

What's New for School Year 2013-14 in Phage Genome Annotation

Created by DJS December, 2013.

The purpose of this document is to target the most prominent changes and/or updates to DNA Master Annotation Guide. Happy Annotating!

Change in final file Format (Section 12.1)

This year, the final file format requested has changed (again). A final .dnam5 file is one that has the following properties (Figure 1).

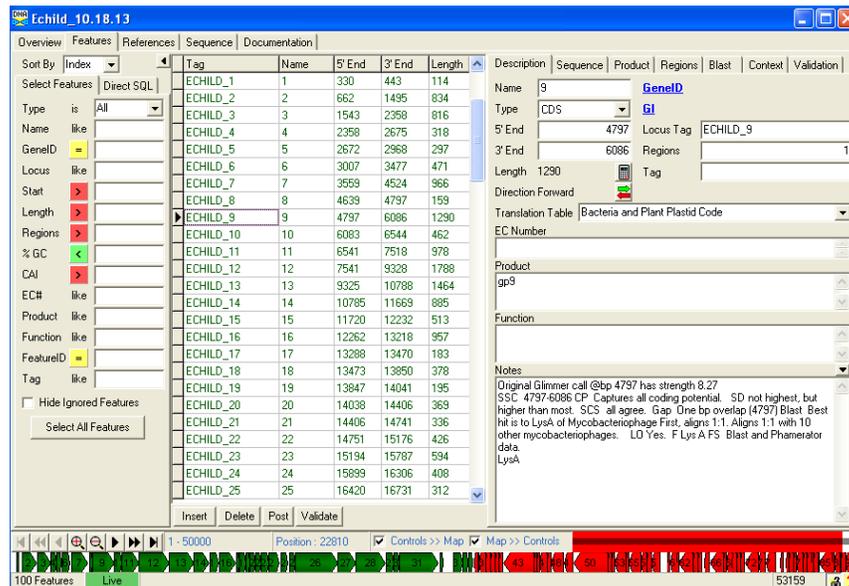


Figure 1

1. It must be named "YourPhageName_Final.dnam5", which will help distinguish it from other versions you may have been working on.
2. **It must contain one entry and set of notes per feature.** That means that if you have merged multiple files, you need to have evaluated the data from each source, come to a decision, and deleted erroneous or repetitive versions of each feature. The notes for each feature should contain **everything** listed in **Section 9.6** about proper documentation of your gene calls. You may have to delete some notes, or even rewrite some notes from scratch to meet this criterion.

3. All features must be validated (**Section 9.3.2**).
4. All features must be re-numbered if necessary (**Section 9.3.3**).
5. All features must be re-BLASTed (**Section 9.3.4**).
6. Any functions are noted in the Notes fields, along with their source (**Section 9.3.3**).
If it is determined that a particular gene has no function, include NKF (no known function) in the notes along with the sources for that determination.

Note: Our request for a different file format is because your final submission file is used to create a final version in Phamerator AND a GenBank file. These formats have differences. The file you submit is modified for both.

Local BLASTp

There are lots of avenues to explore to determine the functions of the genes (blastp at NCBI, Phamerator, publications, HHPred to name a few). Did you know that phagesDB has a BLASTp function? You can BLASTp any protein of your choice against a protein database derived from Phamerator. The protein data at PhagesDB contains the most curated Hatfull lab data. The newest entries tend to have the most complete data. Go to the Home Page of PhagesDB and click the BLAST tab in the top Banner and choose BLASTp (Figure 2). Happy Blasting!



Figure 2

Widen Feature Table

The default for the Feature Table includes Name, 5' End and Length. Right Click on 'Name' and choose "Widen Feature Table" (See Figure 3). In this view, the Feature table includes Tag, Name, 5'End, 3'End, and Length.

