Annotation Notebook Considerations

Just like when doing wet bench work, writing about what you did during annotation is needed. For each gene call you are making a claim and then providing the data that supports that claim. Essentially you are creating an argument for each gene call much like you would provide evidence in wet bench experiments to reach a conclusion. Most importantly, the evidence presented should make it very transparent how you reached the conclusion that you did.

There are a number of different ways that have been used by faculty. Using a PowerPoint slide for each gene may be the best idea that I have seen, but others use Word documents or even Excel documents. The decision of lab notebook formatting is up to you, it is the content and context presented in the notebook that is valuable. When you ultimately submit your files for QC, you will be required to fill out a coversheet that summarizes what you have accomplished during the annotation. The coversheet is still under revision, but will likely ask some basic questions about whether you completed steps expected for every genome. The coversheet will also provide a chance to indicate what genes were problematic. It is really important to document as you go.

When you have completed your annotation, the notebook is the record to document the work that you did to the QC team. So you will want to make this readable to them. We ask that as you annotate your genome you address the following 3 questions for each gene feature (include coordinates) when making your claims:

Is it a gene?
What is its start?
What is its function?

The data that you will want to consider is outlined in the Guiding Principles and includes:

Is it a gene?

- Coding potential: Which program called coding potential, and what did it look like (coordinates, how strong of an example)? (the 0 -1 graphic line in GeneMark, the coordinate data reported (in DNA Master) from Glimmer).
- Synteny: Is this in the right place for this gene?
- Blastp: How does this inform your call?
- Blastn: How does this inform your call? What does this tell you that the Blastp doesn't determine and vice versa?

What is its start:

- Coding potential: Which program called coding potential, and what did it look like (coordinates, how strong of an example)? (the 0 -1 graphic line in GeneMark, whether it is reported (in DNA Master) from Glimmer
- Starterator: Does this provide useful information? Explain.
- Start choices:
  - Did you capture all coding potential?
  - Spacing: what is the relationship to the upstream gene?
  - Ribosomal binding site data, promoter considerations: Is it necessary and how did you apply it
- Blastp: How does this inform your call?
- Blastn: How does this inform your call? What does this tell you that the Blastp doesn't determine and vice versa?

What is its function?

- Synteny: Is this the right place for this gene?
- HHpred: What databases were used, what was informative?
- Blastp: for function assignments. did you blast at phagesDB and NCBI?
- Note: this can inform "Is it a gene". And "what is its start".

A picture is worth a thousand words. As you address the considerations listed above and the Guiding Principles, showing the data will be necessary. Show often!