gene 7:

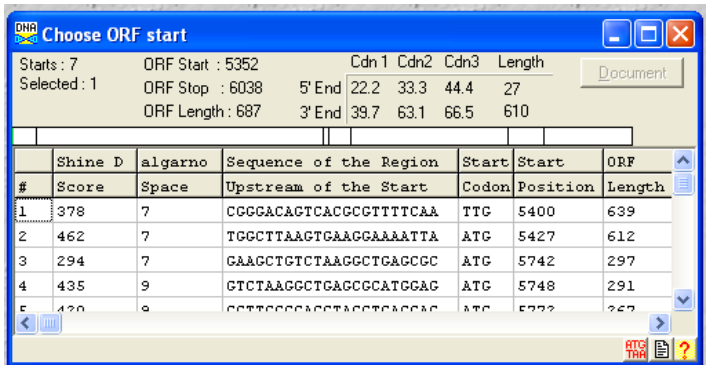
Coding potential: The coding potential for this gene begins around 5450.



Gene 7 has two possible start choices:



The start that Glimmer and GeneMark have selected at 5427 and the TTG start at 5400. Both starts include all the coding potential. GeneMark never calls TTG starts and Glimmer undercalls them, so we must take that into consideration when deciding which start to pick (not a good time to say, "well, all things being equal, we will take the algorithms' call," because TTG starts are **NOT** equal from the point of view of the programs.)



When we look at the two starts in context of gap closing and RBS scores,

the TTG start at 5400 has a score of 378 and the ATG start has a RBS score of 462. However, the size of the gap left between the stop codon of gene 6 is either 24bp or 54bp.

BLAST data: The BLAST tab shows that the longer gene start has been called for the genes already in GenBank (our alignment has a mismatch of 1:10).

Given that 24bp is already a sizable gap for a phage genome and the other GenBank phages use the longer start, we will pick the TTG start at 5400.

Change the gene start on the Description tab, and click the calculator button to recalculate gene length and to post the change to the database.

ReBLAST the gene through DNA Master to make sure that you see the correct alignment.

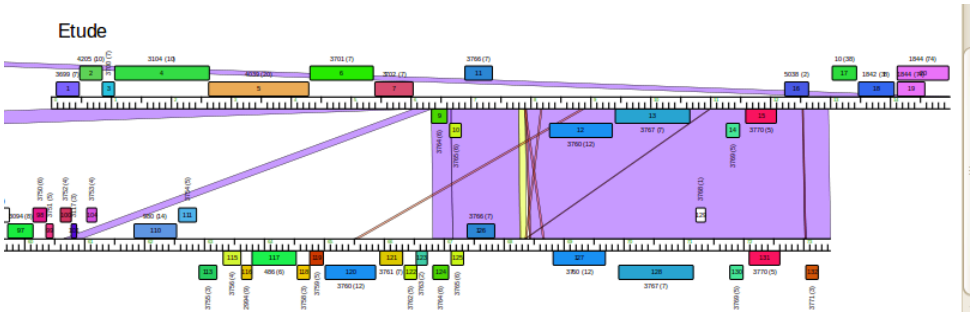
Functional assignment: the NCBI results indicated that this is a likely scaffolding protein in another phage and this is supported by synteny and by the Hatfull Maps.

Gene 8:

Gene 8 is a very small gene (38 res) that exactly matches the beginning of LeBron gene 8 in the BLAST tab alignments. Normally, I would include this gene in an annotation, even though it is very small because it shows that a gene has been truncated and therefore is a good example of genome mosaicism and recombination. Unfortunately, gene 8 also overlaps a tRNA. Generally we do not see coding sequence (CDS) and tRNA overlaps, except for possibly a few bases at the 3' end of both of them when they are transcribed in opposite directions. The positioning of the tRNA, and the truncation of gene 8 suggests that the tRNA interrupted the original gene 8 during some kind of recombination event. I will therefore delete the called gene 8 from the auto-annotation by clicking "Delete" at the bottom of the center column, and then "yes" in the confirmation box.

I am not going to renumber the genes again until I am done with the annotation.

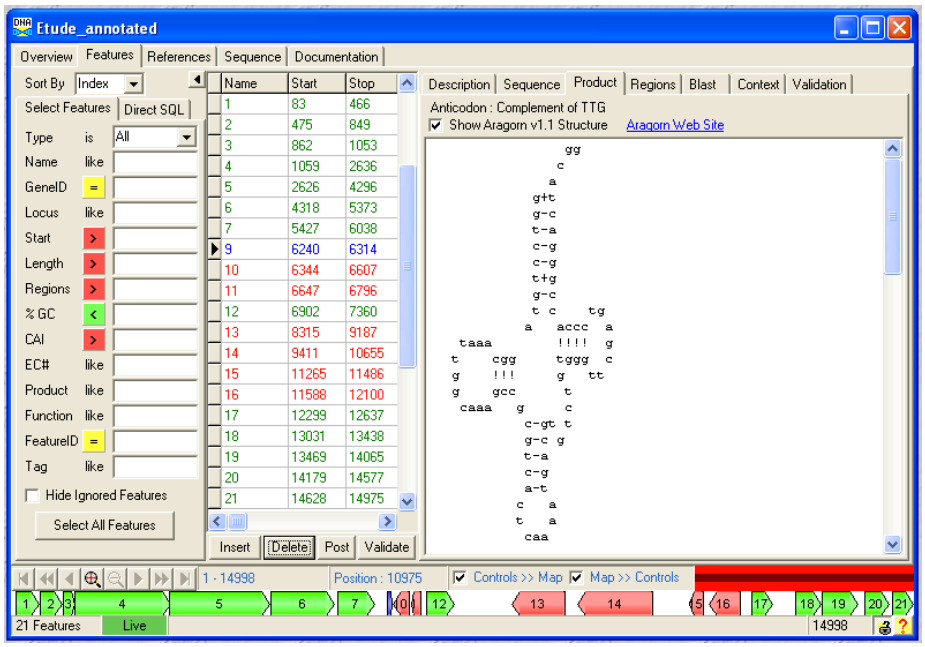
Now we know from our Phamerator alignment that Etude no longer has a high degree of similarity to LeBron after gene 8. But there are some lines in the phamerator map indicating that gene 9 of Etude is similar to something farther to the right in the LeBron genome. Slide the Etude map to the right in relation to LeBron and you will see:



LeBron is still the bottom genome, but you can't see the title any more because it is all the way over at the left end of the genome. From the map above, it looks like the next 8kbp are very similar to LeBron, with the final two kbp having no similarity.