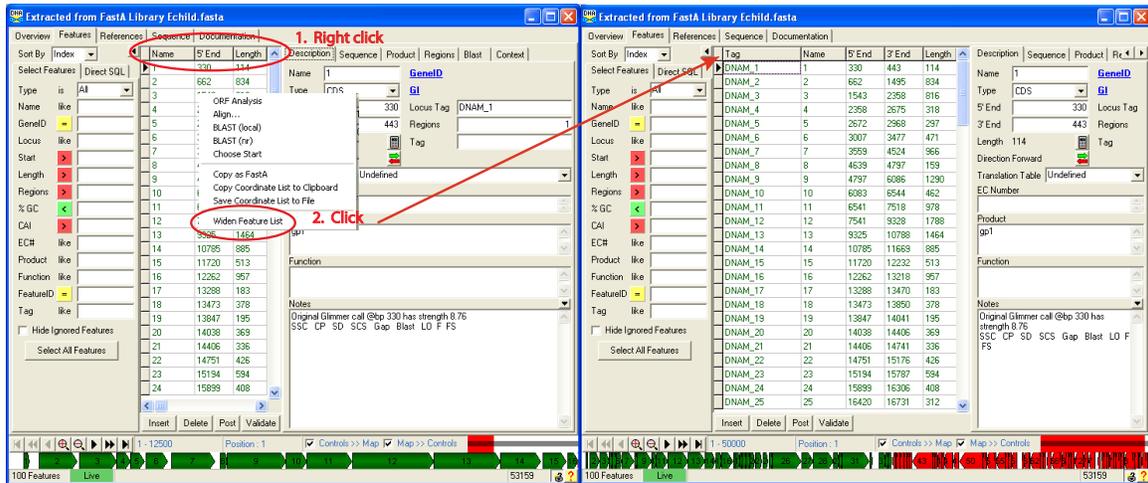


Figure 3

Choosing Start Data

The Choose start window is dramatically new and updated! The algorithms to score Shine Dalgarno sequences have been modified to include lots of choices. These choices are still under evaluation. In the meantime, you can find the "Old DNA Master" data in the SD Scoring Matrix and Spacing Weight Matrix in the upper right corner of the Choose ORF Start window. Look for more information coming soon!

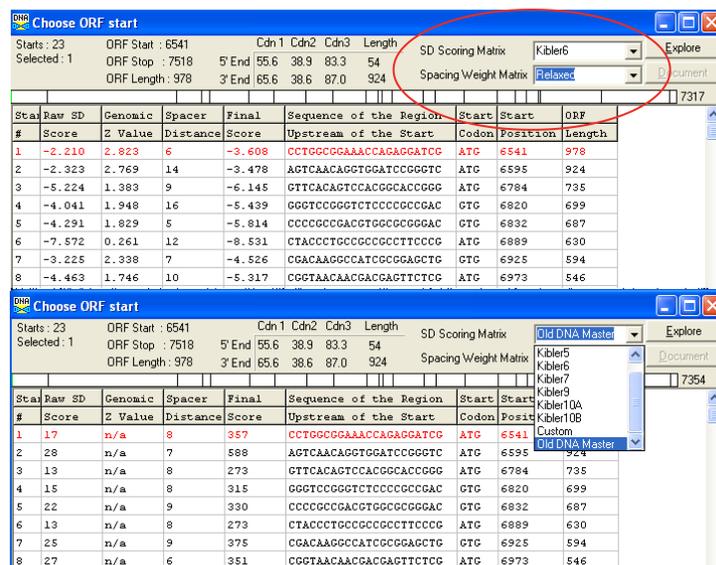


Figure 4

Event Manager

Though this isn't new or different, it is often overlooked. Go to Tools -> Event Manager. This window gives you more information when DNA Master is not working. Use the DNA Master Help Menu to evaluate the various events.

DNA Master Genome Manager and Compare Genomes functions

DNA Master allows you to load phages genomes into your own local database, and then perform several advanced bioinformatic analyses through the "compare genomes" function. The simplest one may be the 'Map comparison' of your phage genome with other genomes of your choice.

1. Load files

In order to compare files you will need to load files into Genome Manager. There are two kinds of files that you will want to load:

a. Current DNA Master files

To load a current file into your genome manager, with your .dnam5 file open, click "Genome -> Add to Database".

b. GenBank files.

To retrieve genomes already in GenBank, click "Tools -> Genome Manager" (Figure 5). The genome manager has a number of tabs, the left-most one labeled "Browse". This tab allows you to view all the files in your local database. The fourth tab from the left is labeled "Retrieve". From this tab, you can search NCBI for GenBank files and import them directly into your local database or open them as .dnam5 files.

