# Where do I click now?
# A reference guide for genome annotation in DNA Master

Getting started in DNA Master; from an auto annotated and BLASTED genome

- **Launch DNA Master** according to the DNA Master Installation Guide
- **Open the correct DNA Master file** (auto annotated and BLASTED according to DNA Master Quick Start Guide) using the Open command in the File menu.
- **Widen the Features table.** In the DNA Master tab menu→ [Features] →Right click "Name, 5'End, Length" row →Widen Features List
- **View Frames**. In DNA Master Tab menu →DNA → Frames
- **Display ORFs on Frames**. Click ORF button in the bottom right corner of Frames window

Navigating in DNA Master

- **Show RBS values**. On Frames window click ORF of interest to highlight it and click RBS button in the bottom right corner.
- **Clear ORF highlights on Frames window**. Click green arrow next to RBS button on bottom right of Frames window.
- **View entire sequence**. In the DNA Master Tab menu → [Sequence]. If ORF is selected it will be highlighted.
- **View DNA sequence of ORF**. Select ORF of interest in Features table → [[Sequence]] in Features tab.
- **View predicted AA sequence of ORF.** Select ORF of interest in Features table→ [[Product]] in Features tab
- **Show BLAST results.** Select ORF of interest → [[Blast]] in Features table. (Primarily for alignment data.)

Recording Notes in DNA Master; "The Big Ten" *[examples entries are in brackets]*

*Note: do NOT delete auto annotation notes*

1. **SSC: Start and Stop Coordinates.** List the coordinates of the first and last nucleotides in the ORF. For reverse sequences list the coordinates from lower to higher followed by (REV). *[34690-34823]*
2. **CP: Coding Potential**. Is coding potential contained within selected start coordinate*? [Yes, all of it.],[No, cp extends beyond 5'start.]*
3. **SD:** SD score of start codon (Kibler 6/ Karlin Medium). Is the Final Score the least negative? Is the Z score greater than 2? Record these numbers*. [Final = -2.112, Z = 3.151, these are the best scores.]*
4. **SCS:** Start Choice Source. Did Glimmer and/or GM call this start? If neither did, why did you choose something else*? [Agrees with GM], [Neither start has best BLAST alignments]*
5. **Gap:** Size of gap or overlap with upstream gene, in base pairs*.* Note: If your gene is transcribed left to right, the upstream gene is to the left of your gene. If your gene is transcribed right to left, the upstream gene is to the right. *[4 bp overlap]*
6. **BLAST:** Refers to N-terminus alignment of closest relatives (not function!). *[Top hit is gp63 of JoeDirt, Q: 1-133, S,1-130. Most other top hits are also q1:s1. One hit was q1:s8 GUmbie.]*
7. **LO:** Is this the longest possible ORF that does not generate a large overlap? If not, comment about your logic*. [Yes, there are two starts upstream. Both create large overlaps.]*
8. **ST:** Starterator. Is a Single Start given? If yes, record 'Single Start' and nucleotide number, if many starts record 'NI' (not informative), or 'NA' (not applicable) for an orpham. *[Single start, nt# 34690)]*
9. **F and FS:** Function and Function Source. Record the putative function and where that information came from. Use the official function nomenclature from the Official functions document. Use 'NKF' for No Known Function*. [ F= NKF, FS =checked BLAST, Phamerator, HHPred]; [F= terminase, FS = Blastp (NCBI) phage JoeDirt gp63, terminase, E=0; Blastp (phagesdb) same; Phamerator, CrimD gp 4; HHPred match to terminase prob = 99% over 80% of the protein]*
10. **Logic:** Discuss why you chose the start site you did over other possible starts, and support your logic by comparing SD scores, coding potential, gap and overlap sizes and BLAST alignments*. [Glimmer originally called the start at 34699. This does have the highest SD score but there is a 25 bp overlap with the upstream gene. We can use a downstream start at 34690 and still have all the coding potential contained, and a good SD score. So, we chose the downstream start. The choice was popular among other annotations because many hits had q1:s1 alignments, showing that they used the same start as we called here.]*

# Refining the Annotation

## Is this a gene?

- Is there coding potential? How strong is the coding potential?
- Does the predicted ORF fill the space?
- How does the predicted ORF fit in the context of the entire genome? Other genomes?

## Positional Annotation

### Is the predicted start site the best?

- What start did Glimmer and Genemark call?
- Is there an overlap? If the overlap is more than four base pairs you should look for a start site that reduces overlap. [It is unusual for an overlap to be more than 4 bp.]
- Is there a gap? Is there coding potential in the gap and possible starts that would include additional coding potential (CP)?
- If multiple start possibilities, that contain all CP, how do the SD scores compare?
    - Is the SD-Z score >2?
    - Is the Final SD Score the highest (least negative)?
- What are the BLAST results for the most likely start choices? Is the alignment beginning at Q:1 and S:1? If not, are there other starts that are available to make alignment 1:1?
- If you change the start to a different ORF start are the BLAST alignments better?
- In Starterator is there only a single start in common for all members of the pham, or are there multiple starts possible?

## Functional Annotation

### Can a putative function be determined?

- BLASTp at NCBI.
    - Do any of the matches with E values < 0.01 with good query coverage have listed functions in context of phage biology?
    - Do parts of the protein match to conserved domains? If so investigate these. What are they and what do they do?
- BLAST at phagesdb using the same criteria as above.
- Look at Phamerator. Do close relatives have functions listed? Are there domains listed?
- Copy and paste the sequence into HHPred to look for 3D structural similarities. Even though amino acid sequences may vary, structural similarities may help predict function.
    - A significant alignment of structures should have a Probability >90%.
    - Check the length of Subject sequence to the length of the match.
- Does the evidence agree?
- If you cannot predict a function, enter “NKF” or No Known Function.

## Enter Notes in DNA Master

- Check for consensus on your gene call.
- Follow the proper format and enter “The Big Ten” into the DNA Master Notes section.