Gene 9 (the tRNA):

Click on the product tab:



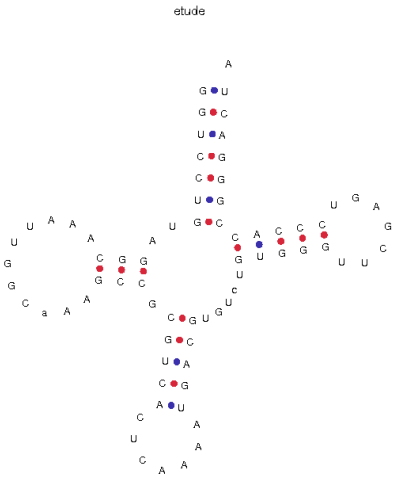
If you check the box at the top marked "Show Aragorn v1.1 Structure" the folded view of the tRNA will appear. Examine the top stem and 3' end of the tRNA: does the stem have seven base pairs? No, it has eight. Is the 3' end sequence after the stemloop NCCA? No, It is NGCT. This means that we will need to look at the tRNA outputs from the other programs and trim the tRNA appropriately.

Web-Based Aragorn:

```
etude
14998 nucleotides in sequence
Mean G+C content = 60.2%

1.

              c
             a
           g+t
           g-c
           t-a
           c-g
           c-g
           t+g
           g-c     tg
          t   cacc  a
   taaa   a    !!!!  g
  t    cgg     gtggg  c
  g    !!!     t    tt
  g    gcc     c
   caaa        t
          g+tg
          c-g
          g-c
          t-a
          c-g
          a-t
         c   a
         t   a
          caa


   tRNA-Leu(caa)
   75 bases, %GC = 56.0
   Sequence [6240,6314]


Primary sequence for tRNA-Leu(caa)
1    .   10    .   20    .   30    .   40    .   50
ggtcctgtaggcaaattggcaaagccgcgtcactcaaaatgacgtgtctg
tgggttcgagtcccaccgggactac
```

The output for Web-based Aragorn is much better: a seven-base pair top stem, and on the 3' end there is a discriminator base (the A), followed by a single C from the CCA. This follows our tRNA rule: the trimming of the CCA part of the sequence once it deviates from CCA.

The tRNAscan SE output is:



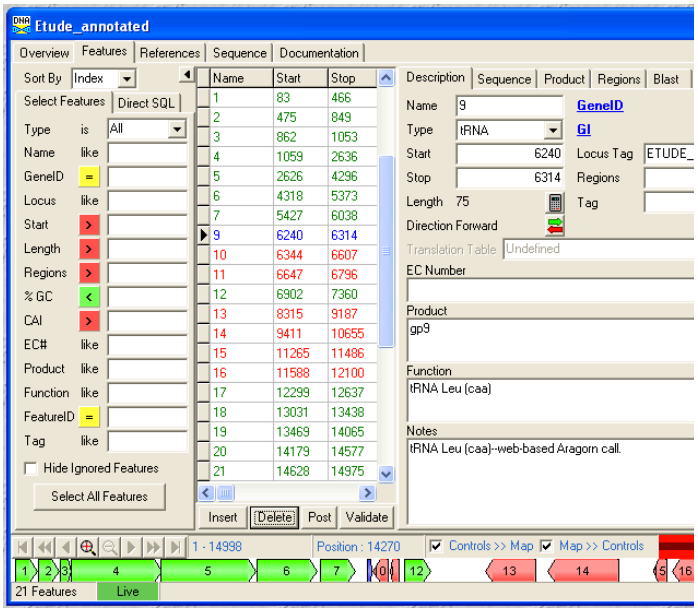
Results

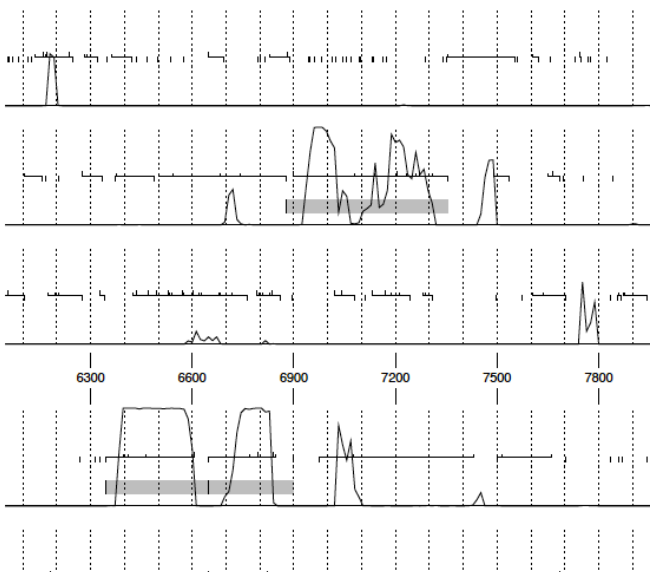
Sequence Name	tRNA #	tRNA Begin	Bounds End	tRNA Type	Anti Codon	Intron Begin	Bounds End	Cove Score
etude	1	6240	6313	Leu	CAA	0	0	58.66

View tRNA

Here the tRNA sequence begins in the same place as the web-based Aragorn output, and ends one base before it does, generating a slightly different secondary structure.

In this case, I prefer the web-based Aragorn, that has the beginning of the CCA sequence that is found at the end of canonical tRNAs. So I will adjust the tRNA coordinates to those and click the calculator button to recalculate the gene length. I will also add the tRNA amino acid and anti-codon to the function and notes fields:

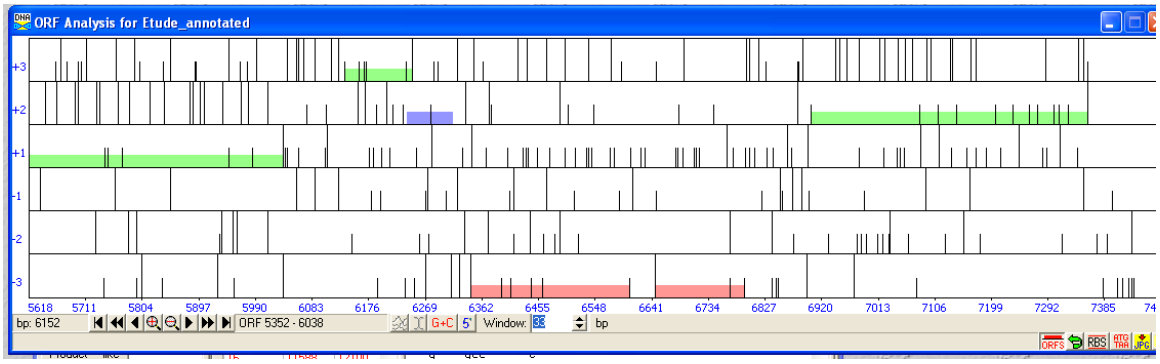




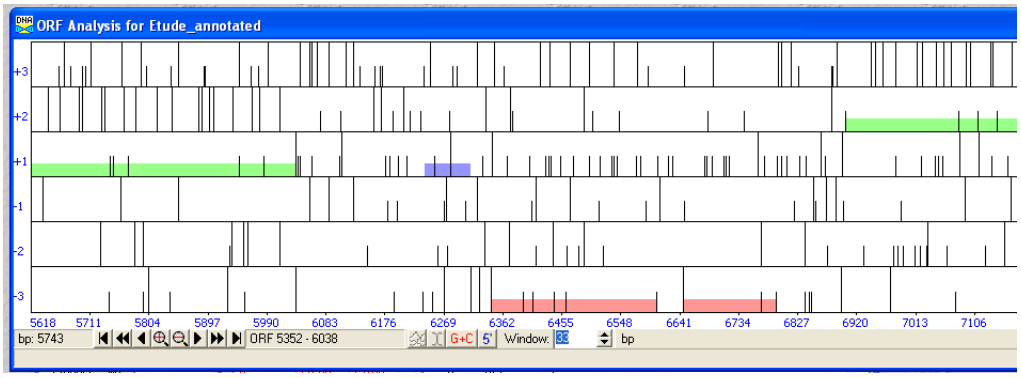
Gene 10: It seems pretty clear from GeneMark TB coding potential, from our Phamerator alignment, and from DNA Master that the next gene is a reverse gene, rather than a forward gene like all the previous ones.