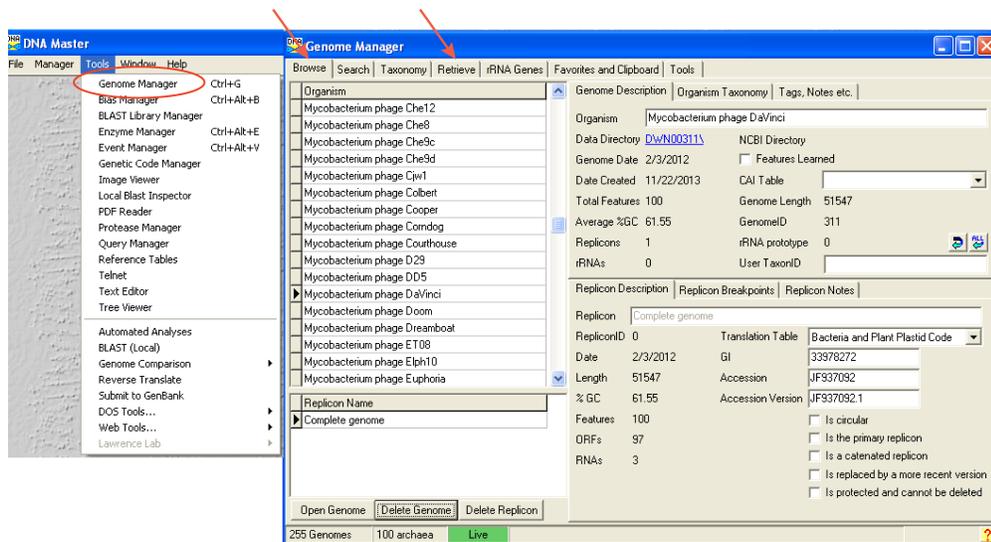


Figure 5

Using Figure 6 as a guide, follow this procedure to load Genome Manager with GenBank files.

- A. Go to Retrieve -> Fetch By accession
- B. Enter the name of the phage of interest. There are a few points about the Mycobacteriophages that you will want to keep in mind: PBI submits mycobacteriophage genomes to GenBank as Mycobacteriophage [DaVinci], when GenBank curates the file as a reference sequence, GenBank names the project Mycobacterium phage DaVinci. You made to look for both. In addition, there can be more than file for any given phage. There can be the submitted file and a the reference sequence file.. We recommend the submitted file. In addition, if the file was revised, there can be additional files.
- C. When you find the files of interest, click the “keep in Mind’ button.
- D. That action places the Accession Number in the left box on the page.
- E. Change the Fetch by Accession action to Save in local database.
- F. Change windows to the Browse window.
- G. You are ready to compare your genomes of interest. You will start by placing those genomes on a Clipboard.

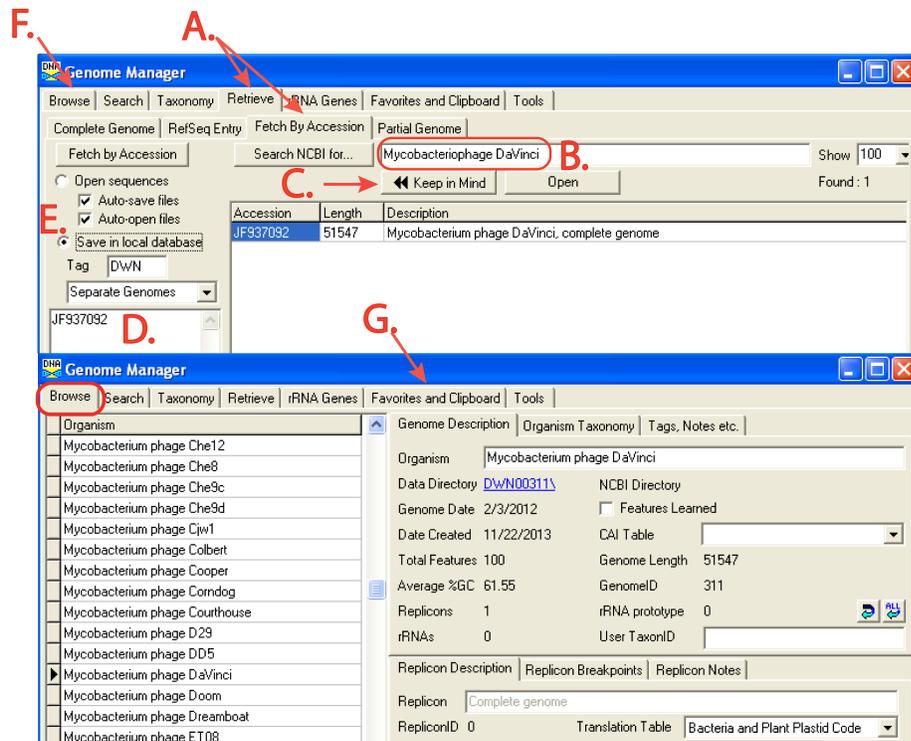


Figure 6

Note: you can add all and any file from NCBI into your Genome Manager. Once files are available you can organize them various ways. We will use the Clipboard for this task.

2. Collect your genomes for comparison

To collect your genomes for comparison, select Favorites and Clipboard from the Genome Manager menu. Once there Add the genomes of interest to the Clipboard (Figure 7). Then you may close that window.

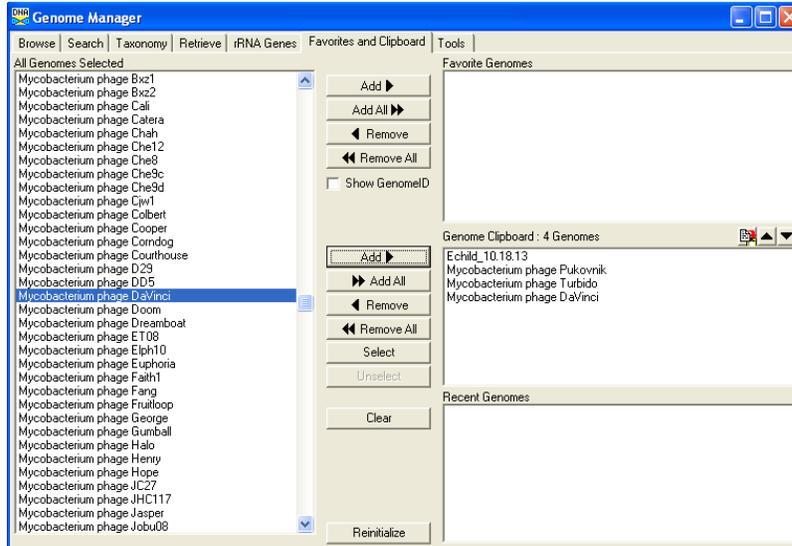


Figure 7

3. Compare Genomes

- a. Preference Settings: To perform genome comparisons between genomes within your genome manager, you first must enter a release code into a field in the program preferences. Click "File→Preferences", and then click the far right tab labeled "Miscellaneous". At the bottom of this window, there is a field labeled "Release Code". Enter "Watson" into this field, and then click "Apply". Then click "OK", and close the window (Figure 8).

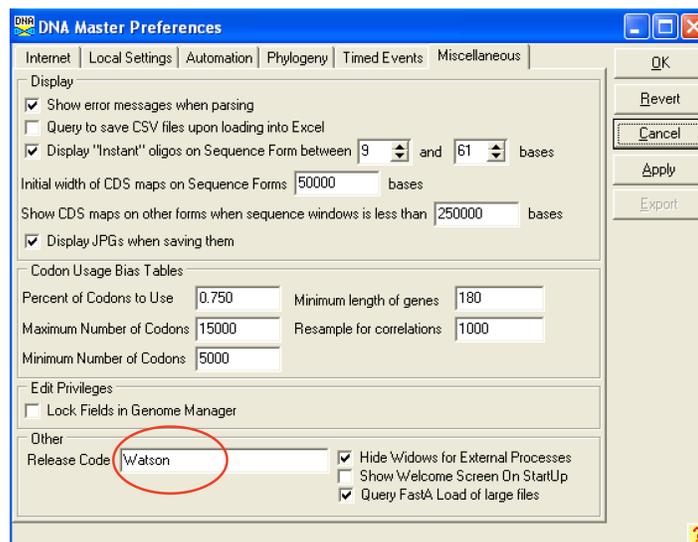


Figure 8

- b. Go to Tools -> Genome Comparison-> Manual from the main menu.
- c. The window that appears should look like the one in Figure 9.
- d. Click Clipboard and the genomes that you selected to place on the clipboard will fill the top right field of this window.