There are actually three separate ORFs in the fourth tier in the GeneMark TB output, the first two corresponding to two different auto-annotated genes, with the final ORF overlapping with the forward ORF in the second frame. Both algorithms decided to call the forward ORF as a gene instead of the third reverse ORF (you can't pick both; they overlap almost completely). The coding potential of the forwards ORF is much better than the third reverse ORF—as shown by the higher, more extensive peaks, so we will also pick that gene when we get to gene 12.

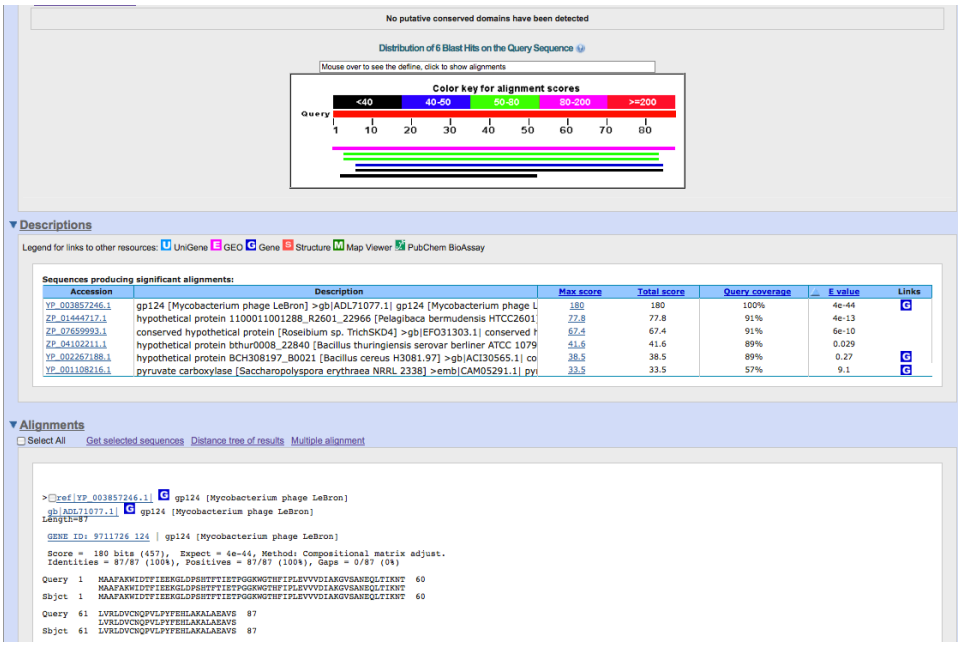
Back to gene 10: when a forward gene and reverse gene meet stop codon to stop codon (or end of tRNA to stop codon as genes 9 and 10 do), it is not necessary to leave much space between the two ends of the genes. Several bases is sufficient. However in the opposite case, when the forward and reverse genes meet start codon to start codon, it is necessary to leave at least 50-60 bases between the two starts. This is because there will be a promoter for the RNA polymerase preceding the start codon of each gene. Since gene 10's stop codon has plenty of room after the tRNA transcript ends, we don't need to worry about this here (See below). We will need to worry about it when we select the starts for both gene 10 and gene 11.



In the frames window, the deleted gene is still highlighted. Click the green "refresh" arrow at the lower right side (next to the ORFs button) to update the frames window.



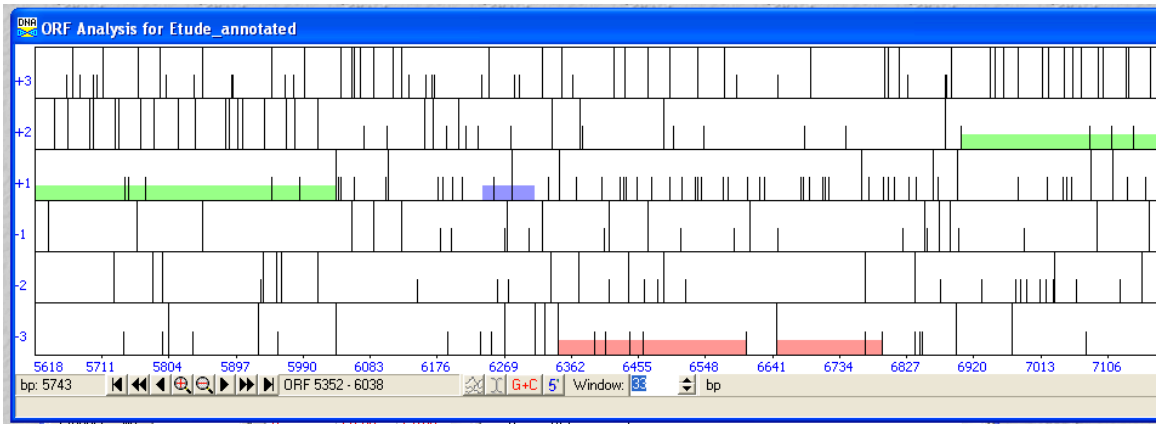
Gene 10 only has one real choice for a start codon, and it was selected by both the GeneMark and Glimmer aligorthms.

BLAST check: 

We still are similar to LeBron, only now we are similar to gene 124. We still align perfectly with the same start codon as selected in the LeBron annotation.

In the annotation notes, when you are calculating the "gap/overlap" number, you should still look at the start codon of gene 10, only now you should compare to the stop codon of gene 11, because we are going in the reverse direction. The reason why we include these gap or overlap numbers is to see how well the start we chose fills out the genome from this gene to the neighboring one. Since we can't change the stop codon, the only way to fill the genome is by changing the start codon. This number provides an extra reference as to how closely the genes are called in your annotation. LeBron gene 124 has no known function.

Genes 11 and 12: As mentioned above, gene 11 is a reverse gene while gene 12 is a forward gene. As they will be "head to head" (so to speak), we need to take care in choosing the starts for each of them that we leave at least 50-60 bases between the two genes.



Currently, if we accept the two GeneMark calls for the genes, we will just barely squeak by with our minimum of 60 bases (6841 to 6902). Notice, however, in the above figure, that it

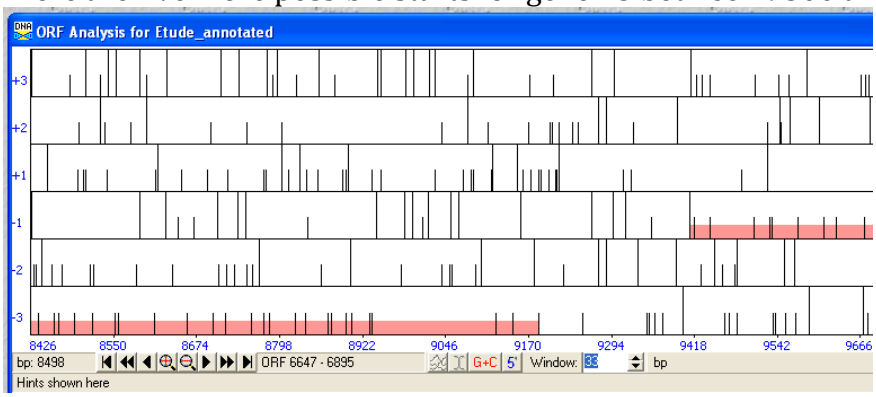
is not possible to extend gene 12 to start any closer to gene 11 (there aren't any more start codons in the same frame any closer to gene 11 than the one already called). On the other hand, Gene 11 has four possible start codons, including the one used in the Glimmer call, which is way back at position 6796. From looking at the GeneMark TB coding potential trace, it is pretty clear that the Glimmer start does not encompass all the coding potential, so we will eliminate this choice as a possible start. We do still have three more starts to check: the one selected by GeneMark and the two immediately after it. If we check the RBS scores for all three starts; we find that the GeneMark start also has the highest RBS score. So we will accept the GeneMark call for gene 11. Gene 12 really only has one possiblility for a start, and it is the one called by both programs above.

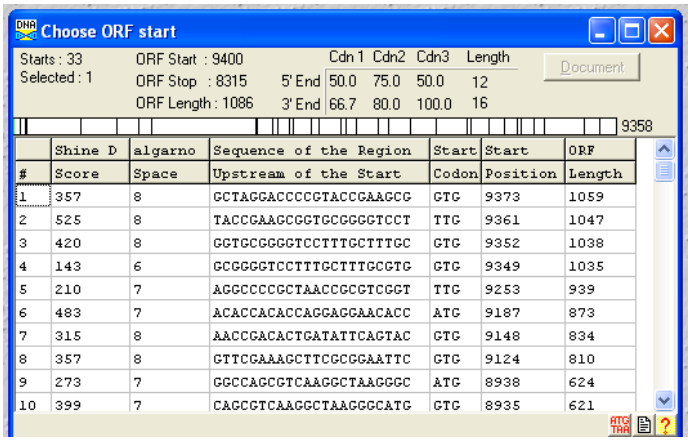
BLAST check: gene 11 aligns with LeBron gene 125 query 1 to sbjct 1. Gene 12 aligns with LeBron gene 126 query 22 to sbjct 22 (which still counts as picking the same start, just the beginnings of the genes are not as similar as the later portions.)

Gene 13: While there are some blips in the GeneMark TB coding potential, none of them align well within an open reading frame (if the peaks did fit better into an ORF I would be likely to include them as genes). So we will leave a fairly large gap between gene 12 and gene 13. This is also what the algorithms suggest.

Both the Glimmer and GeneMark calls suggest that gene 13 begins at postion 9187, but this start leaves a huge gap between gene 14 and 13 (gene 13 is reverse, so we compare its start to the upstream gene). The BLAST alignment suggests that this gene is same length as LeBron (1:1), but shorter than (UPIE 1:59).

There are five more possible starts for gene 13 between 9300 and 9400.



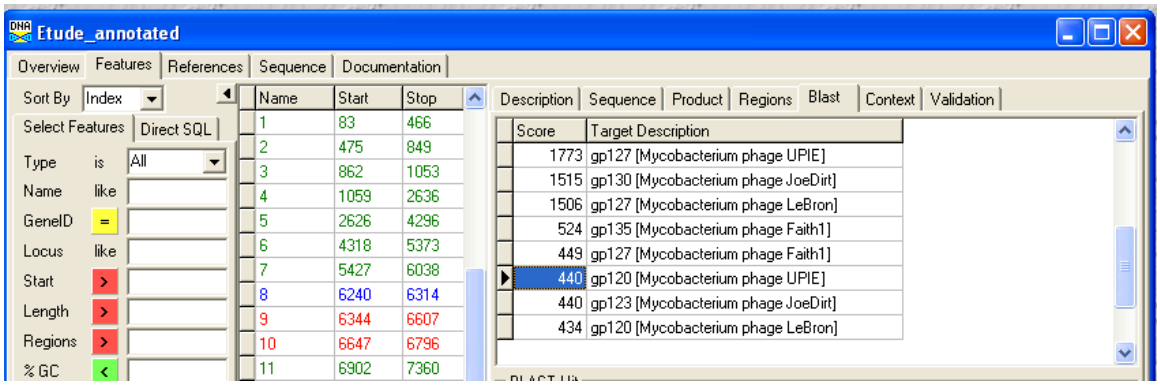


#	Shine D Score	algarno Space	Sequence of the Region Upstream of the Start	Start Codon	Start Position	ORF Length
1	357	8	GCTAGGACCCCGTACCGAAGCG	GTG	9373	1059
2	525	8	TACCGAAGCGGTGCGGGGTCCT	TTG	9361	1047
3	420	8	GGTGCGGGGTCCTTTGCTTTGC	GTG	9352	1038
4	143	6	GCGGGGTCCTTTGCTTTGCGTG	GTG	9349	1035
5	210	7	AGGCCCCGCTAACCGCGTCGGT	TTG	9253	939
6	483	7	ACACCACACCAGGAGGAACACC	ATG	9187	873
7	315	8	AACCGACACTGATATTCAGTAC	GTG	9148	834
8	357	8	GTTCGAAAGCTTCGCGGAATTC	GTG	9124	810
9	273	7	GGCCAGCGTCAAGGCTAAGGGC	ATG	8938	624
10	399	7	CAGCGTCAAGGCTAAGGGCATG	GTG	8935	621