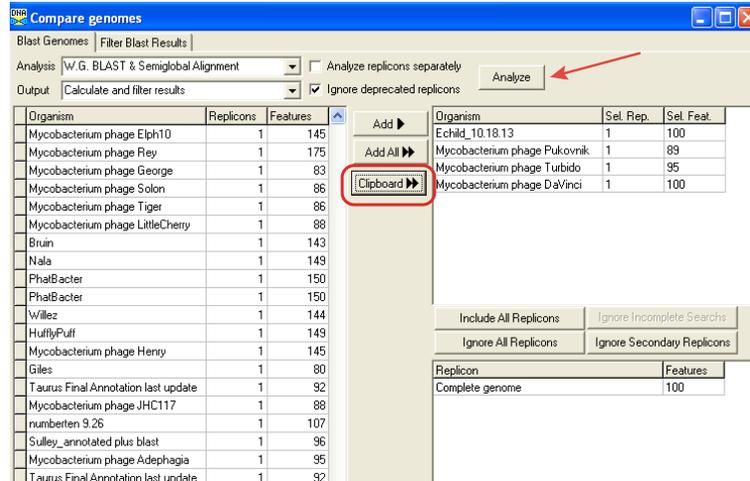


Figure 9

- e. Then click Analyze. This process will take a few minutes (It is dependent on the number of genomes that you have selected. A % completion window appears at the bottom right of the window.
- f. Once complete, a new window replaces the last one with a menu as depicted in Figure 10. Choose Map comparison from this menu. Figure 10 contains 3 separate windows. The last one is a graphical gene comparison map, similar to the maps seen in Phamerator. Comparative analysis is based on W.G. Blast & Semiglobal Alignment.

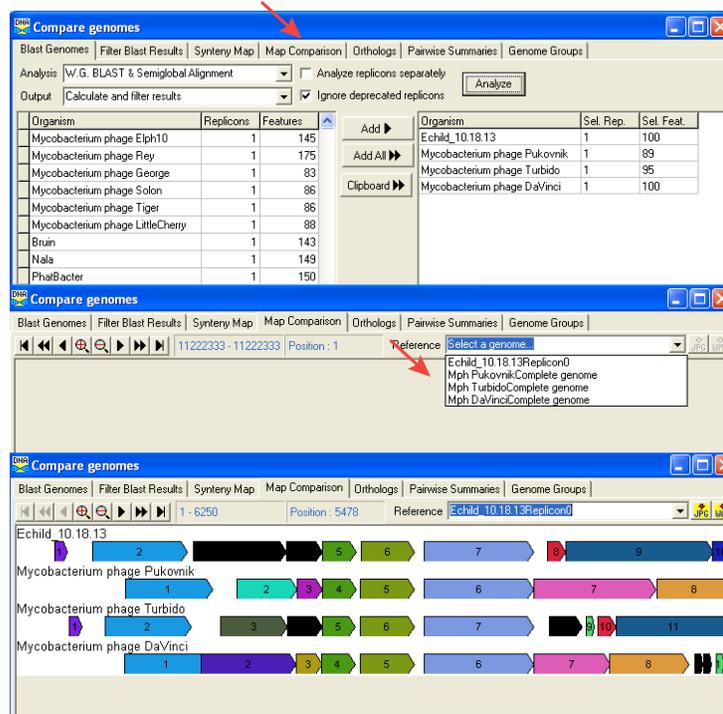
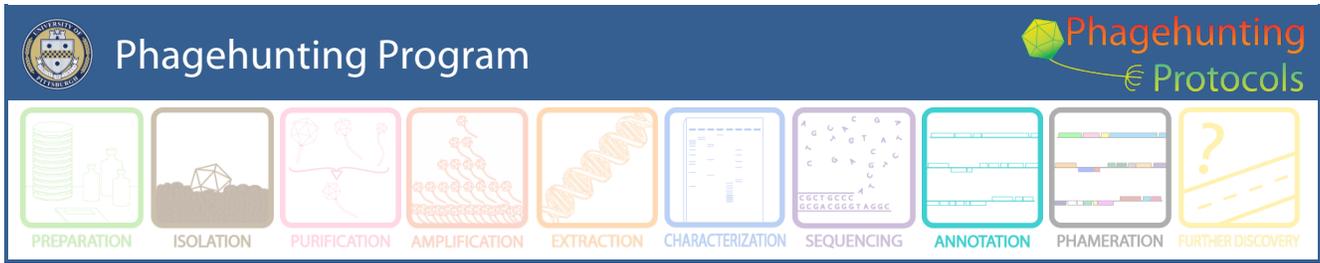


Figure 10

Note: This is a quick and easy way to compare your gene calls with genomes of your choice. It is especially helpful to be able to graphically display what is in a particular genome file. Depending on how many genomes you've added, and how long they are, this process can take quite a bit of time (an hour). For just a few genomes, it will be relatively quick—several minutes. To read more about this, look in the DNA Master Help files for "Genome Comparison".