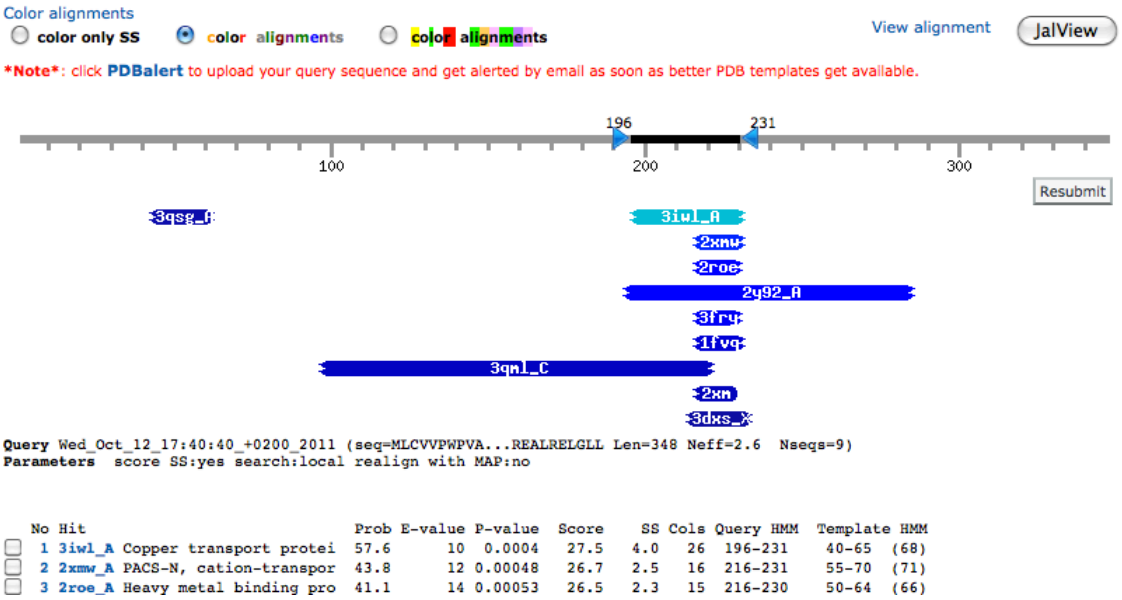
The best RBS score goes to the TTG at 9361. We will pick this start.

BLAST check: while we now match UPIE, we are substantially longer than LeBron,(Query 59 aligns with Sbjct 1)—which is interesting. The two phages seem very similar according to phamerator (all that purple between the genomes), so why wouldn't the LeBron annotators have chosen the longer start that we did? While it is not necessary to match the starts for all genes 1 to 1 with a similar phage in genbank, I was still puzzled. To solve this, I actually loaded the LeBron sequence into the web-based GeneMark TB coding potential site, and examined the two outputs side by side. It turns out that LeBron has a point mutation in this area which causes an extra stop codon in this frame, and the start that we chose for Etude is not a possibility in LeBron. The point mutation is not a large enough sequence difference to show up as another color in phamerator. I am happy to proceed with our start selection.

(Revision note: In the year since I first wrote this case-study, we have annotated and submitted many more Cluster L phages. While the BLAST data a year gave me only two choices to pick from with regards to this start, the current BLAST data from 10/10/11 seems to suggest that the shorter start is more likely. The current BLAST data also reveals something else:



Something that aligns with this gene appears multiple times in the same phage: twice each in UPIE, JoeDirt, LeBron, and Faith1. This is the sign of either a recent recombination and duplication event or of a parasitic element of DNA in a phage genome. There are several types of parasitic elements; the most common being a HNH homing endonuclease encoded by an intein or intron. The easiest way to positively ID an HNH endonuclease is through HHPred. In the case of our gene above, HHPred does not indicate the presence of the HNH domain, and therefore this is more likely the sign of a recombination-duplication event.



Gene 14: Is another reverse gene, this time in the fifth tier in the GeneMark TB coding potential view. We will ignore the peak in the forward third tier because we can't call both of them and the reverse one is larger and lovelier. There is only one possible start codon for gene 14.

BLAST Check: we match LeBron gene 128, with the start codons aligning perfectly.

Gene 15: while there is a teensy little peak in the forward direction in the second tier GeneMark TB coding potential trace nicely centered in an ORF, I am more inclined to omit it for simply being too small. It is also important to get a feel for you phage genome—are all the genes very tightly packed? Do multiple genes have very small blips of coding potential? You need to think about these things when you are making your decision. There is also a trend in most phage genomes for clusters of genes to be transcribed in the same direction, and Etude is no exception. And since we are going from a reverse gene 14 to another reverse gene --supported by the algorithms calls-- we will skip the teeny blip in the second tier.

However, the best way to really be sure would be to add this gene into your annotation, and BLAST the protein sequence to see if there are any similar genes in GenBank.

Gene 16 is the larger peak in the fourth tier after the smaller peak in the fifth tier, then gene 17 is forward again in the second tier(the peak in the top tier does not align well with an ORF).

Back to Gene 15:

There are really only two choices for a start for gene 15; either the GeneMark call or the Glimmer call. The GeneMark call has a higher RBS score so we will pick the GeneMark call—both for RBS score and for being a longer gene.

When we do the BLAST check, we match LeBron gene 130 perfectly at the start.

At this point, it is worth revisiting whether or not we want to include the little forward ORF with the blip of coding potential to see if it aligns with LeBron gene 129. Our gene 15 matches LeBron 128, and our gene 16 matches LeBron 130. Our phamerator alignment indicates that there is still the highest level of sequence similarity between these genomes in this region, but again, a point mutation resulting in the loss of a start or stop codon would not be enough to change the nucleotide identity color in phamerator. And again, it is not necessary to make the genome annotations match, but it is worth looking at the data. When in doubt, BLAST the potential gene to see if there are any matches in GenBank. I will leave it as a exercise to see if LeBron gene 129 exists intact in Etude, and if so, should it be included in the annotation.

Gene 16:

Gene 15 has a stop codon almost immediately to the right the start codon called by Glimmer and Genemark (remember, we are in the reverse direction, so the gene is transcribed right to left). This means that the start codon selected by the two algorithms is already the longest possible start that can be selected for this gene.

The RBS score is nice and high, and the start encompasses all the GeneMark TB coding potential. I am happy to accept the call as is.

BLAST check: we match gene 131 from LeBron with a perfect start codon alignment.

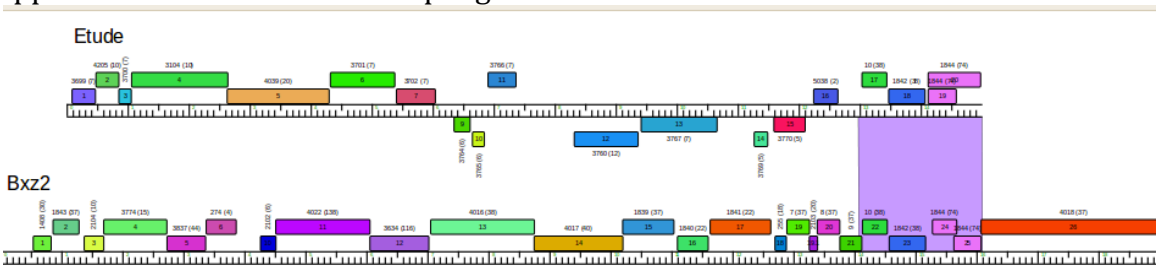
Gene 17: Gene 17 is the peak shown in the fourth tier of the GeneMark TB coding potential. Since we are once again switching from reverse to forward, we need to make sure that we leave at least 60 bases between the two gene calls.