What's New and Newly Recommended in the Annotation Guide

Created by djs November, 2012. Last updated on 1.22.2013, Revised 4.19.2103 Revised 10.22.2013 and still included because it is just good advice!

Phamerator News:

The user manual for Phamerator is located in Help menu. Check it for updates!

We recommend that you use the Notes Template: (This is not new, but we are still recommending it.)

This template can be added to the preferences setting and be added to the **Notes** window of all auto-annotated gene predictions

The suggested template is **SSC: CP: SD: SCS: Gap: Blast: LO: F: FS:**

- **SSC:** Start/stop coordinates. (This may seem redundant because there are "Start" and "Stop" fields that already contain this information, but it serves as a double-check that all changes you made are actually contained in the final file.)
- **CP:** Whether or not your start includes all the coding potential identified by GeneMark.
- **SD:** Whether or not the start has the best SD score of all this ORF's possible starts.
- **SCS** (Start choice source): Whether or not the gene was called by Glimmer and GeneMark, and if the start was called by same.
- **Gap** (or overlap): Any significant gap or overlap with preceding gene (in basepairs).
- **Blast:** The best BLAST match, and the alignment of the gene start with that BLAST match. (For example, "Matches KBG gp32, Query 1 to Subject 1", or "Aligns with Thibault gp45 q3:s45".)
- **LO** (Longest ORF): Whether or not the coordinates you have chosen yield the longest possible gene for that ORF.
- **F** (Function): Gene Function
- **FS** (Function source): source for the function (see **Section 10**). If the function assignment comes from a Hatfull-approved map in the Appendix, please also enter it into the field labeled "Function" directly above the "Notes" field. Otherwise, only enter the putative functional assignment in the Notes.
- Anything else you think is important. In particular if you made a different choice than previous annotators have made in published genomes, and feel very strongly about your choice, this is the place to let us know. **Example:** If your gene start does

not match the published starts of similar genes in GenBank, an explanation of why not. (“Published Thibault gp45 start not present in my sequence” or “Thibault start caused a 200 bp overlap with upstream gene”)

Caution: Do not add hard returns in the template of the **Notes** window because it takes up too much space and is a formatting problem.

GeneMark options:

1. It can be difficult for students to understand the data imported into DNA Master has the same value as the GeneMark *M.Tb* or *M. smegmatis* graphical data. You can get the same data (in a somewhat different graphic output – portrait vs. landscape representation) at http://exon.gatech.edu/heuristic_hmm2.cgi where you can run the GeneMark heuristic model on the web.

2. There is an updated version of GeneMark (2.8) for bacterial models. It can be found at http://exon.gatech.edu/gmhmm2_prok.cgi. This is relatively new and untested. Preliminary data suggests it will evaluate the genomes identically to GeneMark 2.5.

3. When using the web-based GeneMark against a model organism, the model organism you pick is based on the assumption that the phage and host have a relationship. That relationship may not be helpful when calling genes. I recommend running both the *M. smegmatis* and *M.Tb* and compare. One may provide better information than the other. If you do this, please send us a note with your evaluation.

Provisional Cluster Assignments:

No cluster assignment should be made on phagesdb.org before the genome sequence has been completed and evaluated. If you assign a cluster based on other data, it will be considered a provisional assignment.

Sequence Changes:

If you retrieve your sequence from phagesdb.org, your phage sequence will be ‘finished’ and in the correct orientation. If you are working on additional sequences or find that the sequence does need changed, DNA Master can make those changes without losing your database information for that genome. (Which means you won’t lose the work you have done to this point.) You can reverse-complement the genome or add and subtract bases. However, changes made on the screen do not become part of the binary database file UNTIL you post them to that file. Posting in the sequencing window is accomplished by clicking on the **Raw** button.