There are three possible starts for gene 17 before a stop codon appears in the frame, including the Glimmer start:

All three start encompass all the coding potential shown in the GeneMark TB trace; however, the RBS score is best for the second of the three starts; plus this gene call fills our genome gap a bit better without getting in the way of promoter sites. We will pick the second start, at 12242.

BLAST check: this gene aligns with UPIE, but not with LeBron. Even though phamerator shows that this sequence is still similar to LeBron, and the LeBron annotation does not call this gene, we are going to rely on our own data and Glimmer calls. It is possible that LeBron has another point mutation in this region, eliminating this ORF. We can check on the LeBron genome's coding potential in the GeneMark program again, if we want to.

Genes 18-21 The remainder of the genes are no longer similar to LeBron, but instead appear to match the A3 cluster phages.



Gene 18: There is a fairly big (400bp) gap between the end of gene 17 and the beginning of the Glimmer/GeneMark calls for gene 18. While unusual in phage genomes, this is OK. There is no coding potential blips anywhere in the GeneMark TB coding potential traces between the two genes, and while there is an earlier alternate start codon at 12827, its RBS score is low, while the RBS score for the Glimmer/GeneMark call is very high! So we will pick the Glimmer/GeneMark call.

BLAST check: when we BLAST the amino acid sequence using BLASTp, we find that we align perfectly with Bxz2 gene 22, with the Query 1 aligning with Sbjct 1.

Gene 19: We know from the GeneMark TB coding trace that gene 18 and gene 19 are in the same frame, so the stop codon of 18 precludes any start codon for gene 19 earlier in the genome. The start codon selected by Glimmer and GeneMark is already the longest possible gene call for the gene, encompasses all the coding potential, and has a RBS good score. We will accept the algorithms' calls.

BLAST check:

We match the Bxz2 gene 23 perfectly, with a Query 1 aligning to Sbjct 1.

Functional assignment: This is the major tail subunit of the phage.

Gene 20: There are two possible starts to Gene 20, the one called by Glimmer and GeneMark, and the one upstream at 14116 (another TTG). If you check the coding potential GeneMark TB output, you will see that the Glimmer and GeneMark calls do not encompass all the coding potential. When we check the RBS scores, we get a higher score for the TTG start than the Glimmer/GeneMark start. We will pick the 14116 start.

BLAST check: This gene matches Microwolf gene 25 perfectly (aligns Query 1 to Sbjct 1).

Functional assignment: this is the first of the two tail assembly chaperones (the equivalent of G in phage lambda).

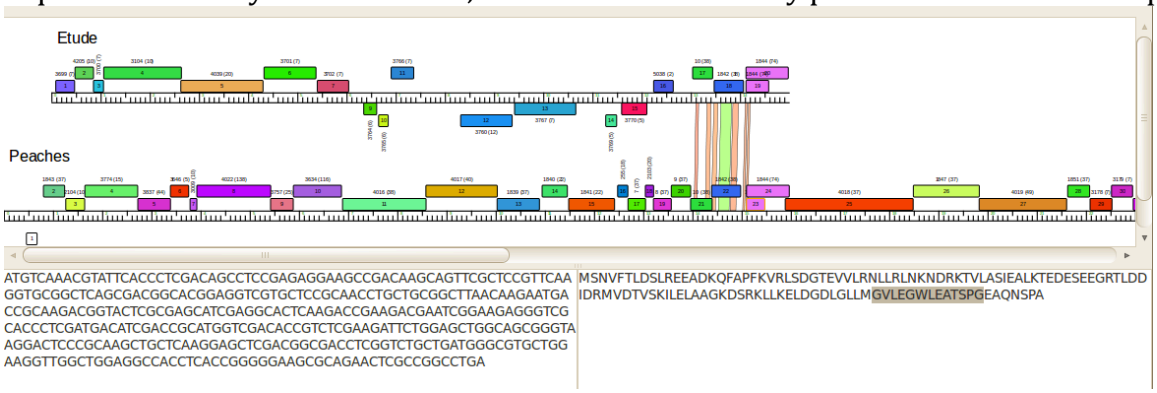
Gene 21: The GeneMark call for this gene overlaps with the end of gene 20, while the Glimmer call leaves a large gap. The GeneMark call encompasses all the coding potential while the Glimmer call does not. This is quite a large overlap.

BLAST check: This gene matches Bxz2 gene 24 perfectly, but begins in the middle of the equivalent gene for Peaches and Eagle. This is because of its function: this is the second of the two tail assembly chaperones, and actually begins at 14116 and then frameshifts into the remainder of this gene, creating the G-T fusion.

Note about this function:

Bxz2 gene 23 is a tail assembly chaperone, (gene "G" in phage lambda), and with its partner gene 24 (gene "T" in lambda) makes a fusion protein that helps assemble the tail of the phage correctly. Both the first protein, the "G" like protein, and the fusion "G-T" like protein are produced, however, the second gene product, the "T"-like protein, is not made on its own but only as part of the fusion. In the flexible tailed phages, the tail assembly chaperones frequently precede the tapemeasure gene (generally the longest gene in the genome), and are characterized by a "slippery sequence" that allows the ribosome to shift translational frame during protein synthesis. The ribosome will "slip" back a base, causing a -1 frameshift. Another way to think about it is that the ribosome reads the same base twice. The slippery sequence is generally rich in As but can begin with Gs as well. There are numerous examples of the G-T frameshift in phamerator (look for any phage genome that has two genes that begin at the same start codon, with one of them being longer than the other and followed by the longest gene in the genome). Notice I put it in the Etude annotation in phamerator already. In the phages that we have studied, the G-T slippery sequence almost always occurrs at the end of the G gene (in our case Etude 20).

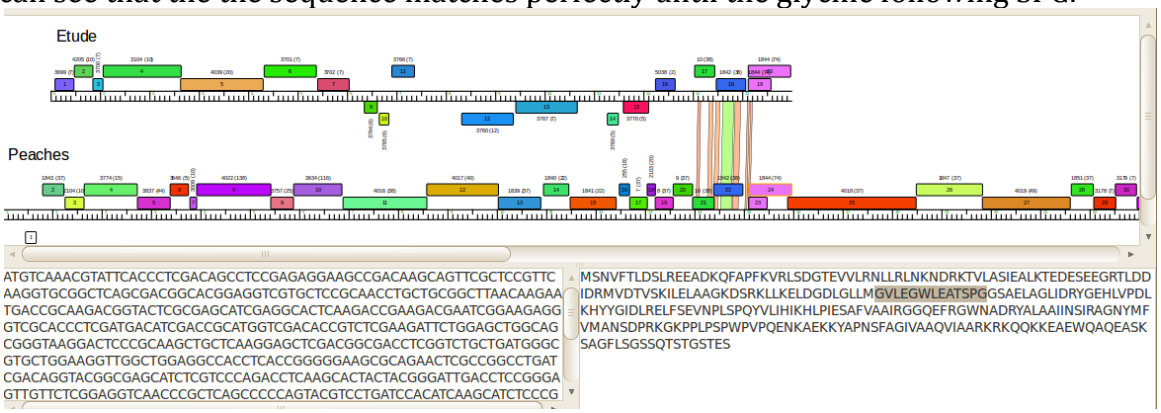
To find the coordinate of the frameshift, the easiest way is to find a closely related phage that already has the frameshift correctly annotated. If we look in our BLAST hits, you will see that phage Peaches is similar to Etude, and has the frameshift already correctly annotated in phamerator. While Peaches and Etude do not have a ton of nucleotide sequence similarity between them, notice the tail assembly proteins are in the same pham.



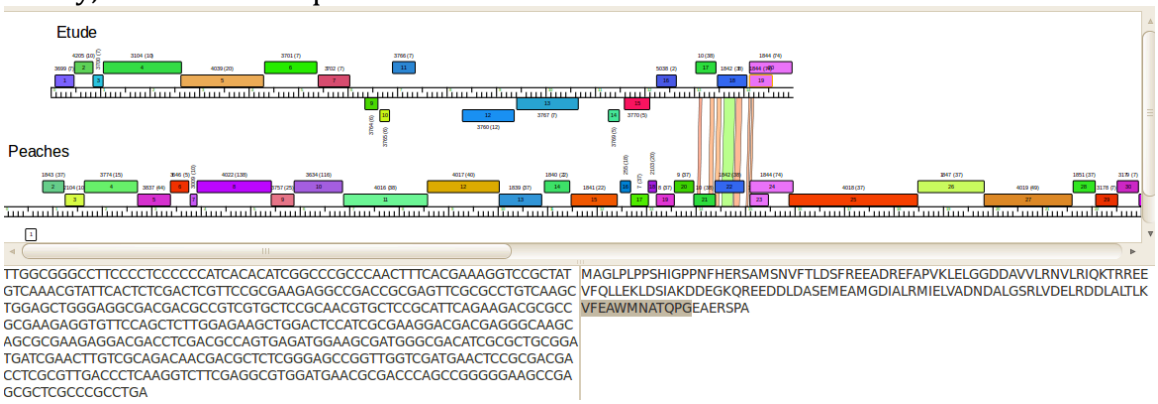
In the picture above, I have clicked on the first, shorter of the two purple Peaches genes (the G equivalent), and then highlighted the C-terminal portion of the protein sequence where frameshifting is likely to occurr.

In this next picture, I have clicked on the longer, correctly annotated frameshifted G-T fusion protein, and highlighted the same amino acids as in the shorter previous gene. We

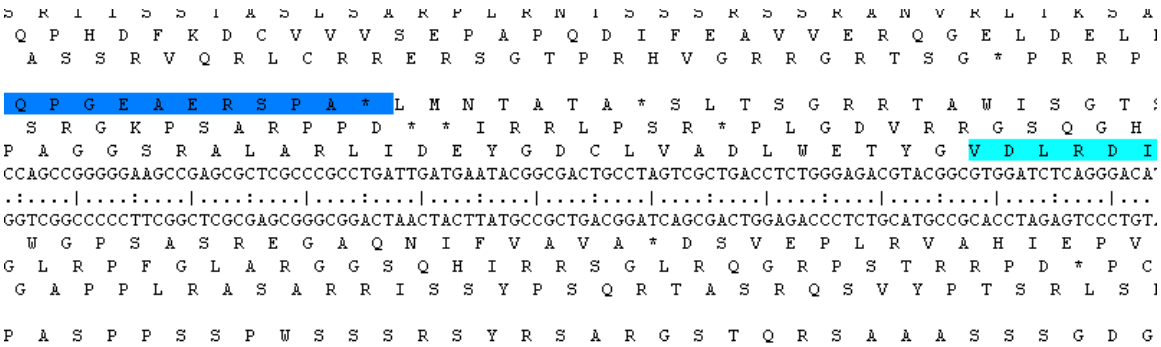
can see that the the sequence matches perfectly until the glycine following SPG.



Finally, I look for the equiavlent area in Etude:



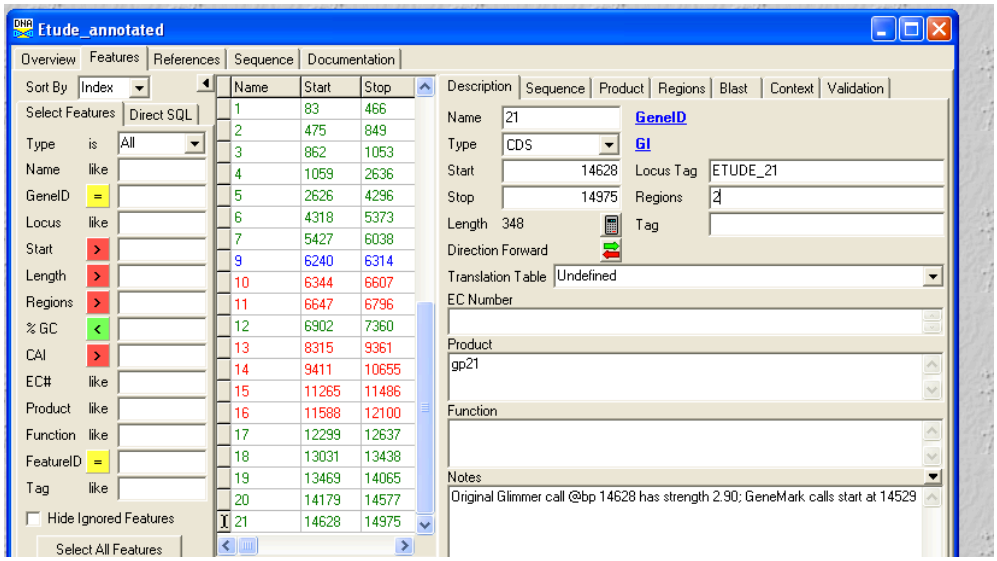
Again, above I have clicked on the first of two tail assembly genes, and then looked at the C-terminal end where frameshifting is likely to occurr. Now, we go back into our six-frame translation of the sequence.



The dark blue bar above above is from gp 20, the first tail assembly protein; while the purple start codon is from the gp 21 call. You can also see here that gene 20 is in the first translational frame while gene 21 is in the third translational frame. (a -1 shift). Now we look for our Peaches-like sequence towards the C-terminal end of gene 20—and it is seven bp from the right-hand end of the picture. The nucleotide sequence for those amino acids is GGGGGAA—a slippery sequence if ever there was one. Now to go from the top frame, gene 20; into the bottom frame to gene 21; we pick a nucleotide that will be included twice: once as the third base in a triplet codon for gene 19; and once as the first base in a triplet codon

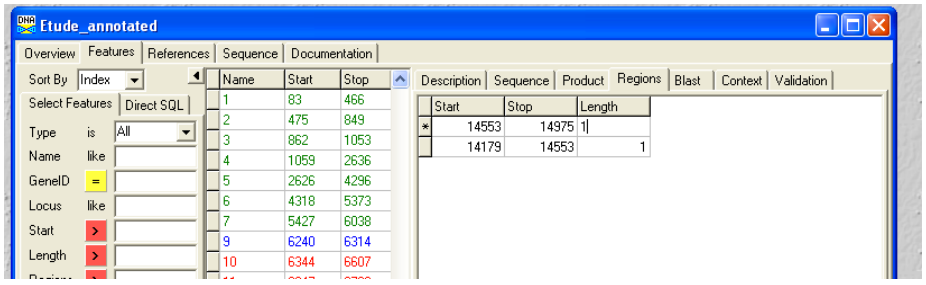
for gene 20. If we use the Peaches annotation as a guide, the amino acid sequence in the fusion protein should go from P-G-E-A to P-G-G-S. This means that the nucleotide that gets counted twice is the third G in the glycine codon of PGEA (and the first G of the second glycine codon in PGGS). This is nucleotide number 14553.

In DNA Master, on the description tab for gene 21, change the number in the regions field from 1 to 2:

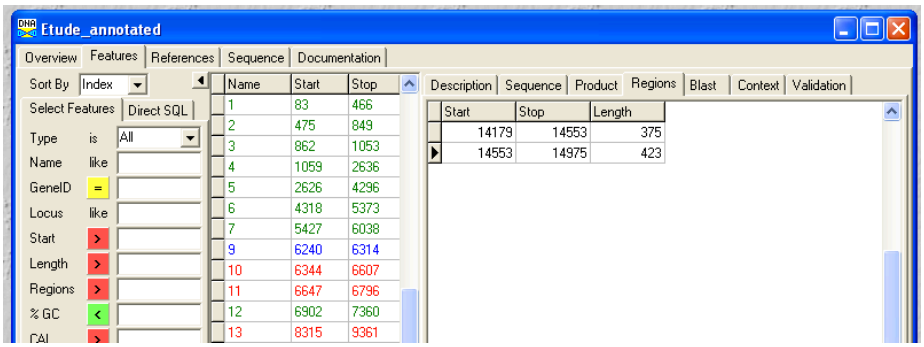


Now click the Regions tab:

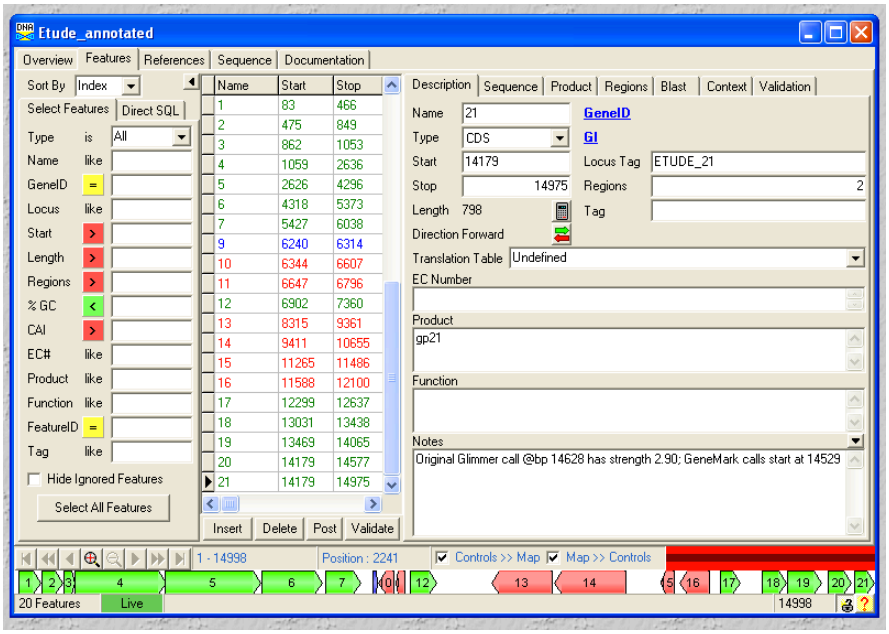
Enter in the coordinates for the upstream regions, followed by any number you like in the lengths field. Click “Insert” at the bottom of the regions tab, and then enter the coordinates of the second region in the fields that appear (they will first appear on top of the first region, but will later be correctly reordered):



click “assign lengths”

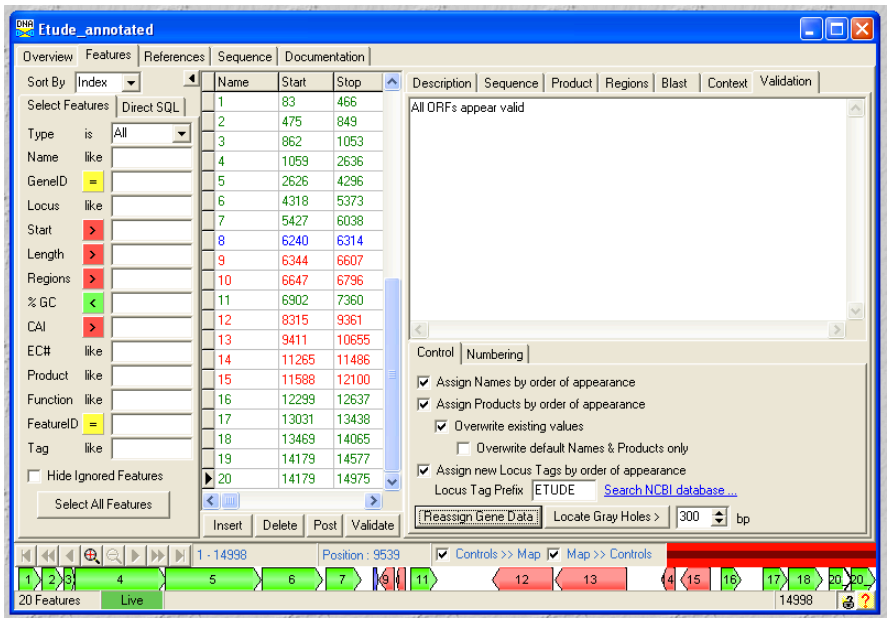


The correct numbers will be calculated. Now return to the description tab, and adjust the start coordinate accordingly.



Enter your gene Notes.

Finally, validate and renumber all your genes.



If you have not been reBLASTing all your genes, now is a good time to delete all the BLAST hits from your file and do a new complete genome BLAST. Then start your QC.

-Review your notes (correct format? All the information?)
-check those gene gaps and overlaps one last time. Did you miss any?
Finally save your final file (yourphagename_final.dnam5 is a good name), and –if this was your real genome--send it off to Pitt for review.